Landscape genomics

Dr Stéphane Joost – Oliver Selmoni (Msc)
Laboratory of Geographic Information Systems (LASIG)
Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Terminology…

MOLECULAR ECOLOGY

Molecular Ecology (2010) **19**, 3760–3772 doi: 10.1111/j.1365-294X.2010.04717.x

Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field

STÉPHANIE MANEL,*† STÉPHANE JOOST,‡ BRYAN K. EPPERSON,§ ROLF HOLDEREGGER,¶ ANDREW STORFER,** MICHAEL S. ROSENBERG,†† KIM T. SCRIBNER,‡‡ AURÉLIE BONIN§§ and MARIE-JOSÉE FORTIN¶¶

Box 2. Clarification of terms

A number of recent terms, including landscape genetics (Manel *et al.* 2003), landscape genomics (Luikart *et al.* 2003; Joost *et al.* 2007), molecular genecology (Hamilton *et al.* 2002; Skot *et al.* 2002), and ecological genomics (Ungerer *et al.* 2008), have recently been introduced to describe studies aimed at understanding the impact of the environment/landscape on genetic response. These are in fact not new research fields, but rather involve the interdisciplinary integration of multiple pre-existing research disciplines, including spatial statistics, landscape ecology, population genetics and molecular biology. These terms were initially introduced to facilitate the discussion of researchers across disciplines; however, the multiplication of similar terms has led to the need for clarification.

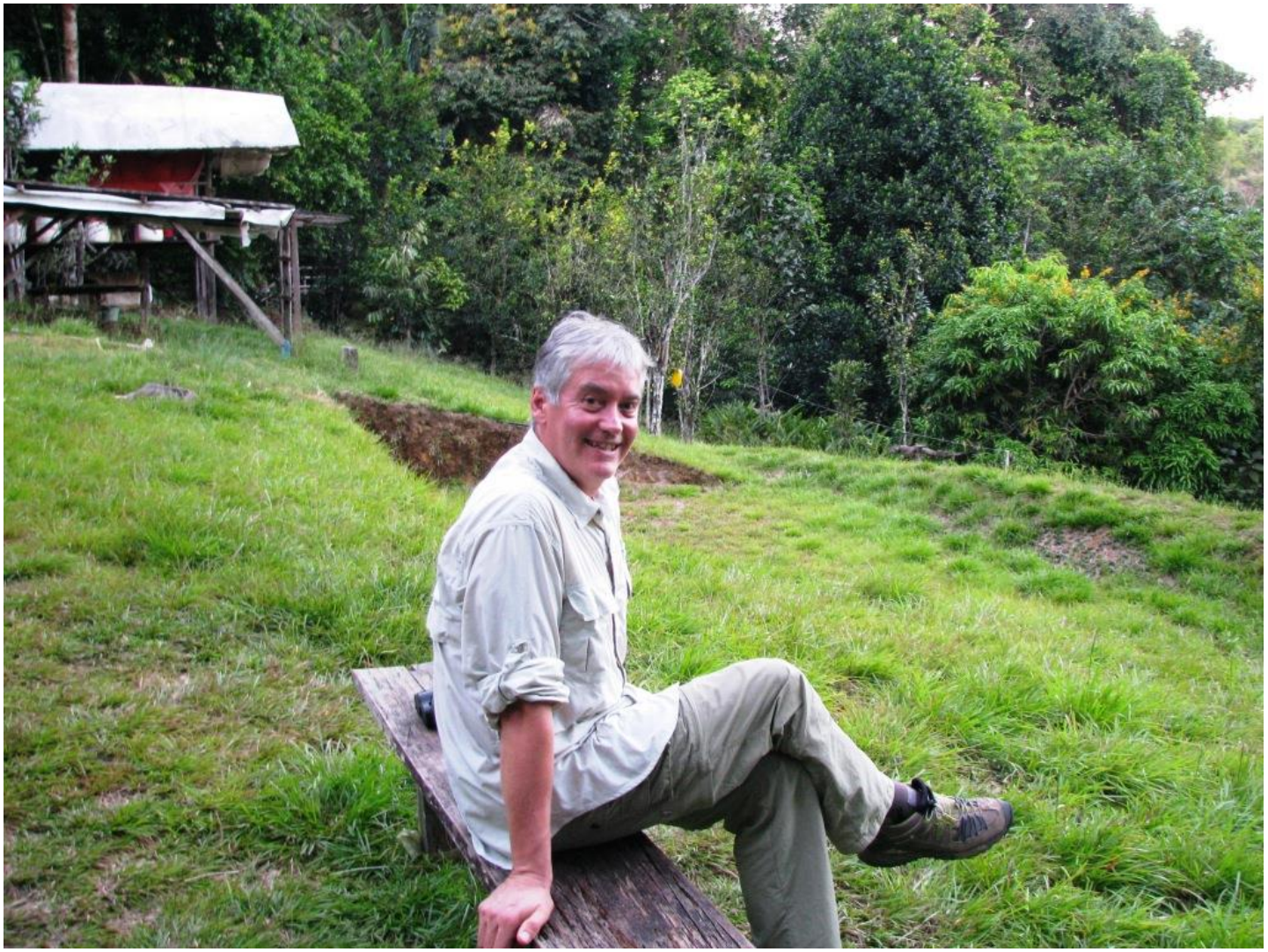
Landscape genetics (Manel *et al.* 2003) aims to provide information about the interaction between landscape features and microevolutionary processes, such as gene flow, genetic drift or selection. Most current applications of landscape genetics focus on gene flow and migration (processes that can either facilitate or constrain local adaptation), i.e. the effect of the environment on the selectively neutral component of genetic diversity (Storfer *et al.* 2007; Anderson *et al.* 2010). However, landscape genetics also aims to correlate allele frequencies with the environment in order to understand the effect of the environment on the adaptive component of genetic diversity (Holderegger *et al.* 2006).

