Population and Landscape Genomics Workshop – Canberra
Practical work in landscape genomics

Leempoel Kevin & Joost Stéphane

Thursday, 27 March 2014

Content

Introduction

Our practical work on landscape genomics will use data on Loblolly pine (*Pinus taeda*) sampled in the US by Eckert lab http://eckertdata.blogspot.ch/, (Eckert, Bower, et al., 2010; Eckert, van Heerwaarden, et al., 2010). The purpose is to compute associations between SNPs data and environmental variables that will be downloaded or computed in a GIS.

This practical was tested on Windows 7 and Linux programs were run in a virtual box using Ubuntu.

We will use a **GIS software** mostly to visualize data (Quantum GIS), another one to produce environmental variables from Digital Elevation Models (DEMs) (SAGA GIS), and a third one to compute spatial statistics (OpenGeoda).

Quantum GIS can be found here:

- For windows: http://www.qgis.org/en/site/forusers/download.html
- For Mac OS: http://www.kyngchaos.com/software/qgis

SAGA GIS can be found here:

- For Windows (recommended) : http://sourceforge.net/projects/saga-gis/files/
- For Mac OS : http://sourceforge.net/apps/trac/saga-gis/wiki/Compiling SAGA on MacOSX Compilation of SAGA GIS for Mac OS is quite complicated. I would recommend to borrow a computer running Windows for this part.

OpenGeoda can be found here:

- For Windows
- For MacOS

Landscape Genomics software

We will use are SamBada - based on multivariate logistic regressions -, LFMM, which considers population structure, and Admixture which computes membership coefficients to populations for each individual.

- **SamBada** can be found here for Linux, Windows and Mac OS: http://lasig.epfl.ch/sambada

- **LFMM** can be found here for Mac OS, Linux and Windows 64 bits (GUI version will be easier but slower):http://membres-timc.imag.fr/Eric.Frichot/lfmm/software.htm

- **Admixture** can be found here for Linux and Mac OS: http://www.genetics.ucla.edu/software/admixture/download.html For Windows users we recommend to install a virtual box using Ubuntu or to borrow a computer running Mac OS or Linux. You could also use STRUCTURE to obtain similar coefficients

- Virtual box: https://www.virtualbox.org/wiki/Downloads

- Ubuntu: http://www.ubuntu.com/download/desktop

- STRUCTURE:http://pritchardlab.stanford.edu/structure_software/release_versions/v2.3.4/html/structure.html

We will also use **R for statistical analyses**. You can install a modified GUI for R such as R studio

- R (Linux, Mac OS, Windows): http://www.r-project.org/
- R Studio (Linux, Mac OS, Windows): https://www.rstudio.com/ide/download/desktop

Following packages should be installed as well:

- RSAGA

Genetic Data

We transformed **genetic data** to PLINK format in both binary (BED) and ordinary (PED) format.
Data can be found here:

- LoblollyPineGeneticData.zip

Environmental Data

Sampling locations with aridity variables can be found here:

- Loblolly Pine coordinates

Several individuals have identical coordinates. In the purpose of visualizing them all in a GIS, we suggest modified coordinates.

- Loblolly Pine Visualization Coordinates

We will use **climatic variables** from **Worldclim** datasets.
In the interest of time, we have created a **subset for our study zone** that you can download from our server : **WorldClim_Subset.zip**
Original datasets can be downloaded either by thiles or for the entire world:

- Entire world: http://www.worldclim.org/current
- By tiles: http://www.worldclim.org/tiles.php

DEMs can be found on Earth Explorer (subscription is mandatory before download):

- http://earthexplorer.usgs.gov/ We will use **GTOPO30** only.

Recommended readings

Papers regarding practical work datasets can be found on Eckert's blog:

- http://eckertdata.blogspot.ch/ at the date of 27 January 2012

We also recommend reading papers and documentation related to the software we will use

SamBada

- Joost S., Bonin A., Bruford M.W., Despres L., Conord C., Erhardt G. & Taberlet P. (2007) A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. Molecular Ecology 16, 3955-69.

LFMM

- Frichot E., Schoville S.D., Bouchard G. & Francois O. (2013) Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. Molecular Biology and Evolution 30, 1687-99.
- LFMM tutorial

Admixture

- http://www.genetics.ucla.edu/software/admixture/publications.html
- http://www.genetics.ucla.edu/software/admixture/admixture-manual.pdf

Visualisation of data

We will first visualize the data in Quantum GIS.

A note on coordinate systems

When we use different types of data in a GIS, we have to make sure that they are expressed in the same coordinate system. The standard global system is **WGS84** and most environmental data and DEMs available are georeferenced in that system. WGS84 is expressed in degrees and not in meters. However, for most applications, we use **local projected coordinate systems** that are often national systems and calibrated to have a good local accuracy.

In our case, samples of Loblolly pines are originally in latitude/longitude (WGS84) and are spread across a large region of the United States. For the simplicity of this practical work, we will use only one projected reference system when needed: **UTM16North**.

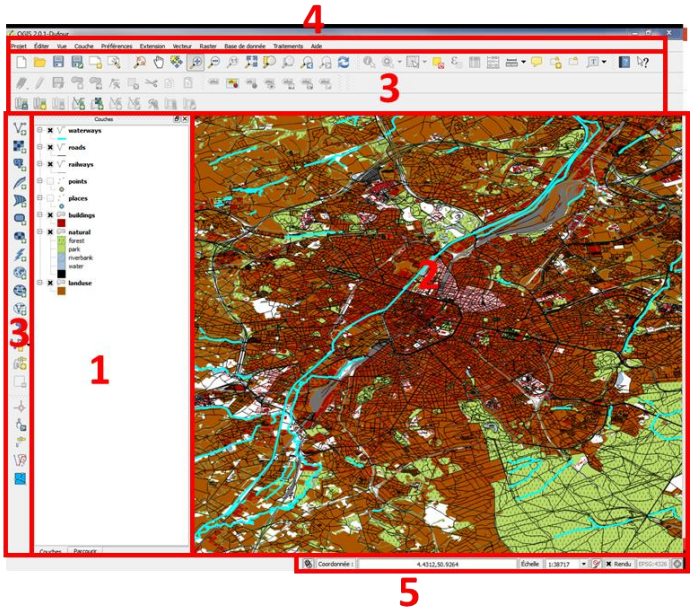
We will use their **EPSG** code, a structured dataset of Coordinate Reference Systems and Coordinate Transformation.

EPSG code for WGS84 : 4326

EPSG code for UTM16North : 26916

Quantum GIS

Open Quantum GIS

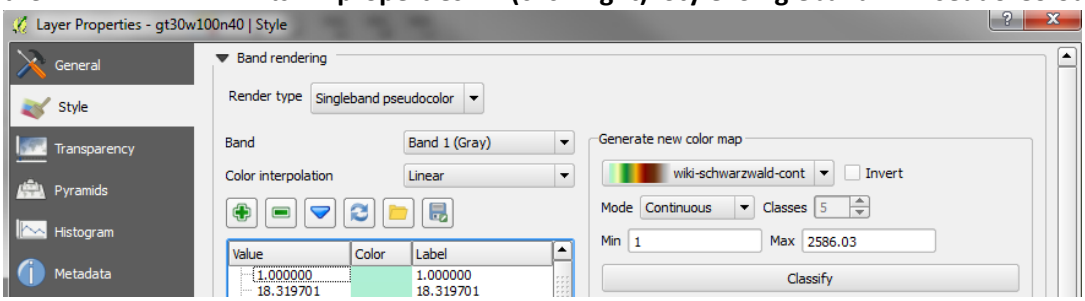


QIGS interface is divided in 5 distinct zones :

1. Map Layers
2. Map view
3. Toolbars
4. Menu bar
5. Status bar

- Map Layers: Shows all data imported or created in the project.
- Map view : Activate or deactivate each layer in the display zone by clicking on its cross  in the layers manager.
- Status bar: displays coordinates and scale.

