

Samβada software

Dr Stéphane Joost – Oliver Selmoni (Msc)

Laboratory of Geographic Information Systems (LASIG)

Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

LABORATORY OF GEOGRAPHIC INFORMATION SYSTEMS **LASIG**

[Home](#) [Research](#) [Teaching](#) [Services](#) [People](#) [Publications](#) [Links](#) **[Software](#)** [LSSR](#) [GIRAPH](#) [Open positions](#)

Share: [!\[\]\(3dfb8d66e81160ad61421a3452093d1b_img.jpg\)](#) [!\[\]\(21ece2018b00c7267b3324c50bbed633_img.jpg\)](#) [!\[\]\(074da87f0b7a74793bdf823413604aae_img.jpg\)](#) [!\[\]\(e3dcb983f6af01f6fe3b18e0a7169676_img.jpg\)](#) [!\[\]\(64236d586c7572d933ce39c4de709b6e_img.jpg\)](#)

Samβada

Samβada is an integrated software for landscape genomic analysis of large datasets. The key features are the study of local adaptation in relationship with environment and the measure of spatial autocorrelation in environmental and molecular datasets.

Releases

Samβada is available on GitHub

<https://github.com/Sylvie/sambada>

The latest release is

release **v0.7.0**

<https://github.com/Sylvie/sambada/releases>

References

- Stucki, S., Orozco-terWengel, P., Forester, B. R., Duruz, S., Colli, L., Masembe, C., ... Joost, S. (2017). High performance computation of landscape genomic models including local indicators of spatial association. Molecular Ecology Resources, 17(5), 1072–1089. doi:10.1111/1755-0998.12629
- Stucki, S. (2014) Développement d'outils de géo-calcul haute performance pour l'identification de régions du génome potentiellement soumises à la sélection naturelle: analyse spatiale de la diversité de panels de polymorphismes nucléotidiques à haute densité (800k) chez Bos taurus et B. indicus en Ouganda, EPFL PhD Thesis no 6014, doi:10.5075/epfl-thesis-6014

Samβada

Pic2Map

KEYWORDS

GIS, Spatial Analysis, Decision-making support, Exploratory Spatial Data Analysis - Landscape genetics, Landscape genomics - Geomedicine, Spatial epidemiology

CONTACT

Secretary

EPFL ENAC IIE LASIG

Batiment GC

Vers le plan d'orientation GC D2 397

Station 18

CH-1015 Lausanne

Tel: +41 21 693 27 55

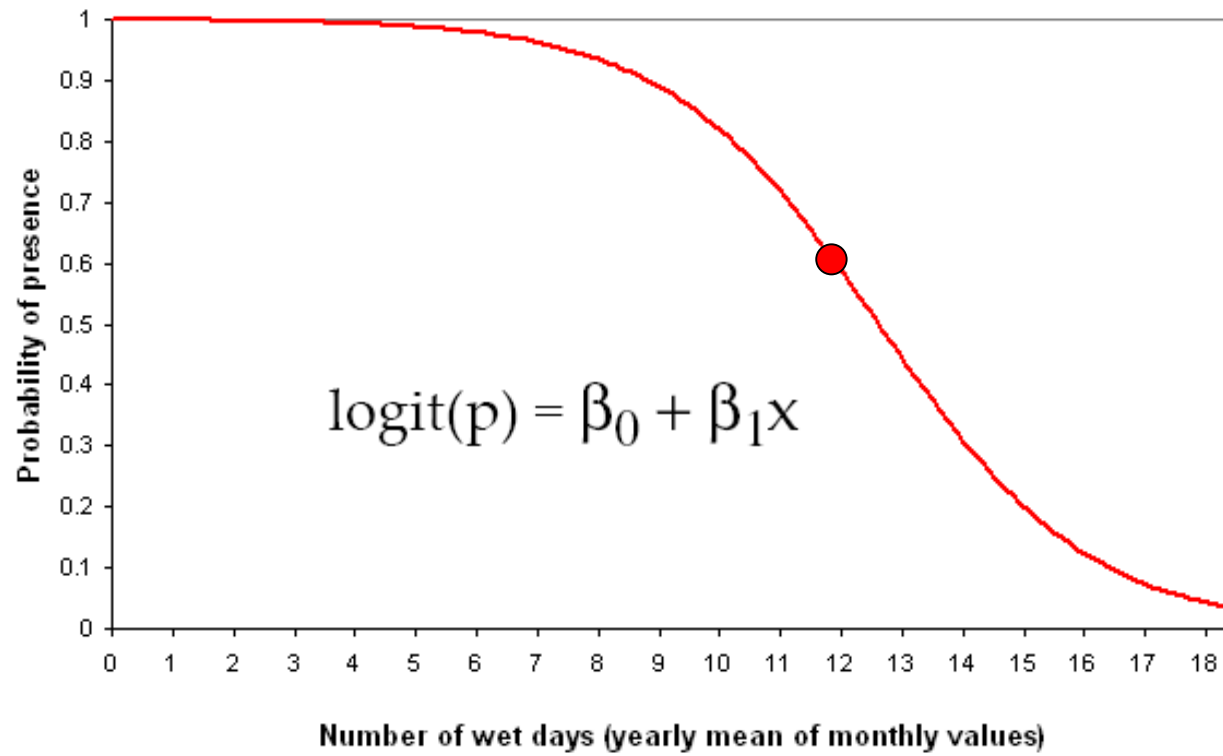
Fax: +41 21 693 57 90

Logistic regression

Individuals			Genetic markers															Environmental variables							
1	farmid	animalid	OARJMP29_allele2_137	OARJMP29_allele2_138	OARJMP29_allele2_140	OARJMP29_allele2_141	OARJMP29_allele2_142	OARJMP29_allele2_143	OARJMP29_allele2_144	OARJMP29_allele2_145	OARJMP29_allele2_146	OARJMP29_allele2_147	OARJMP29_allele2_148	OARJMP29_allele2_149	OARJMP29_allele2_150	OARJMP29_allele2_151	OARJMP29_allele2_152	OARJMP29_allele2_153	OARJMP29_allele2_154	OARJMP29_allele2_155	wndjan	altitude	wndfeb	wndmar	wndapr
1044	PL-4005	OAPLPOM25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.1	22	4.6	5	4.4
1045	PL-4005	OAPLPOM26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.1	22	4.6	5	4.4
1046	PL-4006	OAPLPOM01	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	5.3	153	4.8	4.9	4.3
1047	PL-4006	OAPLPOM15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.3	153	4.8	4.9	4.3
1048	PL-4006	OAPLPOM24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	5.3	153	4.8	4.9	4.3
1049	PL-4007	OAPLPOM05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.3	250	4.8	5	4.5
1050	PL-4007	OAPLPOM16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	5.3	250	4.8	5	4.5
1051	PL-4008	OAPLPOM09	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	5.2	166	4.8	5	4.4
1052	PL-4008	OAPLPOM19	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.2	166	4.8	5	4.4
1053	PL-4008	OAPLPOM20	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	5.2	166	4.8	5	4.4
1054	PL-4009	OAPLPOM10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.5	87	5	5.2	4.6
1055	PL-4009	OAPLPOM21	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.5	87	5	5.2	4.6
1056	PL-4010	OAPLPOM08	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	5.4	208	4.9	5.1	4.5

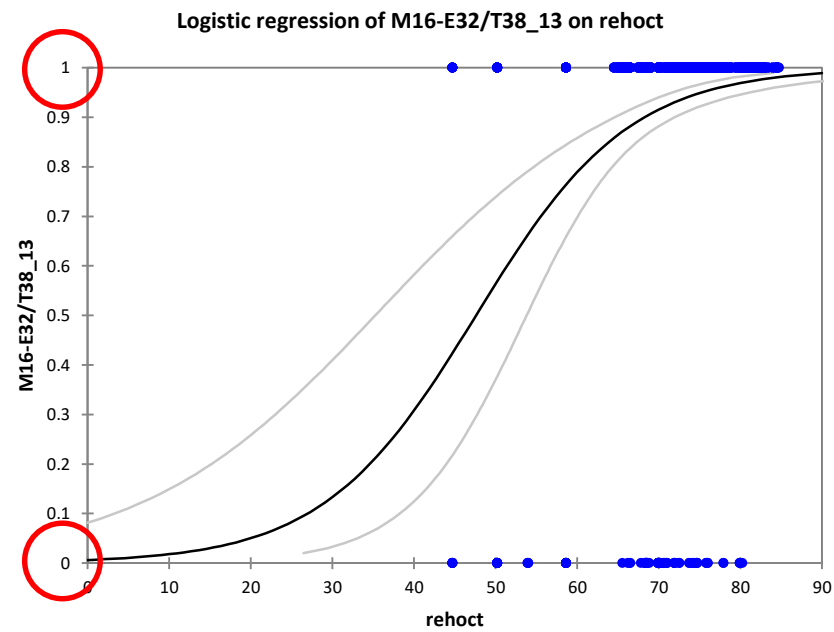
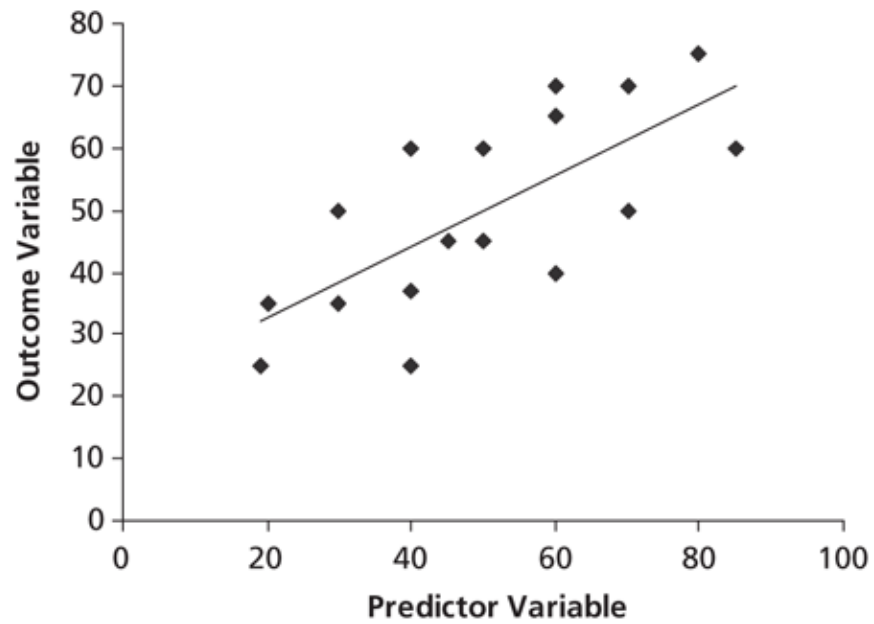
Multiple parallel logistic regressions

Sheep breeds : allele OARJMP29_157



Nominal variable

- We use simple logistic regression when we have one nominal variable with two values (male/female, dead/alive, existing/non-existing) and one measurement variable
- The nominal variable is the dependent variable, and the measurement variable is the independent variable



Probability

- Simple logistic regression is analogous to linear regression, except that the dependent variable is nominal, not a measurement
- One goal is to see whether the probability of getting a particular value of the nominal variable is associated with the measurement variable
- The other goal is to predict the probability of getting a particular value of the nominal variable, given the measurement variable

