Multi-scale landscape genomic models: the role of very high resolution digital elevation models in evolutionary biology

THÈSE N^{O} 6626 (2015)

PRÉSENTÉE LE 26 JUIN 2015

À LA FACULTÉ DE L'ENVIRONNEMENT NATUREL, ARCHITECTURAL ET CONSTRUIT
LABORATOIRE DE SYSTÈMES D'INFORMATION GÉOGRAPHIQUE
PROGRAMME DOCTORAL EN GÉNIE CIVIL ET ENVIRONNEMENT

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Kevin LEEMPOEL

acceptée sur proposition du jury:

Prof. A. Berne, président du jury
Prof. F. Golay, Dr S. Joost, directeurs de thèse
Prof. R. Weibel, rapporteur
Dr P. Orozco ter Wengel, rapporteur
Prof. J. D. Jensen, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2015

The question is, are we happy to suppose that our grandchildren may never be able to see an elephant except in a picture book?

Sir David Attenborough

Remerciements

Voilà, c'est la fin. La fin de 4 années intenses passées en Suisse, et l'occasion de remercier tous ceux qui ont compté pour moi. J'en profite donc pour être le plus exhaustif possible, au risque d'être fort long, mais ce n'est pas tous les jours qu'on termine sa thèse (heureusement d'ailleurs).

Je pense avant tout à mes amis proches, Laurent, Nicolas, Guillaume, Alice, Sarah, et à mon père, ma mère, ma sœur. Je leur dédie cette thèse car, sans leur soutien, je n'aurais jamais trouvé l'énergie pour me lancer dans cette aventure. Vous m'avez souvent manqué (oui ça m'arrive…) et rentrer pour un weekend fut un de mes passe-temps favori avec vous, Xavier, Anne, Emilie, Julie, Cora, Cosme, Jon, Jonas, Sophie, Seb. A défaut d'être reposants, ces weekends furent inoubliables pour moi.

En Suisse, j'ai rapidement fait la connaissance d'un hippie qui fit sa thèse en parallèle avec la mienne. Tim, je n'oublierai jamais l'accueil que tu m'as réservé, tu m'a permis de rencontrer de vrais amis, Cédric, Flo, Hugues, François, Gabi, Xavier, Sylvain, Tristan, Valérie, Laura, Sandrine. Tu m'as aussi trainé en montagne et à la course à pied. Avec un peu de mauvaise foi, je dirais que tu m'as aussi trainé à Sat.

Mes anciens colocs évidemment, Arthur, Julien, et plus tard Filippo. On retiendra plutôt nos Barbecues à nos talents culinaires, les bières sur le balcon aux vidanges envahissantes et le charme du salon à la piaule d'Arthur (et son costume de bain).

A l'EPFL, j'ai rencontré beaucoup d'autres expatriés. Matthew d'abord, à un weekend de ski. Mais c'est surtout autour de bières belges que nous sommes devenus proches. J'ai arrêté de compter les activités qu'il a organisées pour nous. Et nous étions nombreux : Dom, Klas, Elin, Henrike, Clara, Jef, Ksenya, Marina, Sebastian, Mariia et Alina, bien sûr. Alina aura relu ma thèse, me narguant avec son titre de docteur, mais elle aura surtout été une épatante camarade, me soutenant dans les moments difficiles, abusant de son sourire indéfectible pour me remettre de bonne humeur. Alina, tu comptes beaucoup pour moi et j'espère que nous continuerons comme ça longtemps.

Je ne sais pas si le LaSIG est un bon laboratoire, mais en tout cas, il y règne une sacrée ambiance ! Un tout grand merci à Stéphane pour m'avoir cru capable de mener à bien ce travail et pour s'être impliqué si souvent dans mes recherches et publications. Merci à François Golay qui, malgré les innombrables pierres dans son jardin, aura trouvé le temps pour faire vivre ce labo. Enfin, mes nombreux collègues et en particulier Sylvie pour sa gentillesse et son calme éternel, mais également André, Devis, Ivo, Lachance, Magda, Ahmed, Emma, Jessie, Marc, Matthew, Raphaëlle, Estelle, Solange, Robson.

Enfin, je remercie nos nombreux collaborateurs dont j'oublie sûrement une bonne partie : Céline Geiser, Christian Parisod, Stéphanie Manel, Christelle Melodelima, Jonathan Rolland, Dimitri Van de Ville, Daniela Biosa, Benjamin Dauphin, Badr Benjelloun, François Pompanon, Pablo Orozco ter Wengel.

Merci à tous.

Abstract

Environmental heterogeneity is one of the main actors of biodiversity and species adaptation as it exerts a selective pressure on observable characteristics of living organisms. Consequently, local adaptation favours certain genetic variants and, by doing so, leaves a footprint in the genetic heritage across populations. The identification of these adaptive genetic variations is the main objective of landscape genomics and allows, among other things, to study the role of specific regions of the genome in evolutionary processes. Landscape genomics studies also provide essential information for species conservation and for the prediction of migrations due to environmental changes.

To identify these adaptations to the environment, it is necessary to define a study area where the populations are sampled. However, defining the scale of the study area is not a trivial task. In fact, whether the work is carried out at a local or at a broad scale determines the relevance of environmental factors (climate, soil, topography) and the type of signature of selection that will be observed. In addition, the concept of scale in ecology takes into account not only the extent of the study area but also the pattern and density of the geographic distribution of observations, and the spatial resolution of predictors (environmental variables), which is intrinsically linked to the extent. However, *a priori* indications about the relevance of any resolution over another are rare in the literature and it is therefore essential to question this issue.

In this thesis, we propose a multi-scale landscape genomic framework to identify signatures of adaptation to the environment. This multidisciplinary framework lies at the interface between geographic information systems, spatial analysis, environmental modelling, population genetics and computer science. Specifically, we focus on the relevance of variables derived from Digital Elevation Models (DEMs) and on the application of multi-scale analysis aiming to detect signatures of selection.

We applied this analytical framework to three case studies, comprising four species: *Biscutella laevigata* sampled at a local scale in the Alps, *Plantago major* sampled at a regional scale in an urban environment, sheep (*Ovis aries*) and goats (*Capra hircus*) sampled all across the territory of Morocco. In particular, the case of *Biscutella laevigata* allowed us to evaluate the role of topographic features based on Very High Resolution DEMs and to include DEM-derived variables as predictors in association models to study the adaptation of species to their local environment. On the other hand, the case of Moroccan sheep and goats permitted to include for the first time whole genome sequence data within landscape genomic models.

The results revealed several important findings. First, we showed that micro-climate variability is highly dependent on topographic factors at a local scale and that therefore, DEMs are relevant for understanding species adaptation to a mountainous environment. We also demonstrated that it is

essential to consider the scale of spatial representativeness by assessing DEM-derived variables at various spatial resolutions. Indeed, two out of three case studies showed that the models involving topographic variables were sensitive to changes in resolution.

At the same time, we used independent methods to detect the signatures of selection in order to improve the robustness of these detections. In the case of sheep and goats, we demonstrated the usefulness of high-density genetic data by identifying "peaks of selection" within specific regions of several chromosomes. We then used these regions to identify genes potentially subject to selection to formulate hypotheses about the possible relationship between the genotype and a phenotype.

Finally, we took advantage of the spatial dimension of our case studies and measured the spatial autocorrelation (SA) of the frequency of genetic markers. Our results indicate that SA is stronger for genetic markers subject to selection as compared to neutral markers. These results contribute to a controversial debate on spatial dependence between samples. Indeed, although SA is a source of false associations in statistical approaches, it is also a natural phenomenon that inevitably induces autocorrelation at the level of genetic variants present in significant associations.

In summary, we used several landscape genetics approaches to understand the role of environmental factors in the local adaptation of various species. Our findings mainly provide an important contribution to the understanding and use of scale in landscape genomics, also useful in landscape ecology.

Keywords: Digital Elevation Models, Multi-scale analysis, Very High Spatial Resolution, Landscape Genomics, Whole Genome Sequence Data, Generalized Linear Model, Geographic Information Systems, Spatial Autocorrelation, Temperature and Humidity Loggers, Local Adaptation, *Biscutella Laevigata*, *Plantago Major*, *Capra Hircus*, *Ovis Aries, Nextgen*

Résumé

L'hétérogénéité environnementale est un des principaux acteurs de la biodiversité et de l'adaptation des espèces en exerçant une pression de sélection sur les caractères d'une espèce ou d'une population. Par conséquent, l'adaptation locale favorise certains variants génétiques et, ce faisant, laisse une empreinte dans le patrimoine génétique des populations. L'identification de ces variations génétiques adaptatives est l'objectif principal de la génomique environnementale et permet, entre autres, d'étudier le rôle du génome dans les processus évolutifs. La génomique environnementale fournit également des informations essentielles pour la conservation des espèces et pour la prédiction des flux migratoires liés aux changements environnementaux.

Afin d'identifier cette adaptation à l'environnement, il est nécessaire de définir une zone d'étude au sein de laquelle sont échantillonnés des individus. Pourtant, délimiter l'échelle de la zone d'étude n'est pas évident. En effet, travailler à une large échelle ou à une échelle locale détermine la pertinence des facteurs environnementaux (climat, sol, topographie) et celle des signatures de sélection qui seront observées. De plus, le concept d'échelle en écologie tient compte non seulement de l'étendue de la zone d'étude mais également de la forme et de la densité de la distribution géographique des observations, et de la résolution spatiale des variables environnementales, qui est intrinsèquement liée à l'étendue. Or, des informations disponibles a priori sur la pertinence de l'utilisation d'une résolution par rapport à une autre sont rarement fournies dans la littérature, et il est donc fondamental d'étudier cette question.

Dans cette thèse, nous proposons un cadre d'analyse multi-échelle en génomique environnementale pour identifier des signatures de sélection naturelle dans le but de comprendre le phénomène évolutif de l'adaptation à l'environnement local. Ce cadre d'analyse pluridisciplinaire se situe à l'interface entre les systèmes d'information géographique, l'analyse spatiale, la modélisation environnementale, la génétique des populations et l'informatique. Plus particulièrement, nous nous concentrons sur la pertinence des variables calculée à partir de modèles numériques d'altitude (MNA) – dont une partie à très haute résolution – et sur l'application de l'analyse multi-échelle dans le but de détecter des signatures de sélection.

Nous avons appliqué ce cadre d'analyse à trois cas d'études, comprenant quatre espèces : *Biscutella laevigata* échantillonnée à une échelle locale dans les Alpes, *Plantago major* à une échelle régionale dans un environnement urbain, la chèvre (*Capra hircus*) et le mouton (*Ovis aries*) au Maroc. En particulier, le cas de *Biscutella laevigata* nous a permis d'évaluer le rôle de la topographie dans l'adaptation à l'environnement, grâce à un MNA à très haute résolution; et le cas des chèvres et moutons échantillonnés au Maroc a permis d'utiliser pour la première fois des données génétiques issues du séquençage intégral de 320 individus dans des modèles de génomique environnementale.

Les résultats révèlent plusieurs découvertes importantes. Premièrement, nous montrons que la variabilité microclimatique est fortement dépendante des facteurs topographiques à une échelle locale et que, par conséquent, les MNA sont pertinents pour comprendre l'adaptation des espèces à un environnement montagneux. Nous démontrons également qu'il est fondamental de considérer l'échelle de représentativité spatiale, en évaluant les variables issues d'un MNA à différentes résolutions spatiales. En effet, deux des trois cas d'études montrent que les modèles impliquant des variables topographiques sont sensibles à des changements de résolution.

D'autre part, nous avons utilisé des méthodes indépendantes pour détecter les signatures de sélection, et ce afin d'améliorer la robustesse de ces détections. Dans le cas des chèvres et moutons, nous avons mis en évidence l'utilité des données génétique de très haute densité en identifiant des « pics de sélection » au sein de certaines régions des chromosomes. Nous pouvons alors utiliser ces régions pour identifier des gènes potentiellement soumis à la sélection et pour formuler des hypothèses à propos des relations possibles entre le génotype et un phénotype.

Enfin, nous avons tiré parti de la dimension géographique de ces cas d'étude en mesurant l'autocorrélation spatiale (AS) de la fréquence des marqueurs génétiques. Nos résultats soulignent que l'AS est plus forte pour les marqueurs génétiques soumis à la sélection que pour les marqueurs neutres. Ces résultats contribuent à un débat controversé sur la dépendance spatiale entre les échantillons. En effet, bien que l'AS soit la cause d'associations fausses dans les approches statistiques, elle est surtout un phénomène naturel qui induit inévitablement une autocorrélation au niveau des variants génétiques présents dans les associations significatives.

En résumé, nous avons utilisés plusieurs approches de génétique environnementale pour comprendre le rôle des facteurs topo-climatiques dans les processus d'adaptation locale de plusieurs espèces. Nos conclusions fournissent une contribution importante à la compréhension et à l'utilisation de l'échelle en génomique environnementale, et en écologie spatiale de manière générale.

Mots-clés: Modèle numérique d'altitude, analyse multi-échelle, très haute résolution, génomique environnementale, séquençage intégral, adaptation locale, modèle linéaire généralisé, systèmes d'information géographique, autocorrélation spatiale, capteurs de température et humidité, *Biscutella Laevigata*, *Plantago Major*, *Capra Hircus*, *Ovis Aries, Nextgen*

TABLE OF CONTENT

Environmental heterogeneity shapes adaptation at a local scale

Problematics

Detecting signatures of local adaptation

Environmental heterogeneity is known to be one of the main drivers of species diversity and local adaptation (Darwin & Wallace 1858). Environment exerts a selective pressure on phenotypic traits of living organisms, and by doing so, leaves a footprint of adaptation in the genetic information across populations. Finding these footprints is the goal of many studies that aim to identify genes related to local adaptation and to understand their function. Even though most phenotypic traits are determined by multiple genetic variations, identifying genomic regions potentially under selection helps to localise ecologically meaningful traits (Tiffin & Ross-Ibarra 2014). More importantly, genetic variations also provide essential clues for conservation practices, by discriminating independent groups, localising boundaries in the landscape (Allendorf et al. 2010) as well as by forecasting migrations and adaptations of populations to environmental changes, such as habitat fragmentation or climate change (Segelbacher et al. 2010; Manel & Holderegger 2013).

Mechanisms of adaption can be studied using population genetics approaches. According to population genetics theory (Fisher 1930), it is expected that most of the genetic diversity – genetic differences between individuals of one or several populations – is neutral in an evolutionary sense, meaning that these variations do not affect the survival of individuals (Kimura 1968). At the same time, variations under positive or balancing selection are instead assumed to show respectively higher or lower genetic divergence between populations. These variations under positive, negative or balancing selection are thus typically considered to arise as outliers in a response to environmental pressure, whereas neutral variations are influenced by demographic forces and population history in a similar way (Lewontin & Krakauer 1973; Beaumont & Nichols 1996; Luikart *et al.* 2003). However, despite the power of population genetics approaches to detect environment-related genetic variants, they often fail to provide insights on causes of natural selection that led to adaptation, such as environmental pressures for example (Schoville *et al.* 2012).

In order to identify the environmental factors that are responsible for local adaptation, we need to combine our knowledge on genetic variations with information on the landscape. More importantly, we want to understand how do the major characteristics of the landscape, such as climate factors, topography, habitat fragmentation and landscape structure shape populations through demographic or selection processes (Manel *et al.* 2003; Holderegger & Wagner 2008). These are the main goals of a discipline called landscape genetics. In this discipline, one goal is to

use the environment to screen genetic variations for potential signs of adaptation, without necessarily looking at variations in observable traits. One of the main assets of landscape genetics is that it takes advantage of the geographic component of both genetic and environmental data and makes use of Geographic Information Systems (GIS) to explore and analyse data spatially (Joost 2006). Landscape genetics thus hold a great promise to answer multiple fundamental questions on local adaptation, but could also be useful in conservation perspectives, for example to analyse gene flow in space and time or to locate evolutionary significant habitats (Joost *et al.* 2010). However, this direction still remains mostly unexplored and there are only few examples of conservation management where landscape genetics has been applied in practice (Segelbacher *et al.* 2010).