1. Open the file of **Loblolly Pine coordinates** (Loblolly Pine ENVdata .csv) using **add a text delimited file** button  , choose the appropriate text delimiter and define X and Y columns if it isn't automatic. It will then ask you to choose a coordinate system, which is here the standard latitude/longitude projection called **WGS84**.
2. When the file is displayed in your map layers, save your point layer as a **shapefile** (shp) by making a right click on the layer in the map layers window and click on **save as**. Choose the path where you want to save it, otherwise there is no change to do.
3. **Download** the **GTOPO30** DEM from http://earthexplorer.usgs.gov/ website for our study area. It will be available as a tile in the coordinate system WGS84.
4. You can add now a background DEM layer using **add a raster layer** button  and open the file **gt30w100n40.tif**. However, a DEM seen like this is not very informative, so we will compute a hillshade model to better understand were samples are.
5. Go to **Raster>Terrain Analysis>Hillshade** to compute a Hillshade raster layer and add it to the project.
6. We can now reorganize the different layers to help us in the visualisation. Loblolly Pine point layer should be on top of the Map Layers and either Hillshade or DEM should have a transparency of 30-50% to visualize both layers. You can also change the colour scheme of the DEM in its **properties (click-right)>Style>Singleband Pseudo-Colours.**



We now have a good idea of where the samples are and we can now acquire environmental information.

7. You can save your project and close QGIS for the moment.

Acquiring environmental information at sample's locations

We will first compute several environment variables from a digital elevation model (DEM)

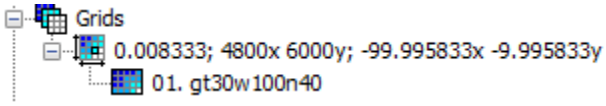
Computing Digital Elevation Model environmental variables

Create a new project in **SAGA GIS** (**File<Project<New project**).

8. Click on **Load**, and open the shapefile you created with QGIS and the DEM (select **All files** instead of **All recognized files** to open a .tif).

The numbers shown in your Data workspace are :

- the cell size;
- the # of cells on x and y;
- The coordinates of the southwest corner.

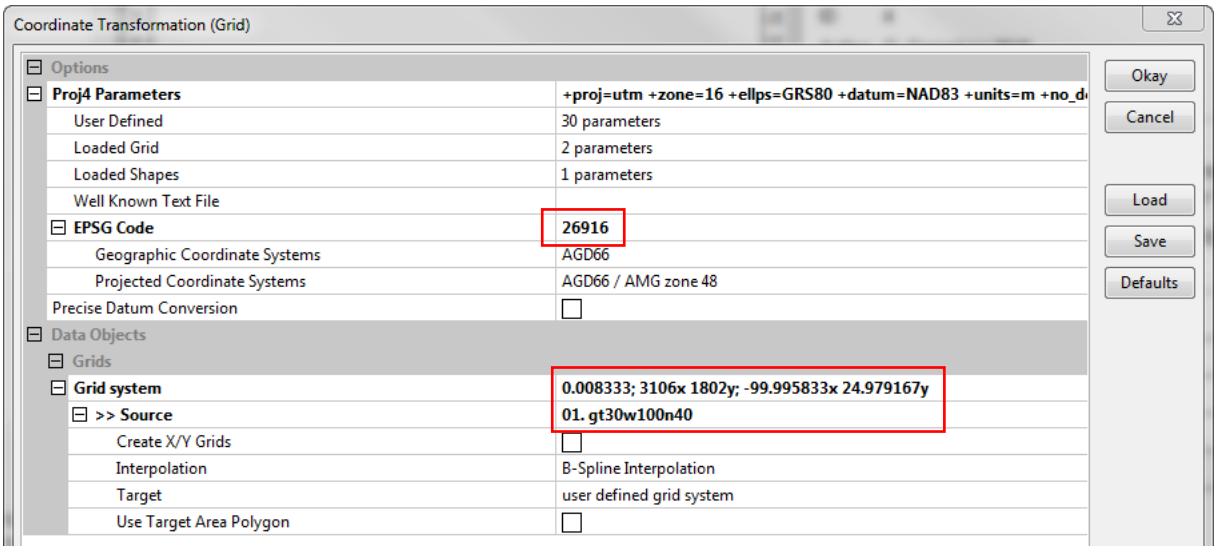


The area depicted is too big for our purpose, so we will cut it at the size of our area.

9. We will thus use a **shapefile of the US boundaries**, you can download it here : http://geocommons.com/overlays/21424 Load it in SAGA GIS
10. To cut the DEM, use the function **Shapes – Grid > Clip Grid with Polygon** and cut it with US boundaries shapefile.
11. Check that all our Loblolly pines are in the area of the DEM by overlaying the DEM and the loblolly pine shapefile.
12. **Save your project**. I gives you also the opportunity to save the subset of the DEM.

The DEM provided is based on GTOPO30 data, expressed here in degrees, so we have to reproject it from WGS84 to UTM16N to be in meters.

13. In **Projection – Proj. 4>Coordinates transformation (Grid)**, select your DEM and type the correct projection system : EPSG **26916** and **press enter.** It wil change the first line of the window. Then select your DEM in **Grid system** and click okay.



Local Morphometry

Now that the DEM is expressed in meters, we can compute DEM variables.

SAGA offers a very useful toolbox to quickly compute terrain attributes .

14. Click on **Terrain Analysis – Morphometry > Slope, Aspect, Curvature** .

Choose your **grid system** and select your **DEM** raster. Only **Slope, Aspect** and **Curvature** output options should be set to **[create]**. The method is "**Fit 2.Degree Polynom (Zevenbergen & Thorne 1987)**" (default value).

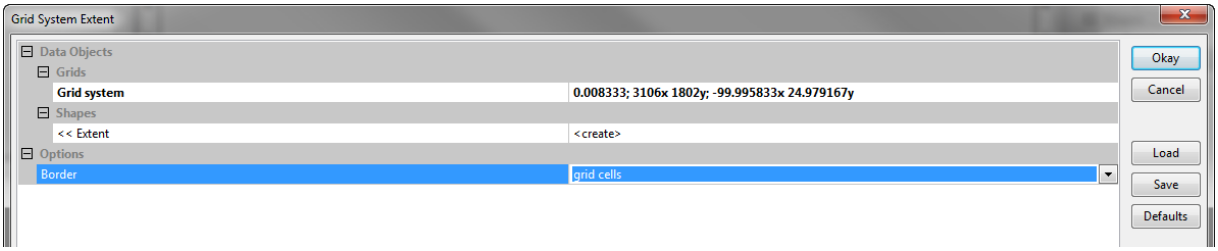
Potential Incoming Solar Radiation

Computing solar radiation variables requires having layers of latitude and longitude.

Therefore we will use the DEM we cut in **WGS84** to extract latitude and longitude values of each pixel and convert them later to **UTM16North**.

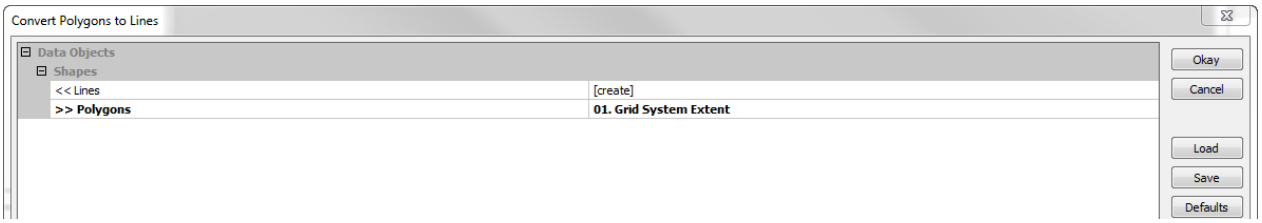
First, we need to create a polygon at the size of the system extent.

15. Go to the module **Shape – Grid > Grid System Extent**. Select the DEM in WGS84 and select **grid cells** for border.