Expected value

- In any regression, a key parameter is the conditional mean $E(Y|x)$ (*“the expected value of Y , given x ”*)
- This is the expected value of the Y variable given the value of the independent x variable
- In a linear regression, this quantity is expressed as: $E(Y|x) = \beta_0 + \beta_1 x$
- This expression implies that $E(Y|x)$ may take on any value between $-\infty$ and $+\infty$

Bernoulli distribution

- In the case of a binary variable the result of observations is either a «success» or a «failure» (1 or 0, Bernoulli distribution)
- $E(Y|x)$ is expressing a probability. The probability of success is $p = P(Y=1)$
- The probability of failure is $P(Y=0) = 1-p$
- Whatever the value of x , the expected value will range between 0 and 1

Inadmissible values

- It is not possible to use standard linear regression to calculate a function expressing a relationship between a binomial dependent variable and a quantitative independent variable
- The first reason is that the predicted values will become greater than 1 and less than 0 when moving far enough on the x-axis (x ranges between $-\infty$ and $+\infty$) and such values are theoretically inadmissible

A range between 0 and 1

- Then one assumption of regression is that the variance of Y is constant across values of x what cannot be the case with a binary variable
- Finally, the error is not normally distributed as Y only takes «0» and «1» values
- We need to find a function that relates the independent variable x to the rolling mean of the bivariate dependent variable $P(\hat{Y})$
- ... and which limits predicted values in a range between 0 and 1

Null hypothesis

- The null hypothesis is that the probability of a particular value of the nominal variable is not associated with the value of the measurement variable
- In other words, the curve describing the relationship between the measurement variable and the probability of the nominal variable has a slope of zero


Odds

- In our case, the Y variable is the probability for a maker to exist for given values of environmental variables
- This probability may take values from 0 to 1
- The limited range of this probability would present problems if used directly in a linear regression
- So instead we use the odds

= the likelihood that the event will take place

Calculating odds

- The odds = $\frac{p}{1-p}$
- If the observed probability of marker M1 to be present is 0.6
- the odds of M1 is
$$\frac{0.6}{(1 - 0.6)} = 1.5$$
- This can be expressed as
“1.5 to 1” odds for marker M1 to be present



ID	M1	EnvVar
1	1	16
2	1	16
3	1	20
4	0	16
5	1	17
6	0	13
7	0	12
8	1	16
9	1	18
10	0	10

Natural log

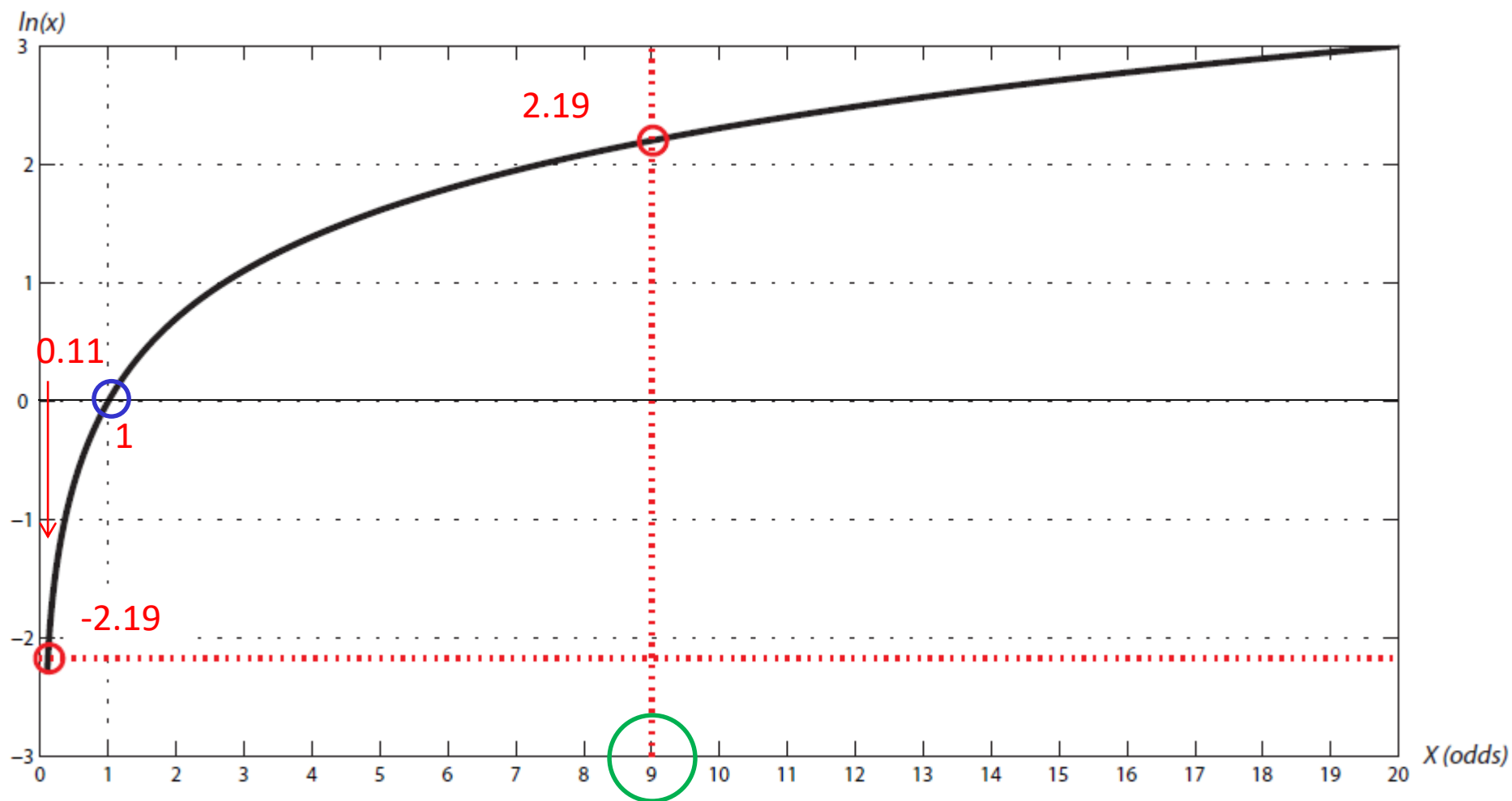
- Odds = $\beta_0 + \beta_1 x$
- Taking the natural log of the odds makes the variable more suitable for a regression, so the result of a logistic regression is an equation that looks like :
- $\ln \frac{p}{(1-p)} = \beta_0 + \beta_1 x$

Why natural log ?

- Let us consider a probability of 0.9 for the marker to exist
- $\frac{0.9}{(1 - 0.9)} = 9$, this is an odds of 9 to 1
- Now, the odds for the marker of not existing
- $\frac{0.1}{(1 - 0.1)} = 0.11$
- It should be the opposite odds, what the value of 0.11 does not express compared to 9

Properties of the natural logarithm

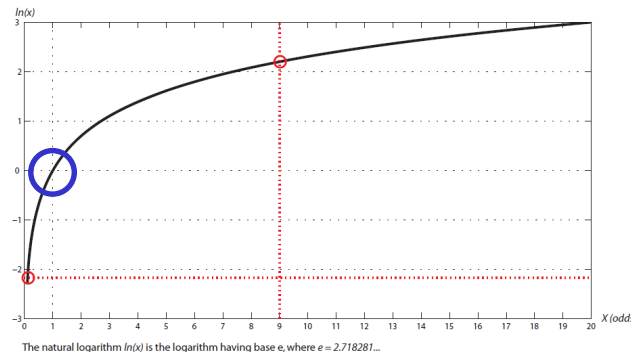
- This is where the properties of the natural logarithm are used, to express this asymmetry
- Indeed, $\ln(9) = 2.19$ and $\ln(0.11) = -2.19$
- It means that the \ln odds for our marker to exist for a given value of x is exactly opposite to the \ln odds of not existing



The natural logarithm $\ln(x)$ is the logarithm having base e , where $e = 2.718281...$

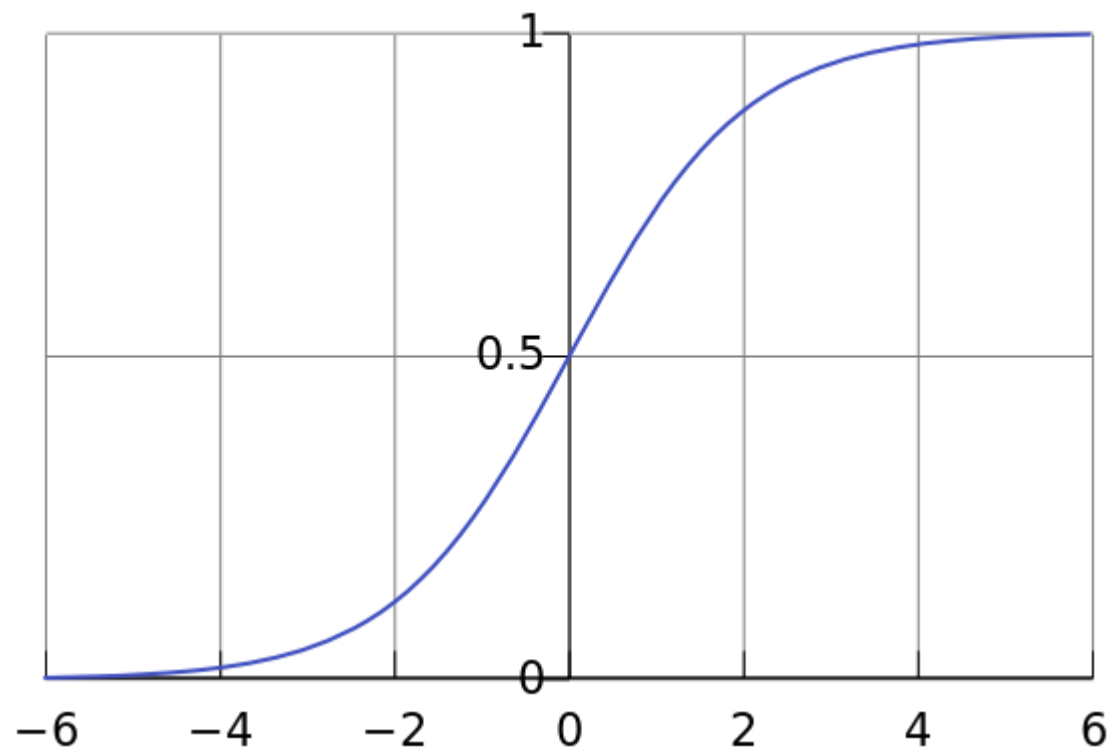
Observations

- the natural logarithm is zero when x is 1
- when x is larger than 1, the natural logarithm curves up slowly
- when x is less than 1, the natural logarithm is less than 0, and decreases rapidly (vertical asymptote) as x approaches 0



Consequences

- if $p = 0.5$, the odds are $\frac{0.5}{0.5} = 1$,
and $\ln(1) = 0$
- if $p > 0.5$, $\ln \left(\frac{p}{1-p} \right)$ is positive
- if $p < 0.5$, $\ln \left(\frac{p}{1-p} \right)$ is negative