The operational scale of adaptation and the relevance of environmental data

When studying local adaptation, it is important to define how local a study should be and how local we do expect adaptation to operate. The overall area encompassed by the study, commonly referred to as extent, is a key component of scale and often subject to debate. In fact, it was long believed that a high rate of gene flow - the transfer of alleles or genes from one population to another - would only allow local adaptation at relatively large extents. However, numerous recent studies have shown the opposite, highlighting that local adaptation is common and that local populations could represent important evolutionary units (Hereford 2009; Richardson *et al.* 2014). Because local adaptation is more widespread than previously thought, the genetic diversity that translates into functional traits also remains underappreciated. In addition, conservation efforts should target smaller habitats in order to preserve the full range of evolutionary units. Indeed, local adaptation could create more resilience to abrupt human disturbances, for example resistance to pests or herbicide in agriculture.

However, defining the ecological scale of a study (i.e. the grain and extent to be used) raises important questions regarding the relevance of environmental variables. If local adaptation does occur at a local scale, how fine should the resolution of our environmental variables be? While it is obvious that common climatic variables (the most used environmental data, e.g. WorldClim with 1km spatial resolution) such as temperature or precipitation are applicable to adaptation studies at a regional scale, it is not clear whether they are still relevant locally since other factors could play a more important role but also because their spatial resolution might be too coarse. In addition, the range of scales at which local adaptation operates still remains unknown and depends on many factors such as demography, mobility or dispersal, life history and generation time (Hall & Beissinger 2014). In addition, fieldwork data are only available at narrow temporal scales while adaptation occurs over much longer periods of time (Landguth *et al.* 2010). It also means that, in most cases, it is difficult to obtain fieldwork data for wide-ranging species because of the distance they travel. Until now, most studies that investigate environmental influence on evolutionary processes use available large-scale environmental datasets and do not assess several sources or spatial resolutions. Yet, at a local scale, other factors such as topographic structure are driving environmental conditions encountered by plants (Körner 2003). Typically, variations in topography are derived from Digital Elevation Models (DEMs) and thus, it is crucial to define the appropriate spatial resolution of environmental variables. Therefore, it would be relevant to apply

a multi-resolution approach starting for example with a resolution finer than the average home range of the organism (Anderson *et al.* 2010) and then assessing coarser resolutions. However, using different resolutions raises other concerns regarding the scale at which topographical data represents at best the natural phenomenon studied. Indeed, correlations between environmental features and proxies from DEMs are valid at one scale but may significantly change with resolution (Levin 1992). In fact, most of these variables were developed to approximate features of the terrain and the relation may not hold anymore at a much finer scale, and thus cannot be used to provide an insight on local adaptation.

Objectives

This thesis proposes a multi-scale landscape genomic framework to identify signatures of selection. Particularly, it focuses on the relevance of Very High Resolution Digital Elevation Models (VHR DEMs) and on the application of a multi-scale analysis to detect local adaptation to environment. We expected that VHR DEMs would refine association models in landscape genomics and identify potential candidates of selection. The central question of this work can be expressed as follows:

What can we learn on the adaptation of species to environment by applying a multi-scale landscape genomic framework?

To answer this main question, we explore three case studies that differ between each other on their scale (or extent), species, topography, genetic data and sampling scheme. With their analysis and comparison, we hope to provide a solid ground to answer this main question. Therefore, for each of the following case studies, we analyse the two main components of scale – spatial resolution and extent - through a series of sub-questions. With resolution, we aim to understand what level of detail is necessary in topographic-related variables to detect local adaptation. Our goal is not to find the resolution suitable to any case, but to evaluate how sensitive the model is to a change of resolution and to assess the possibility of deriving an optimal resolution.

To answer this main objective, we ask the following sub-questions:

1. How important is topography to model micro-habitat conditions encountered by plants and to detect signatures of selection?

To answer this question, we assess in the first case study (Chapter 4) whether fine scale topography obtained from a very high resolution DEM can model ecologically relevant features such as temperature, humidity or snow cover to settle their usefulness at a local scale. In a second phase, we perform correlations between these DEM variables and genetic data of an alpine plant to find out if signatures of selections related to topographic heterogeneity can be found at such local scale.

2. Is very high resolution necessary to model micro-habitat conditions encountered by plants?

It is often expected that a higher precision should bring results that are more accurate, but it should not be forgotten that a high amount of details might blur the output signal. In each case study, we produced DEM-derived variables at different nested scales, and establish the same correlations in a multi-scale framework, thus specifying the level of detail at which we detect the most significant signals of local adaptation.

3. How does the relevance of DEM-derived variable vary in function of the extent of the study site and of the mobility of the species?

While we might expect that DEM-derived variables are relevant for local scale studies in alpine plants, their relevance at larger extents and in other environments is unknown. In fact, different scales of study may highlight different environmental variables and spatial distribution patterns. In addition, different species show various dispersal distances or mobility and thus are likely to adapt to climate variability or to different topographical conditions. In Chapters 5 and 6, we apply the same workflow as for the local scale of Chapter 4, but this time at a regional and large scale respectively.

4. Are significant associations identified by means of correlative methods also detected by population genetics approaches?

One may expect that strong patterns of selection, associated with particular genomic loci, could be detected by several independent approaches. However, because these approaches differ by their prerequisites, such as inclusion of population structure, spatial autocorrelation or environmental variables, they appear to be more or less conservative than others are. Therefore, for each case study, we compared loci under selection identified by different approaches in order to understand why these approaches do or do not detect the same genetic markers.

5. How can spatial autocorrelation contribute to the analysis of signatures of selection?

Spatial autocorrelation is at the same time a natural component of landscape features and a source of spurious correlations. However, its quantification in adaptation to environment has been largely omitted in most real case studies. We assessed for each case study the benefits of global and local spatial autocorrelation measurements, and evaluated their fluctuation over continuous distance.

6. How does whole genome sequencing improve the detection of signatures of adaptation?

The main advantage of using WGS data is the ability to assess the exact position of the signatures of adaptation within the genome, and benefit from the large online databases to search for gene functionality and evaluate their concordance with associated environmental variable. In reverse, when the detected gene has not been described yet, future studies may be facilitated by knowing a potential actor of selection on specific genes. In the third case study (Chapter 6), we aimed to detect signatures of selection from independent approaches by comparing their detections on chromosomes and evaluating gene functions of commonly found genetic markers.

Methodology

To facilitate comparisons, each of the three case studies is analysed in a similar way. For each of them, we first obtained relevant environmental variables in order to correlate them with genetic data. To do so, in addition to retrieving climatic data from existing databases, we computed DEM-derived variables at multiple scales and extracted their values at sampling locations. Afterwards, we compared these correlations with population genetic approach, in order to discover common detections and understand discrepancies. We also analysed the spatial and genetic structure of the population in order to characterize the genetic dataset and to identify barriers to gene flow. Finally, it was essential to regroup information in graphs. Indeed, since we use several methods of detection, and because our data are embedded within a geographic context, we thought the best way to analyse the results was to produce graphs grouping general information per method, per association, and for comparisons between methods. Details on this workflow can be found at the beginning of Chapter 3.

The three case studies we propose differ mainly by their extent; starting from a local scale for an alpine plant, to a regional scale for a plant in an urban environment, finishing at a large scale for two domesticated mammal. However, they also differ from each other by several parameters, and thus, do not necessarily use the same variables. In fact, the studied species are different and the spatial distribution of samples is not comparable. In addition, certain methods of detection are not applicable in every case. For these reasons, while we consolidate common methods used in Chapter 3, we analyse each case study separately. In each case study chapter, we provide a description of the general context and purpose, specific parameters of methods applied, results and a short interpretation. Hereunder, we describe the motivations proper to each case study and summarize their characteristics in Table 1.1.

Table 1.1 Summary of the three case studies. Description of the key parameters regarding scale, genetic and environmental data

Species	*Biscutella laevigata*	*Plantago major*	*Ovis aries (OA) & Capra hircus (CH)*
Study site	Les Rochers-de-Naye (Switzerland)	Geneva (Switzerland)	Morocco
Scale	Local	Regional	Large
Study area	1.5 x 0.4 km	13 x 23 km	1160 x 906 km
# of samples	361	464	161 OA & 161 CH
Genetic markers	AFLP	SNPs	SNPs (WGS)
# of loci	266	464	1.7 and 1.8 million
DEM resolution range (m)	0.5 – 8	2 – 64	90 – 2880
Climatic variables	Temperature and Humidity loggers	Swiss Eco-climatic GIS data (Zimmermann & Kienast 1999)	WorldClim
Climatic variables window size (m)	/	25 – 825	1000 – 33000

1.1.1 Relevance of DEM-derived variables to detect local adaptation of the alpine herb *Biscutella laevigata L* in les Rochers-de-Naye (Swiss Alps).

Observing adaptation of alpine plants at a local scale requires fine scale environmental data. However, while it is often expected that a higher precision should bring more accurate results, it should not be forgotten that a high amount of details might blur the output signal. Our first case study takes advantage of very high resolution, not only to model microhabitat conditions with a fine resolution, but also to evaluate the scale dependency of associations between genetic markers and topographic variables.

B. laevigata is a perennial alpine plant that occurs in small patches in warm and dry areas. Our study zone is situated at « les Rochers-de-Naye » (N46°26‘00‘‘ E6°58‘50‘‘) where *B. laevigata* forms a natural hybrid zone between closely related lineages. This case study is unique for several reasons: first, most of the individuals live close to the ridge and in aggregates. Second, observing adaptation at such a local scale requires acquiring fine scale environmental data. Therefore, we acquired and gathered different types of variables for this study (climatic variables, DEM-derived variables, snow cover, and infrared red aerial image) to detect local adaptation of *B. laevigata* at a very fine scale.

The first part of the study investigates the usefulness of DEMs as a surrogate to important climatic variables and highlights their ability to model micro-habitat conditions as those encountered by plants (Leempoel *et al. accepted*; Körner 2003). We modelled the relationship between primary as well as secondary VHR DEM-derived environmental variables (e.g. direct solar radiation, wetness index, vector ruggedness measure) and climatic variables measured in the field. To evaluate further the influence of the spatial resolution, VHR DEM-derived variables showing spatial resolutions of 0.5, 1, 2 and 4 meters were used to assess the goodness-of-fit and the significance of the models. The second part investigates on one hand the genetic dispersal and population structure of the sampled plants and on the other hand attempts to correlate local environmental data with genetic variation. A multi-scale approach was applied here as well on DEMs in order to evaluate the scale dependency of these correlations.

1.1.2 Local adaptation of *Plantago major* L. in the urban environment of Geneva

Green spaces and biodiversity in general have an important direct and indirect economic value. However, the recent fragmentation and anthropogenic pressure on these habitats is increasing and modifies the connectivity and the way species adapt to urban environment. Therefore, the analysis of their population structure, gene flow and local adaptation is primordial in a perspective of conservation (Cushman *et al.* 2006). In addition, despite being well studied in natural habitats, fragmentation is rarely assessed in urban environments where it is known to be faster (Di Giulio *et al.* 2009).

As part of the UrbanGene project on Geneva biodiversity, 464 individuals of the plant *Plantago major* were sampled along five transects departing from the city centre. *P. major* grows in a wide variety of habitat (lawns, along roadsides, areas disturbed by humans) and is naturally present in urban environments. It often grows in compacted or disturbed soils, making it important for soil rehabilitation.

The purpose of this study is to assess whether climate, topography or urban characteristics are important to understand both the population structure and local adaptation to the environment of *P. major*. In fact, the sampled plants are encountered in habitats differencing 1) in topography, with several hills surrounding the city; 2) in climate, with the mountain massifs of the French Alps and the Jura creating local variability of climatic conditions; and 3) in urbanization density from the city of Geneva. These factors can influence both the connectivity of the population and its local adaptation.

1.1.3 Identification of signature of selection in Moroccan sheep and goats using whole genome sequencing and multi-resolution environmental data

Domesticated species are submitted to both human and environmental pressure, leading to locally adapted populations. However, local breeds are nowadays threatened by industrial breeds with better production values but little evaluation of their capacity to adapt to new habitats. This case study is part of the project NEXTGEN (http://NextGen.epfl.ch/), which is the first

project aiming at a comparative analysis of whole genome data and high density sequencing for sheep, goats and cattle. Our purpose in this case is to evaluate the local adaptation of sheep and goats in Morocco, by assessing breeds uniqueness and their adaptive genetic resources. In fact, identification of both neutral and adaptive variation would highlight the necessity of keeping these resources.

For this purpose, sheep (*Ovis aries L.*) and goats (*Capra hircus L.*) were uniformly sampled across Morocco in order to encounter a wide range of environmental conditions as well as to assess the geographic structure of the population. Because the study area is large and that the mobility of these species is considerable, we expect that climatic variables would play a more important role than topography in the detection of signatures of selection. In addition, with two species sampled in a similar way, we may identify similar genomic regions associated with the same environmental variables. Using whole genome sequencing, this project is the first to accomplish a reference genome for goats and opens new perspective in the detection of candidate SNPs and genes that are under selective pressure. Identification of genomic regions under selection is likely to allow us investigating or discovering genes located at, or nearby, detected genomic variations.

Outline of the following chapters

In the following chapters, we will present the state of research in adaptive landscape genetics in Chapter 2, emphasising the theoretical background in population and landscape genetics as well as the recent progresses in the acquisition of environmental data. In Data and Methods (Chapter 3), We present the different environmental datasets we used, the acquisition of VHR DEMs and the methods used to detect signatures of selection. The next three chapters are the three cases studies above-mentioned containing a detailed description of the case studies and their results, as well as an interpretation of their results. Finally, a general discussion is provided in Chapter 7, comparing the three case studies and discussing the hypothesis we made.

Chapter 2 State of research

2.1.1 Discovering potential signatures of selection

Natural selection can occur when individuals are subject to an environmental pressure. This process can either favour individuals who are more adapted to their environment than others or disadvantage those endowed with unfavourable traits (Darwin & Wallace 1858). Concretely, this capacity for adaptation translates in variations between individuals in the genetic code. However, these genetic variants, or polymorphism, are not necessarily related to adaptation. In fact, most variants are neutral in the evolutionary sense, which means that having one variant or another is not beneficial or disadvantageous for those who carry it. What differs between adaptive and neutral, or non-adaptive, traits is the evolution of their frequency through time in a population. In fact, if a variant is highly beneficial, future generations will inherit this variant and it will eventually become fixed in the population (Figure 2.1). On the other hand, demographic processes such as random genetic drift or migration mainly determine the neutral genetic background. In other words, several individuals in a given population will carry identical DNA sequences at certain genomic regions where adaptive pressure is exerted, but random variations instead in the rest of their genome. This is what evolutionary biologists have been trying to quantify for decades: the contribution of natural selection to the overall genetic diversity of a population (Nielsen 2005). Studying adaptation thus involves identifying genomic regions under selection that stand out from the neutral genetic background (Allendorf *et al.* 2010). Many methods have been developed for this purpose over the past decades. They are mainly extended versions of the Fst-outlier approach from Lewontin & Krakauer (1973) and differ mainly in the demographic models implemented, the statistical approach (frequentist or Bayesian) and whether selection is explicitly included in the model (Savolainen *et al.* 2013).