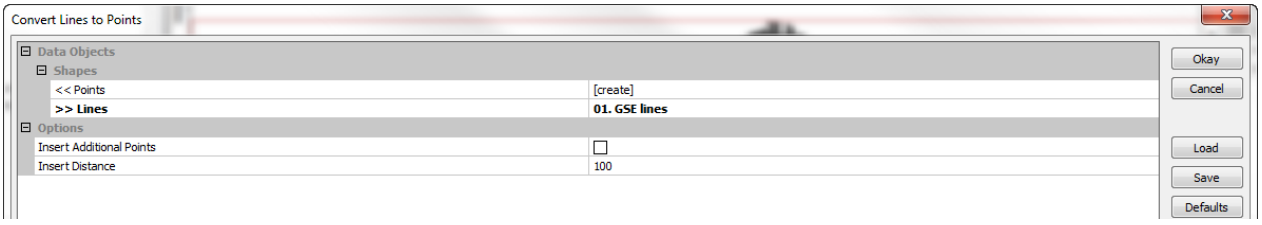


Now we have to convert the corners of the polygon to points, but we cannot do that directly so we'll first convert the polygon to lines and then the lines to points

16. Go to the module **Shape – lines > Convert Polygons to Lines**



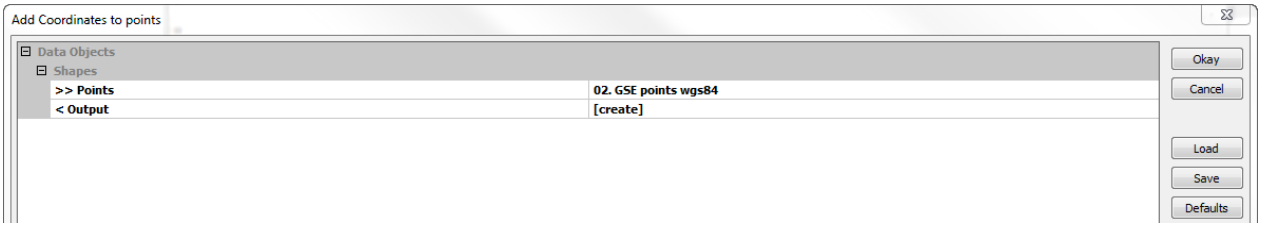
17. Go to the module **Shape – points > Convert Lines to Points**



If you do a **map with your DEM and your points in WGS84**, the points should be at the 4 corners of the DEMs' minimal and maximal extents.

We can now add the coordinates to points.

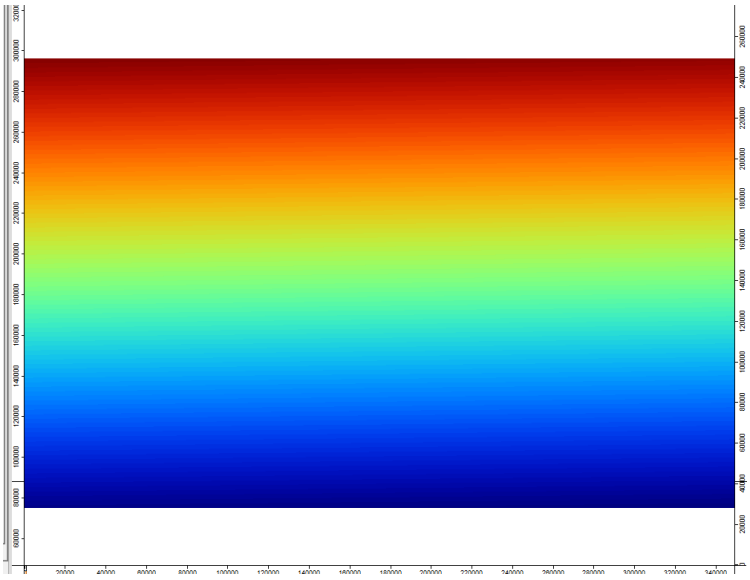
18. Go to the module **Shape – points > Add coordinates to Points**
    Select your transformed point shapefile (lat/long) and create a new one. The lat/long coordinates will be added as attributes.



Now, we have to produce a grid of X(long) and a grid of Y(lat) at the same extent as the DEM.

19. Go to the module **Grid – Gridding > Natural Neighbour**
    Select your **system extent point shapefile** and the attribute **X**. In the **Options**, the **target Grid** should be set to **grid**. Click **Okay**. A new will open, asking you to select the grid system. Select the grid system of your DEM in **WGS84**.
20. Do the same with attribute **Y** and rename it DEM lat.

The resulting **DEM lat** grid should look like this.

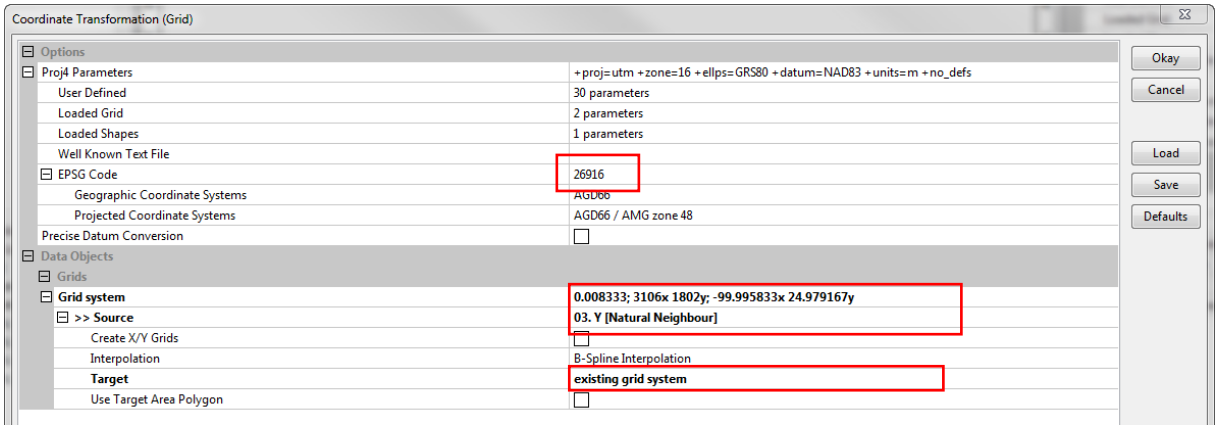


That's it; you now have your latitude and longitude layers expressed in degrees but georeferenced in WGS84 coordinate system. The computation of the lat and long grid layers may take a while and use A LOT of memory.

Now, we just have to **transform these two layers** to our projected coordinate system **UTM16N**. Before you do that, make sure that lat and long layers are in the WGS84 coordinates system (**see description tab of the raster**). If not, use **Projection – Proj. 4>Set Coordinate reference system**.

We will use the same module of transformation we used earlier except that this time we want them to be in the same grid system as our transformed DEM.

21. In **Projection – Proj. 4>Coordinates transformation (Grid)**, type the correct projection system : EPSG **26916** and **press enter.** It wil change the first line of the window. Then select your Lat raster in **Grid system** and change **Target to existing Grid system** and click okay. A new window will open, asking you to choose the grid system, select the one in UTM16N.



In order to calculate the Potential Incoming Solar Radiation, we need to compute first the skyview factor

22. **Terrain analysis – Lighting, Visibility>Sky View Factor**

Leave the parameters to default.

We can now compute **Potential Incoming Solar Radiation**. By specifying a latitude and longitude grid, the usual latitude choice and planetary binding will disappear.

23. **Terrain analysis – Lighting, Visibility>Potential Incoming Solar Radiation.** Select your DEM, the sky view factor you computed and the two layers of latitude/longitude given in the data. In addition to direct and diffuse, select **Total Insolation** and switch to **[create]**. Select a period of 1 day and a time resolution of 1h for 21 June.

It will take some time.