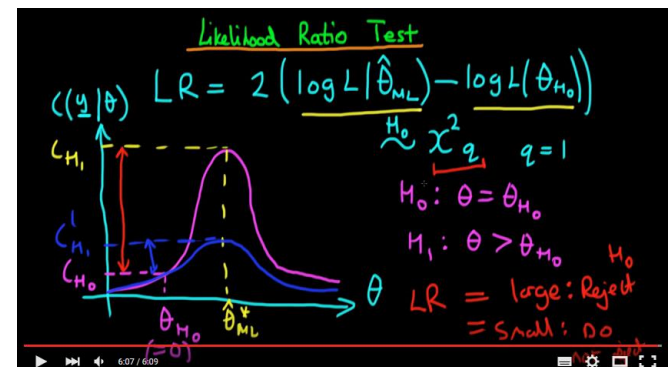
Significance of the models

- The likelihood ratio (G) uses the difference between the probability of obtaining the observed results under the logistic model ...
- ... and the probability of obtaining the observed results in a model with no relationship between the independent and dependent variables

https://www.youtube.com/watch?v=Tn5y2i_MqQ8

- Wald test:

<https://www.youtube.com/watch?v=TFKbyXAfr1M>



Multiple comparisons

- The multiple comparisons or multiple testing problem occurs when one considers a set of statistical inferences simultaneously
- The more inferences are made, the more likely erroneous inferences are to occur.
- If multiple hypotheses are tested, the chance of a rare event increases and the likelihood of incorrectly rejecting a null hypothesis increases (false positives)

Bonferroni

- The Bonferroni correction compensates for that increase by testing each individual hypothesis at a significance level of α/m
- Where α is the desired overall alpha level and m is the number of hypotheses
- E.g. if a trial is testing $m = 20$ hypotheses with a desired $\alpha = 0.05$, then the Bonferroni correction would test each individual hypothesis at $\alpha = 0.05 / 20 = 0.0025$

False Discovery Rate (FDR)

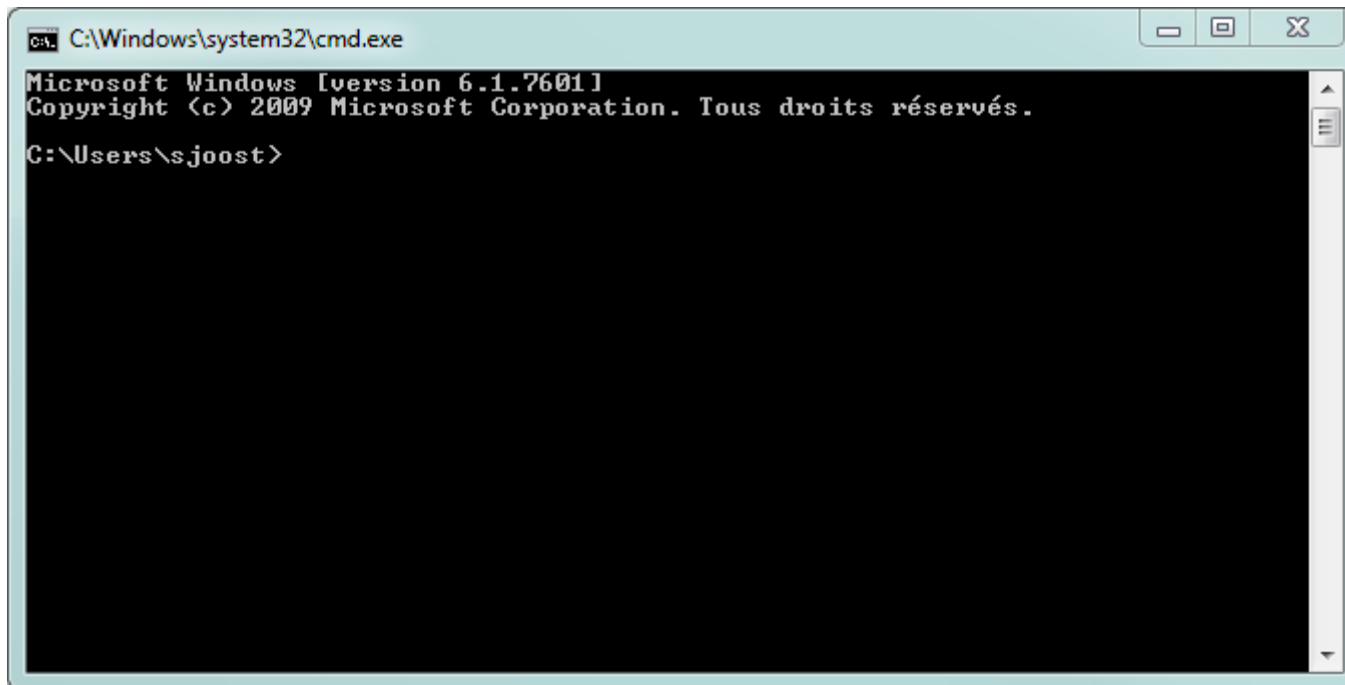
- False discovery rate (FDR) is a method of addressing the rate of type I errors (false positives) in null hypothesis testing when conducting multiple comparisons
- FDR-controlling procedures provide less stringent control of Type I errors than Bonferroni

Q values

- Based on p-values
- The minimum false discovery rate at which the test may be significant
- FDR procedures by Benjamini, Hochberg, Storey and Tibshirani, and others...

Sambada

- Run Sambada from a shell
- E.g. Windows command console
- cmd in the start menu



Three programs are available:

Samβada processes univariate and multivariate logistic models for the landscape genomics analysis and optionally measures the spatial autocorrelation in environmental and molecular datasets;

Supervision can split molecular data in blocks in order to run the analysis on several computers, and can merge the results afterwards;

RecodePLINK can translate molecular data from PLINK's to Samβada's format.

The user must provide **Samβada** with a parameter file to set up the analysis as well as environmental and molecular data. The workflow is summarised on fig. 2 and the data format is presented in the next section.

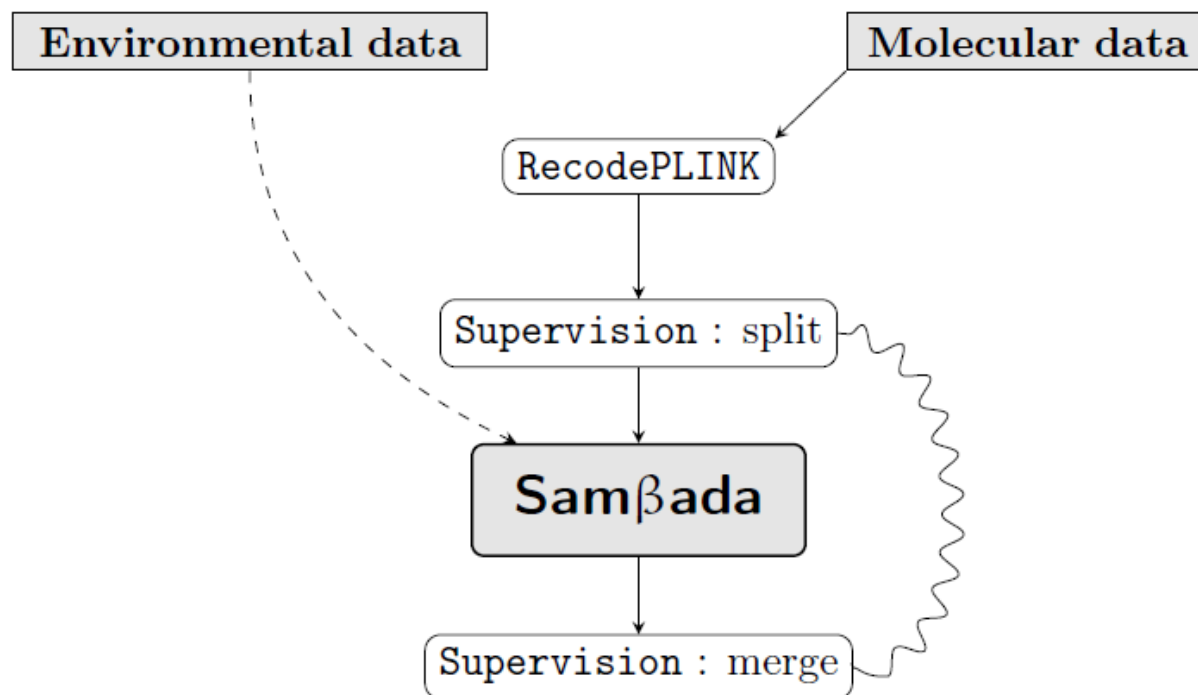


Figure 2 – Workflow of analysis. Rectangles stand for data and round-cornered figures stand for programs. Grey elements are mandatory and white ones are optional. **Samβada** computes correlative models and spatial autocorrelation. The two other features are optional: **Supervision** enables distributed computing while **RecodePLINK** transforms .ped/.map files to comply with **Samβada**’s format. Arrows show processing order; **Samβada** input consists in environmental data (dashed line) and molecular data (solid line). The zigzag line indicates that **Supervision** is used before and after the main analysis.

Data format

- Sambada's input consist of molecular and environmental data
- They can be provided as a single or two separate files
- Files may have any name and extension
- Each line provides information for an individual
- Each column contains an environmental variable or a binary molecular marker
- Information about data format and analysis
- Options are specified separately in the **parameter file**

Data organization

- The header line is optional
- The column separator is up to the user
- Sample names (identifiers) are optional,
- Some columns may be excluded from the analysis (for instance phenotypical information stored with the environmental data)
- If there is a single data file, environmental data must be provided in the first columns, and molecular data in the last ones
- Sample names and geographic coordinates are considered as environmental data
- If data is split between two files, samples must be in the same order in both files
- Missing data can be coded as any character **string**, for instance NaN or ?

Examples

NAME	ENV1	ENV2	ENV3	ENV4	ENV5
ID1	46	972	236	230	132
ID2	32	987	238	232	83
ID3	32	987	238	232	83
ID4	32	987	NaN	232	83
ID5	32	987	238	232	83
ID6	35	1021	235	230	87

Figure 3 – Example of environmental file (`env-data.txt`).

NAME	M4	M7	M8	M9	M16	M17	M18
ID1	1	1	1	0	1	1	1
ID2	0	0	0	0	0	0	NaN
ID3	0	1	0	0	0	0	1
ID4	0	0	1	1	0	1	1
ID5	0	0	1	0	0	1	0
ID6	0	1	0	0	0	1	0

Figure 4 – Example of molecular file (`mol-data.txt`).

Example

NAME	ENV1	ENV2	ENV3	ENV4	ENV5	M4	M7	M8	M9	M16	M17	M18
ID1	46	972	236	230	132	1	1	1	0	1	1	1
ID2	32	987	238	232	83	0	0	0	0	0	0	NaN
ID3	32	987	238	232	83	0	1	0	0	0	0	1
ID4	32	987	NaN	232	83	0	0	1	1	0	1	1
ID5	32	987	238	232	83	0	0	1	0	0	1	0
ID6	35	1021	235	230	87	0	1	0	0	0	1	0

Figure 5 – Example of combined file for environmental and molecular data, corresponding to fig. 3 and 4 (combo-data.txt). Environmental data must be provided in the first columns (left part of the tabular), and molecular data in the last columns (right part). The identifier, if any, is considered as an environmental variable.