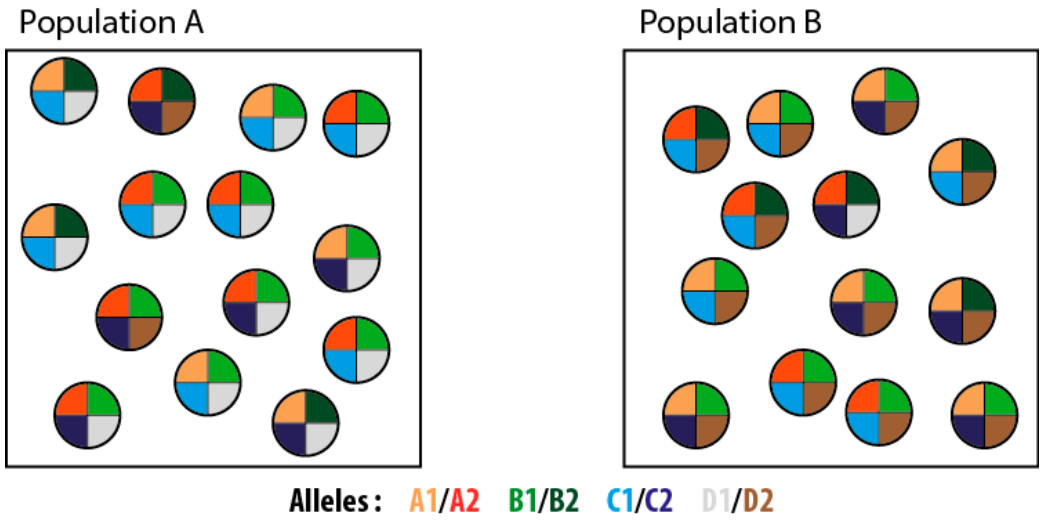


Figure 2.1 Example of outlier detection in population genetics. Each circle represents an individual and each quarter a locus. Colours represent different variants, or alleles. Differences in allele frequencies between two populations are used to create an expected range of differences in frequencies for neutral loci. One allele of a locus (brown) has a higher frequency in population B. Therefore, locus D is detected as an outlier and is a candidate to natural selection.

Environmental features structure genetic variation at both the population and individual levels. Studying how landscape elements affect the distribution of neutral and adaptive genetic variations is in fact the primary goal of landscape genetics (LG), a combination of population genetics (PG) and landscape ecology. LG was originally defined by Manel *et al.* (2003) as a scientific discipline that "aims to provide information about the interaction between landscape features and micro-evolutionary processes, such as gene flow, genetic drift or selection". The purpose is thus to identify genetic discontinuities and their correlations with environmental features such as barriers or environmental clines, and determine to what extent the landscape is involved in the distribution of functional adaptive variation (Schwartz *et al.* 2009; Lowry 2010). Recent studies have also tested isolation by resistance – the permeability trough different habitats – or environment, by partitioning landscape and environmental factors (Hall & Beissinger 2014).

Most LG studies focused on gene flow and on the variation of the genetic structure with respect to habitat connectivity (Storfer *et al.* 2006; Holderegger & Wagner 2008; Manel *et al.* 2010a). However, as emphasized by Lowry (2010), landscape genetics is lacking methodological approaches that would explain how landscape features influence adaptive genetic variations and aim to correlate allele frequencies with the environment in order to understand its effect on the adaptive component of genetic diversity (Holderegger & Wagner 2008). Such methods test for correlation between alleles and environmental variables, where the significant models provide an insight on natural selection (Figure 2.2; Joost *et al.* 2007). The main advantage of this approach is the possibility to screen the genome for selection signals without any a priori hypothesis about

specific loci. Such correlative approaches can be applied either to an individual or population-based design sampled across a heterogeneous landscape, and may rely on the same genetic markers as in large genome scans.

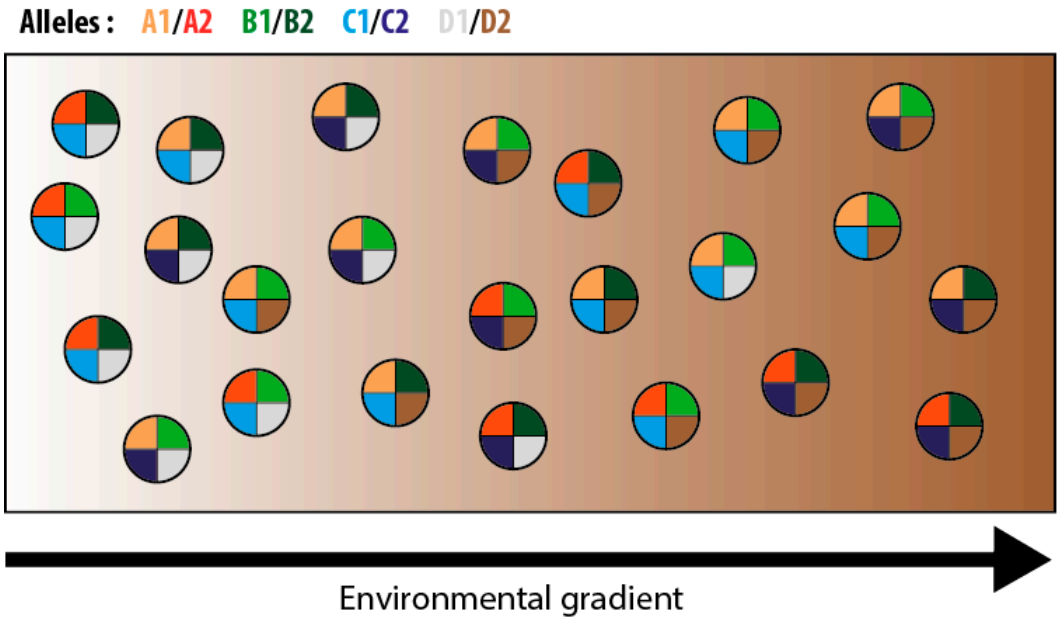


Figure 2.2 Detection of loci under selection using correlative approaches. Each circle represents an individual and each quarter - a locus. Colours represent different variants, or alleles. Here we illustrate how an environmental gradient exerts a selection pressure on grey/brown locus. The brown variant is present close to the brown extremity and grey variant close to the grey extremity of the gradient. Other loci are not affected by the gradient, their distribution depends thus mostly on gene flow.

Benefits of including the geographic component of genetic data

Because genetic information is embedded in a geographic context, geographic information is an important component of landscape genetics. It provides a view of genetic diversity and natural selection processes that complement information obtained from population genetics models. For this purpose, Geographic Information Systems (GIS) are essential to landscape genetics. By using GIS, one can precisely relate genetic variation (within sampled individuals) with geographic coordinates to visualize spatial genetic patterns. Based on these patterns, one can further generate *post hoc* analysis regarding the causes of genetic boundaries. In fact, one of the main advantages of GIS is the possibility to overlay genetic information with layers of physical barriers, landcover or topographical maps (Manel & Holderegger 2013), or to understand the distribution of neutral genetic variation and gene flow. Beyond these arguments, landscape genetics can benefit from exploratory spatial data analysis (ESDA), which permits to identify atypical locations (spatial outliers), clusters or patterns of association. In addition, ESDA is particularly helpful when its tools are interactive, allowing the user to dynamically connect samples on a map to histograms, boxplots or Moran's scatterplot, such as in GeoDa (Anselin *et al.* 2006). Among this collection of methods, spatial autocorrelation is the most important because almost all geographic phenomena show

values similarities with location similarity (Anselin 1998). Spatial autocorrelation, as measured by Moran's I and Local Indicators of Spatial Association (LISA), allow us to identify and localise spatial autocorrelation patterns. For example, Moran's I can be used to estimate the scale of gene flow (Hall & Beissinger 2014). However, spatial autocorrelation is also a paradox: When spatial autocorrelation is observed in regression models, independence assumptions for errors are violated in standard statistical tests (developed in section 3.5.3).

But above all, the main interest of GIS in the context of correlative approaches is to extract values of environmental variables from the geographic coordinates of individuals. Spatial coincidence of these variables can be either acquired by direct measurement in the field or extracted from interpolated and remote sensing data. It is in this GIS environment that Joost *et al.* (2007) proposed the spatial analysis method (SAM) to detect adaptive loci. SAM performs logistic regression in order to identify non-random distribution of genetic variants and their relationship with environmental variables. It has been originally used in the large pine weevil (*Hylobius abietis*) and in sheep (*Ovis aries*) to identify commonly detected loci with PG methods, highlighting potential drivers of selection. Several other recent studies used spatial patterns of selection to study adaptation (Manel *et al.* 2010b; Poncet *et al.* 2010). The advantages of correlative approaches can be summarized as follows: i) independence from population genetic assumptions such as Hardy-Weinberg, ii) operational unit at the individual or population level, iii) identification of candidate environmental variables of selection, iv) quantification of correlations with genetic markers (Joost *et al.* 2007).

The trouble with population structure

Regardless of the outlier detection methods used, the study of the population structure is a major challenge. Indeed, patterns of selection can be similar to several demographic effects, thus leading to false detections of adaptation, also named false positives. Many methods in fact assume that neutral regions of the genome will freely move between populations via gene flow, while loci under selection will show higher genomic divergence across habitats. Such structuring demographic processes are bottlenecks or drift for example (Figure 2.3). However, these demographic processes can lead to patterns that are similar to selection. In fact, several studies have shown that both types of outlier detection methods show a high rate of false positives, especially correlative approaches (Pérez-Figueroa *et al.* 2010; De Mita *et al.* 2013). It is thus essential to analyse the genetic structure of populations in addition to underlying selection, and we rather talk about signatures of selection or loci possibly under selection because of this confounding effect (Nielsen 2005; Lowry 2010).

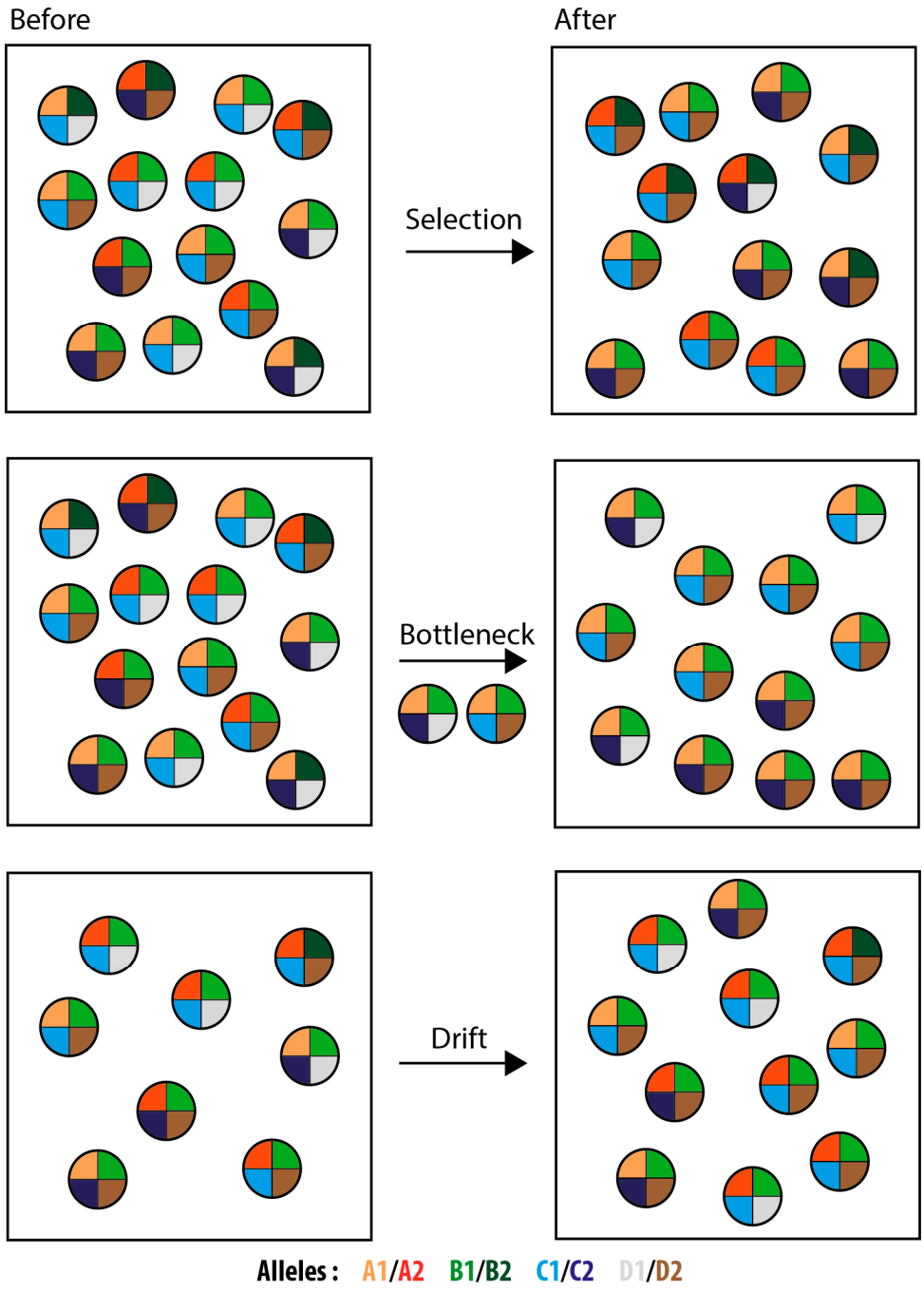


Figure 2.3 Schematic interpretations of different demographic processes within a population that lead to patterns similar to selection. Each circle represents an individual and each quarter a locus. Colours represent different variants, or alleles. Both bottleneck and drift can produce similar patterns on the brown variant as when it is under selection. Bottleneck involves a drastic reduction in population size and recolonization by few individuals, thus reducing genetic diversity. Genetic drift is the change of allele frequency due to stochastic processes. Alleles are here eventually lost or fixed, especially with small or moderate population sizes.

Differences in sampling design

It is not trivial to define what a population is. Indeed, there is no consensus on a definition of populations and there are numerous versions existing in the literature (Waples & Gaggiotti 2006). More importantly, defining a population can yield bias by creating subjective groups of individuals, which, in turn, could have management and conservation implications (Allendorf *et al.* 2010; Segelbacher *et al.* 2010). Landscape genetics tends to use more and more individuals as the operational unit and therefore does not require discrete populations to be defined in advance (Manel *et al.* 2003). This certainly eliminates the population assignment bias, but of course, creates an incompatibility with PG studies. In addition, a good sampling strategy in LG requires individuals to be sampled continuously through the geographic area of the study and not only in several previously-identified discrete locations. Finally, individual-based sampling is essential to fine-scale genetic studies in order to detect discontinuities and localize environmental features that serve as barriers to gene flow (Allendorf *et al.* 2010). It results in the fact that LG can use other statistical approaches. It is indeed important to consider spatial dependence between sampled individuals and do the sampling at a geographical scale small enough to test for spatial autocorrelation in genetic data (Hall & Beissinger 2014). Manel *et al.* (2003) insisted on the necessity to use specific statistical tests for LG data such as spatial autocorrelation and correlograms, interpolations, clustering approaches, PCA or mantel tests.

Another difference between PG and LG is that the sampling design should be stratified across important environmental variables. Response to environment in this case is the main criteria and is often incompatible with PG sampling strategies (Manel *et al.* 2010a). Sampling scheme should match the spatial distribution of the population, whether it is clustered or continuous along a gradient. Otherwise, it can fail to detect genetic differences caused by landscape features (Manel *et al.* 2012a).

Recent advancements in landscape genetics

Landscape genetics evolved with the technical advancement of modern genomics but especially with the development and adoption of statistical and modelling approaches. The former have in fact flourished and became more complex. Landscape genetics tools have started to include population structure, landscape and historical ecology, niche modelling and conservation (Petren 2013; Manel & Holderegger 2013). These new methods aim to correct the high rate of false positives and to process large datasets in a reasonable amount of time (Manel & Segelbacher 2009). Among them we can cite Bayenv and LFMM (Coop *et al.* 2010; Frichot *et al.* 2013), some of which are detailed later on. Criticism has also been expressed concerning the Mantel test, one of the most used methods for correlating environmental features and genetic similarity. Mantel's elevated type 1 error rates and the non-independence of response and predictor variables was observed in multiple independent studies (Balkenhol *et al.* 2009; Manel & Holderegger 2013). Later it was even suggested by certain groups to stop using mantel test and prefer linear correlations, regressions and canonical analysis (Legendre & Fortin 2010; Bolliger *et al.* 2014). They also sug-

gested incorporating the covariance of allele frequencies, unmeasured environmental variation with Moran's Eigenvector maps or demographic effects in mixed effect models (Bolker *et al.* 2009; Manel *et al.* 2010b).