Few markers → adaptive landscape genetics

Landscape genomics (Luikart *et al.* 2003; Joost *et al.* 2007) uses correlation studies between the genomic data and the environment to identify genes either potentially linked to candidate genes or the genes themselves under selection. Landscape genomics is included in landscape genetics, but refers more specifically to the use of the future large amount of genetic data due to high-throughput sequencing. Landscape genomics is thus at the interface of bioinformatics, genomics, spatial statistics and landscape ecology.

Molecular genecology (Hamilton *et al.* 2002) is the study of geographical clines in the frequencies of alleles and their relationship to ecological clines in environmental conditions. Its objectives are largely the same as for the other research fields listed above.

Ecological genomics (Ungerer *et al.* 2008) integrates over several disciplines and seeks to understand the genetic mechanisms underlying responses of organisms to their natural environment. It is broader than landscape genetics and genomics, since it further includes experimental and laboratory approaches.



Review TRENDS in Ecology and Evolution Vol.18 No.4 April 2003

Landscape genetics: combining landscape ecology and population genetics

Stéphanie Manel[1], Michael K. Schwartz[2], Gordon Luikart[1] and Pierre Taberlet[1]

[1]Laboratoire d'Ecologie Alpine, Equipe Génomique des Populations et Biodiversité, UMR CNRS 5553, BP 53, Université Joseph Fourier, 38041 Grenoble Cedex 9, France
[2]Rocky Mountain Research Station, US Forest Service, 800 E. Beckwith, Missoula, MT 59801, USA

THE POWER AND PROMISE OF POPULATION GENOMICS: FROM GENOTYPING TO GENOME TYPING

Gordon Luikart, Phillip R. England, David Tallmon, Steve Jordan and Pierre Taberlet

Population genomics has the potential to improve studies of evolutionary genetics, molecular ecology and conservation biology, by facilitating the identification of adaptive molecular variation and by improving the estimation of important parameters such as population size, migration rates and phylogenetic relationships. There has been much excitement in the recent literature about the identification of adaptive molecular variation using the population-genomic approach. However, the most useful contribution of the genomics model to population genetics will be improving inferences about population demography and evolutionary history.

NATURE REVIEWS | GENETICS

VOLUME 4 | DECEMBER 2003 | 981

In the references…

18. Manel, S., Schwartz, M., Luikart, G. & Taberlet, P. Landscape genetics: combining landscape ecology and population genetics. *Trends Ecol. Evol.* **18**, 189–197 (2003).
This article summarizes the statistical approaches that are available for relating spatial variation in population-genetic patterns to spatial variation in environmental patterns, This article and the population-genomic concepts discussed here show the feasibility of a 'landscape genomic' approach using association studies between the genome and environments.
19. Waples, R. S. Genetic methods for estimating the effective size of cetacean populations. *Report of the International Whaling Commission (Special Issue)* **13**, 279–300 (1991).
20. Yang, Z. Likelihood and Bayes estimation of ancestral population size in hominoids using data from multiple loci. *Genetics* **162**, 1811–1823 (2002).

Molecular Ecology (2007) **16**, 3955–3969 doi: 10.1111/j.1365-294X.2007.03442.x

A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation

S. JOOST,*† A. BONIN,‡ M. W. BRUFORD,§ L. DESPRÉS,‡ C. CONORD,‡ G. ERHARDT¶ and P. TABERLET‡**

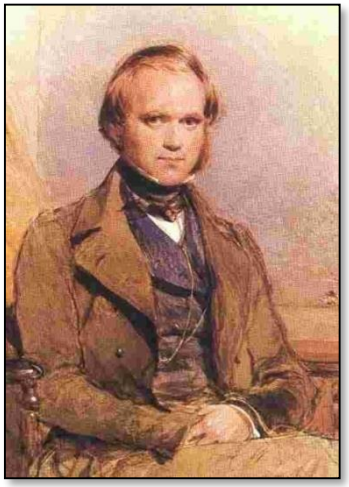
*Istituto di Zootecnica, Università Cattolica del S.Cuore, via E. Parmense 84, 29100 Piacenza, Italy, †Laboratoire de Systèmes d'Information Géographique, Ecole Polytechnique Fédérale de Lausanne (EPFL), Bâtiment GC, Station 18, 1015 Lausanne, Switzerland, ‡Laboratoire d'Ecologie Alpine, CNRS-UMR 5553, Université Joseph Fourier, BP 53, 38041 Grenoble Cedex 09, France; §Cardiff School of Biosciences, Main Building, Museum Avenue, Cardiff CF10 3TL, UK, ¶Department of Animal Breeding and Genetics, Justus-Liebig-University of Giessen, Ludwigstrasse 21B, 35390 Giessen, Germany