Parameter file

- The parameter file contains one line per parameter
- Parameters can be specified in any order
- Each line begins with the name of the current parameter followed by the values separated by **spaces**
- Some parameters are mandatory
- Any line beginning with a hash character (#) will be ignored

```

HEADERS YES
WORDDELIM " "
* NUMVARENV 24
* NUMMARK 120103
* NUMINDIV 804
IDINDIV short_name ID_indiv
SPATIAL longitude latitude SPHERICAL NEAREST 20
AUTOCORR BOTH MARK 1000
* DIMMAX 1
* SAVETYPE END BEST 0.01

```

Figure 6 – Example of a parameter file for setting up Samβada’s analysis. Each line contains an option for the computation, those marked with a sign in the margin are mandatory. The line order has no influence. In this example, the two first lines indicate that data files contain a header line and that columns are separated by spaces. The next lines state the number of environmental variables, the number of molecular markers and the number of individuals/samples. The option IDINDIV indicates which columns contain identifiers of individuals; here environmental and molecular data are recorded in two separated files. The next two lines address the measure of the spatial autocorrelation, with the coordinates names, which are spherical, the weighting scheme and the bandwidth; here the 20 nearest neighbours are taken into account. The analysis will include both global and local autocorrelation (BOTH) of molecular markers (MARK) and the significance will be assessed with 1,000 permutations. The next option means that the detection of selection signatures will rely on univariate models (DIMMAX 1). The last line indicates that results will be stored at the end of the process, that only significant models with a significant parent will be stored and that the threshold for significance is set to 1% (before Bonferroni’s correction).

```

HEADERS YES
NUMVARENV 6
NUMMARK 8
NUMINDIV 6
IDINDIV NAME
DIMMAX 1
SAVETYPE END ALL 0.01

```



*Parameter file for
separate env and
molecular data*

```

HEADERS YES
NUMVARENV 6
NUMMARK 7
NUMINDIV 6
IDINDIV NAME
DIMMAX 1
SAVETYPE END ALL 0.01

```



*Parameter file for
a unique matrix
with env and
molecular data*

Program launch

- The command changes slightly if there are two separated input files
 - *Sambada parameterFile envFile molecularFile*
`Sambada param.txt env-data.txt mol-data.txt`
- Sambada is launched as follows if environmental and molecular data are stored in the same file
 - *Sambada parameterFile dataFile*
`Sambada param-combo.txt combo-data.txt`

Note on program's names

- The actual names of the compiled programs also contain the version number and the name of the target operating system
- When running the commands, please use the names of the programs found in the directory “Binaries”
- For instance:
- Sambada → `Sambada-v0.5.3-Win64.exe`
- Supervision → `Supervision-v0.5.3-OSX`
- RecodePLINK → `RecodePlink-v0.5.3-Ubuntu`
- Hint: Most terminal/console interpreters enable auto-completion of names by hitting the “TAB” key.

Functions

- The first line shows the parameter name, whether it is mandatory, the list of possible values (or the expected type) and the default value. The paragraph is completed by a description of the option.

Data files and format

INPUTFILE	Optional	(string)	-
-----------	----------	----------	---

Name(s) of the data file(s). If there are two files, indicate first the environmental file then the molecular file.
This information may also be given as an argument to the program.

OUTPUTFILE	Optional	(string)	-
	Base name(s) for the results file(s). If this option is omitted, the output files will be named after the molecular input file. The different output files are distinguished by adding suffixes (“-Out-”, “-AS-”, ...), thus the input files are untouched. With this option, the results can be saved in a different folder than the data.		
HEADERS	Optional	Yes / No	No
	Presence or absence of variable names. If present, they are read on the first line of the data file, otherwise the environmental variables are labelled P1, P2, P3,... and molecular markers are labelled M1, M2, M3,....		
WORDDELIM	Optional	(char)	‘ ’
	Word delimiter, it must be a single character. This option applies to both molecular and environmental data, while the parameter file is assumed to be space-separated.		
LOG	Optional	<i>1 value, see descr.</i>	BOTH
	Location of log information.		
	1 {	TERMINAL	print the log on the standard output,
		CONSOLE	
		FILE	writes the log in a file with the suffix “-log”,
		BOTH	uses both methods.

Data size

NUMVARENV	Mandatory	(int)	-
Number of columns with environmental variables, including ignored variables and the column of identifiers (if any). If there is one input file, this counts the number of columns that don't concern molecular data ² . If there are two input files, NUMVARENV counts the total number of columns in the environmental data file.			
NUMMARK	Mandatory	(int)	-
Number of columns with molecular data, including ignored data and identifiers if applicable. If there is one input file, this counts the number of columns concerning molecular markers ² . If there are two input files, NUMMARK counts the total number of columns in the molecular data file. For distributed analysis, this parameter must indicate the number of markers for the current block of data followed by the total number of markers ³ .			
NUMINDIV	Mandatory	(int)	-
Number of samples included in the data file(s).			

Active and inactive columns

IDINDIV	Optional	(string or int)	-
Name(s) of the column(s) containing sample identifiers ⁴ . These optional identifiers are used to label samples in the output files for local spatial autocorrelation (otherwise the line numbers are used). If there are two data files, two names (or numbers) can be provided, the first one is for the environmental data and the second one is for molecular data. The identifier columns are automatically set as inactive. Moreover if this option is specified with two data files, the two identifiers must match on each line. (Sample must be in the same order in each file.)			
COLSUPENV	Optional	(string or int)	-
Name(s) of the column(s) in the environmental data to be excluded from the analysis ⁴ . These columns are set as inactive. (For instance, COLSUPENV can indicate columns such as the sampling date or the name of the area.)			
COLSUPMARK	Optional	(string or int)	-
Name(s) of the column(s) in the molecular data to be excluded from the analysis ⁴ . These columns are set as inactive.			

SUBSETVARENV Optional (string or int) -

Name(s) of the column(s) in the environmental data to be included in the analysis while the other columns are set as inactive. The different options cumulate: the active columns are those listed here, minus those specified with COLSUPENV as well as IDINDIV.

SUBSETMARK Optional (string or int) -

Name(s) of the column(s) in the molecular data to be included in the analysis while the other columns are set as inactive. The different options cumulate: the active columns are those listed here, minus those specified with COLSUPMARK as well as IDINDIV.

Logistic model and results storage.

DIMMAX	Mandatory	(int)	-
	Maximum number of environmental variables included in the logistic models. The models with less parameters are computed as well. Use 1 for univariate models, 2 for univariate and bivariate models, ... Please refer to section 5.1.5 for more information on including prior knowledge in multivariate models.		
SAVETYPE	Mandatory	3 values, see descr.	-
	Saving method and model selection.		
	1 $\left\{ \begin{array}{l} \text{REAL} \\ \text{END} \end{array} \right.$	Storage mode: REAL saves results during processing, END writes them upon completion of computation. The second option enables sorting the models according to their Wald scores before saving.	
	2 $\left\{ \begin{array}{l} \text{ALL} \\ \text{SIGNIF} \\ \text{BEST} \end{array} \right.$	Model selection: ALL saves all models, SIGNIF saves significant models (according to the G and Wald scores) and BEST saves significant models with at least a significant parent.	
	3 $\left\{ (\text{double}) \right.$	Significance threshold (p -value) for options SIGNIF and BEST. The Bonferroni correction is applied on this threshold.	

Example: SAVETYPE END BEST 0.01

UNCONVERGEDMODELS Optional

Yes / No

No

This option controls the back-up of unconverged models. If enabled, these models are saved in a separate file with the suffix “-unconvergedModels”.

Spatial autocorrelation

SPATIAL	Optional	5 values, see descr.	-
1	{ (string or int)	Column name (or number) for longitude.	
2	{ (string or int)	Column name (or number) for latitude.	
3	{ SPHERICAL CARTESIAN	Type of coordinates (spherical or projected).	
4	{ DISTANCE GAUSSIAN BISQUARE NEAREST	Type of weighting scheme, see fig. 9	
5	{ (double or int)	Bandwidth of the weighting function • Cases DISTANCE, GAUSSIAN, BISQUARE: Input type is (double). Units are in [m] for SPHERICAL coordinates; for CARTESIAN coordinates, units match those of the samples' positions. • Case NEAREST: Input type is (int).	

Example: SPATIAL X Y CARTESIAN BISQUARE 120

AUTOCORR	Optional	<i>3 values, see descr.</i>	-
	This entry requires the specification of SPATIAL.		
1	{ GLOBAL LOCAL BOTH	Type of indices to compute: Moran's I for the global spatial autocorrelation, LISA for the local one.	
2	{ ENV MARK BOTH	Variables for the analysis.	
3	{ (int)	Number of permutations for computing the pseudo p -values (default=99).	

Example: AUTOCORR GLOBAL BOTH 999

SHAPEFILE	Optional	YES / NO	NO
-----------	----------	----------	----

With this option, the LISA are saved as a shapefile (in addition to the usual output). This format is composed of three files: .shp, .shx and .bdf. These files can be loaded together in any GIS software to map the local autocorrelation. This entry requires the specification of SPATIAL.

Output

- Sambada produces several output files. To illustrate the naming scheme, let us assume that the molecular data file is named [data.ext](#)
- If the log is saved for future reference, the corresponding file is named [data-log.ext](#)
- For logistic regressions, there is one file for constant models, which are not sorted
- There is also one file per distinct number of parameters (univariate, bivariate, trivariate models and so on)
- In these files, models are sorted according to their Wald scores
- Results files are named as follows: constant models are saved in the file [data-Out-0.ext](#), univariate models in the file [data-Out-1.ext](#), bivariate models the file [data-Out-2.ext](#), etc.

Constant models

Marker	Loglikelihood	AverageProb	Beta_0	NumError
Hapmap43437-BTA-101873_AA	-228.2100569	0.082089552	-2.414289083	0
Hapmap43437-BTA-101873_AG	-542.450042	0.404228856	-0.387875415	0
Hapmap43437-BTA-101873_GG	-556.9893006	0.513681592	0.054740033	0
ARS-BFGL-NGS-16466_AA	-44.84132815	0.009950249	-4.600157644	0
ARS-BFGL-NGS-16466_AG	-389.8189189	0.189054726	-1.456164041	0
ARS-BFGL-NGS-16466_GG	-401.2120224	0.800995025	1.392524911	0
Hapmap34944-BES1_Contig627_1906_AA	-456.4590694	0.254975124	-1.072251619	0
Hapmap34944-BES1_Contig627_1906_AC	-555.856645	0.470149254	-0.119545151	0
Hapmap34944-BES1_Contig627_1906_CC	-472.7907257	0.274875622	-0.970024485	0

Figure 11 – Exemple of Samβada's results for constant models, there is one marker per line. The first column is the name of the molecular marker, here the locus name combined with the allele name. The following columns are the log-likelihood, the frequency of the marker, the estimate of parameter β_0 for the logistic model and the error code (0 if success). Constant models are not sorted and thus are in the same order as the markers in the input file. When considered markers are SNPs like here, there are three binary markers per locus.