The emerging number of different methods that aim to detect loci under selection highlights the necessity to compare them systematically especially since they produce conflicting results (Bolliger *et al.* 2014). Only few studies have applied, as far as we know, both types of approaches (i.e. correlative and population genetic approaches) to the same dataset. In addition, previous studies used only a limited number of methods on simulations (Pérez-Figueroa *et al.* 2010; De Mita *et al.* 2013; Jones *et al.* 2013) or empirical data (Bonin *et al.*, 2006; Parisod & Joost, 2010).

A final important aspect regarding the detection of loci under selection is a rapid increase in the genetic markers available. Indeed, we now have to deal with datasets of ≈100k SNPs for hundreds of individuals and, consequently, computation time of statistical tests becomes a limiting factor. However, genomics is not only an increase of dataset sizes, it also increase the quality and density of information. Particularly, when the species or a related species as been fully sequenced, we can easily identify the location of each single nucleotide polymorphism (SNP) within genome, judge if it falls into a coding region and thus identify if the SNP has a direct functional role (Schwartz *et al.* 2009). However, most LG studies have so far used neutral genetic markers such as microsatellites (Bolliger *et al.* 2014) or Amplified Fragment Length Polymorphisms (AFLPs), with the disadvantage of not providing information on their genomic localisation. Today, SNPs are the most promising genetic data and allow us to quickly spot regions of the chromosomes where selected loci are identified (Manel & Segelbacher 2009). In addition, genomics is increasing the precision of approaches that require neutral loci by genotyping thousands of them or more and easily excluding those under selection (Allendorf *et al.* 2010). Nevertheless, the increasing size of datasets is thus one of the main advantages of simple correlative approaches because they are capable of identifying outlier loci from the vast genomic background in a few hours, in contrast to complex methods requiring a neutral simulation model. These recent developments in the detection of adaptive loci in large datasets are grouped under the term landscape genomics (Luikart *et al.* 2003; Joost *et al.* 2007; Schwartz *et al.* 2009; Guillot *et al.* 2014).

2.1.2 Recovering relevant environmental information

One way to collect environmental data is to measure directly in the field, but this is costly and time-consuming, especially for large-scale studies. In addition, it often involves measurements at specific time periods that are not necessarily the most appropriate for studies on evolutionary processes. Therefore, most studies depend on climatic variables interpolated at large geographical scales from weather stations distributed across territories of interest, such as the WorldClim dataset (Figure 2.4; Hijmans *et al.* 2005; http://www.worldclim.org/current). These data are often delivered in continuous grids and their spatial resolution typically varies between 1km and 10 km. This inevitably results in a multi-scale problem when integrated in GIS, as generalization and aggregation of the data will occur (Joost *et al.* 2010).

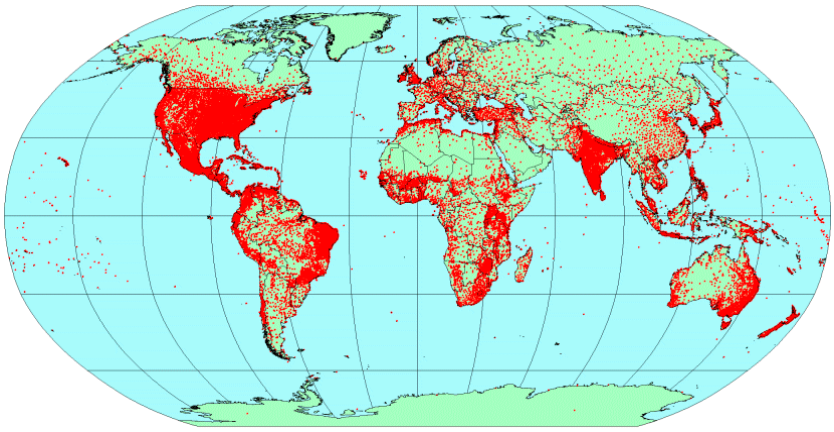


Figure 2.4 Global distribution of climate stations (red dots) used in WorldClim datasets to interpolate precipitation. From http://www.worldclim.org/methods

Environmental variables over large grids can also be retrieved through the interpolation of long-term weather station data combined with DEMs in regression models. The main advantage of this approach is the possibility of detecting the heterogeneity with respect to topography (Figure 2.5). In order to compute accurate regression models from such combined datasets, it is preferable to work with weather stations encountering different climatic regimes (Zimmermann & Kienast 1999). Datasets at 25m resolution are important examples of this method and their data have been proven useful to many studies, including studies in landscape genetics (Manel *et al.* 2010b; Parisod & Joost 2010; Jones *et al.* 2013; Fischer *et al.* 2013).

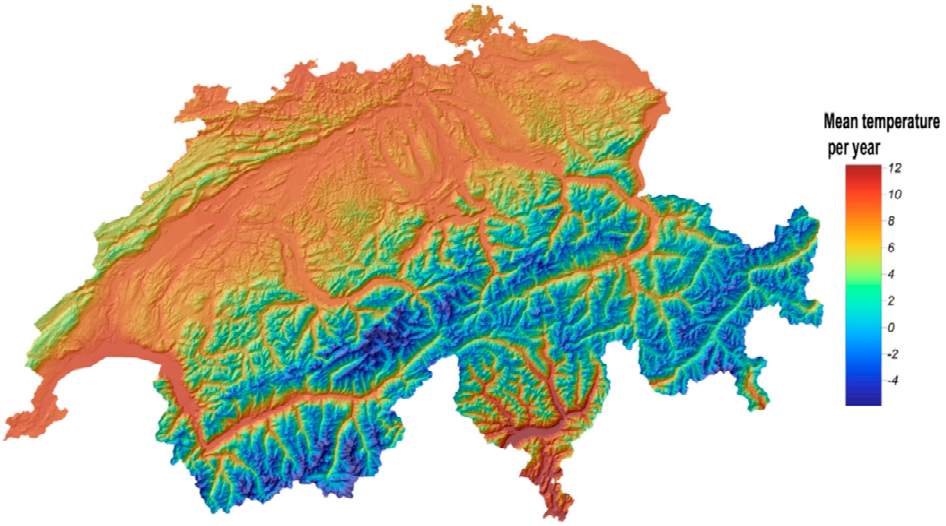


Figure 2.5 Example of a product from the Swiss Eco-Climatic GIS data dataset (http://www.unil.ch/ecospat/en/home/menuguid/tools--data/data.html). This map shows the mean annual temperature interpolated over Switzerland using weather station data and a digital elevation model.

Alternatively, environmental data can be obtained from Digital Elevation Models (DEMs) (Figure 2.6). The most common use of DEMs in landscape and evolutionary ecology consists in retrieving altitude or computing primary terrain attributes (i.e. slope, aspect and curvature), which underlie biophysical processes at local or regional scales, especially in mountainous areas (Guisan & Zimmermann 2000; Kozak *et al.* 2008; Manel *et al.* 2010a). DEMs appeared in the 50s, have become increasingly popular in the 1980s but were available only for some countries (Miller & Laflamme 1958; Moore *et al.* 1991). However, they had the disadvantages of being costly and had a coarse resolution. Since, their increasing accuracy, resolution and availability turned them into accessible indicators of topographic variability, though not necessarily those with the highest predictive potential (Guisan & Zimmermann 2000; Lassueur *et al.* 2006; Kalbermatten *et al.* 2012) .

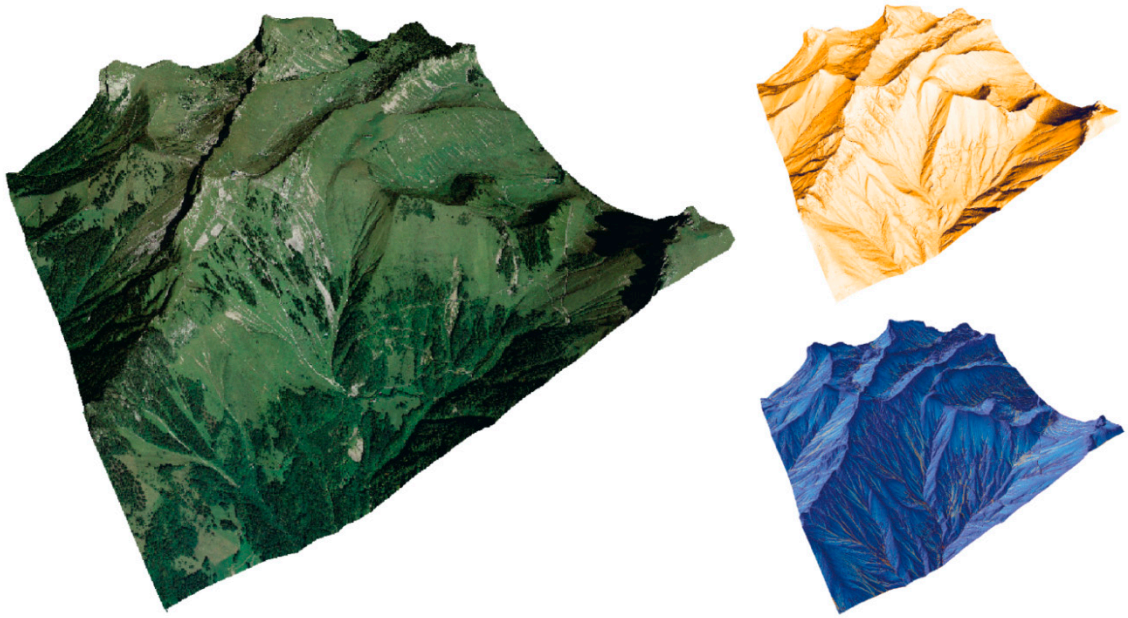


Figure 2.6 Example of a High Resolution Digital Elevation Model and several variables derived exclusively from the DEM. The first image shows an aerial image draped on a 3D representation of the DEM. The two images on the right are the Total insolation in September (top-right) and the catchment area (bottom-right)

A large variety of DEM-derived variables can be computed. Conventionally, primary terrain attributes are calculated from the directional derivatives of the topographic surface (Wilson & Gallant 2000, Chapter I; Böhner *et al.* 2002). In many studies, primary attributes have been used as proxies for factors such as solar radiation (Fu & Rich 2002), evapotranspiration (Guisan & Zimmermann 2000), overland and subsurface flow (Broxton *et al.* 2009), soil water content (Moore *et al.* 1991), snowmelt (Lyon *et al.* 2008), wind, erosion/deposition rate, or soil characteristics (Wilson & Gallant 2000). However, more complex variables have been developed over the last two decades to model hydrological processes, solar radiation or local morphometry directly (Wilson & Gallant 2000; Kalbermatten *et al.* 2012). Named secondary topographic attributes, they are a combination of primary attributes and more complex neighbourhood: solar radiation for example combines slope, aspect, sunshine duration and adjacent relief. A higher predictive power of secondary topographic attributes such as wetness indices (Beven & Kirkby 1979), stream power (Moore *et al.* 1991), terrain ruggedness (Riley *et al.* 1999) or temperature (Wilson & Gallant 2000) may be of

particular interest for assessing ecological patterns related to specific processes at a landscape scale. For example, Böhner & Selige (2006) used two secondary topographic attributes - a wetness index and a solifluction index - to predict soil pH and snow cover. Secondary topographic attributes have also be developed for specific ecological purposes, such as differentiating bighorn sheep habitats across different mountain ranges using the Vector Ruggedness Measure (VRM) developed by Sappington et al. (2007). They found that the correlation between the commonly used Terrain Ruggedness Index and slope was indeed very high at their study site and the production of VRM improved the description of relief heterogeneity as well as of habitat differentiation. Despite these convincing examples, DEM-derived variables remain underexploited and a better understanding of their ecological relevance is thus necessary for a broader usage in studies dealing with natural landscapes (Manel *et al.* 2010a; *Leempoel et al.* accepted, Appendix II).

Although newly developed DEMs come with finer resolution and higher accuracy, their relevance to provide results that are more accurate is unknown. In particular, to what extent high resolution likely evidence micro-relief and related micro-climate physical phenomena that may not be grasped at coarser resolutions remains poorly known (Levin 1992; Marceau & Hay 1999; Cavazzi *et al.* 2013). Indeed, no consensus has emerged yet on the benefits and drawback of high resolution and this is well illustrated by the multi-resolution approaches of Pradervand *et al.* (2013) that did hardly improve species distribution models of alpine plants at a regional scale, although the distribution of some plants known to live in microhabitats was significantly better predicted. In addition, the accuracy of sampling's georeferencing could result in inappropriate extractions of environmental values at sampling positions thus creating artefacts in the models.

Therefore, evaluating the influence of scale on computation of environmental variables is essential. In fact, geomorphological structures naturally constitute a continuum multi-scale and characterizing landscape processes at a single scale is far too simple and requires a multi-resolution analysis (Wilson & Gallant 2000). However, geomorphometric variables (e.g. slope, aspect, watersheds, wetness indices etc.) are complicate to interpret because they are often tested on DEM at one relatively coarse resolution in order to correlate with particular feature of the topography, but these relations may not hold at finer resolutions (Gallant & Hutchinson 1996; Marceau & Hay 1999).

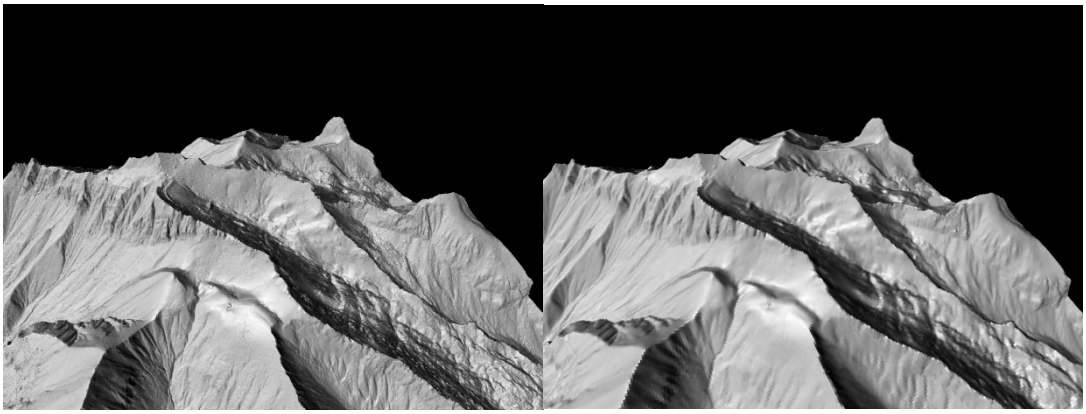


Figure 2.7 A Visual analysis of a shaded DEM at different scales shows different features. The first image is generated on the basis of the initial DEM at 0.5m resolution while the second is taken from the generalized DEM at 8m.

As mentioned above, high-resolution images do not imply better results but at least they must contain information from larger scales (Figure 2.7). Authors have thus proposed generalisation techniques in order to evaluate the impact of resolution on the computation of variables. For example, Wood (1996) proposed a multi-scale version of the first and second derivatives of the surface by enlarging the local neighbourhood. Consequently, the computation of these local indicators translates to more general features of the DEM. More recently, Kalbermatten (2010) proposed a multi-scale analysis framework of DEMs in the frequency domain in order to visualise and extract topographic features at different nested scales. For this purpose, Fourier transforms could not be used due to their stationarity in space. Therefore, Kalbermatten (2010) explored wavelet transform to perform the signal processing tasks. This scale-driven process delimitates scale intervals, successively filtering the low-pass information contained in the DEM and reconstructing high-resolution images using the high pass coefficients only. The transform obtained gives a high positive value when the local signal matches the wave in shape and dimension and a low value when it does not. While Wavelets have been used in landscape ecology to characterize landscape structure at different scales (Dale & Mah 1998), it was the first time they were applied for visual analysis of topographic structures (Kalbermatten *et al.* 2012). Related to signal processing of DEMs, Kalbermatten *et al.* (2012) also proposes a series of terrain indicators obtained from structure tensors (i.e. energy, coherency and orientation). He evaluated their usefulness on the same case study in a multi-scale context (Figure 2.8). While this application is not entirely relevant for our purpose, the DEM generalization we used is based on these development, more precisely on their scaling function.