24. Do the same for 21 December and save all your layers with appropriate names.

Exporting the data

The coordinates of your loblolly pine shapefile are still in WGS84, therefore if we want to extract the values of our DEM variables, we need to transform them as well in **UTM16N**.

If necessary, assign the correct projection system to your shapefile (Lat Long) in **Projection – Proj. 4>Set coordinate reference system**. Select your shapefile in Data Objects>Shapes.

25. In **Projection – Proj. 4>Coordinates transformation (Shapes)**, select your shapefile and type the correct EPSG code.

Now we can transfer the values of these rasters to the points.

26. Use **Shape – Grid<Add grid_values to points**. Interpolation method should be set to "**nearest neighbour**".
27. Save the shapefile you just created and export also the table as a text file using **Import/Export - Tables < Export Text Table**.

28. Save your project and close it.

Retrieving environmental information from major internet sources

We will use data from Worldclim and download them directly from their website. However, these files are heavy and we will not ask you to download them all and extract about 100Go of data. In addition, because of their size, we cannot open them all in SAGA GIS at once. It means that we have to cut them at the size of our study zone and you don't want to do it manually for 50 rasters.

Therefore, we decided to give the raster already cut, so you will be able to open most of them at once and extract their values at sampling locations.

You can found these rasters here: https://documents.epfl.ch/groups/l/la/landscape-genetics/www/LandscapeGenomicsWorkshop/

For information, I cut them using a script in RSAGA that you can re-use for other purposes.

```
library(RSAGA)

setwd("C:/Data")

myenv <- rsaga.env(workspace=getwd(), path="C:/Program Files/SAGA-GIS")

filename='bio_'

for (i in 1:19){

 rsaga.geoprocessor("io_gdal",0,list(GRIDS=paste('Worldclim/', filename, i, '.sgrd', sep=""),

                              FILES=paste('Worldclim/', filename, i, '.bil',sep="")))

 rsaga.geoprocessor("shapes_grid",7,list(OUTPUT=paste('Worldclim/', filename, i, 'TEMP.sgrd', sep=""),

                                 INPUT=paste('Worldclim/', filename, i, '.sgrd', sep=""),

                                 POLYGONS= paste('Worldclim/','Grid System Extent Loblolly.shp',sep="")))

 rsaga.geoprocessor("grid_tools",15,list(INPUT=paste('Worldclim/', filename, i, 'TEMP.sgrd', sep=""),

                                 RESULT=paste('Worldclim/', filename, i, '_USA_WGS84.sgrd', sep=""),

                                 METHOD='0', OLD='-9999', NEW='-99999'))

}
```

Create a new project in SAGA GIS (**File<Project<New project**).

29. Click on **Load**, and open your loblolly pine shapefile and the rasters of climatic data.

The shapefile and the climatic rasters are all in Lat/Long – WGS84, so no coordinate transformations are needed.

30. Transfer the values of these rasters to the points using **Shape – Grid<Add grid_values to points**. Interpolation method should be set to "nearest neighbour".
31. Save the shapefile you just created and export also the table as a text file using **Import/Export - Tables < Export Text Table**.
32. Save your project and close it.

Correlation coefficients to eliminate multi collinearity

We now have plenty of variables at our sampling locations. What we want to do now is to merge both tables and evaluate if some are highly correlated or not.

33. Merge the tables with DEM and climatic variables in one file (using Excel or R for example)
34. And use R to compute correlations coefficients between the environmental variables at sampling locations. You can use the function **cor(x, type="spearman")** to get correlation coefficients.

The idea here is to get rid of variables that are highly correlated so suppress every variables that have a correlation coefficient higher than abs(0.8).

35. You should try to have a dataset with not more than 10 environmental variables (otherwise computation time in the next steps will take too long).

Identifying loci under selection with SAMβada

We will now use SAMbada in order to assess if some markers are associated with an environmental variable. What we are also interested in is to measure how much population structure can explain the spatial repartition of genetic markers. Therefore we will first use admixture to estimate population structure and add membership coefficients as variables in multivariate logistic regressions in SAMbada.

Computing membership coefficients with Admixture

Admixture computes membership coefficients for each individual of the datasets based on SNPs data. However, we have to estimate in advance the number of populations. Therefore, we will ask Admixture to compute these coefficients for several K (number of populations) and use their likelihood to find the best K.

36. In the console, go to the directory where you downloaded admixture **cd C:\Data\Loblolly** (for example)
37. And type a command to compute coefficients using our binary PLINK file for several Ks.

For example : sudo ./admixture32 loblollypine-binary.bed 2

Depending on if you are working in linux or mac directly or in a virtual box, you may have to add **sudo ./** at the beginning. Also, depending you are working on an **x86** or **x64** environment, you'll have to use **admixture32** and **admixture** respectively. **loblollypine-binary.bed** is the binary PLINK file containing the SNPs and **2** is the number of populations.

You can use a loop to try several values of K.

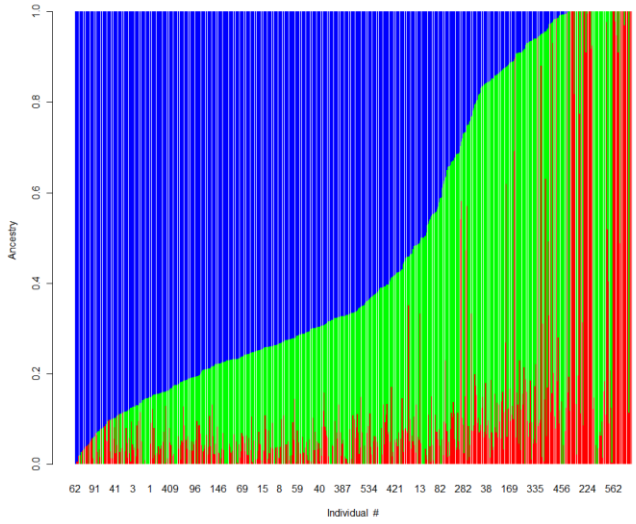
38. **For K in 1 2 3 4 5; do admixture --cv loblollypine-binary.bed $K | tee log${K}.out; done**
39. When it's finished, you can type **grep "K=" *.out**, to show which K is the most appropriate by using the one with the lowest error.

Please refer to the manual for more options.

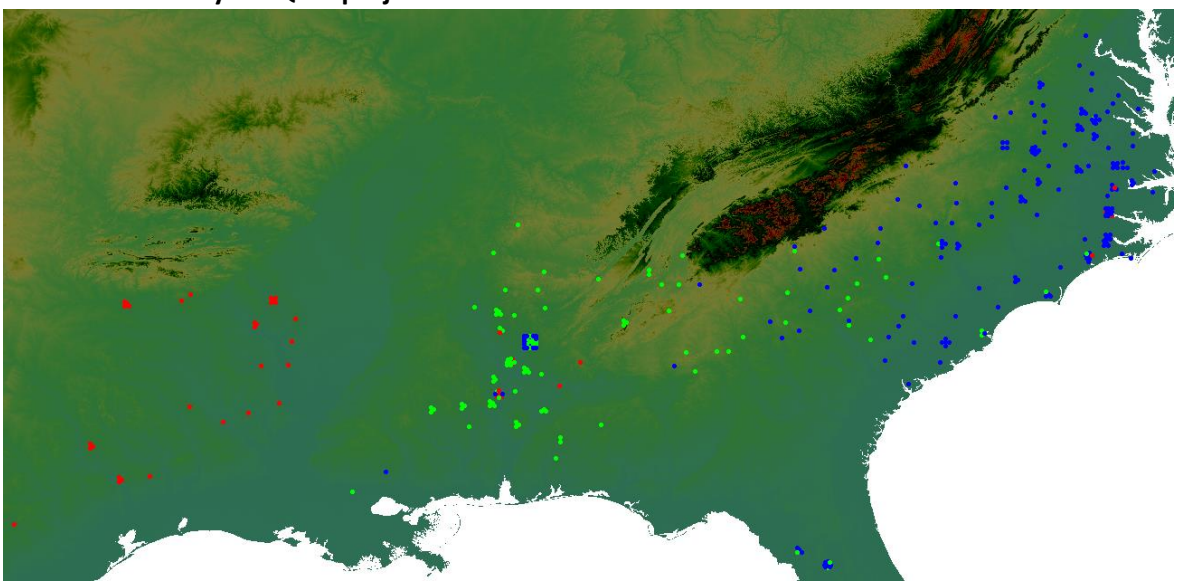
For each value of K, there is a .P (allele frequencies of the inferred ancestral populations) and .Q (ancestry fractions) output file.