Natural selection and adaptation of species to their local environment

Goal

- Show how GIScience methods and data may contribute to :
  - The advancement of our understanding of mechanisms controlling the evolution of species (adaptation to local environment in particular)

Signatures of natural selection



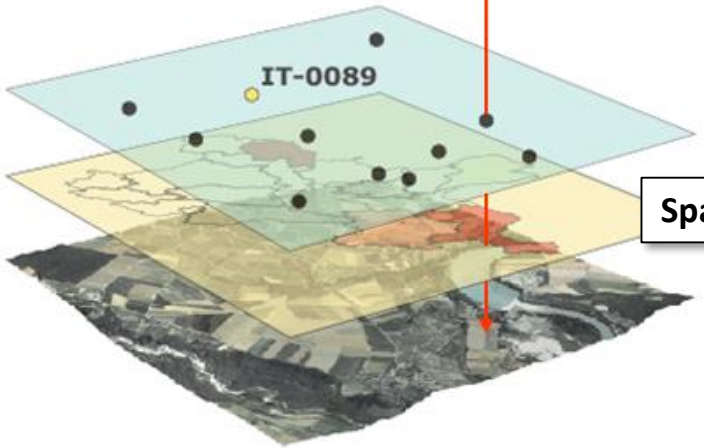
Why looking for signatures of natural selection ?

- Any genetic marker selected by an environmental variable may be associated with a gene
- The function of this gene may be discovered by means of clues provided by this environmental variable (e.g. temperature)

The study of the effects of natural selection may improve our understanding of the genetic mechanisms of evolution

These signatures permit to identify regions of the genome with a particular adaptive value
→ Important in conservation biology to establish priorities among endangered populations

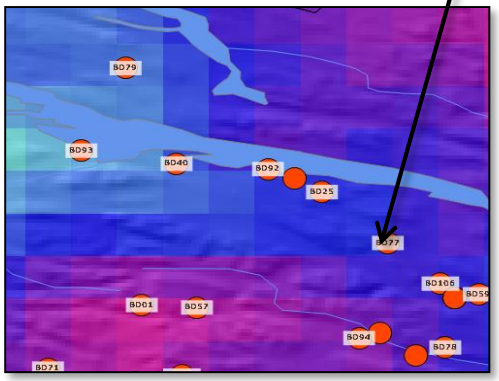
Association models

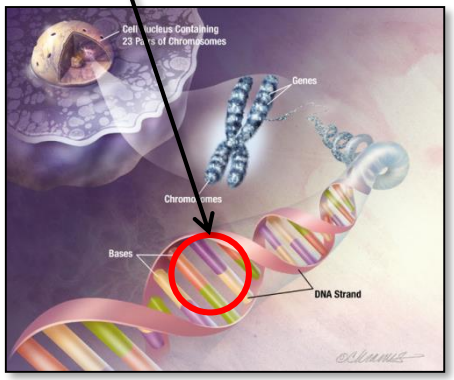


GENETICS

Spatial coincidence

ENVIRONMENT

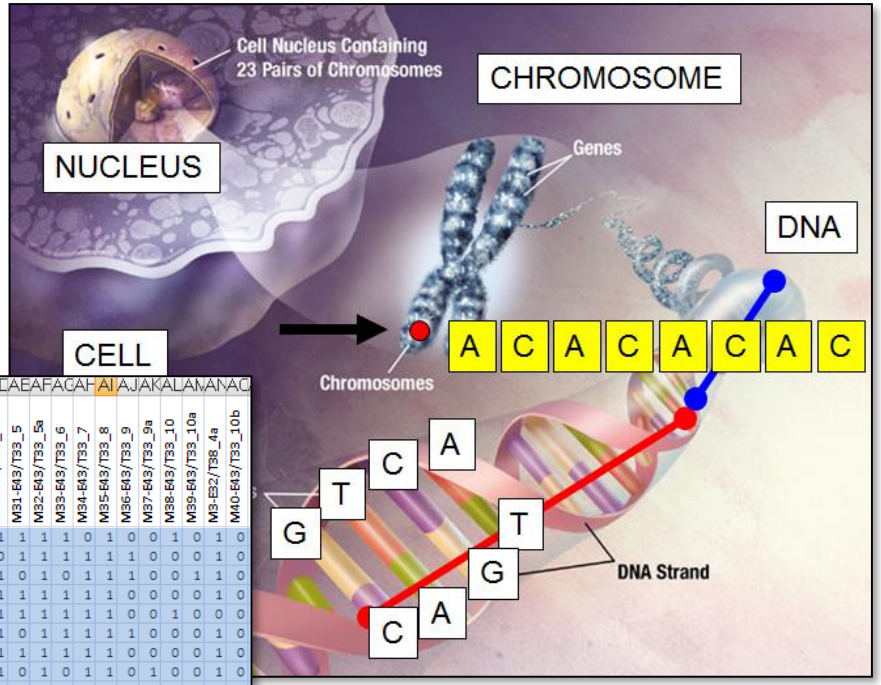




Landscape genomics

Association ?