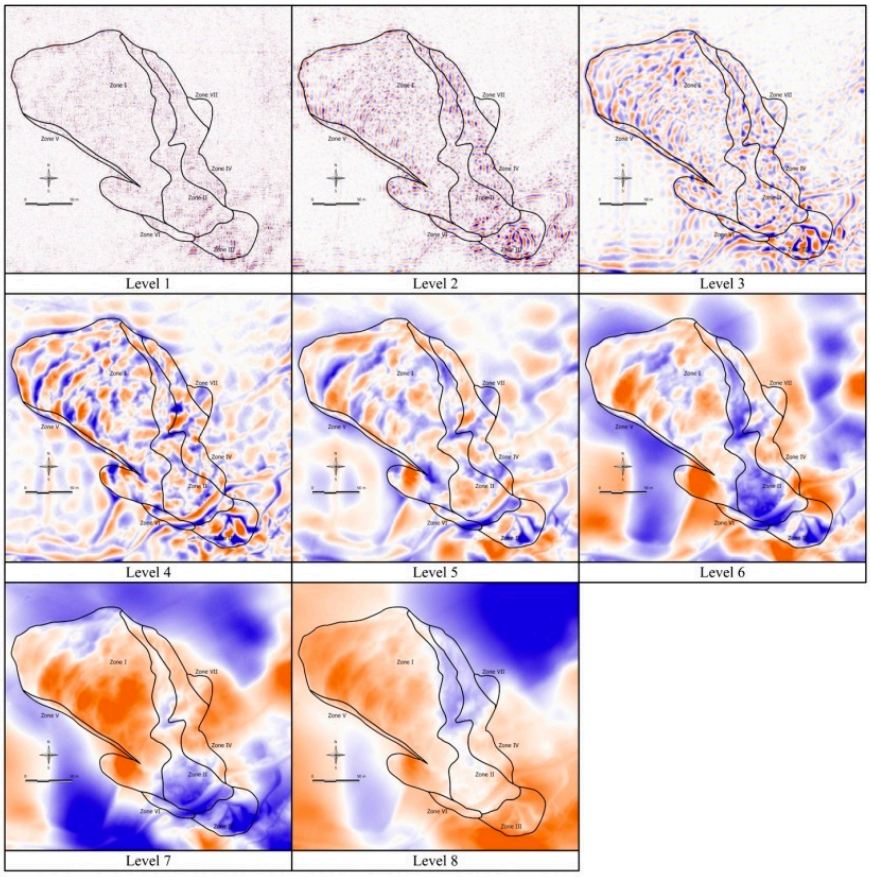


Figure 2.8 Example of multi-scale analysis using the wavelet transform framework on a DEM featuring a landslide from (Kalbermatten *et al.* 2012) *. Each image corresponds to a different level of decomposition where the low-pass coefficient is set to zero and only the high-pass coefficients are used to produce a high-resolution image of concavity and convexity. Taken from* Kalbermatten (2010)

2.1.3 Finding how local is local adaptation

Local adaptation has always been a major focus in evolutionary biology and several definitions exist to describe this process. Local adaptation happens in a geographic area within which movement of individuals regularly occurs (Richardson *et al.* 2014) and a population can be considered locally adapted when it shows a higher fitness than any other populations introduced in this habitat (Kawecki & Ebert 2004).

So far, most studies on local adaptation have focused on identifying phenotypic variation but did not try to explain the underlying processes at the genetic level, which are thus poorly understood (Savolainen *et al.* 2013). In addition, many studies have a particular hypothesis concerning the driver of local adaptation, and others directly assess candidate genes (Savolainen 2011; Manel & Holderegger 2013). However, it is true that obtaining big genetic datasets is costly and thus makes it difficult to identify loci involved in polygenic quantitative traits. In fact, the few that performed such analyses at the genetic level used model species (Fournier-Level *et al.* 2011; Fraser 2013).

For a long time, it was assumed that a high rate of gene flow would prevent local adaptation (Richardson *et al.* 2014). From the theoretical point of view, if populations exchange migrants,

local adaptation is the result of a balance between selection and migration, otherwise they would cancel each other out. Therefore, signatures of selection, despite gene flow, informs us on the strength of selection (Kawecki & Ebert 2004). In addition to gene flow, temporal variations in habitat conditions also act against local adaptation while in contrast, spatial heterogeneity favours the maintenance of polymorphism (Hastings 1983; Kawecki & Ebert 2004).

However, recent theories and examples show the opposite (Richardson *et al.* 2014). Local adaptation is more widespread than thought and it takes place in a wide range of species. Several studies have shown evidences of fine scale adaptation in plants, mice, fish, snails, frogs and many others within a range of distances varying from dozen of meters to several kilometres (Skelly 2004; Vekemans & Hardy 2004; Parisod & Bonvin 2008; Kavanagh *et al.* 2010; Richardson *et al.* 2014). In addition, Many studies have shown genetic variations based on plants phenology along latitudinal clines (Savolainen *et al.* 2013), showing that natural selection can vary spatially and lead to local adaptation along an environmental gradient. It was also observed that clines variation are often stronger in randomly mating species, thus expecting selection to be more efficient (Savolainen 2011). Local adaptation was also shown to arise rapidly and must be due to ongoing or recent spatially varying selection related to differences in environmental conditions (Kawecki & Ebert 2004 and Figure 2.9). For example, many plants species that recolonized inhabitable areas after the last glaciation 10000 years ago show clines in phenology (Savolainen *et al.* 2013).

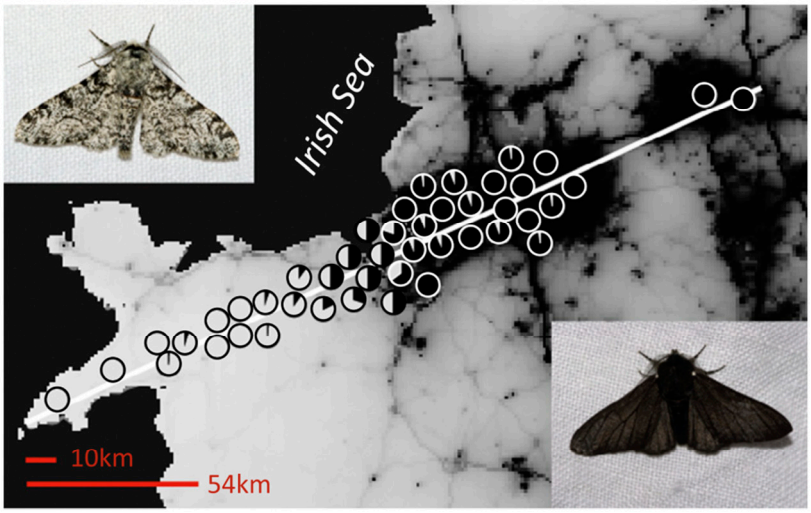


Figure 2.9 Example of a micro-geographic adaptation to a rapid environmental change. The peppered moth, Biston betularia, has a light morph and a dark morph. The map shows the adaptation area, with the colour representing the level of pollution. Circles represent the proportion of moths sampled at a site that were either light or dark. From (Saccheri *et al.* 2008; Richardson *et al.* 2014)

These studies show that local adaptation is frequent but few assessed their findings at different scales. Only a couple of them have considered different extents of study area to study local adaptation. For example, Manel *et al.* (2010b) were able to show that temperature and precipitation are the two main environmental variables that drive adaptation in *Arabis alpina* sampled in the European Alps across different extents.

Finally, among the mechanisms of local adaptation cited in Richardson *et al.* (2014), one is particularly relevant to the topic of this thesis, which is that micro-geographic adaptation often occurs in populations exposed to spatially autocorrelated selection (Figure 2.9). In fact, selective environment are often clustered distributed along clines, leading to positive spatial autocorrelation when effective gene flow occurs from populations facing similar selection.

Chapter 3 Data and Methods

In this chapter, we describe the workflow and methods that were commonly applied to each of the three case studies. Their study areas differ from each other in terms of extent, topography and sampling strategy, thus requiring different environmental datasets and different resolutions. We also describe the methods that were used to evaluate and describe the genetic datasets, to assess population structure and to identify signatures of selection. Finally, we explain the graphical results that were produced in a similar way in each case study.

The workflow is the same for each case study (Figure 3.1). It is summarized here-under and detailed in each case study chapters.

1. *Extracting environmental information*

First, we recovered climatic data from known datasets and extracted their values at sampling locations (e.g. WorldClim). The second type of variables is derived from DEMs that were either acquired from global or local databases. Afterwards, we computed series of DEM-derived variables and extracted their values at sampling locations. To evaluate how much resolution can affect our results, it is crucial to produce corresponding variables at different resolutions either by using a multi-scale approach (for DEM variables) or by using increasing window sizes (for climatic variables). Combining these two sets of variables creates a large dataset in which redundancy between variables must be evaluated and eventually removed some variables before further analyses. Therefore, a subset of variables must be selected based on a maximum threshold of collinearity.

2. *Spatial and genetic structure of the dataset*

We analysed the spatial and genetic structure of each dataset. In fact, it is important to know if the sampling locations are clustered or not, if they belong to different habitats, if gene flow is high and if the genetic structure indicates separate populations. All this information is essential when evaluating the relevance of adaptation signals.

3. *Identification of genetic markers under selection*

Next, markers potentially under selection are identified through several methods. We used two correlative approaches (Samβada and LFMM) to correlate the presence/absence of genetic markers with environmental variables mentioned above. In addition, one population genetic approach (BayeScan) was used, when applicable, to detect differences in allele frequencies between populations.

4. *Graphical illustration of results*

Finally, it was essential to regroup information in graphs. Indeed, since several methods of detection are used and because each dataset is embedded within a geographic context, we thought the best way to analyse the results was to produce graphs representing general information per method, per genetic marker, and graphs for comparisons between methods.

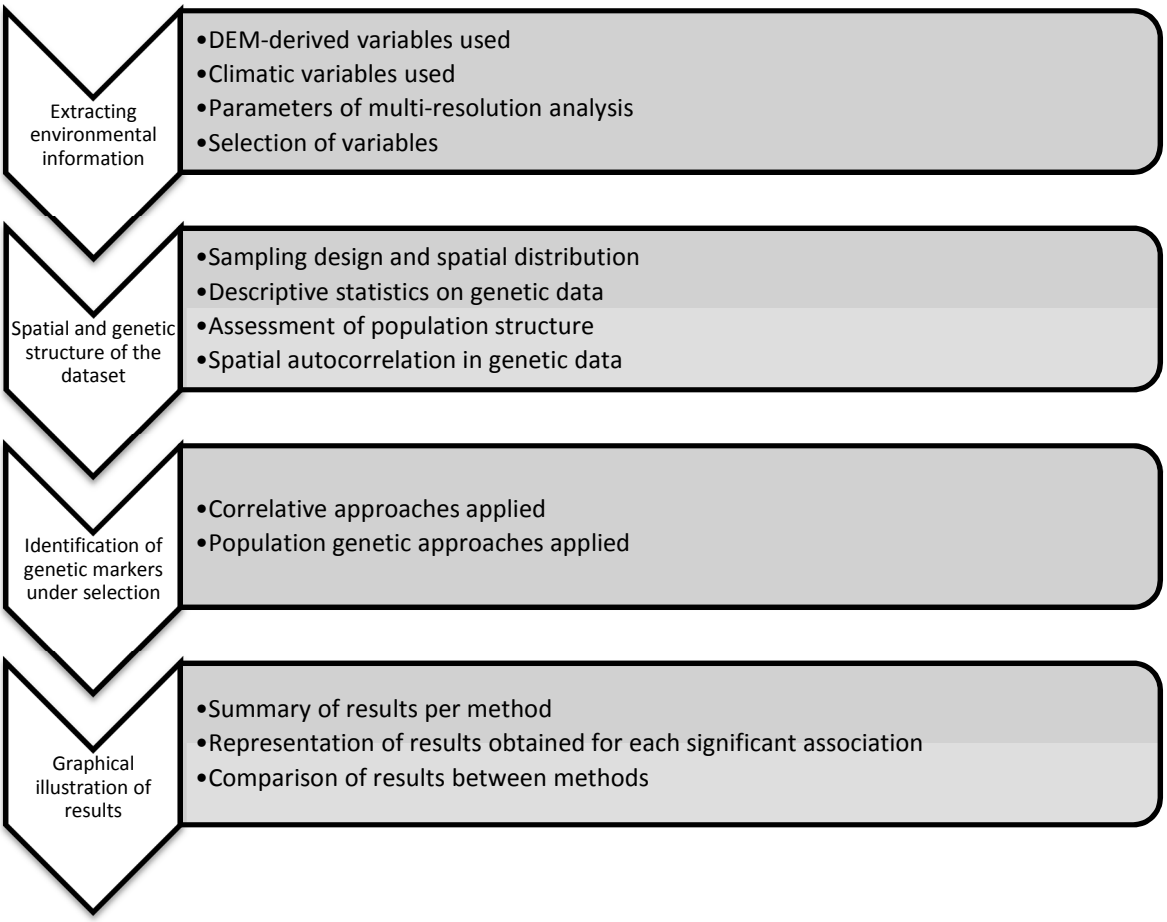


Figure 3.1 Workflow applied to each case study

3.1 Digital Elevation Models

The first section of this sub-section details the methods applied to Digital Elevation Models (DEMs) and the relevance of their use for the different case studies. As mentioned above, each case study requires different DEMs at different resolutions due to the extent of the study area and to the sampling strategy applied. We explain why we decided to acquire very high resolution (VHR) DEMs for the local scale study (*B. laevigata* case study) and develop the treatment and validation of this VHR DEM. Next, we detail multi-scale methods that were applied to each DEM to obtain a continuous representation of topographic features. Finally, we describe each computed DEM-derived variable, and give the parameters used for their computation.

3.1.1 Existing DEMs

The *Sheep & Goats* case study shows the largest extent with samples separated by several kilometres. In this case, existing DEMs with a moderate resolution are sufficient to render on local topography.

To date, existing DEMs with a worldwide coverage have a coarse resolution and limited accuracy. Among them, we can cite the SRTM (Shuttle Radar Topography Mission), which is based on radar interferometry, and the ASTER GDEM, based on stereo-photogrammetry. These models have a resolution of ≈90m and ≈30m respectively and have a poor vertical accuracy (≈15m) (Tachikawa et al. 2011). We decided to use the SRTM instead of the ASTER for several reasons: first, the ASTER model is known to show more artefacts than the SRTM (Tachikawa *et al.* 2011); second, since the study area is quite large, we wanted to avoid long computation time and thus preferred to use a coarser resolution; Finally, SRTM and ASTER GDEM have similar levels of accuracy. The SRTM (void filled) over the territory of Morocco was retrieved in April 2014 from http://earthexplorer.usgs.gov (courtesy of the U.S. Geological Survey).

For the *P. major* case study in Geneva, however, we had to look for models with a higher resolution, since our study area is small (15 x15 km) and our samples are located close to each other (tens to hundreds of meters). National or local administrations typically provide adequate moderate-to-high resolution DEMs. In Switzerland, the federal office of topography (swisstopo) provides a DEM at 25m resolution, based on contour lines of national maps, and a 2m model based on LIDAR (Light Detection And Ranging) with an accuracy of 0.5m; © 2013 swisstopo (JD100064).

However, the administration of the state of Geneva acquired in 2009 a 1m resolution model, with a vertical accuracy of 15cm (http://ge.ch/sitg/sitg_catalog/sitg_services?service_id=27&page=1). This model covers an area slightly larger than the canton territory. However, some individuals were situated outside of this DEM's limits (see Figure 5.7). We had a look at DEMs from the French national geographic institute (IGN) to cover these areas but the resolution of their models is too coarse (25m) for a multi-scale study at that scale. We were thus not able to extract DEM variables values at these sampling locations, which were thus not included in the analysis.