Results for univariate models

Marker	Env_1	Loglikelihood	Gscore	WaldScore	NumError	Efron	McFadden	McFaddenAdj	CoxSnell	Nagelkerke	AIC	BIC	Beta_0	Beta_1
Hapmap41074-BTA-73520_AA	prec7	-443.11	208.53	151.72	0	0.25	0.19	0.19	0.23	0.10	890.22	912.98	-2.04	0.03
ARS-BFGL-NGS-113888_GG	prec7	-441.73	208.67	151.70	0	0.25	0.19	0.19	0.23	0.10	887.47	910.23	-2.02	0.03
Hapmap41762-BTA-117570_GG	prec7	-435.96	202.93	148.43	0	0.24	0.19	0.19	0.22	0.10	875.92	898.68	-1.86	0.03
ARS-BFGL-NGS-46098_GG	prec7	-440.04	200.82	147.60	0	0.24	0.19	0.18	0.22	0.10	884.07	906.83	-1.88	0.03
ARS-BFGL-NGS-113888_GG	latitude	-449.13	193.89	146.89	0	0.23	0.18	0.17	0.21	0.09	902.25	925.01	-0.73	0.86
Hapmap41074-BTA-73520_AA	latitude	-450.81	193.13	146.61	0	0.23	0.18	0.17	0.21	0.09	905.62	928.38	-0.75	0.85
Hapmap41762-BTA-117570_GG	latitude	-444.40	186.04	141.99	0	0.21	0.17	0.17	0.21	0.09	892.80	915.56	-0.57	0.84
ARS-BFGL-NGS-113888_GG	prec6	-455.48	181.19	138.85	0	0.21	0.17	0.16	0.20	0.09	914.95	937.71	-2.22	0.03
Hapmap41074-BTA-73520_AA	prec6	-457.38	179.99	138.13	0	0.21	0.16	0.16	0.20	0.09	918.77	941.53	-2.23	0.03
ARS-BFGL-NGS-46098_GG	latitude	-451.22	178.45	138.11	0	0.21	0.17	0.16	0.20	0.09	906.44	929.20	-0.59	0.82
Hapmap41813-BTA-27442_AA	prec7	-462.30	179.89	137.52	0	0.22	0.16	0.16	0.20	0.08	928.60	951.36	-1.92	0.03
ARS-BFGL-NGS-46098_GG	prec6	-451.51	177.87	137.27	0	0.21	0.16	0.16	0.20	0.09	907.03	929.78	-2.11	0.03
BTA-73516-no-rs_AA	prec7	-460.18	177.43	136.04	0	0.21	0.16	0.16	0.20	0.08	924.35	947.11	-1.83	0.03
Hapmap41813-BTA-27442_AA	latitude	-469.89	164.71	130.98	0	0.20	0.15	0.15	0.19	0.08	943.77	966.53	-0.76	0.76
Hapmap41762-BTA-117570_GG	prec6	-454.17	166.51	130.97	0	0.20	0.15	0.15	0.19	0.08	912.33	935.09	-1.96	0.03
ARS-BFGL-NGS-46098_GG	longitude	-458.86	163.18	130.95	0	0.18	0.15	0.15	0.18	0.08	921.72	944.48	-23.95	0.76
Hapmap41074-BTA-73520_AA	bio7	-457.07	180.61	129.73	0	0.21	0.16	0.16	0.20	0.09	918.14	940.90	-11.85	0.08
ARS-BFGL-NGS-113888_GG	bio7	-456.32	179.50	128.90	0	0.20	0.16	0.16	0.20	0.09	916.64	939.40	-11.82	0.08
BTA-73516-no-rs_AA	latitude	-468.36	161.06	128.61	0	0.19	0.15	0.14	0.18	0.08	940.72	963.48	-0.67	0.76
Hapmap28985-BTA-73836_CC	prec6	-457.78	157.45	125.68	0	0.19	0.15	0.14	0.18	0.08	919.57	942.33	1.87	-0.03
Hapmap31863-BTA-27454_GG	prec7	-474.85	155.28	123.46	0	0.19	0.14	0.14	0.18	0.07	953.70	976.43	-1.91	0.02
ARS-BFGL-NGS-46098_GG	bio7	-456.70	167.50	121.71	0	0.20	0.15	0.15	0.19	0.08	917.39	940.15	-11.35	0.08
BTA-73516-no-rs_AA	prec6	-474.90	147.99	119.50	0	0.17	0.13	0.13	0.17	0.07	953.79	976.55	-1.97	0.03
Hapmap41762-BTA-117570_GG	bio7	-460.77	153.30	113.69	0	0.18	0.14	0.14	0.17	0.07	925.54	948.30	-10.71	0.07
Hapmap28985-BTA-73836_GG	bio3	-381.27	160.94	111.21	0	0.21	0.17	0.17	0.18	0.10	766.54	789.30	19.98	-0.26
ARS-BFGL-NGS-113888_GG	bio3	-471.77	148.61	106.51	0	0.17	0.14	0.13	0.17	0.07	947.53	970.29	20.21	-0.24

Figure 12 – Example of Samβada's results for univariate models, there is one marker per line. The first column is the name of the molecular marker, here the locus name combined with the allele name. The second column is the name of the environmental variable. The following columns are the log-likelihood, G score, Wald score and the error code (0 if success). The five next columns are goodness-of-fit measures for the regression (pseudo- R^2). The analysis includes the AIC (*Akaike information criterion*) and BIC (*Bayesian information criterion*) as well. The two last column contain the parameters β for the regression, one constant parameter and one corresponding to the environmental variable. Results file for multivariate models contain additional columns for environmental variables (Env_2, Env_3, ...) and for regression parameters (Beta_2, Beta_3, ...).

Spatial autocorrelation results

- Results are stored separately for environmental data and molecular markers
- In each case, there are three output files
 - The first one is named [Data-AS-Env.ext](#) (or [Data-AS-Mark.ext](#)) and stores Moran's I and local indicators of spatial association
 - The second file is either named [Data-AS-Env-Sim.ext](#) (or [Data-AS-Mark-Sim.ext](#)) and stores the simulated values of the global Moran's I for each variable
 - The third file is named [Data-AS-Env-pVal.ext](#) (or [Data-AS-Mark-pVal.ext](#)) and stores the pseudo p-values for the permutations-based significance tests

List of possible errors

- 0 Success
- 1 Exponential divergence ($X\beta$ is diverging)
- 2 Singular matrix (impossible to invert the information matrix)
- 3 Too large β (divergence)
- 4 Maximal number of iteration number reached without convergence
- 5 Monomorphic marker (appears in the output file for constant models)
- 6 Significant model with non-significant parents (multivariate analysis with option **SIGNIF**)

Soon: R-Sambada

The functions of the package include pre-processing, running of sambada and post-processing.

Preprocessing

- Relying on the package SNPRelate it accepts various formats (plink bed, plink ped, vgf, gds).
- According to user-defined thresholds the dataset is filtered for Minor Allele Frequency (MAF), Linkage Disequilibrium (LD) and Missing Rate.
- Pipeline to create an environmental dataset out of a file containing the sample location. The program will download climatic and altitudinal variables from global databases (wordclim, SRTM) choosing the required tiles according to the location of samples.
- A csv file containing the ID of the sample, its location and the associated environmental variable is created.
- The final “environmental file” is elaborated : redundant environmental variables are removed (according to a user-defined threshold).
- Population structure is assessed using the PCA-based implementation in SNPRelate. Sambada deals with population variables in a similar way like environmental variables (independent, explanatory variables).

R-Sambada

Running Sambada

- The C++ sambada code is included in the R package and invoked with an R function.
- Supervision module is included to manage parallel processing (HPC) using the R-package Foreach and DoParallel.

R-Sambada

Postprocessing

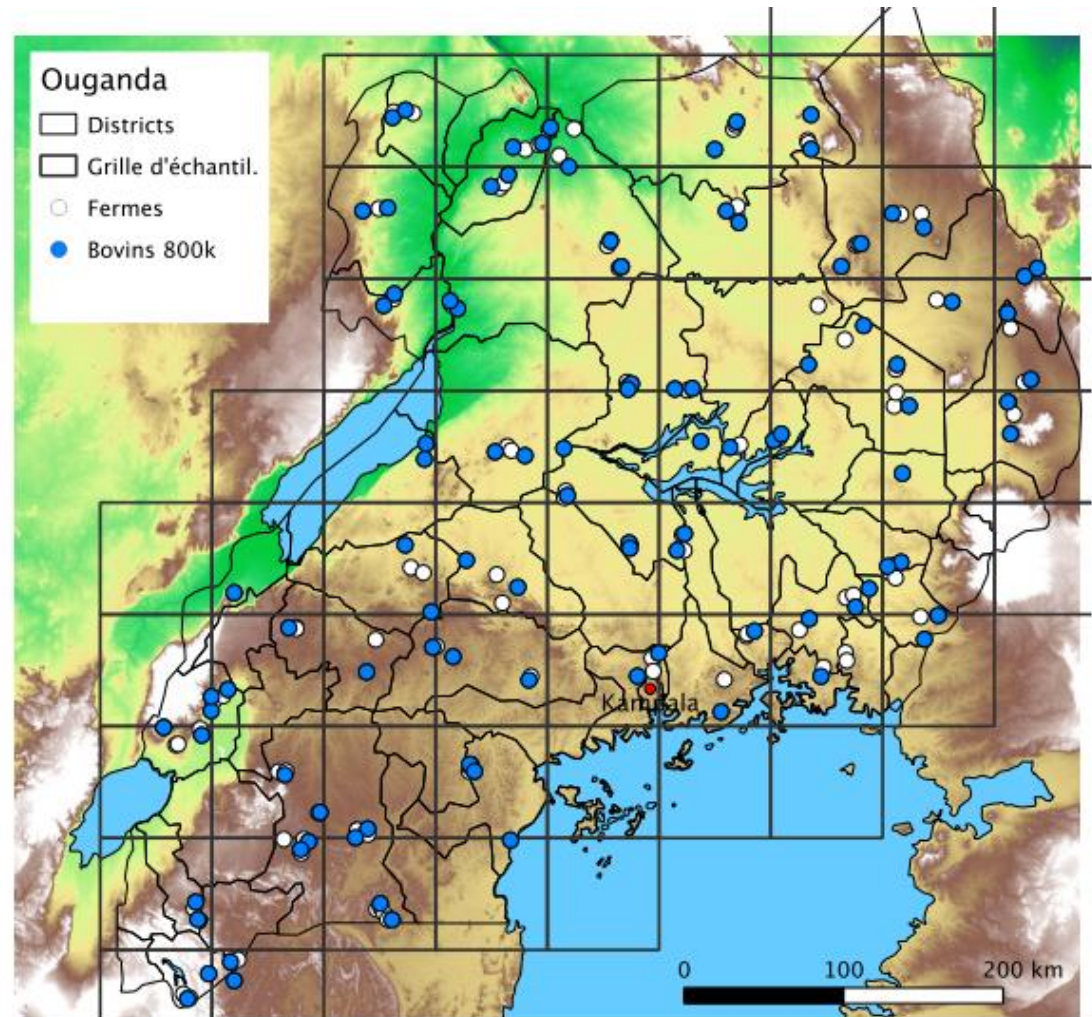
- q-values calculation based on Storey method and sorting out of models
- manhattan plot of all environmental variables and list of the most significant markers.
- connection to the *ensemble* database
- Generation of geographic maps to show the distribution of the marker and associated environmental variables
- interactive interface to specify an environmental variables and a chromosome: opens a local web-browser displaying the manhattan plot of the chosen region. The user can interactively click on a point, which will provide the name of the marker, its position and nearby genes. pvalue of the model marker x environmental variables is provided.