Complete the genetic matrix…



GEO

	A	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO
			M100-E45/T32_37	M101-E45/T32_38	M102-E45/T32_39	M10-E32/T38_9	M11-E32/T38_10	M12-E32/T38_11	M13-E32/T38_11a	M14-E32/T38_11b	M15-E32/T38_12	M16-E32/T38_13	M17-E32/T38_14	M18-E32/T38_15	M19-E32/T38_17	M1-E32/T38_3a	M20-E32/T38_17a	M21-E32/T38_19	M22-E32/T38_20a	M23-E32/T38_21	M24-E32/T38_21b	M25-E32/T38_22	M26-E32/T38_22a	M27-E43/T33_1	M28-E43/T33_2	M29-E43/T33_3a	M2-E32/T38_4	M30-E43/T33_4	M31-E43/T33_5	M32-E43/T33_5a	M33-E43/T33_6	M34-E43/T33_7	M35-E43/T33_8	M36-E43/T33_9	M37-E43/T33_9a	M38-E43/T33_10	M39-E43/T33_10a	M3-E32/T38_4a	M40-E43/T33_10b
1	farmid	ANIMALID	longitude	latitude																																				
2	PT-0015	CHPOBRA3	-8.2676	41.6848	1	1	1	0	1	1	0	0	0	1	1	0	1	0	1	0	1	0	0	0	0	1	0	0	0	1	1	1	1	0	1	0	0	1	0	1	0
3	PT-0022	CHPOBRA26	-7.9273	41.5847	1	1	1	0	1	1	0	0	0	1	1	0	1	0	1	0	1	0	0	1	0	1	0	0	0	0	1	1	1	1	1	1	0	0	0	1	0
4	PT-0023	CHPOBRA28	-7.9273	41.5847	1	1	0	0	1	1	0	0	0	1	1	0	1	0	1	0	1	0	0	1	0	1	0	0	0	1	0	1	0	1	1	1	0	0	1	1	0
5	PT-0023	CHPOBRA29	-7.9273	41.5847	1	1	1	1	1	1	0	0	0	1	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	1	0
6	PT-0021	CHPOBRA23	-7.8426	41.395	0	0	0	1	1	1	0	0	0	1	1	0	1	0	1	0	1	0	0	0	0	1	0	0	0	1	1	1	1	1	1	0	0	1	0	0	0
7	PT-0021	CHPOBRA25	-7.8426	41.395	1	1	1	1	1	1	0	0	0	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	0	0	0	1	0
8	PT-0017	CHPOBRA10	-7.8269	41.4732	1	1	1	1	1	1	1	0	0	0	1	0	1	0	1	0	1	0	0	0	0	1	0	0	0	1	1	1	1	1	1	0	0	0	0	1	0
9	PT-0017	CHPOBRA11	-7.8269	41.4732	1	1	0	0	1	1	0	0	0	1	0	1	1	0	1	0	1	0	0	0	0	1	0	0	0	1	0	1	0	1	1	0	1	0	0	1	0
10	PT-0017	CHPOBRA12	-7.8269	41.4732	1	0	0	Na	1	1	0	0	0	0	1	0	1	0	1	0	1	0	0	0	0	1	1	0	0	1	1	1	1	1	1	0	0	0	0	1	0
11	PT-0020	CHPOBRA21	-7.8215	41.4235	0	1	0	1	1	1	1	0	0	0	1	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	1	0
12	PT-0020	CHPOBRA22	-7.8215	41.4235	1	1	0	1	1	1	0	0	0	1	1	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	1	0
13	PT-0019	CHPOBRA18	-7.7811	41.439	1	1	1	1	1	1	0	0	0	1	1	0	1	0	1	0	1	0	0	0	0	1	0	0	0	1	1	1	1	1	1	0	0	0	0	1	0
14	PT-0019	CHPOBRA19	-7.7811	41.439	1	1	1	1	1	1	0	0	0	1	1	0	1	0	1	0	1	0	0	0	0	1	0	0	0	1	1	1	1	1	1	0	0	0	0	1	0
15	PT-0018	CHPOBRA13	-7.78	41.4379	1	1	1	1	1	1	0	0	0	1	1	0	1	0	1	0	1	0	0	0	0	1	0	0	0	1	1	1	1	1	1	0	0	0	0	1	0
16	PT-0018	CHPOBRA14	-7.78	41.4379	1	1	0	0	1	1	0	0	0	1	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	1	0
17	PT-0018	CHPOBRA15	-7.78	41.4379	1	1	0	1	1	1	0	0	0	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	1	0
18	PT-0016	CHPOBRA5	-7.7291	41.4752	1	1	1	0	1	1	0	0	0	1	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	1	0
19	PT-0016	CHPOBRA6	-7.7291	41.4752	1	0	0	1	1	1	0	0	0	1	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	1	0
20	PT-0016	CHPOBRA8	-7.7291	41.4752	1	0	0	1	1	1	0	0	0	1	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	1	0
21	ES-0060	CHSPFLR35	-6.00263	37.7014	1	0	0	1	1	1	0	0	0	1	1	0	1	0	1	0	1	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	0	0	0	1	0
22	ES-0060	CHSPFLR37	-6.00263	37.7014	1	0	0	0	1	1	0	0	0	1	1	0	1	0	1	0	1	0	0	1	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	1	0
23	ES-0060	CHSPFLR38	-6.00263	37.7014	1	1	0	0	1	1	0	0	0	1	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	0	0	0	1	0
24	ES-0060	CHSPFLR39	-6.00263	37.7014	1	1	1	0	1	1	0	0	0	1	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	1	0	1	0	0
25	ES-0060	CHSPFLR40	-6.00263	37.7014	1	1	0	1	1	1	0	0	0	1	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	1	1	1	0	0	1	0
26	ES-0059	CHSPFLR27	-5.79995	37.9703	1	1	0	0	1	1	1	0	0	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	0	1	0	1	0
27	ES-0059	CHSPFLR28	-5.79995	37.9703	1	1	0	1	1	1	1	0	0	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	1	1	1	1	0	0	0
28	ES-0059	CHSPFLR29	-5.79995	37.9703	1	1	1	1	1	1	1	Na	Na	1	1	0	1	Na	Na	0	1	0	0	0	Na	0	0	0	Na	1	1	1	1	0	1	1	0	0	0	Na	0