We have two options to visualize **population structure (.Q output file)**.

40. One is to use the R script detailed in the manual to create a **bar plot graph** where each color represents the membership of each individual to a population.



41. The second is to modify .Q file in order to **add columns of coordinates** and **a column** in which is **indicated to which population** (so an integer number) the individual is most likely to belong. The purpose is to create a map in which we can observe where populations are situated. However, you may have notice in the data that **several individuals have the same coordinates** and cannot therefore be all represented on the map. Use the **coordinates for visualization** in the modified .Q file. With these coordinates, individuals with identical coordinates were displaced around their original coordinate.
42. **Add this table to your QGIS project.**



43. The other interest of membership coefficients is to include them in multivariate logistic models. To do so, include in your table of environmental variables one or two columns of membership coefficients from Admixture. For example, in the figure above, we found that one population is isolated from the two others. Therefore, the membership coefficients to this cluster can be added as a variable. For the other two populations, only one column of coefficient is enough since they are highly correlated due to their isolation from the third population.

For information, if you have several clearly distinct population, it may be recommend to do a PCA on the coefficients and use one or two axis from the PCA as population structure variable in the logistic models.

Transforming data from PLINK or others to SAMβada and LFMM

SAMbada can work with genetic and environmental data in distinct files. However it needs to convert genetic data to its own format. To do so, it integrates a module that transforms PLINK data to SAMbada data.

44. In the console, select the folder where SAMbada is using the command **cd C:\Data\Loblolly** (for example) Then type the following line
45. **.\recodePlink.exe 622 3082 .\loblollypine.map Lob_GEN.txt** Where 622 is the number of individuals and 3082 the number of genetic markers. **Lob_GEN.txt** is the output file. Data recoding is fast.

```
622 ?ú 3082
.\loblollypine.map .\loblollypine.ped .\loblollypine.map
Temps de lecture: 1
622 6 1
```

Multivariate models in SAMβada

We can now modify the parameter file to inform SAMbada on the content of our dataset and define which models we want to compute.

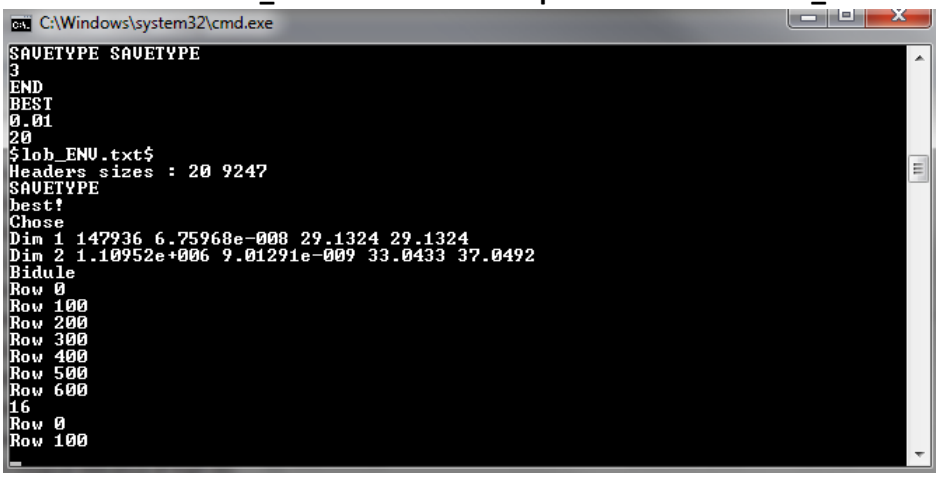
The structure of the parameter file is as follow:

```
INPUTFILE lob_ENV.txt lob_GEN.txt

HEADERS YES

NUMVARENV 20

NUMMARK 9247

NUMINDIV 622

IDINDIV ID_indiv

COLSUPENV state lat long

DIMMAX 2

SAVETYPE END BEST 0.01
```

Values are delimited by spaces; therefore there cannot be any spaces in the names of the variables. We suggest to get rid of the column **county** as several spaces are present in this column. Otherwise you can specify another text delimiter. Please **refer to the manual for details**.

DIMMAX refers to the maximum number of variables for each model. In in the interest of time, we suggest not to use more than 2. The last line means that your output file "…Out-1" and "…Out-2" only contains the significant correlations at 99%. Models are ranked in decreasing order for the G-score (the most significant is thus the first one). If you have selected to write all models in the output file, you can open the output table in excel and apply the function **CHIDIST** on G and Wald scores to obtain their p-values.

46. In the console type the following line (check filenames) to start computation of models in SAMbada: **WIN64_Sambada.v09.exe parameters.txt lob_ENV.txt lob_GEN.txt**



It will take around 20 minutes to perform computations. Some errors wil occur. Its perfectly normal, some markers are constant (either only 0 or only 1) and some models will not converge.

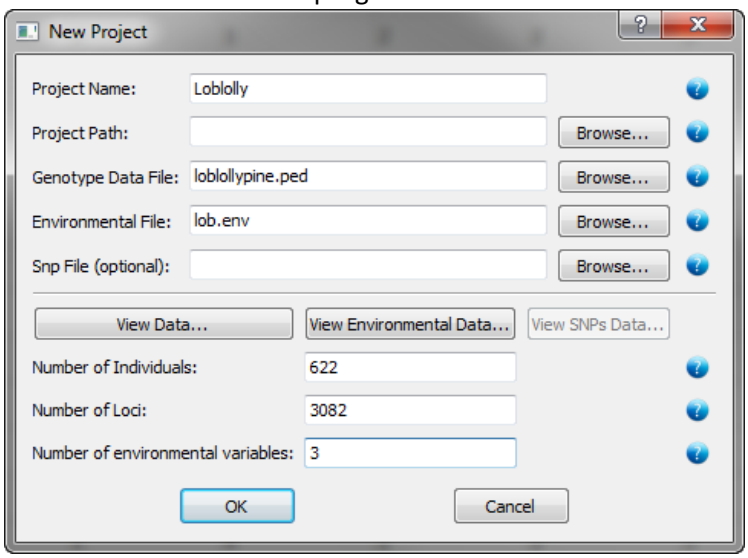
47. Have a look at **univariate and bivariate results** to analyse how many markers are significant and which environmental variables have a significant explanatory power.

Identifying loci under selection with approaches considering population structure directly

LFMM

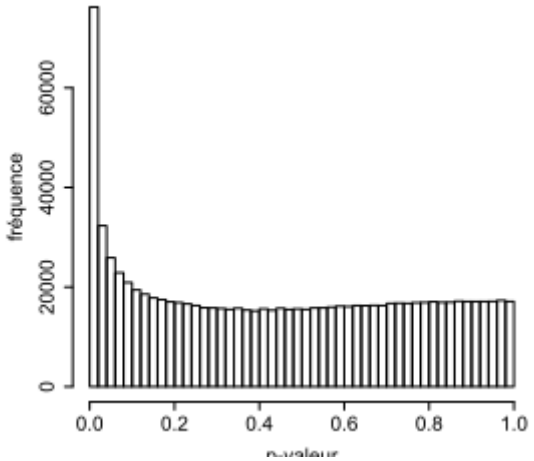
LFMM performs univariate logistic regression but includes latent factors that estimate population structure and therefore tries to suppress false positives.