Example: Bos Taurus & Bos Indicus in Ouganda

FP7 NEXTGEN project

804 individuals,
Illumina 50k beadchip, 41'215 SNPs
= 120'102 genotypes
23 Environmental variables:

- Worldclim
- DEM variables from SRTM



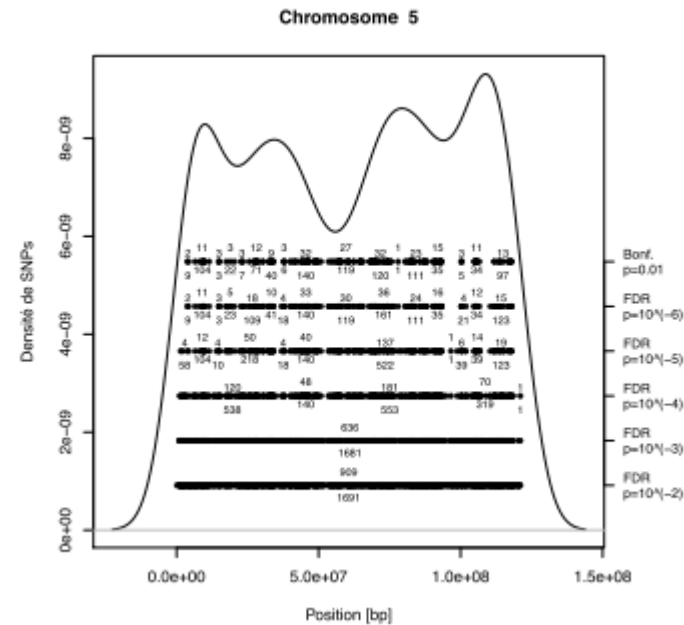
Results

2'699'510 models processed

Different significance thresholds
can be applied

With a bonferroni correction
applied at $\alpha=0.01$:

12'782 significant associations



Map of detected loci on chromosome 5

Marqueur	Chr.	Position [Mbp]	Env_1	Score G	Score Wald	AIC	BIC	Corrélation
ARS-BFGL-NGS-113888_GG	5	48.30	prec7	208.67	151.70	887.47	910.23	+
ARS-BFGL-NGS-113888_GG	5	48.30	latitude	193.89	146.89	902.25	925.01	+
ARS-BFGL-NGS-113888_GG	5	48.30	prec6	181.19	138.85	914.95	937.71	+
ARS-BFGL-NGS-113888_GG	5	48.30	bio7	179.50	128.90	916.64	939.40	+
Hapmap41074-BTA-73520_AA	5	48.40	prec7	208.53	151.72	890.22	912.98	+
Hapmap41074-BTA-73520_AA	5	48.40	latitude	193.13	146.61	905.62	928.38	+
Hapmap41074-BTA-73520_AA	5	48.40	bio7	180.61	129.73	918.14	940.90	+
Hapmap41074-BTA-73520_AA	5	48.40	prec6	179.99	138.13	918.77	941.53	+
Hapmap41762-BTA-117570_GG	5	18.90	prec7	202.93	148.43	875.92	898.68	+
Hapmap41762-BTA-117570_GG	5	18.90	latitude	186.04	141.99	892.80	915.56	+
Hapmap41762-BTA-117570_GG	5	18.90	prec6	166.51	130.97	912.33	935.09	+
ARS-BFGL-NGS-46098_GG	20	3.00	prec7	200.82	147.60	884.07	906.83	+
ARS-BFGL-NGS-46098_GG	20	3.00	latitude	178.45	138.11	906.44	929.20	+
ARS-BFGL-NGS-46098_GG	20	3.00	prec6	177.87	137.27	907.03	929.78	+
ARS-BFGL-NGS-46098_GG	20	3.00	bio7	167.50	121.71	917.39	940.15	+
ARS-BFGL-NGS-46098_GG	20	3.00	longitude	163.18	130.95	921.72	944.48	+

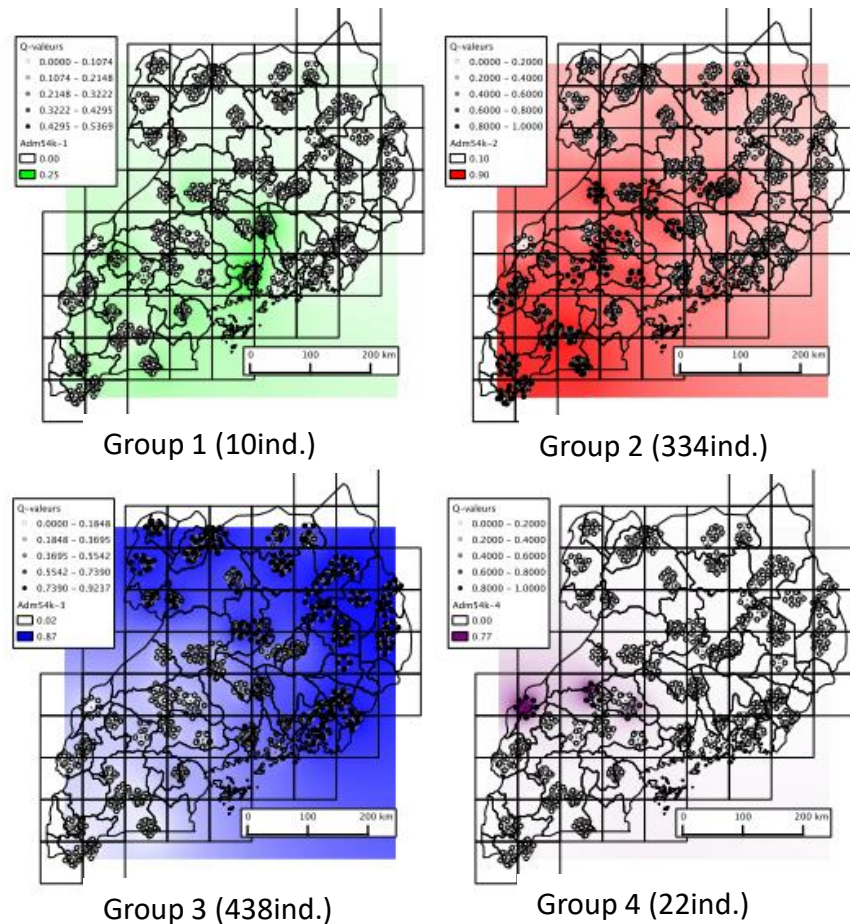
Examples of significant associations from univariate models in Sambada

Admixture

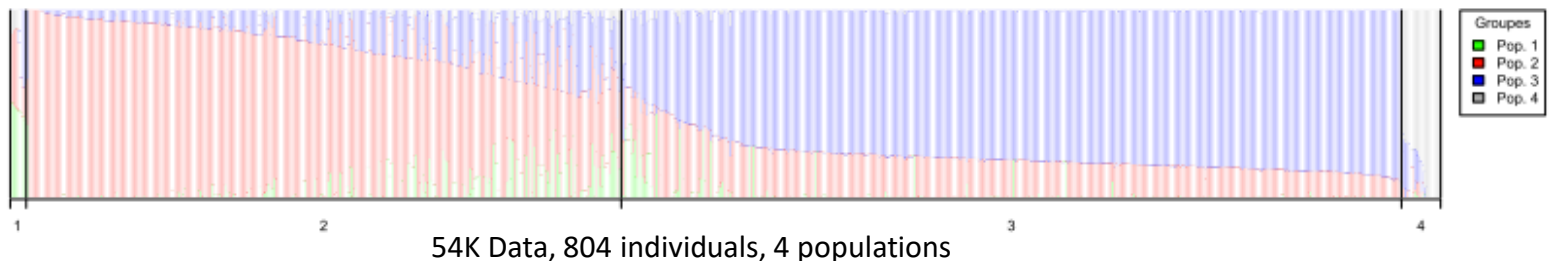
Maximum likelihood estimation of individual ancestries from SNP genotype datasets

User must define the estimated number of populations (K)

Cross-validation procedure to estimate the best K



Maps of genetic cluster from Admixture



Compared analysis

BayEnv, LFMM, Sambada

Knowing population structure

Neutral marker

Frequency will covary between populations due to shared demographic history.

BayEnv

Compute covariance between populations from

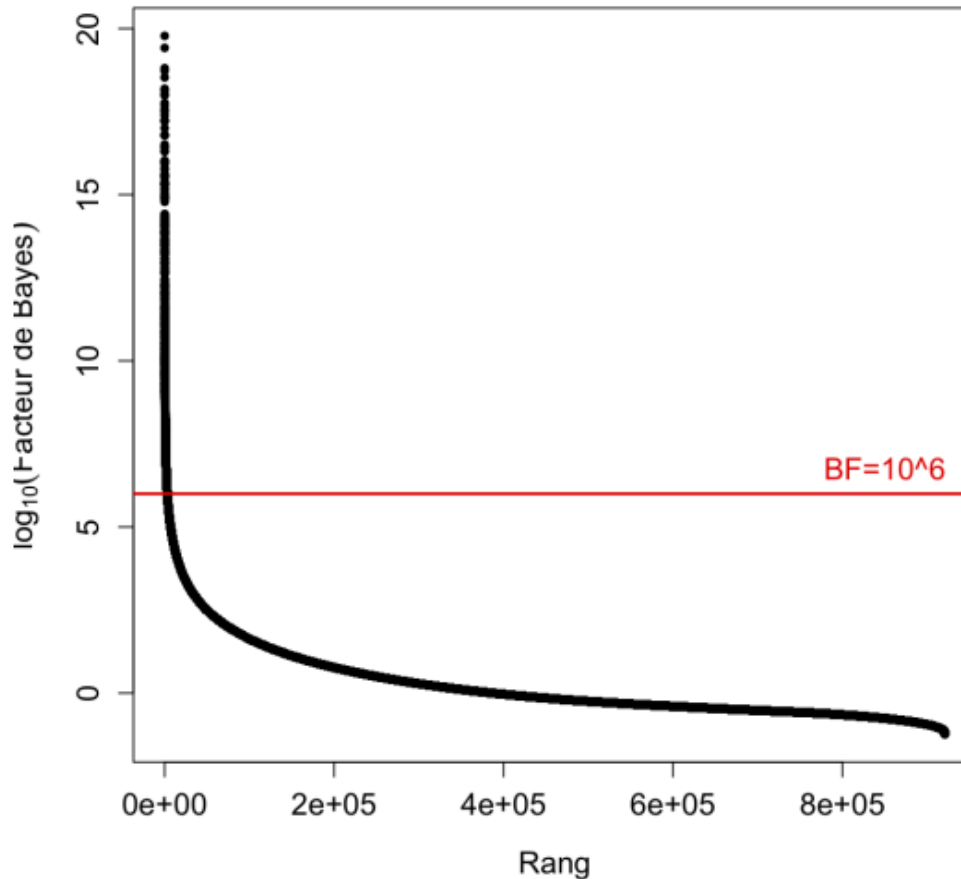
- population structure
- set of neutral loci

Build a robust null model.

Use this model to detect adaptive loci.