GENETICS

LECA Laboratoire d'Ecologie Alpine

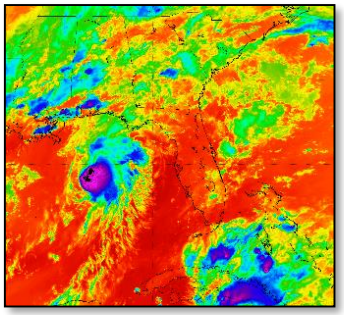
CARDIFF UNIVERSITY PRIFYSGOL CAERDYDD School of Biosciences

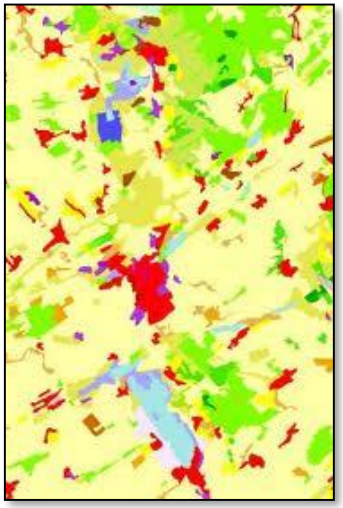
EMBL-EBI European Bioinformatics Institute

…with environmental variables characterizing sampling locations

GEO | **GENETICS** | **ENVIRONMENT**

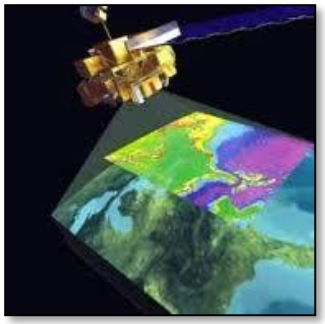
1	farmid	animalid	OARJMP29_allele2_137	OARJMP29_allele2_139	OARJMP29_allele2_141	OARJMP29_allele2_143	e2_145	e2_147	e2_149	e2_151	e2_153	e2_155	e2_157	OARJMP29_allele2_159	OARJMP29_allele2_161	OARJMP29_allele2_163	OARJMP29_allele2_165	OARJMP29_allele2_167	wndjan	altitude	wndfeb	wndmar	wndapr
1044	PL-4005	OAPLPOM25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	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	5.1	22	4.6	5	4.4
1046	PL-4006	OAPLPOM01	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	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	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	5.3	250	4.8	5	4.5
1050	PL-4007	OAPLPOM16	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	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	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	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	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	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	5.4	208	4.9	5.1	4.5