Finally, a finer resolution was necessary for the *B. laevigata* case study. Indeed, a steep cliff located close to sampling locations is a potential source of artefacts in the DEM. Therefore, we considered that swisstopo model at 2m resolution might show significant artefacts because the model processing is automatized and controlled solely over large areas. In addition, because this case study focuses on a local scale, the resolution of 2m might not be accurate enough to grasp microhabitat variability. Therefore, we decided to acquire two very high resolution DEMs using two distinct approaches. The first DEM we acquired using a drone (abbreviated RPOD05). The main advantage of using drones for model acquisition is that they are cheap and easy to manipulate. However, resulting digital-photogrammetry models are typically less precise than LIDAR data and, because they cannot access the terrain under the canopy, these models are not appropriate for forested areas. The second DEM we acquired is a high-resolution Light Detection And Ranging (LIDAR) model (Heli05). LIDAR models have the advantages of being precise and enable the coverage of either a surface or a terrain by filtering the point cloud. However, they remain fairly expensive.

3.1.2 Acquisition and comparison of Very High Resolution DEMs

The RPOD05 model was acquired by the R-pod research group (HEIG-VD) in September 2011 (http://www.r-pod.ch) with the help of the Sensefly drone (http://www.sensefly.com/). The purpose was to capture series of images and build a 3D model using stereo-photogrammetry. The drone thus captured images every 5 seconds according to a flight plan over the sampling area. Afterwards, the research group uploaded these images in the software Pix4D, which uses image matching techniques, to find common points between images and produce a DEM as well as a rectified aerial image, or orthophoto. They obtained a Ground Sample Distance (GSD) of 7.5cm and a longitudinal and lateral covering of 82% and 40% respectively. However, not all flight lines were performed due to strong wind, which resulted in a limited overlap over the ridge. The model acquired has a spatial resolution of 0.5m and an estimated altitude accuracy of 1m.

The HELI05 model was acquired by the Helimap Company using a LIDAR scanner in October 2011 (http://www.helimap.ch/). The LIDAR sensor was mounted on a helicopter, which allowed investigators to produce a very detailed model while flying close to the ground at a low speed. To reduce the costs, we decided to treat the LIDAR point cloud in our lab.

Processing

The laser point cloud obtained from Helimap was filtered in Terrascan (http://www.terrasolid.fi/en/products/terrascan). This software allowed us to visualise and classify points, as well as to produce DEMs with different methods. We first analysed the initial classification of points as obtained with default parameters (Figure 3.2).

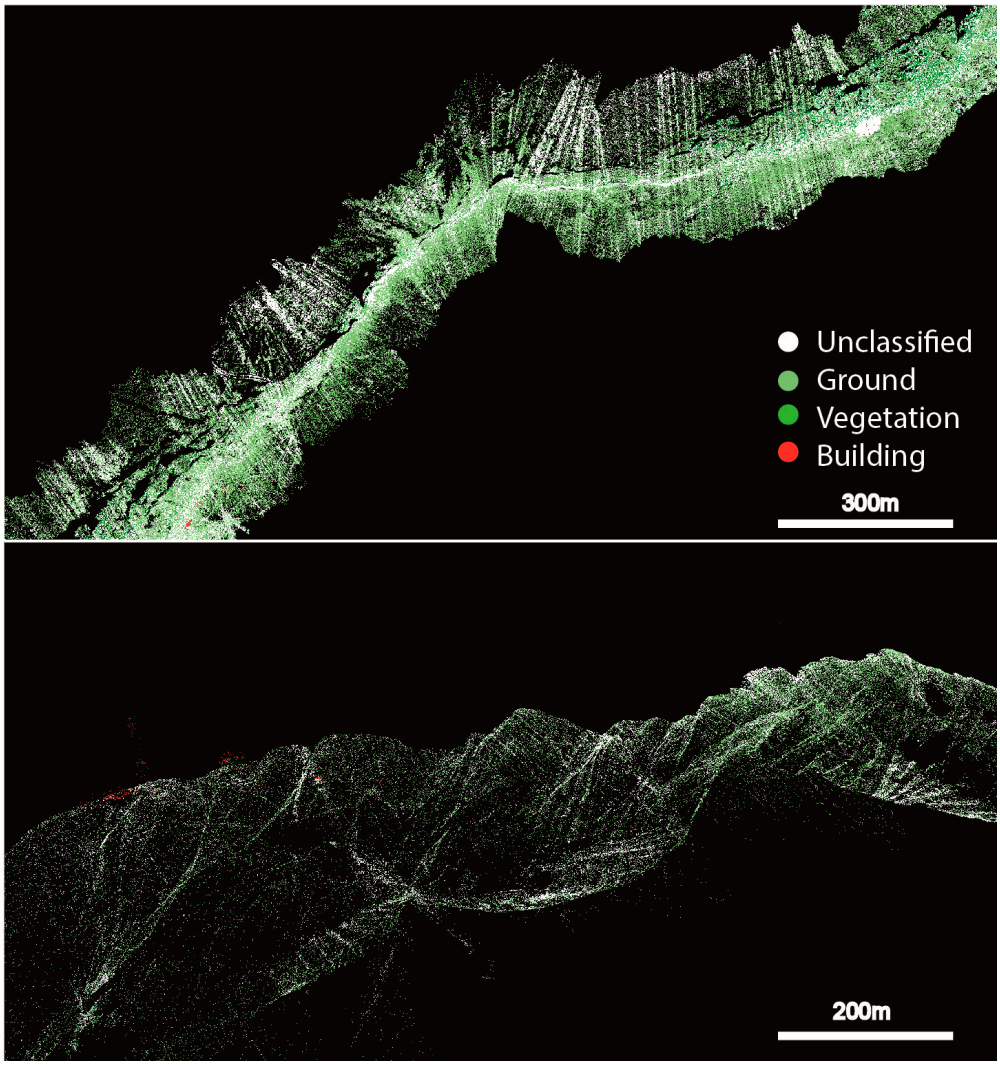


Figure 3.2 Illustration of the LIDAR point cloud on the ridge of "Les Rochers-de-Naye" obtained from Helimap with points classified by category. Both images show the initial classification of points. Top view with oriented to the north (top), 3D view (bottom)

We first noted the presence of void zones (no data) on the northern side of the ridge that are due to the obstruction of the LIDAR sight by the steep cliff. However, this does not cause a problem for the analysis since these voids are far enough from the sampling locations, meaning voids will not interfere in the computation of variables, at least at high resolution. These voids are filled in a further step.

Then we observed that many points are considered as unclassified even though some could belong to the ground class. This problem is often observed in LIDAR models including steep topography and can be solved by changing the classification parameters. Therefore, we performed a reclassification of ground points from the unclassified class to increase the precision of the terrain model. After discussion with the staff of Helimap company, we decided to use the following parameters for this reclassification: Classification maximums: Terrain angle: 90°; Iteration angle: 9° to plane; Iteration distance: 0.2m to plane. Classification Options: Reduce iteration angle when edge length < 0.5m.

In addition, we performed a visual inspection of the ground class and corrected major errors of reclassification by hand.

After these processing steps in Terrascan, we exported the ground class as a separate point cloud in order to estimate the density of points per pixel in SAGA GIS (see section 3.1.4). We wanted to make sure that there were enough points per pixel to guarantee a sufficient precision of the final DEM with at least 3 points per pixel. Initially, we aimed to have a DEM with a resolution of 25cm, but as can be seen in Figure 3.4, the density of such a model is too low close to some sampling locations. Therefore, we opted for a model at 0.5m resolution.

In Figure 3.3, we can see that by choosing the 0.5m resolution, we drastically reduced the amount of pixels with one or two PPP and increased higher densities, thus expanding the histogram. Similarly, in Figure 3.4, we can see that that the ridge is covered more densely.

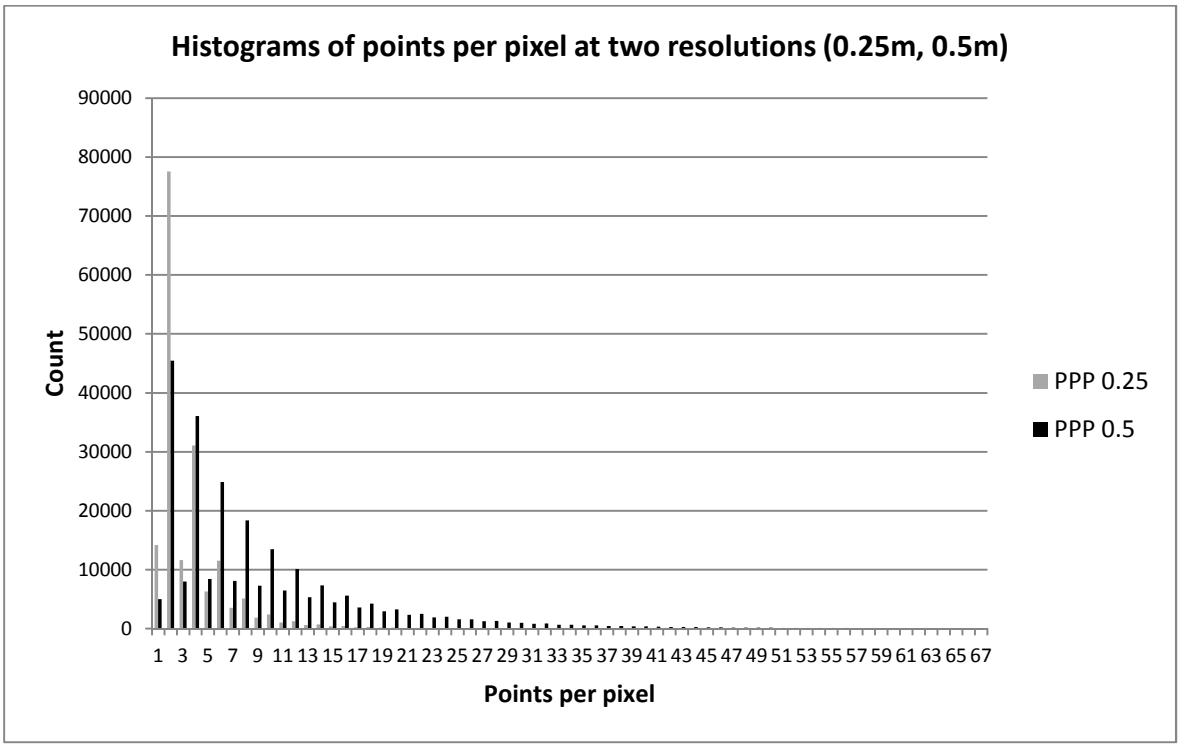


Figure 3.3 Histograms of points per pixel at two resolutions. Comparison of HELIMAP LIDAR point cloud density at 0.25m (grey) and 0.5m (black) resolution. For an expected density of 3 points per pixel, this figure shows that a resolution of 0.25 does not provide results accurate enough.

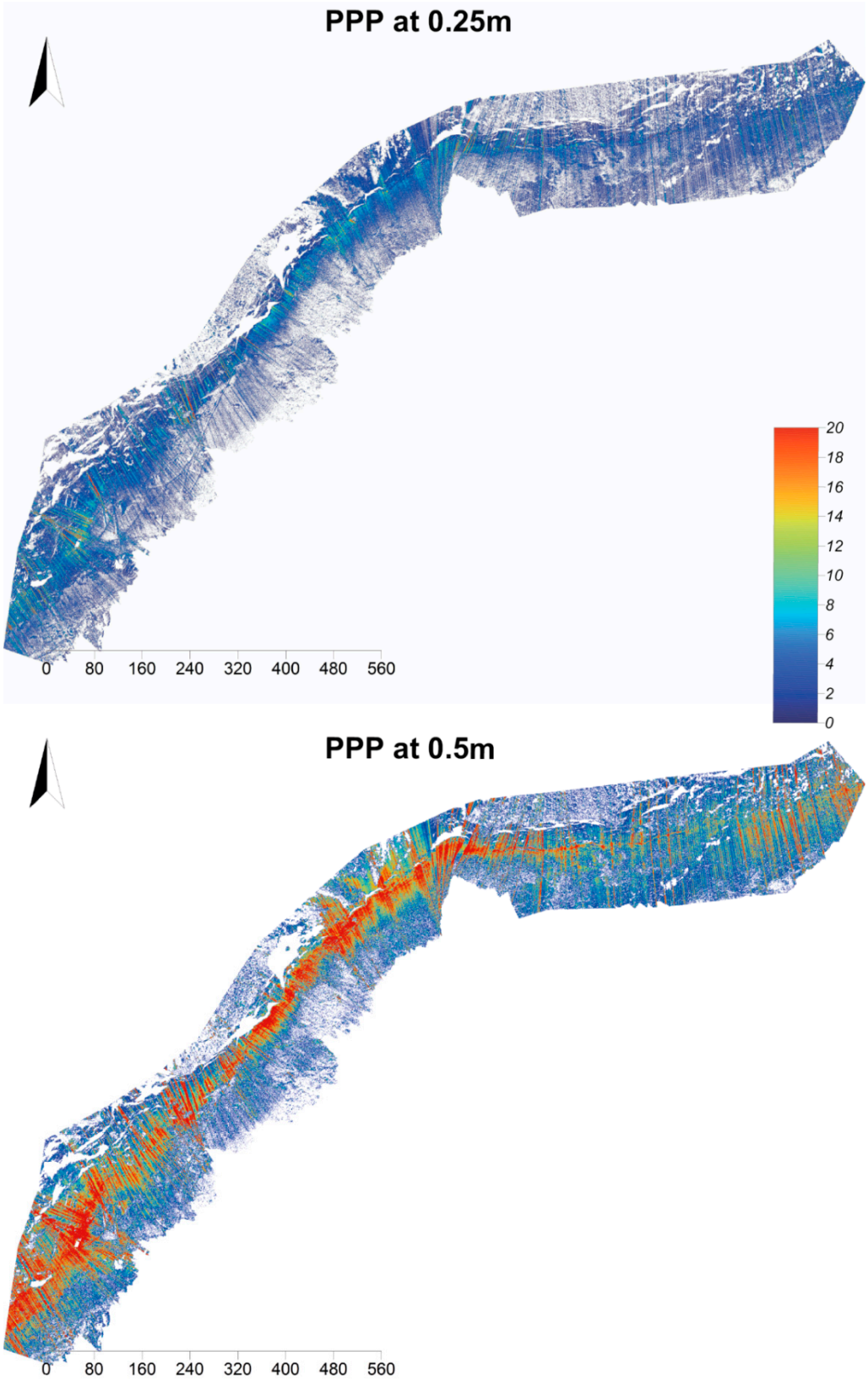


Figure 3.4 Points per pixel at two resolutions (0.25m, 0.5m) on the reclassified LIDAR point cloud obtained from Helimap. Colour scale represents the density and is valid for both images. A density of 0.5m (bottom) shows a better coverage in general and especially for the ridge.

The next step of the analysis workflow consists in exporting the DEM from the point cloud. Again, we used Terrascan to perform this task and chose to export an average per pixel rather than generating a TIN, in order to retain most of the local variability (Figure 3.5).

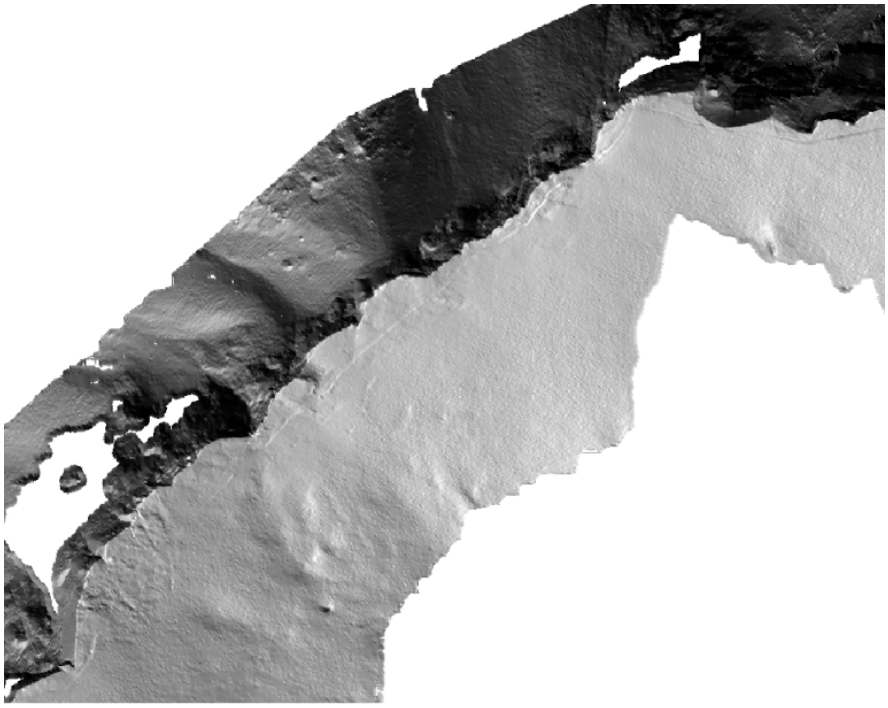


Figure 3.5 Zoom on the ridge from the DEM obtained with average parameter (fill 5pixels). Raw output from Terrascan at a resolution of 0.5m. Voids will be filled later on with a lower resolution model.