48. **Open the GUI** of LFMM and create a project. You can use the **PLINK file** for genetic data. Environmental file cannot contain headers, has to finish by **.env** and has to be space delimited. Inform also the program on the number of individuals, loci and variables



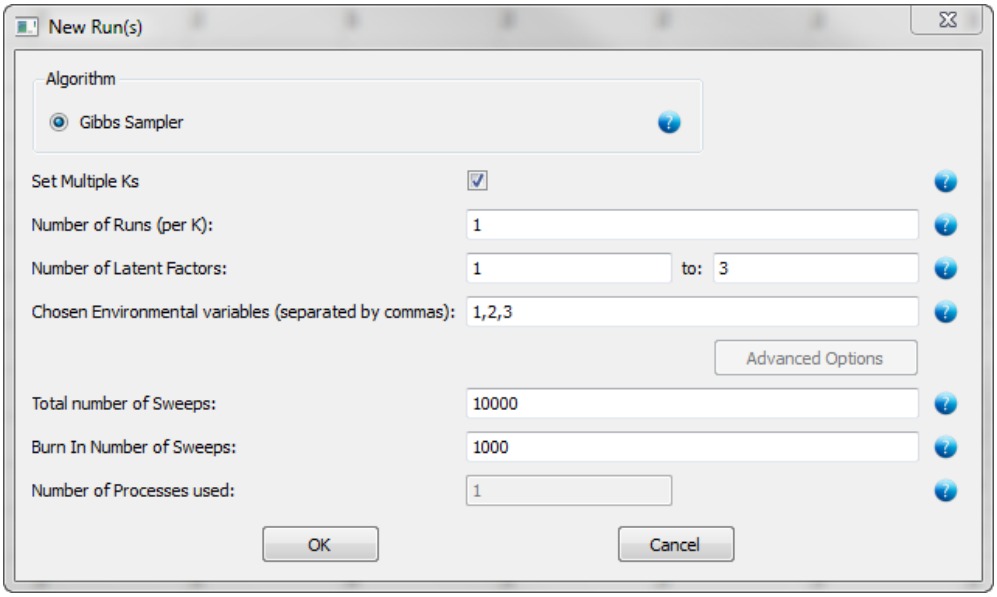
LFMM is slower than SAMbada so we suggest that you identify in SAMbada univariate models the variables that are often significant and with the highest G scores. **Try to focus on 3 or 4 variables** in LFMM, otherwise you will not have time to analyse the results.

The parameter K in LFMM is not the number of populations; it is more related to the number of principal components of population structure. For example if you found 4 populations, K will most likely be equal to 2. Authors recommend trying several values of K to obtain robust information on loci under selection. You could try K values from 1 to 3 for example. What they also suggest is to look at the distribution of p-values for each K. Indeed, the most appropriate K is the one for which p-values distribution looks like this.



This can be done in R using histograms.

49. **Launch** **Runs**



Comparative analysis of detected loci

With the results from multivariate SAMbada and LFMM, try to identify commonly detected loci from both methods. It is interesting to rank them for G or Wald scores from SAMbada for example and look if those detected by LFMM or other methods are also on top.

50. Based on these results, create a **new table** where you only put **markers that seem interesting** such as those detected by both methods or those significant from the multivariate analysis in SAMbada. Add also the **explanatory variables** for these models. Because we are interested in the spatial repartition of these markers, use only their **original coordinates**.

Spatial structure analysis of detected loci

Indicator of spatial autocorrelation using Univariate LISA (Local Indicator of Spatial Association)

Global spatial autocorrelation

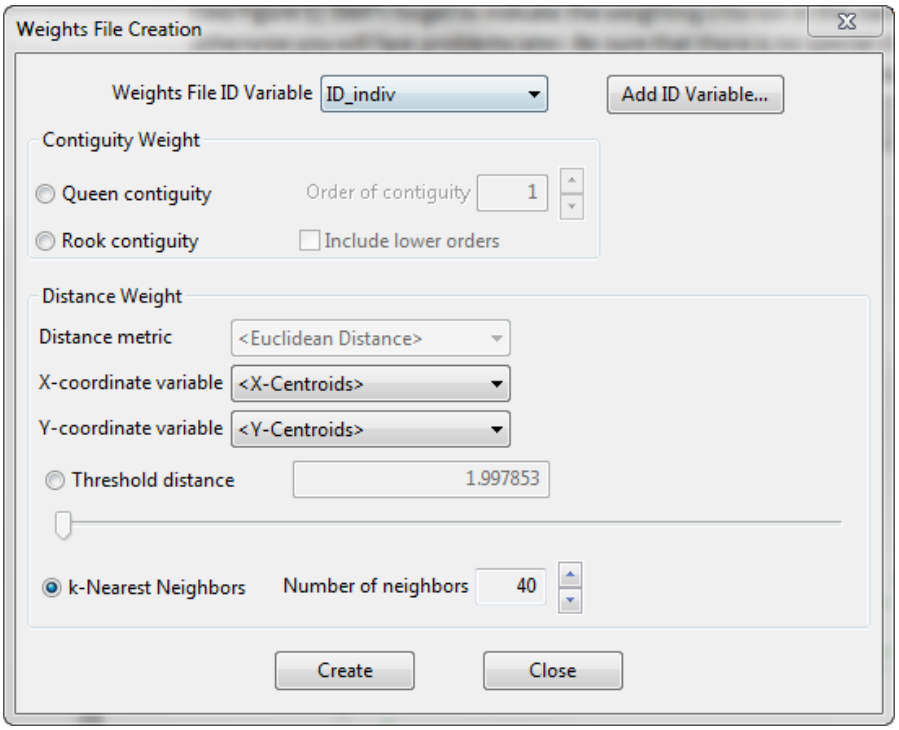
51. **Check your table for missing** values (often "?" for loci) and **replace them by nothing**.
52. **Convert** the **table** you created to a **shapefile in QGIS**.
53. Open **OpenGeoda** software. Click on the left icon in the tool bar to **open the shape file**. Make the map window larger.

We will first assess the global level of spatial autocorrelation of allele frequencies with different weighting criteria (or spatial lags). We first need to create weighting files in order to define the neighbourhood of each sampling point. In other words, with global spatial autocorrelation we want to compare the allele frequency at each sampling location with the mean allele frequency at sampling points included in a given neighbourhood.

Here we will use several numbers of neighbours (20, 40, 60 and 80) to manually build a spatial correlogram providing a Moran's I for each configuration.

54. Select **Tools > Weight > Create** to create the 4 corresponding weighting files. Use **Id_indiv as the ID variable**. Then select **k-Nearest neighbours** and indicate **20** in the corresponding field (see Figure 1). Don't forget to indicate the weighting criterion in the name of the file, otherwise you will face problems later. Be sure that there is no special character in the path to the folder where you save your weighting files, otherwise OpenGeoda won't work.
55. Repeat the task with **40, 60, 80 neighbours**.

A good idea is to open one of the weighting files to understand how it is structured. After the header, you have the list of the 20 nearest points of each of the 600 individuals in the data set, represented by their ID. The distance is indicated in decimal degrees. To calculate the global Moran's I, for each sampling point the software will calculate the mean frequency of a marker M among the 20 nearest neighbors listed in this file, and compare it to the real frequency. When processing a univariate regression of marker M's frequency on the weighted M's frequency, we obtain a slope and this slope is Moran's I (see Anselin's publications in the references for more details).



Now we will calculate the Moran's I for each weighting configuration, and for your selection of SNPs.

56. Select **Space > Univariate Moran's I** and select your marker, on the next window select a weighting file. You will obtain a **Moran scatterplot** showing standardized (centered and reduced) values of your marker frequencies vs weighted frequencies of your marker. Moran's I is an indicator of spatial autocorrelation over the whole area under study and for this particular criterion (20K) only.