[Coop et al., 2010, Günther and Coop, 2012]

Significance threshold



BayEnv provides a Bayes factor for each pair of SNP x environmental variable. No p value available. The empirical threshold (0.1%) is based on the 10th highest Bayes score among 10'000 of the neutral loci (method suggest by Graham Coop) .

Bayenv's drawbacks

- Requires a defined population structure
- To assess associations between environmental variables and markers, a separate file has to be created for each SNP
- Computation time required is very long
- Requires a defined set of neutral markers

Results Bayenv & Sambada

Données	Nombre SNPs détectés			Comp. premiers SNPs détectés			
	Communs	BayEnv	Samβada	Nb SNPs comparés	Communs	Modèles BayEnv	Modèles Samβada
54k	387	400	2'500	100	65	862	352
800k	34	2'083	57	50	2	50	87
800ksub	0	91	7	-	-	-	-

Table 7.19 – Comparaison des résultats fournis par BayEnv et Samβada. Les modèles BayEnv ont été sélectionnés sur la base de leur facteur de Bayes (voir table 7.18). La partie de gauche compte les SNPs communs ainsi que ceux détectés par chaque programme. La partie de droite compte le nombre de SNPs communs parmi les 100 ou 50 SNPs ayant les scores les plus élevés (facteur de Bayes ou score *G*). Les deux dernières colonnes indiquent combien de modèles doivent être considérés pour obtenir 100 ou 50 SNPs différents.

Inferring population structure

Latent Factor Mixed Models

Inferring both environmental and population influence will lower false positive rate. [Frichot et al., 2013]

$$G_{il} = \mathbf{x}_i^T \boldsymbol{\beta}_l + U_i^T V_l$$

G_{il}	number of the derived allele at the locus, $G_{il} = \{0, 1, 2\}$
\mathbf{x}_i	vector of environmental variables
$\boldsymbol{\beta}$	vector of $q + 1$ coefficients
U_i^T, V_l	vectors of length K


There are K latent factors (\sim populations)



$U^T V$ models genetic variation not explained by environmental factors.

(U individual effects, V loci effects)


LFMM (LEA package)

- Expresses the genotype with a linear mixed model and includes latent factors
- Latent factors represent the part of genetic variability that is not explained by the environment
- LFMM provides un z score and a p-value for each model
- Easy to use and rather fast

 New Project




Project Name:




Project Path:

Browse...




Genotype Data File:

Browse...




Environmental File:

Browse...



Snp File (optional):

Browse...




View Data...


View Environmental Data...

View SNPs Data...


Number of Individuals:



Number of Loci:



Number of environmental variables:



OK

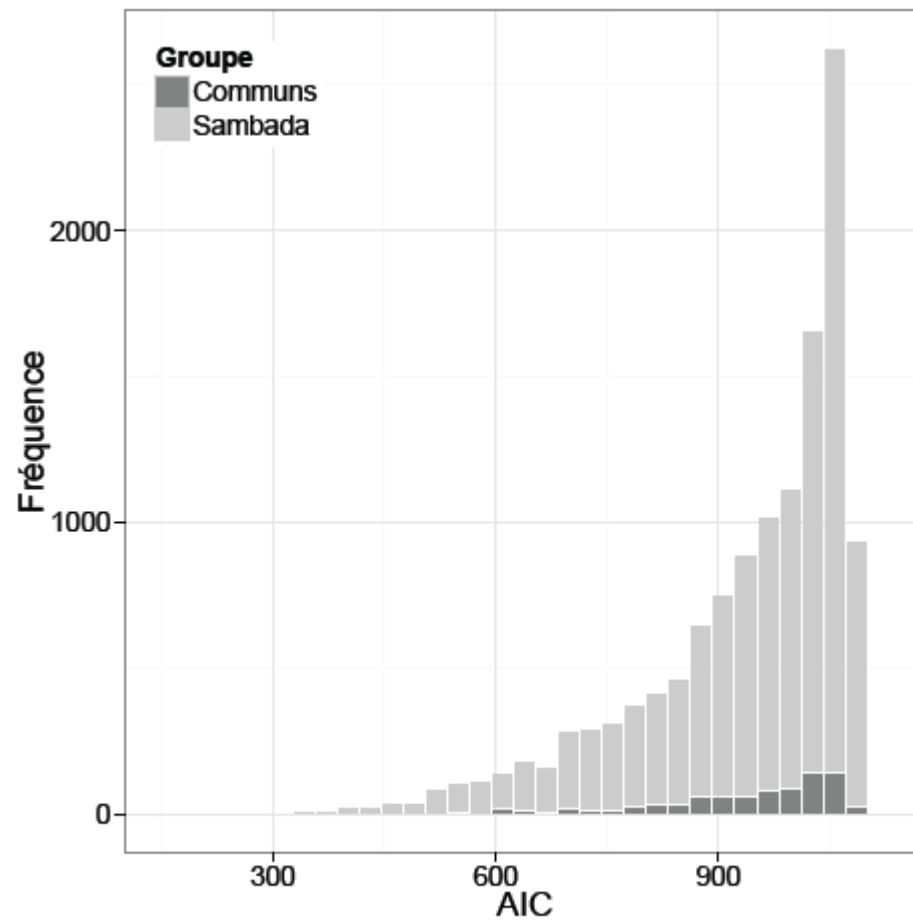
Cancel

Results LFMM & Sambada

Données	Nombre de modèles détectés			Nombre de loci détectés		
	Communs	LFMM	Samβada	Communs	LFMM	Samβada
54k	227	303	12'782	128	245	2'500
800k	0	0	104	0	0	57
800ksub	5	24	12	4	14	7

Table 7.24 – Comparaison des résultats de LFMM et de Samβada pour les trois jeux de données. La partie de gauche compte les modèles communs ainsi que ceux détectés par chaque programme. Les SNPs sont comptés à droite. La p -valeur est fixée 0.01 avant la correction de Bonferroni, pour les deux méthodes.

AIC: LFMM & Sambada



(a) Distribution de l'AIC

Bivariate models with Sambada

Membership coefficients from Admixture are used as co-variates

Marker	Env_1	Env_2	Gscore	WaldScore	AIC	BIC
Hapmap41074-BTA-73520_AA	prec5	prec4	101.12	88.28	928.18	962.32
	longitude	bio3	76.22	62.14	877.27	911.41
Hapmap28985-BTA-73836_CC	bio12	prec11	99.97	87.96	932.75	966.89
	bio12	bio15	79.67	68.40	962.24	996.37
ARS-BFGL-NGS-113888_GG	prec5	prec4	98.45	86.19	925.81	959.95
	longitude	bio3	76.09	61.88	873.44	907.58
	bio15	prec5	75.89	67.71	944.63	978.77
ARS-BFGL-NGS-46098_GG	prec5	prec4	94.67	82.94	906.51	940.65
Hapmap41813-BTA-27442_AA	prec5	prec4	90.65	80.66	960.47	994.61
BTA-73516-no-rs_AA	prec5	prec4	81.35	73.15	964.03	998.17

Marker	Env_1	Env_2	Gscore	WaldScore	AIC	BIC
Hapmap28985-BTA-73836_GG	bio3	ankole	64.70	48.59	703.84	737.98
ARS-BFGL-NGS-106520_AA	bio3	ankole	53.15	44.25	773.58	807.71
BTA-73842-no-rs_GG	bio3	ankole	47.96	40.98	793.94	827.91
Hapmap28985-BTA-73836_GG	latitude	ankole	40.39	37.86	740.25	774.39

See details in

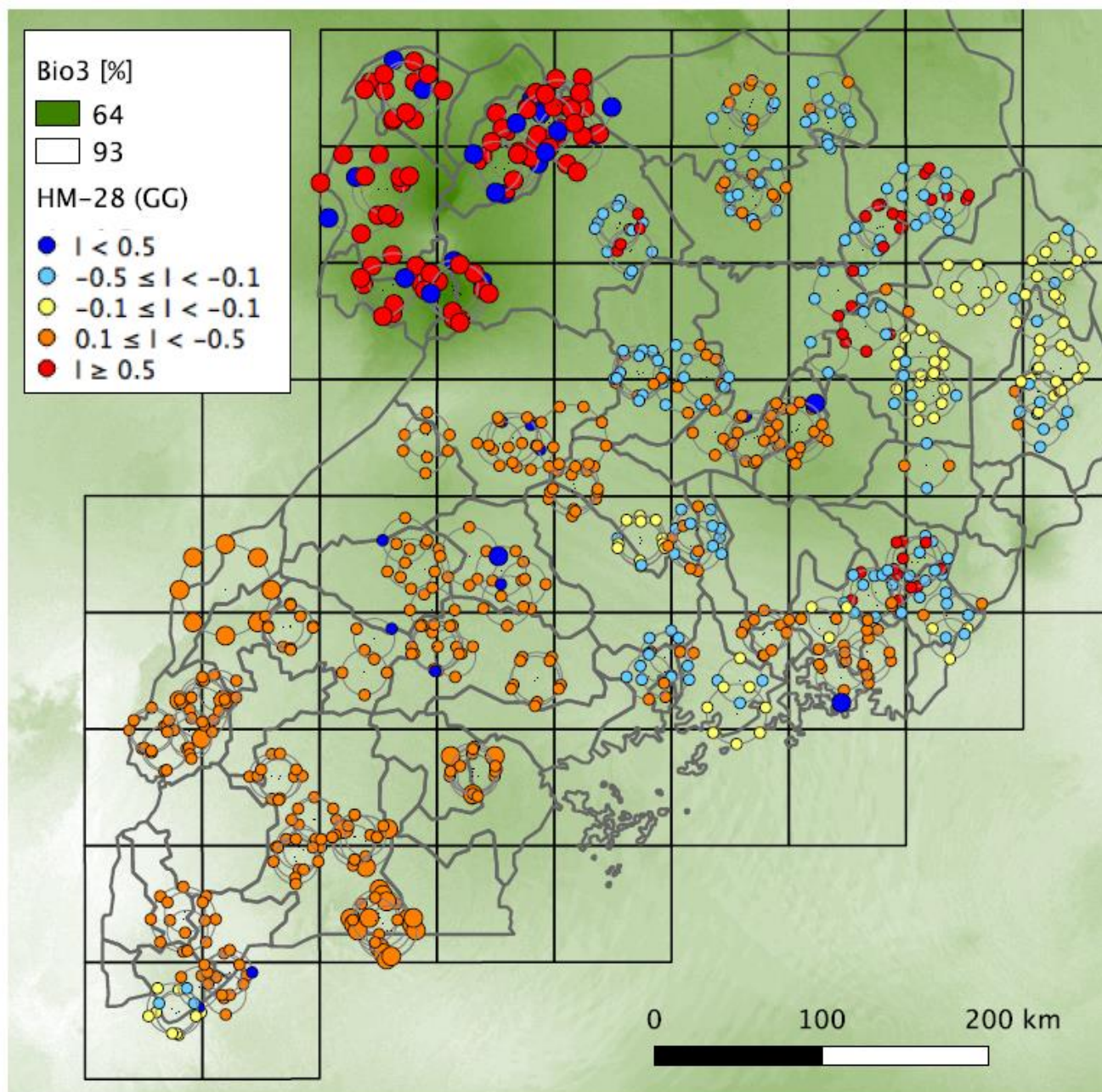
Stucki, S., Orozco-terWengel, P., Forester, B. R., Duruz, S., Colli, L., Masembe, C., ... Joost, S. (2017). High performance computation of landscape genomic models including local indicators of spatial association. *Molecular Ecology Resources*, 17(5), 1072–1089.