The last step of the workflow consisted in filling the voids of the Helimap DEM. To do so, we first resampled the DEM obtained from the state of Vaud. The latter model completes the one from swisstopo (2m resolution) by adding flight lines over 2000m altitude and by increasing the quality of the model itself (five laser pulses instead of one per m² in mountainous areas) (Kleiner *et al.* 2010; http://www.vd.ch/index.php?id=49899). The enhanced model has a resolution of 1m and is resampled at the resolution of the Helimap DEM (0.5m) using the SAGA Multilevel B-Spline Interpolation from Grid. Then we combined the two DEMs with a priority to the Helimap DEM. The resulting filled DEM is hereafter named Heli05 (Figure 3.6).

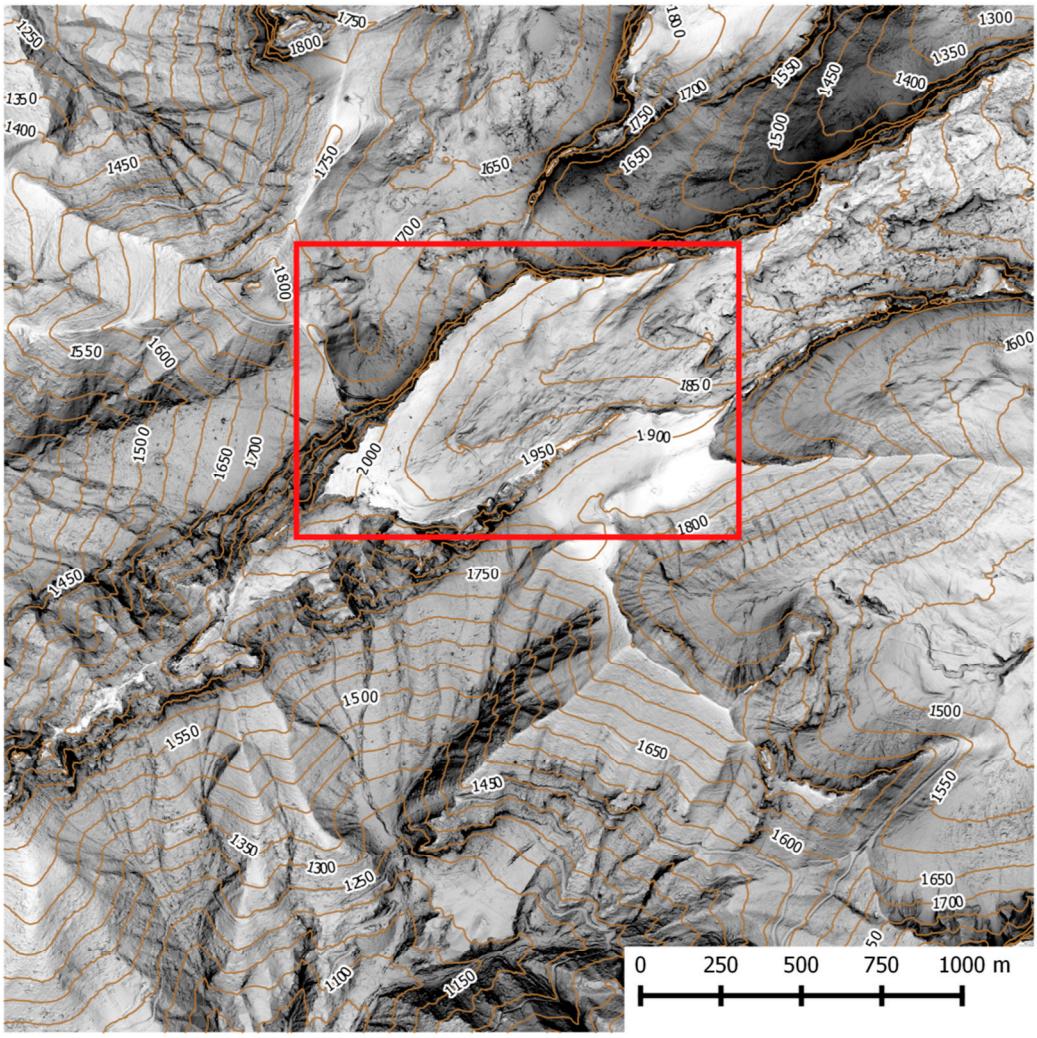


Figure 3.6 Filled DEM at a resolution of 0.5m (Heli05). Hillshade with contour lines every 50 meters. This DEM was used further on to compute DEM-derived variables for the B. laevigata case study. Sampling area is highlighted in the red box.

Comparison between HR and VHR DEMs for the *B. laevigata* case study

To select the most appropriate DEM among the three proposed (i.e. Swisstopo model at 2m – ST2, RPOD05, HELI05), we first compared 511 precise GPS measurements of altitude to altitude values extracted from the DEMs. Afterwards, we quantified the delineation of the ridge versus the geolocation of the sampled plants. All sampling points were geo-referenced with a Differential GPS unit (DGPS) offering a horizontal accuracy of ~2-3cm and a vertical accuracy of ~3-4cm (TOP-CON-HIPer Pro).

Table 3.1 shows that the variability in altitude measurement between DGPS and DEM is globally much higher for the RPOD05 model than for HELI05. The only few high errors in Heli05 can be explained by the position of certain samples. Indeed, some plants were sampled on the other side of the ridge where the slope is very steep; this can lead to a difference between DEM and DGPS

measurements (Figure 3.7). The standard deviation of Heli05 drastically decreased when we took these points out (STD = 0.12).

Table 3.1 Differences in altitude measurement between DGPS coordinates and three candidate DEMs. DPGS measurements were obtained along the ridge at sampling locations of plants.

	DGPS - Heli05	DGPS - Rpod05	DGPS - Swisstopo2
Average	-0.20	0.42	2.50
Min	-0.61	-1.54	-1.43
Max	6.49	35.53	47.32
STD	0.67	3.75	8.27

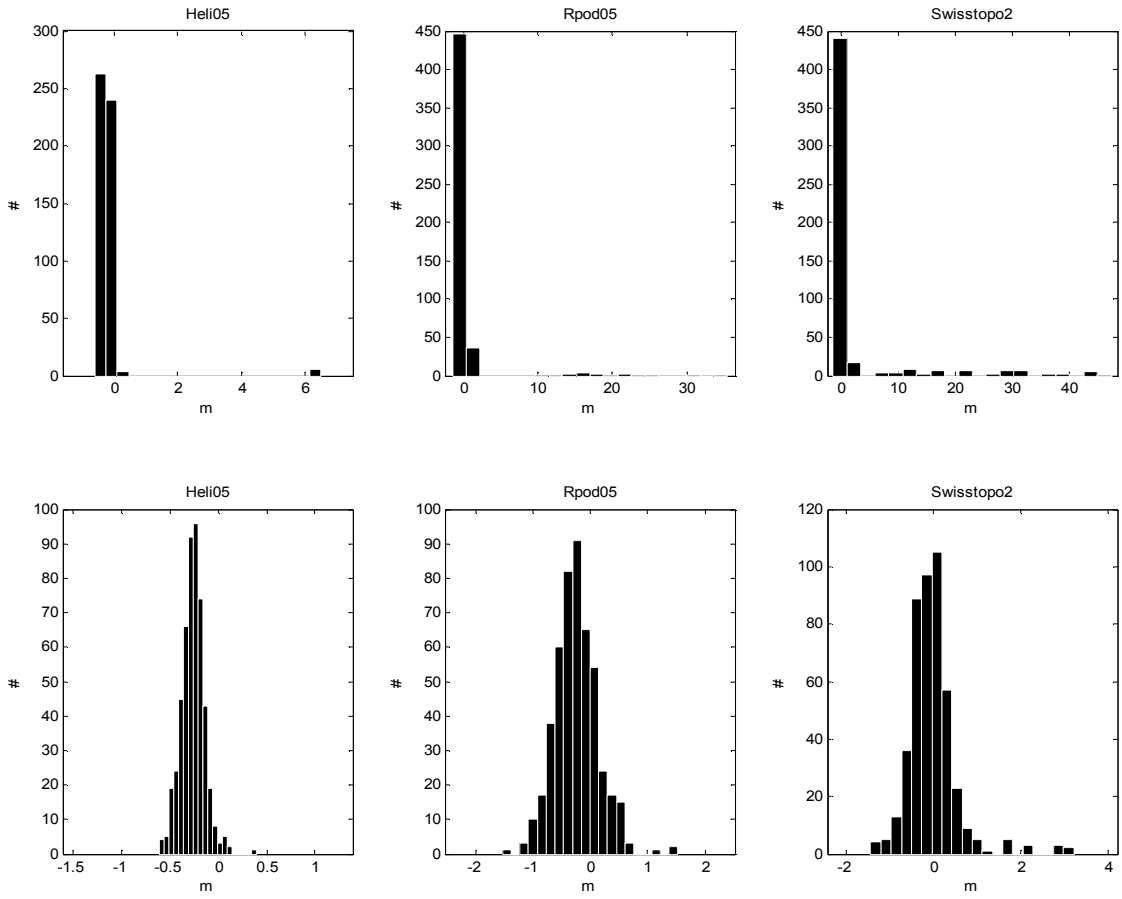


Figure 3.7 Histograms of differences in altitude measurement between DGPS coordinates and the three candidate DEMs. Important differences can be found for all models (top). More precise estimation of the distribution of errors is represented on the bottom histograms.

Next, we looked at the delineation of the ridge by computing an orientation variable for each model and overlaid the DGPS coordinates of sampled plants. We know from field observations that these points are located mostly on the southern-eastern side of the ridge and based on that we can judge how well the ridge is delineated in each DEM. Figure 3.8 illustrates well that Heli05 performs a much better delineation of the ridge than the other DEMs.

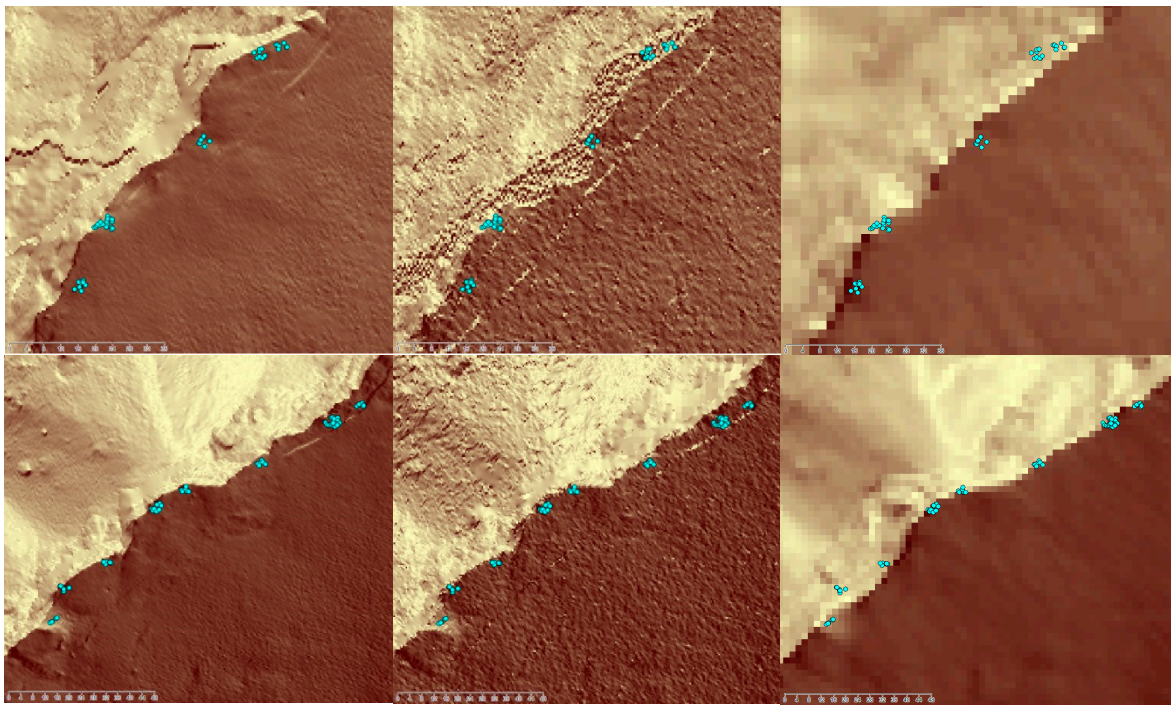


Figure 3.8 Orientation (Aspect) variable computed in SAGA GIS for each DEM at two different locations (top, bottom) of divergences between DGPS coordinates and DEMs: Heli05 (left), Rpod05 (middle), Swisstopo2 (right)

On the basis of these analyses, we concluded that Heli05 is the most precise model out of three proposed and gives the best delineation of the ridge (Leempoel & Joost 2012). Therefore, we used it to compute DEM-derived variables for the *B. laevigata* case study.

3.1.3 Multi-scale analysis

The purpose of a multi-scale analysis is to understand how the resolution of a DEM influences the computation of derived variables and thus, indirectly, how it will affect the significance of their correlations with genetic data. In other words, a multi-resolution analysis is likely to let us know how important microhabitat is, and what level of detail is necessary to detect local adaptation. The purpose of such an analysis is not to determine an optimal resolution throughout models, but to evaluate how sensitive a model is to a change of resolution and how many significant associations would be missed if only one resolution is considered.

Many methods have been proposed to apply a multi-scale generalization of DEMs, and more widely to images (Gallant & Hutchinson 1996; Wood 1996). Wavelet transforms, for example, are signal-processing techniques that focus on the compression and noise reduction of multidimensional signals (Figure 3.9). The work of Kalbermatten (2010) focused on the visual analysis of

DEMs' multi-scale decomposition by using wavelet transforms. His purpose was to analyse DEMs in the frequency domain and to benefit from a multi-dimensional model while keeping the large-scale geometry of the model. He demonstrated that simply averaging pixels provides a poor approximation of the general shape compared with other methods such as B-splines (Kalbermatten 2010). His work thus showed that a wavelet transform pipeline was a clever way to generalize topography and demonstrated the usefulness of B-splines, a generalization of Bezier curve, to model arbitrary functions, such as DEMs.