esa

European Environment Agency



NOAA

New sensors and DEMs

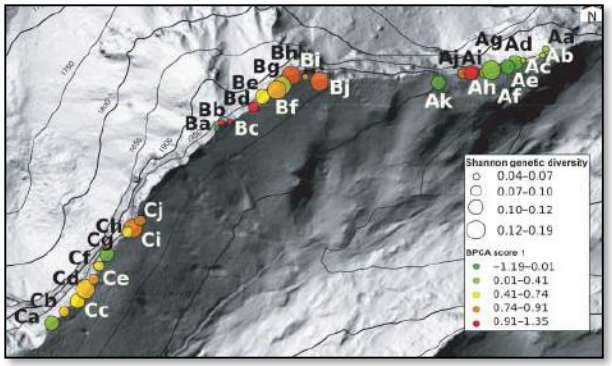
MOLECULAR ECOLOGY

Molecular Ecology (2010) 19, 3760–3772 doi: 10.1111/j.1365-294X.2010.04717.x

Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field

STÉPHANIE MANEL,*† STÉPHANE JOOST,‡ BRYAN K. EPPERSON,§ ROLF HOLDEREGGER,¶ ANDREW STORFER,** MICHAEL S. ROSENBERG,†† KIM T. SCRIBNER,‡‡ AURÉLIE BONIN§§ and MARIE-JOSÉE FORTIN¶¶

Laboratoire Population Environnement Développement, UMR 151 UP/IRD, Université de Provence, 3 place Victor Hugo,



Contents lists available at SciVerse ScienceDirect

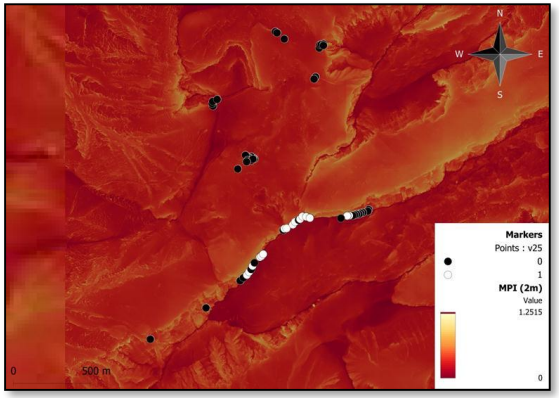
Geomorphology

journal homepage: www.elsevier.com/locate/geomorph

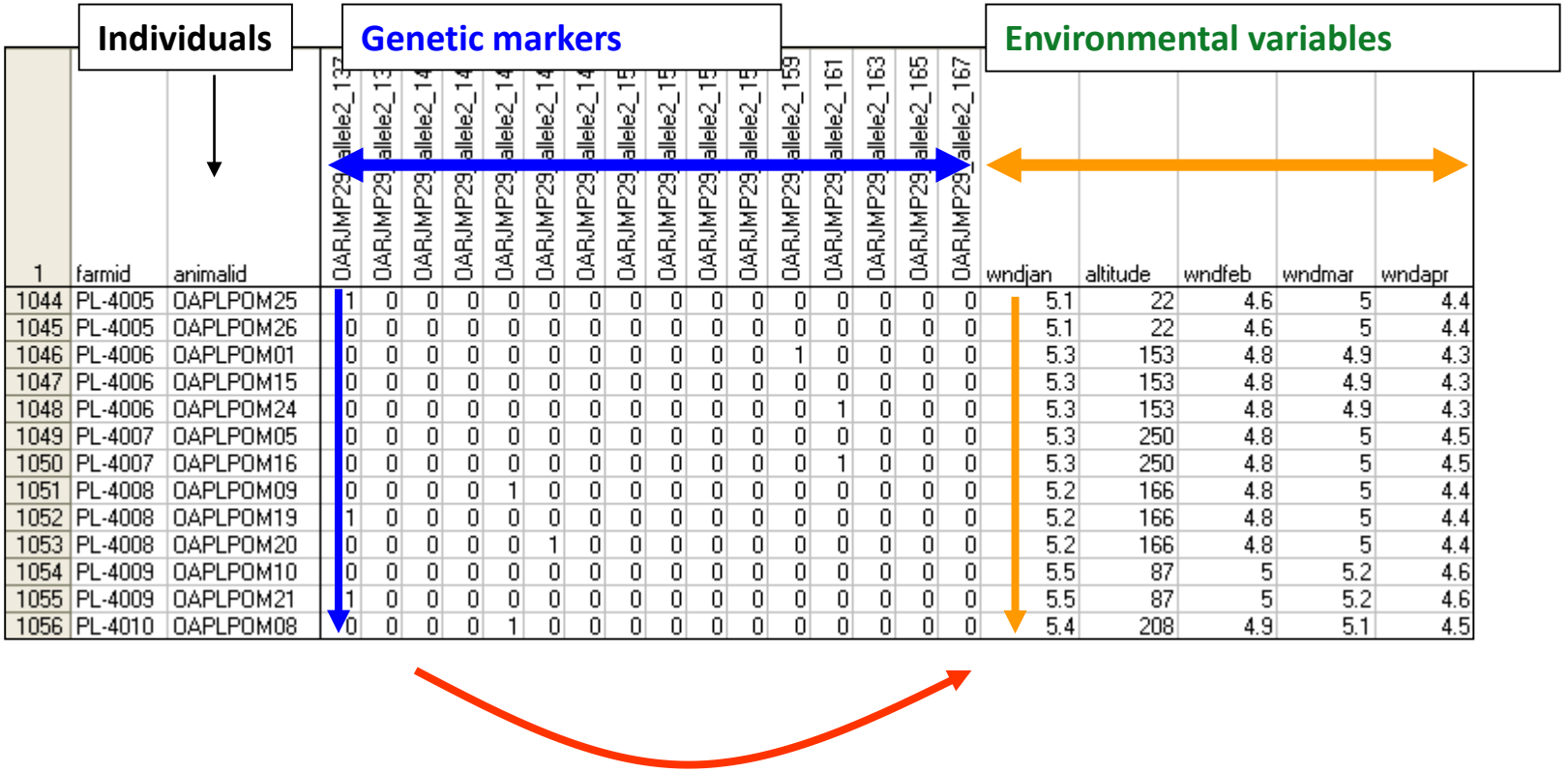
Multiscale analysis of geomorphological and geological features in high resolution digital elevation models using the wavelet transform