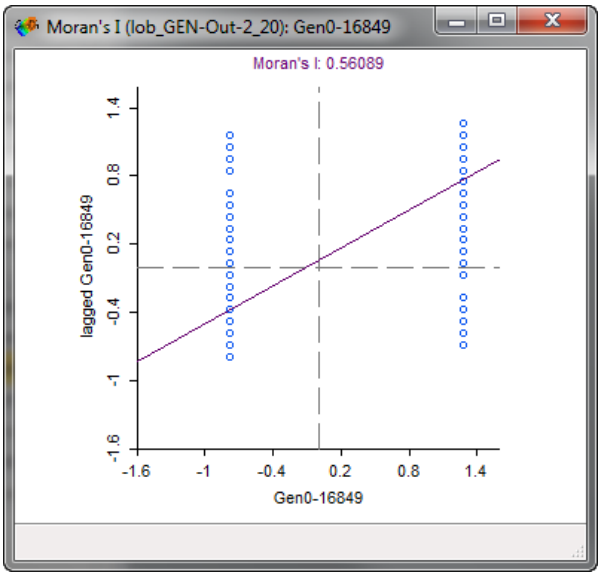


Figure 3: Moran's scatterplot for a marker and for K = 20 nearest neighbors . 0.56 is the slope and the value of Moran's I.

An important element to check is the significance of the global Moran's I obtained. The significance level is calculated on the basis of random permutations between all sampling points in the dataset, using Monte-Carlo simulations. The Moran's I obtained for the real observed geographic configuration is compared to the Moran's I calculated on the basis of many other spatial configurations obtained by means of n permutations. These permutations mean that the attribute values (the allele frequencies) at the sampling points are exchanged at random (e.g. allele frequency of sampling location 1 goes to sampling location 18, allele frequency of sampling location 4 goes to sampling location 14, etc.). The number of possible spatial configurations is the number of spatial objects factorial (n!). In our case it would be 622!.

57. To do this, **right-click on the Moran scattergram and select "Randomization" and then "999"** permutations. On the histogram you will observe that no simulated spatial configuration is larger than the Moran's I measured for the observed situation (the reality measured in the field). Read the explanations in the legend of Figure 4.

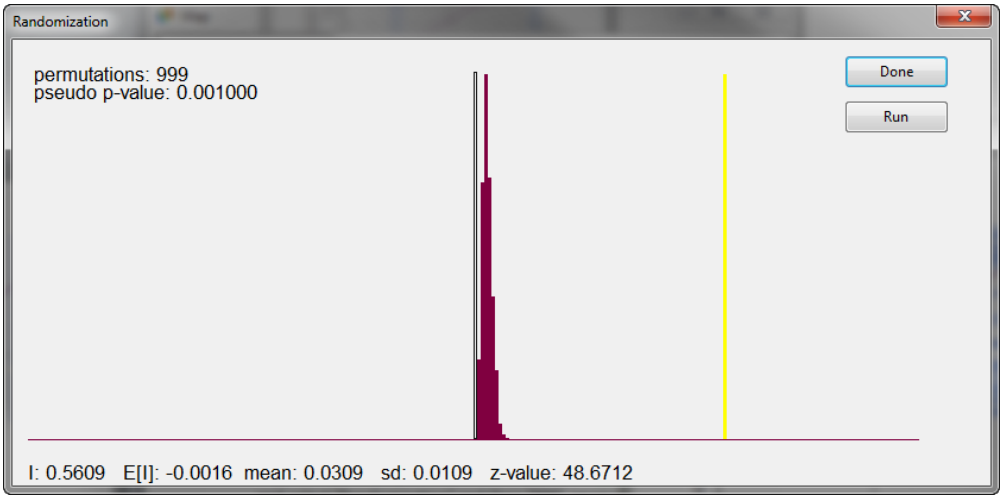


Figure: In the randomization process, the histogram shows the value of the observed Moran's I with the yellow bar (observed situation), for one loci with K = 20 nearest neighbours. Each time you press the "Run" button, the software runs another set of n permutations. In this case, we ask you to run 999 permutations. The histogram shows the statistical distribution of these 999 Moran's I calculated when the "Run" button was pressed (it means 999 out of 622! configurations). The histogram shows the frequency in Y, and classes of Moran's I in X. Here we can read that the yellow bar (observed Moran's I) = 0.56. The p-value is calculated as the number of simulated Moran's I being larger or equal to the observed Moran's I + 1 divided by the total number of random permutations + 1. Here we have 1/1000 = 0.001. This Moran's I is significant.

58. **Try several times to run 999 permutations**, and check the variation of the Index, to be sure that no spatial configuration will generate a Moran's I larger than the one corresponding to the observed situation.
59. Calculate **Moran's I for each selected marker and for each weighting configuration** and check the significance level. Then fill a table in Excel (as shown in Figure 7) with all Moran's I for you selected markers and build a spatial correlogram for each of them.

Moran's I							
	k=20	k=25	k=30	k=35	k=40	k=45	k=50
M1	0.2774						
M2	0.2295						
M3	0.0775						
Significance level for 999 permutations							
	k=20	k=25	k=30	k=35	k=40	k=45	k=50
M1	Yes, 0.001						
M2	Yes, 0.001						
M3	No, 0.013						

Figure: Table to fill in order to calculate spatial correlograms. For that purpose you will use the upper table named Moran's I. The role of the other table is to collect significance values and to mention if the global spatial autocorrelation measured is significant or not.

B. Univariate local spatial autocorrelation (LISA)

Choose one of your markers and calculate a local index of spatial association (LISA, see Anselin 1995 to read how these indices are calculated and to understand the details of the difference with global spatial autocorrelation).

60. Go to **Space>Univariate LISA**, then **select your loci, and a weighting file** (40 nearest neighbors). Then on the window tick all 4 options.

The significance map (Figure 10) shows the significant sampling locations in green, from dark green (most significant) to light green (less significant). Sampling locations in white are not significant. For the later points, it means that when carrying out random permutations (Monte-Carlo simulations), there is always at least one spatial configuration showing a highest Moran's I than the real situation.

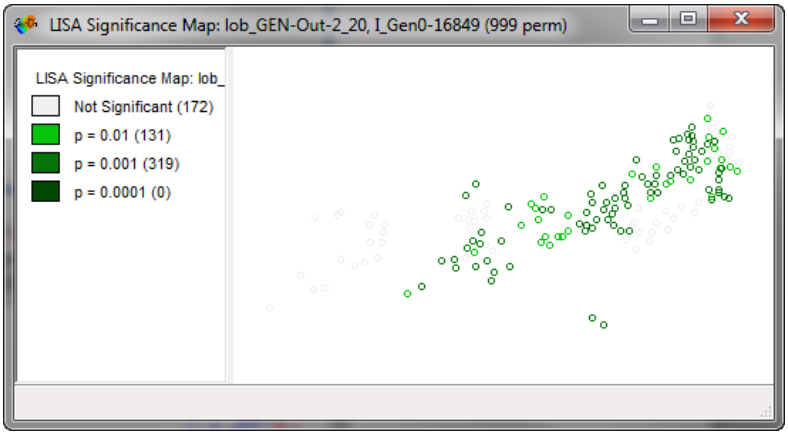


Figure: LISA significance map. The colour scale shows light to dark green sampling points. Dark green are the most significant sampling points. White sampling points (almost invisible here) are not significant (there is no spatial dependence, space is neutral).