	Loci	Chr.	Pos. [Mbp]	Samβada	BayEnv	LFMM	Arlequin	Détections
1	ARS-BFGL-NGS-113888	5	48.32	1	1	0	0	2
2	Hapmap41074-BTA-73520	5	48.35	1	1	0	0	2
3	Hapmap41762-BTA-117570	5	18.94	1	1	0	0	2
4	ARS-BFGL-NGS-46098	20	2.95	1	1	0	0	2
5	Hapmap41813-BTA-27442	5	49.04	1	1	0	0	2
6	BTA-73516-no-rs	5	48.75	1	1	0	0	2
7	Hapmap28985-BTA-73836	5	70.34	1	1	1	1	4
8	Hapmap31863-BTA-27454	5	48.99	1	1	0	0	2
9	ARS-BFGL-NGS-106520	5	70.20	1	1	1	1	4
10	BTA-73842-no-rs	5	70.18	1	1	1	1	4
11	Hapmap50523-BTA-98407	5	46.74	1	1	0	0	2
12	BTB-01400776	20	2.70	1	1	0	0	2
13	Hapmap23956-BTA-36867	15	47.20	1	1	0	0	2
14	ARS-BFGL-NGS-10586	2	128.64	1	1	0	0	2
15	ARS-BFGL-NGS-43694	5	49.65	1	1	0	0	2
16	BTA-122374-no-rs	14	16.44	1	1	0	0	2
17	BTB-01356178	20	2.49	1	1	0	0	2
18	ARS-BFGL-NGS-94862	11	103.53	1	1	1	0	3
19	BTA-108359-no-rs	14	16.31	1	1	0	0	2
20	ARS-BFGL-NGS-15960	5	28.02	1	1	0	0	2

Table 7.32 – Liste des SNPs détectés par Samβada correspondant aux modèles ayant les plus hauts scores *G* pour les données 54k. Les loci sont identifiés par leur nom, leur chromosome et la position qu'ils y occupent, en millions de paires de bases. Les colonnes suivantes indiquent quelles méthodes les ont détectés et la dernière indique le nombre de ces détections. Les loci en caractères gras sont les découvertes communes aux quatre méthodes.

Sambada and spatial autocorrelation

- Spatial autocorrelation indices to relativize these many SNPs identified
- LISA = Local Indices of Spatial Association



(b) HM-28

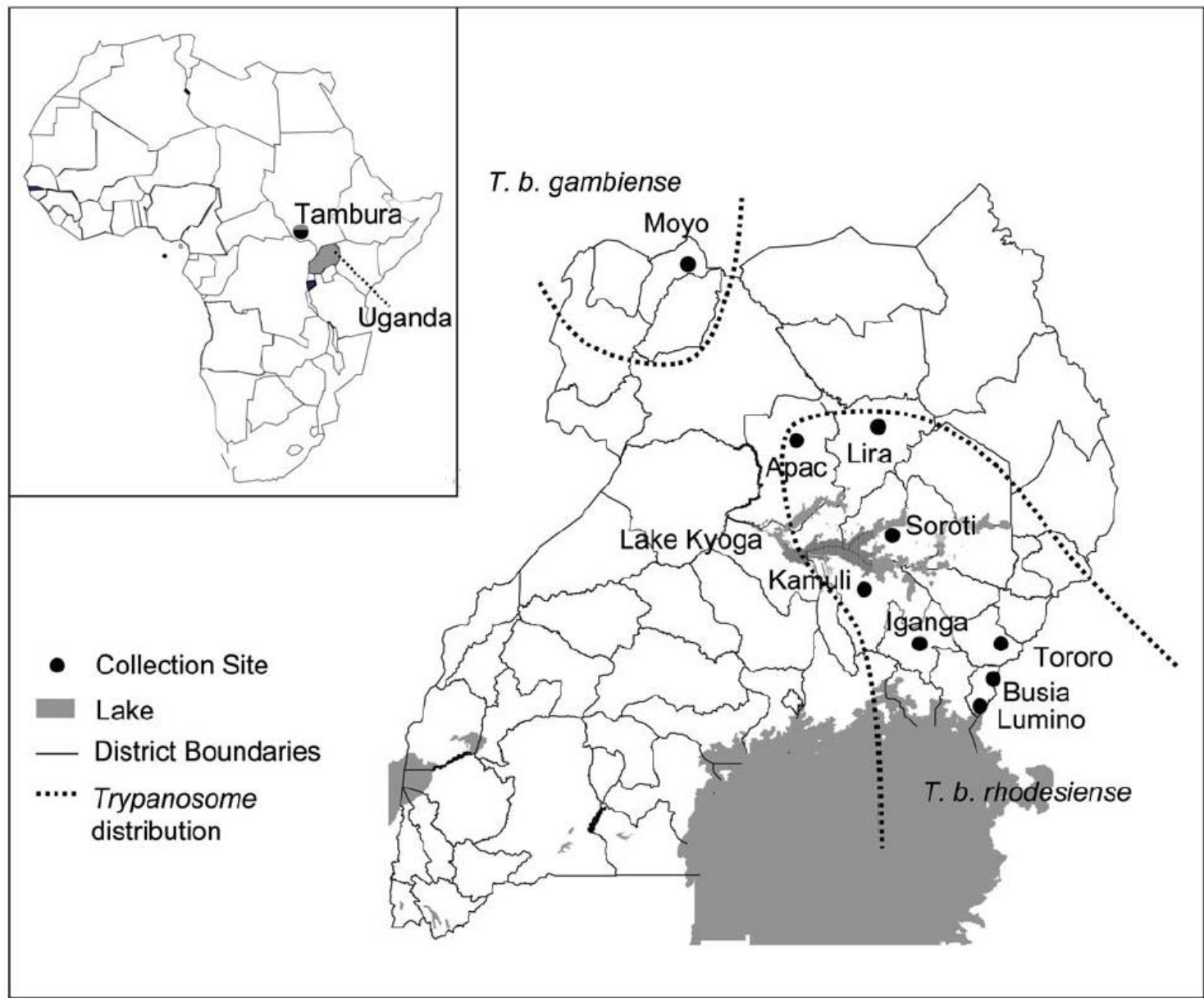


Figure 8.1 – Carte de prévalence de *Trypanosoma brucei gambiense* et *Trypanosoma brucei rhodesiense* en Ouganda. Ces deux vers parasites sont transmis par la mouche Tsé-tsé (*Glossina fuscipes fuscipes*) et provoquent la maladie du sommeil chez l'humain.

Méthode	Taille des données moléculaires (loci)	Populations	Données	Très grand volume de données
BayEnv	Petite (\rightarrow 100k)	Distinctes	Par pop.	Très long, gestion des fichiers problématique
Arlequin	Moyenne (\rightarrow 1M)			Peut traiter les données moléculaire par blocs
LFMM		Mélangées	Par indiv.	Peut traiter chaque variable env. séparément, calcul multitâche
Samβada				Grande (\rightarrow 1M, bivarié)
CoreSAM	Grande (\rightarrow 20M, univarié)			

Table 9.2 – Vue synoptique des approches applicables en fonction du type de traitement. Des données individuelles peuvent être agrégées par populations, l’opération inverse est rarement possible. Les recommandations sur les tailles des jeux de données sont des estimations basées sur les analyses présentées dans ce travail.

Multivariate models with Sambada

- In the multivariate approach, several environmental variables can be used at the same time to model the presence of each genotype
- In this case, the selection procedure is similar to a forward stepwise regression (Dobson & Barnett 2008)
- Both G and Wald tests refer to a null model to build the null hypothesis.
- The current model is compared to the constant model (the same as in the univariate case) using multivariate χ^2 statistics
- While rejecting the null hypothesis in this configuration would indicate that at least one parameter in the model is statistically significant, it would not provide information about which parameter is relevant to the model
- Therefore, SAMBADA assesses parameter significance in multivariate models with either a Wald test applied to each parameter separately (except the constant parameter) or with G tests excluding a parameter at a time: model selection is based on simpler models nested in the current one

Population structure

- Multivariate models allow the inclusion of preexisting knowledge, provided the data constitutes a continuous variable
- If population structure was analysed beforehand and can be represented as a coefficient of membership for each individual, this information can be included in the modelling
- For models involving both an environmental variable and this coefficient, the selection procedure will assess whether the environmental variable is associated with the genotype while taking into account the possible effect of admixture
- Like LFMM

Bivariate modelss

- The multivariate analysis to take population structure into account consists in bivariate models along with their corresponding univariate and constant models
- To this end, a variable 'population structure' is defined by performing a principal component analysis (PCA) on the genetic data
- The univariate models involving the 'population structure' variable are used as 'null models'
- They are used to assess the significance of bivariate models involving the 'population structure' variable + one environmental variable
- See pages 17-18 of the Sambada documentation

LABORATORY OF GEOGRAPHIC INFORMATION SYSTEMS **LASIG**

[Home](#)
[Research](#)
[Teaching](#)
[Services](#)
[People](#)
[Publications](#)
[Links](#)
[Software](#)
[LSSR](#)
[GIRAPH](#)

Share: [f](#) [t](#) [in](#) [g+](#) [e](#)

Samβada

Samβada

Pic2Map

Releases

Documentation

- [Sambadoc-v0.5.3.pdf](#)

Software

- [Sambada v0.5.3 for MacOS](#)
- [Sambada v0.5.3 for Ubuntu](#)
- [Sambada v0.5.3 for Windows 32 bits](#)
- [Sambada v0.5.3 for Windows 64 bits](#)

Source code

- [sambada-v0.5.3-src.zip](#) source code as .zip file
- [sambada-v0.5.3-src.tar.gz](#) source code as .tar.gz file

References

- Stucki, S., Orozco-terWengel, P., Forester, B. R., Duruz, S., Colli, L., Masembe, C., ... Joost, S. (2017). [High performance computation of landscape genomic models including local indicators of spatial association](#). Molecular Ecology Resources, 17(5), 1072–1089. doi:10.1111/1755-0998.12629
- Stucki, S. (2014). Développement d'outils de géo-calcul haute performance pour l'identification de régions du génome potentiellement soumises à la sélection naturelle: analyse spatiale de la diversité de panels de polymorphismes nucléotidiques à haute densité (800k) chez Bos taurus et B. indicus en Ouganda, EPFL PhD Thesis no 6014, doi:10.5075/epfl-thesis-6014

KEYWORDS

GIS, Spatial Analysis, Decision-making support, Exploratory Spatial Data Analysis, Landscape genetics, Landscape genomics, Spatial epidemiology

CONTACT

Secretary

EPFL ENAC IIE LASIG

Batiment GC

Vers le plan d'orientation GC D2 397

Station 18

CH-1015 Lausanne

Tel: +41 21 693 27 55

Fax: +41 21 693 57 90

QUICK LINKS

[MOOC LASIG - Introduction aux Systèmes d'Information Géographique](#)

[Urban planning and public health \(LASIG, CHUV, HUG\)](#)