Michael Kalbermatten [a], Dimitri Van De Ville [b,c], Pascal Turberg [d], Devis Tuia [a], Stéphane Joost [a,*]

[a] Laboratory of Geographical Information Systems (LASIG), Ecole Polytechnique Fédérale de Lausanne (EPFL), Station 18, CH-1015 Lausanne, Switzerland
[b] Medical Image Processing Lab, Ecole Polytechnique Fédérale de Lausanne (EPFL), Station 17, CH-1015 Lausanne, Switzerland
[c] University of Geneva, Rue Gabrielle-Perret-Gentil 4, CH-1211 Geneva, Switzerland
[d] Laboratory of Engineering and Environmental Geology (GEOLEP), Ecole Polytechnique Fédérale de Lausanne (EPFL), Station 18, CH-1015 Lausanne, Switzerland



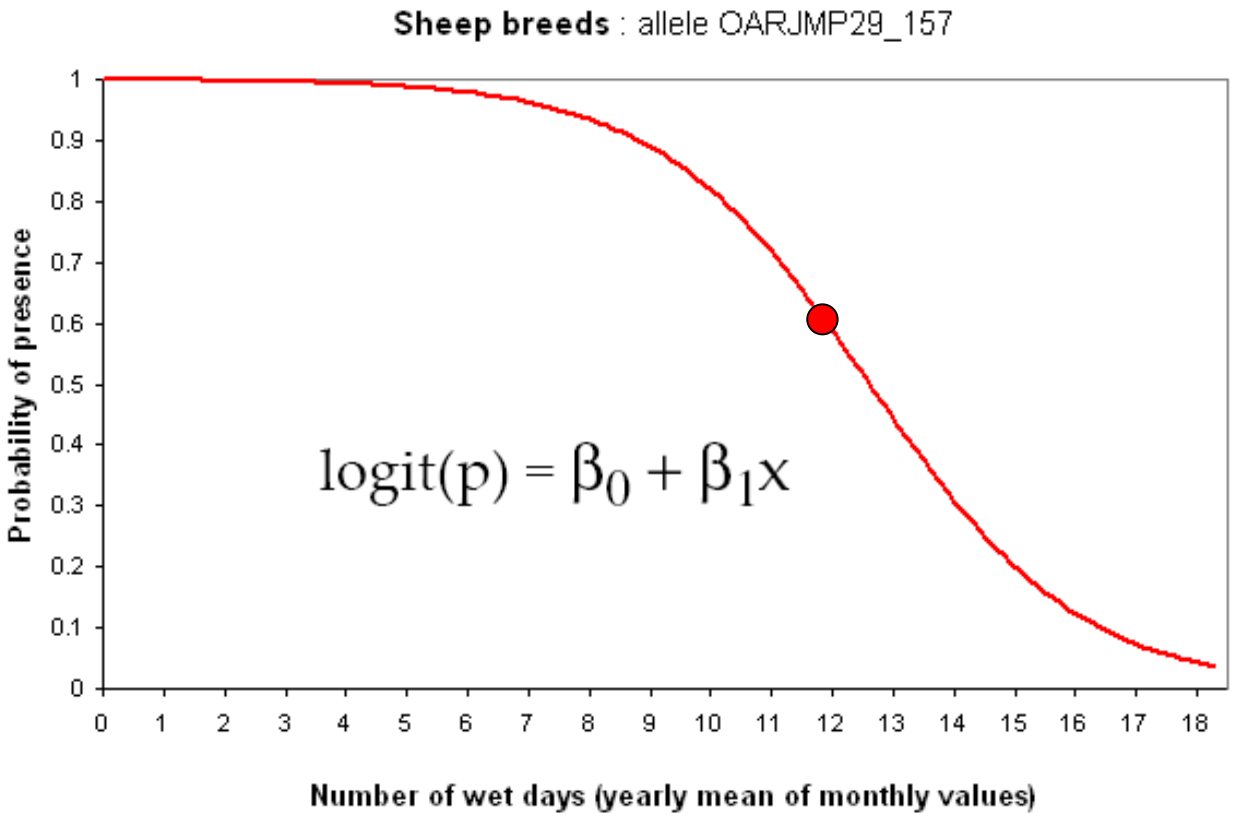
Logistic regression



1	farmid	animalid	OARJMP29_allele2_137	OARJMP29_allele2_13	OARJMP29_allele2_14	OARJMP29_allele2_14	OARJMP29_allele2_14	OARJMP29_allele2_14	OARJMP29_allele2_14	OARJMP29_allele2_15	OARJMP29_allele2_15	OARJMP29_allele2_15	OARJMP29_allele2_15	OARJMP29_allele2_159	OARJMP29_allele2_161	OARJMP29_allele2_163	OARJMP29_allele2_165	OARJMP29_allele2_167	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	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	5.1	22	4.6	5	4.4
1046	PL-4006	OAPLPOM01	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	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	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	5.3	250	4.8	5	4.5
1050	PL-4007	OAPLPOM16	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	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	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	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	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	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	5.4	208	4.9	5.1	4.5

Multiple parallel logistic regressions

Significance of the models



- Does a model including the tested environmental variable significantly explain more variance than a model with a constant only ?

Software developed at EPFL

- **MatSAM v1** (Matlab)
  Univariate logistic regressions on quantitative data, Wald and likelihood ratio (G) statistical tests
- **MatSAM v2** (Matlab)
  Univariate logistic regressions on qualitative data, with pseudo R^2
- **Samβada** (C++) – *Dr Sylvie Stucki*
  Univariate and multivariate logistic regressions, intelligent selection of significant models, pseudo R^2, Moran's I, GWR, HPC capacities

Most recent developments

- Samβada is a C++ development, fast software
- Scythe C++ library
- Samβada processes uni- and multivariate models
- Samβada integrates spatial statistics
- Analysis of the spatial autocorrelation of alleles frequency
- Supports distinction between signatures of selection and demographic processes
- Useful support to identify false positives

R-Sambada

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

Preprocessing

The first function of this category is used to prepare the genomic file. 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 Disiquilibrium (LD) and Missing Rate. Rather than blindly pruning the dataset, an interactive mode will first show the distribution of the considered criterion (e.g. MAF, LD) to illustrate de proportion of lost SNPs if the defined-threshold is applied, thus allowing the user to change the thresholds if too many SNPs are discarded. Then the filtered dataset is transformed into a format that sambada accepts. A second function provides the user with a pipeline to create an environmental dataset out of a file containing the sample location. Unless a raster with an environmental variable is also provided, the program will download climatic and altitudinal variables from global databases (wordclim, SRTM) choosing the needed tiles according to the location