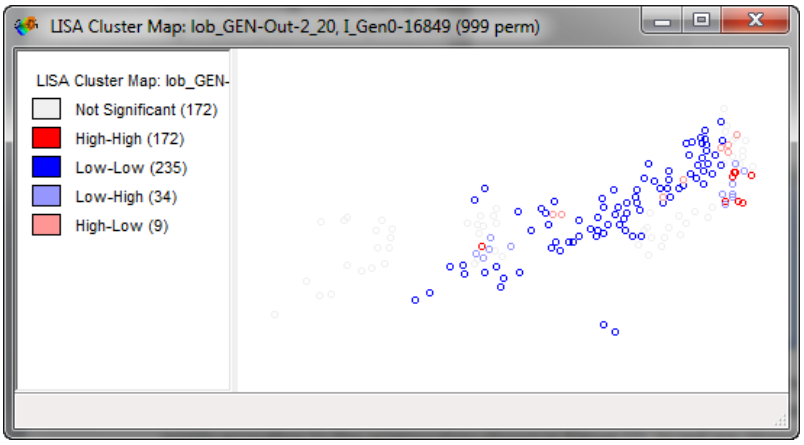
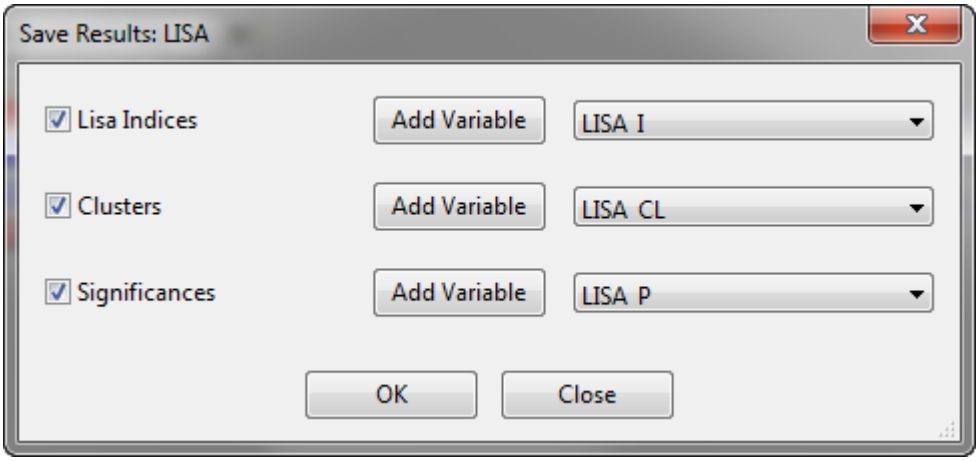


Figure: LISA cluster map. We see only significant sampling points as shown in previous Figure, where allele frequency depends on a space (spatial dependence). Here colors are attributed to sampling points according to the relationships shown in Figure 12. In red are shown High-High correlations: high allele frequency values correlated with high weighted allele frequency values located in the upper right square of the Moran's scattergram. Low-Low are also positive spatial autocorrelation. In light blue are shown low allele frequency values correlated with high weighted allele frequency values and the opposite for light red.

Bivariate LISA of the most relevant associations

Carry out a multivariate LISA analysis on the same marker.

61. In the menu **Space**, choose **Bivariate Local Moran's I**. Use one of the significant environmental variable, and **keep the 40 nearest neighbors** as weighting file. Provide a significance map, a cluster map and a Moran's scattergram. Bivariate LISA is a sort of local correlation coefficient between your marker and the environmental variable. It permits to analyse how the correlation varies in the landscape.
62. **Save the bivariate results by right-clicking on the clusters map and save the LISA variables**.



63. **Save your modified shapefile**. We will use this shapefile in QGIS to produce a final map of these results.
64. You may want to use visualization coordinates to see all LISA indices on a map. To do so you can **convert your LISA shapefile to a table in SAGA GIS or Quantum GIS** or **open the .dbf** of your LISA shapefile with excel, **change the coordinates** and **save it as a table**. Then again **open it in QGIS and convert it to shapefile.**

Creating a results map in QGIS

65. **Open your QGIS project and load your shapefile with LISA indices.**
66. Change the style of your data and make either **graduated symbols for LISA indices or Categories of LISA clusters**.

Exporting a map as a .pdf or .jpg requires using the print composer; it is not possible to directly include a legend and a scale bar in the map view.

However, it is easier to use the print composer if your map view contains the layers you want to export (activate the desired layers, superposition order, transparency, colours).

Click on the print composer and add a new print composer  .

67. Click to **add a new map** and drag it on the entire sheet. 

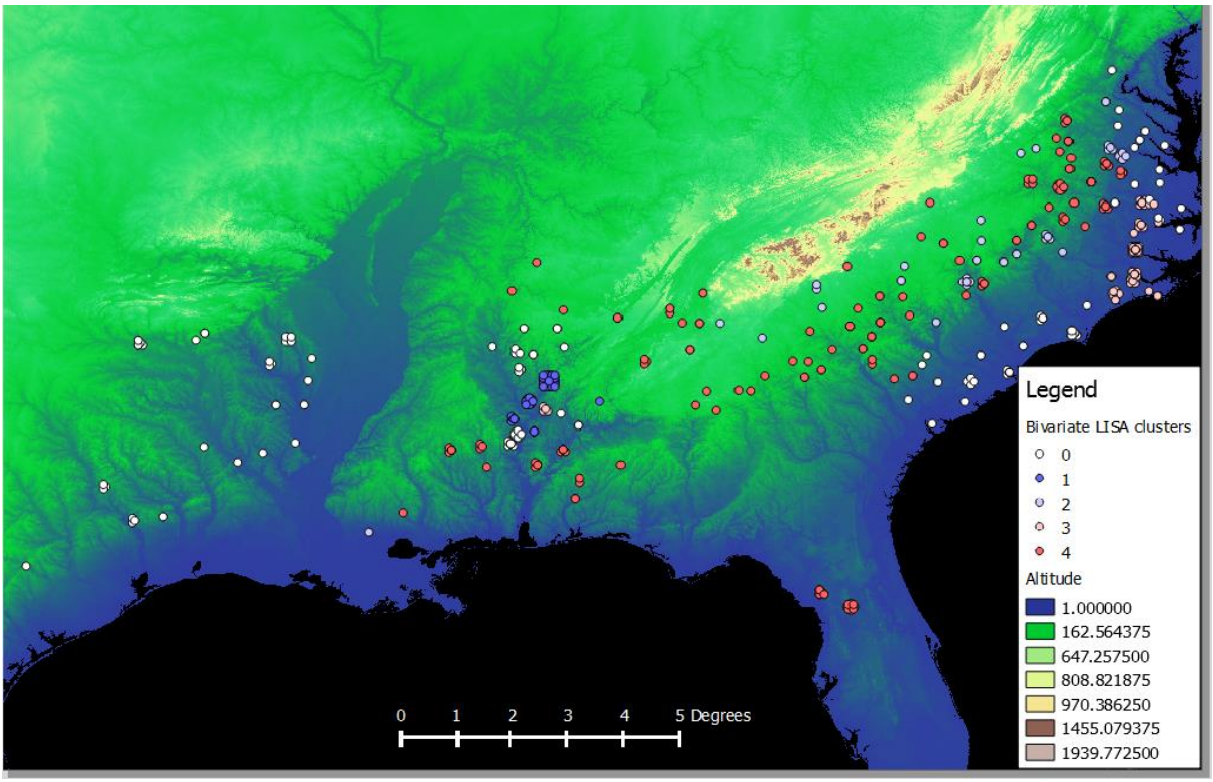
On the pannel on the right, you can change the scale, the scale and the extent of the map in object properties. You can also drag the content in the window with the **move item content** button  .

We will now add a legend and a scale bar to the map.

68. Click on **add a new legend**  then click on the map where you want to place it. Corners of the legend object allow you to resize it.

In the object properties of the legend you can change the elements that are shown in the legend, modify their names, play with the transparency, color and frame of the legend background.

69. To add a scale bar, click on **add a new scale**  . Againg you can modify all its propoerties in the object properties such as defining scale shape, its segments, font size etc. You can also suppress its background by deactivating its frame and putting transparency to zero.
70. You can now export your map in .pdf, .png or other formats.


Legend
Bivariate LISA clusters
0
1
2
3
4
Altitude
1.000000
162.564375
647.257500
808.821875
970.386250
1455.079375
1939.772500
0 1 2 3 4 5 Degrees