of samples. Then a csv file containing the ID of the sample, its location and the associated environmental variable is created. Finally, a third function will prepare the final “environmental file” according to sambada standard. First, too-correlated environmental variables are removed (according to a user-defined threshold). Then the population structure is assessed. It is important to understand here that sambada treats population variables in a similar way as environmental variables, thus explaining the inclusion of population variables in the environmental file. The population structure is assessed using the PCA-based implementation in SNPRelate. If the user prefers to have a membership coefficient to a population rather than the principal components, the result of the PCA can be processed using clustering algorithm. The membership coefficient is then computed as the distance to the cluster centroid. An interactive mode allows the user to see the changes in the clustering according to the chosen k-number of clusters.

R-Sambada

Running Sambada

The C++ sambada code is included in the R package and invoked with an R function. Interestingly, sambada had from the start a module called supervision that was used to split the input file into several files, so that several sambada processes could be started in parallel. Then supervision was again called at the end to merge all result files. The whole chain was tedious, so that in fact supervision, despite its evident benefit, was rarely used. In this R-package we include supervision in the processing chain and run sambada in parallel using the R-package foreach and doParallel. Furthermore, while in sambada population structure could be accounted for by considering it like any other environmental variable. This required a bit of post-processing to adjust the scores and filter out multivariate models that did not contain any population variable. In this release of sambada, population structure can be assessed by specifying in the input parameters the name of the column containing population structure.

R-Sambada

Postprocessing

Sambada computes G-score and Wald-score but does not compute p-values, because the computing of p-values for all models would be time-consuming. Consistently, the first function of the postprocessing includes the computation of p-values for all kept model. Alongside pvalues, the q-values are also calculated based on Storey method. A function can then be called to print a global summary report. This will plot the manhattan plot of all environmental variables and list the most significant markers. Then a connection to the ensemble database is established to list the genes in the nearby region of these markers. Some geographic maps are also provided, showing the distribution of the marker and the environmental variable. The goal with this report is to identify interesting region to further investigate. Once this is done, the user can start an interactive window by specifying an environmental variables and a chromosome (several chromosomes can be handled if the number of kept markers is not too high). This operation relies on the package shiny which opens a local web-browser page first 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. Additionally, the pvalue of the marker with other environmental variables is given. A geographic map is also provided, showing the distribution of the marker, the environmental variable and the population structure. Finally, if the user is only interested in printing a map, it can use the mapping function to plot the geographic distribution of a marker, an environmental variable, a population structure or the spatial autocorrelation of a marker.