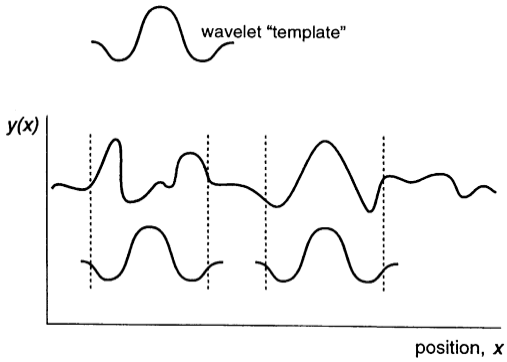


Figure 3.9 Pattern analysis with the wavelet transform. The wavelet template can show a poor match (left) or a good match (right) and thus provide negative or positive values respectively. Changing the scale of the wavelet implies a modification of the window size, while the shape of the wavelet remains intact. Taken from Dale & Mah (1998b)

In Kalbermatten's work, wavelets are used to produce low-pass coefficients (DEM at a coarser resolution) as well as high-pass coefficients containing details that were not passed to the low-pass coefficients. Unlike Kalbermatten (2010), we are not interested in the illustration of structures from different scales at high resolution. Instead, we need to use the low-pass coefficients only to produce variables at different resolutions. Therefore, the calculation of the high pass is unnecessary and we looked for a simpler procedure. After discussion with Dimitri Van De Ville (Medical Image Processing Lab, EPFL), we decided to take advantage of the Gaussian Pyramid algorithm implemented in MATLAB (MATLAB Version 12b. Natick, Massachusetts: The MathWorks Inc., 2010.), as it is known to approximate well cubic b-splines (http://www.mathworks.fr/fr/help/images/ref/impyramid.html).

The code we used can be found in 0. It imports a DEM as a georeferenced Geotiff and exports DEMs at coarser resolutions by automatically updating the cell size and corners of the projection parameters.

Illustration of the Gaussian pyramid using DEM profile cuts

To understand how different generalisation techniques work, we analysed a profile cut at different resolutions and compared the results obtained with the original model (Heli05), with average resampling and with the Gaussian pyramid (Figure 3.10).

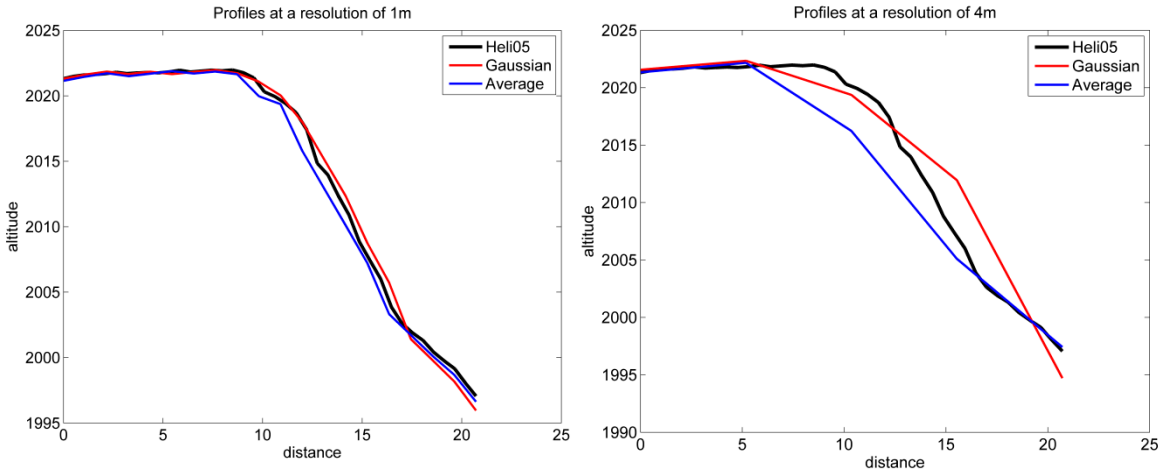


Figure 3.10 Profile cuts on the ridge of the Heli05 DEM. The original profile (thick black line) is compared to an average (blue) and to the Gaussian Pyramid result (red) at 1m (left) and 4m resolution (right).These figures show that a Gaussian pyramid provides a better approximation of the general shape of the profile that averaging pixels.

3.1.4 DEM-derived variables

SAGA GIS

We used SAGA GIS (Böhner *et al.* 2006) to compute environmental variables from DEMs. SAGA GIS is a fast and robust open source software that provides a large diversity of DEM-derived variables. In addition to simple terrain attributes, a large variety of secondary attributes related to morphometry can be computed, like solar radiation and hydrology (see Table 3.2). Although most of the proposed variables come with literature references, SAGA GIS does not provide an updated reference manual and, thus, often lacks a description of specific parameters for each variable. Importantly, SAGA GIS can be accessed from the command console as well as from the R package RSAGA (Brenning 2008), which facilitates automated computation of variables. Therefore, we were able to apply iteratively the same script to all DEMs but also to all case studies. This script can be found in 0.

Many DEM variables can be computed or accessed. Among them, we chose the following:

Altitude (alt). Altitude often correlates with many other variables such as precipitation, temperature, oxygen concentration etc. Even though altitude is often used in landscape genetics, it is difficult to interpret because of its correlations with other crucial variables.

Primary attributes

Primary attributes (**slope, aspect, curvature**) are the simplest variables that can be computed and are often used as proxies for water flow, snow movements, erosion or solar radiation. They are

usually built based on a 3x3 moving window. For this reason, we chose to compute them using the method of Zevenbergen & Thorne (1987). In each case, we computed Slope, Aspect (orientation), Curvature, Plan Curvature and Profile Curvature. Because Aspect ranges from zero to 360 degrees, it was converted in two variables: **Eastness (Sine of Aspect) and Northness (Cosine of Aspect)**.

Morphometric variables.

Several secondary attributes related to morphometry were computed with SAGA GIS.

The **Downslope Distance Gradient (DDG)** quantifies downslope controls on local drainage. It is thus related to the slope but was specifically designed to accurately model the underground water (Hjerdt *et al.* 2004). This variable is typically recommended for the computation of the topographic wetness index, but we also used it as an independent variable.

The **Morphometric Protection Index (MPI)** expresses the protection of a point from the surrounding relief (Yokoyama *et al.* 2002). It is based on the maximum angle found at zenith or at nadir from the point, over a defined radius. It is a good proxy for protection from wind for example.

The **Terrain ruggedness index (TRI)** is a quantitative measure of topographic heterogeneity (Böhner & Antonić 2009). However, it was shown to be highly correlated with slope when slope is steep. Sappington *et al.* (2007) proposed an alternative, the Vector Ruggedness Measure (VRM), which also quantifies rugosity but can be applied to any area. These variables might be related to stone density and be a proxy for certain soil characteristics (e.g. soil porosity).

Finally, the **Sky-view factor (SVF)** expresses the ratio of the radiation received by a planar surface over the radiation emitted by the entire hemispheric environment (Häntzschel *et al.* 2005). In other words, it quantifies the percentage of a hemisphere centred on each point that is obstructed by the surrounding landscape. In addition of being a pre-requisite for the processing of solar radiation variables, it is also a proxy to quantify humidity or protection from wind.

Solar radiation variables

Solar radiation modelling depends on slope, orientation, sky view factor and also takes into account adjacent relief (Wilson & Gallant 2000; Böhner & Antonić 2009). In SAGA GIS, it is possible to compute several variables related to insolation, such as **direct (Di), diffuse (Df) and total insolation (Ti)**, duration of insolation, sunrise and sunset. The user must define the latitude of the grid, either by giving an average latitude (46° for 1st and 2nd case studies) or by creating two grids of latitude and longitude values (for the *Sheep & Goats* case study at a large scale).

The output depends on the chosen time range. Indeed, it is possible to compute solar radiation for a specific hour, day or a range of days. In addition, one must define a time step of computation (0.5h by default). We decided to define two different days instead of monthly averages, the 21st of June and the 21st December.

Hydrology variables

SAGA GIS offers to compute many variables related to hydrology but most of them are oriented towards the modelling of watershed basins rather than providing indices. Here we focused on the **topographic wetness index (TWI)**, which is the logarithm of the ratio between the **catchment area (CA)** and the tangent of slope (Beven & Kirkby 1979). It quantifies the topographic control of hydrological processes and can be computed in two ways. One way is to use the **specific catchment area (SCA)** instead of the CA and the DDG instead of the slope, as proposed in Hjerdt *et al.* (2004). This, however, requires additional prior computation. Particularly, we had to fill the sinks of the DEM to obtain a DEM were each pixel can flow into another one. For this we used the Fill Sinks algorithm from Wang & Liu (2006). Then we used the "Flow width and Specific Catchment Area" algorithm to get the SCA (Gruber & Peckham 2009) and to create the TWI. The second way to compute TWI is simpler. One can do it using the **SAGA Wetness Index (SWI)**, where the computation pipeline is embedded within one algorithm and produces outputs of Catchment Area**, Catchment Slope (CSlo), Modified Catchment Area (MCA)** and the SWI (Böhner & Selige 2006).

We computed 26 DEM variables in total. The description and details of the parameters used to compute them are given in Table 3.2.

Table 3.2 Description of the parameters used to calculate DEM-derived variables at each resolution

	Variable	Abbre-viation	Units	Parameters
	Altitude	Alt	m	
Primary attributes	Slope	Slo	radians	Method = (Zevenbergen & Thorne 1987)
	Eastness (Sine of Aspect)	Eas	radians	
	Northness (Cosine of Aspect)	Nor	radians	
	Profile curvature	Vcu	radians/m	
	Plan curvature	Hcu	radians/m	
	Curvature	Cu	radians/m	
Secondary attributes	Downslope distance gradient	ddg	radians	Vertical distance = 5m
	Morphometric pro-tection index	mpi	no unit; Value is negative when the point is not pro-tected and positive when it is.	Radius = 1 pixel
	Terrain ruggedness index	tri	no unit	Radius = 1 pixel
	Vector Ruggedness Measure	vrm	no unit	Radius = 1 pixel
	Visible Sky	vis	no unit	Max search radius = 10000; Method = sectors; Number of sectors = 8
	Sky-view factor	svf	no unit	
	Diffuse Solar radia-tion in June	Df6	kwh/m²	Latitude=46°; Time Period=30 day; Time resolution=0.5h; Time Span=5 days; Day of year=01/06 -> 30/06; Atmospheric ef-fects=Height of Atmosphere and Vapour pressure
	Direct Solar radiation in June	Di6	kwh/m²	
	Total Solar radiation in June	Ti6	kwh/m²	
	Total Catchment Area	TCa	m²	Method = Multiple Flow Direc-tion
	Specific Catchment Area	SCa	m²/m	
	Flow Width	FW	m	
	Topographic Wetness Index	TWI	SCa/ddg	Area Conversion = No conver-sion; Method = Standard
	Modified Catchment Area	Mca	m²	Suction = 10; Type of Area = square root of CA; Minimum Slope = 0; Offset Slope = 0.1; Slope Weighting = 1
	Catchment Slope	Cslo	%	
	Topographic Wetness Index (SAGA)	SWI	MCa/Slo	

Structure tensor variables

In addition to his work on multi-scale visualisation of DEMs, Kalbermatten (2010) also interpreted wavelet coefficients directly from the gradient of a Laplace-filter. From these wavelet coefficients, it is then possible to create structure tensors with a local Gaussian window of 3x3 and to generate three related variables (Van De Ville *et al.* 2008). These variables had never been used in terrain analysis before and we decided to use them to assess if they correlate to different terrain features (Figure 3.11). We computed these three variables in ImageJ software (Bethesda, USA) using the OrientationJ plugin (Rezakhaniha *et al.* 2012). We included them in our environmental dataset in order to estimate their correlation with other DEM-derived variables and later to include them in association models with genetic markers.

These variables are the following:

Energy: Energy of the local gradient.

Coherency: coherency of the pixel regarding its neighbourhood. This is the ratio between the mean square magnitude of the gradient and the magnitude of the orientation vector. Range [0 1]. A large coherency corresponds to a dominant orientation in the neighbourhood.

Orientation: local orientation of the pixel (degrees).

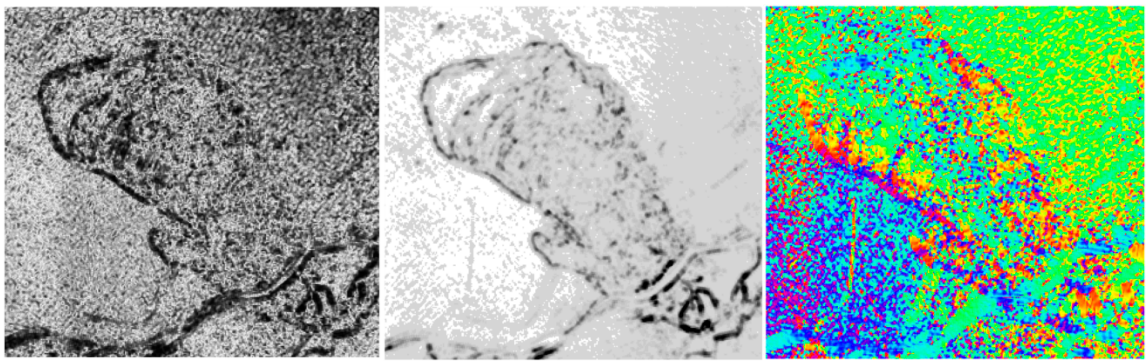


Figure 3.11 Example of structure tensor results for the first decomposition level of a landslide in Switzerland. Coherency (left), energy (centre) and orientation (right). From (Kalbermatten 2010)

3.2 Climatic variables

The next section details climatic variables that were extracted from either existing global or regional datasets. However, for the *B. laevigata* case study, a separate chapter is devoted to the acquisition and evaluation of climatic data (Chapter 4.1.2).

Several climatic datasets are available worldwide. Among them, the CRU (New *et al.* 2002) and the WorldClim (Hijmans *et al.* 2005; http://www.worldclim.org/current) are the most used and are both products of interpolations between weather stations dispersed on each continent (see Figure 2.4). In addition WorldClim dataset is based on several climatic databases gathered over 30

years and on an elevation model (SRTM). For the Sheep & Goats case study, we used the WorldClim dataset because it shows a higher resolution than the CRU dataset (Table 3.3).

Table 3.3 List of variables from the WorldClim dataset used in the Sheep & Goats case study. Variables tmin, tmean, tmax and prec are available for each month of the year.

Variable	Description	Units
tmin	Minimal temperature	°C x 10
tmean	Mean temperature	°C x 10
tmax	Maximal temperature	°C x 10
prec	Precipitation	mm
BIO1	Annual Mean Temperature	°C x 10
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))	°C x 10
BIO3	Isothermality (BIO2/BIO7) (* 100)	°C x 10
BIO4	Temperature Seasonality (standard deviation *100)	°C x 10
BIO5	Max Temperature of Warmest Month	°C x 10
BIO6	Min Temperature of Coldest Month	°C x 10
BIO7	Temperature Annual Range (BIO5-BIO6)	°C x 10
BIO8	Mean Temperature of Wettest Quarter	°C x 10
BIO9	Mean Temperature of Driest Quarter	°C x 10
BIO10	Mean Temperature of Warmest Quarter	°C x 10
BIO11	Mean Temperature of Coldest Quarter	°C x 10
BIO12	Annual Precipitation	mm
BIO13	Precipitation of Wettest Month	mm
BIO14	Precipitation of Driest Month	mm
BIO15	Precipitation Seasonality (Coefficient of Variation)	mm
BIO16	Precipitation of Wettest Quarter	mm
BIO17	Precipitation of Driest Quarter	mm
BIO18	Precipitation of Warmest Quarter	mm
BIO19	Precipitation of Coldest Quarter	mm

Regional or local studies require higher resolution climatic datasets. Indeed, the WorldClim dataset, with its resolution of ≈1km, is not appropriate for the *B. laevigata* and *P. major* case studies (extents of a few kilometres). In addition, when the terrain is rugged like in Switzerland, we expect interpolations models base on weather station data to take topography into account. For this purpose, Zimmermann & Kienast (1999) proposed a set of climatic variables over Switzerland, named Swiss Eco-Climatic GIS data, produced through interpolations and regressions. Some of these variables are based only on measures of mean temperature, precipitation or cloudiness at different locations and elevations in Switzerland, some are based on DEMs only, such as direct solar radiation, and others are a combination of both (Table 3.4). We included these variables in the *P. major* case study.

Table 3.4 Swiss eco-climatic variables used in the P. major case study

Variable	Description	
clou	monthly mean cloudiness	1/10 %
ddeg	annual degree-days with various threshold limits	day*deg
gamst	monthly continentality indices (seasonality of climate)	unitless
prec	monthly mean precipitation sum (1961-1990)	1/10mm
pday	# of precipitation days per growing season	#day
etpt	monthly potential evapotranspiration measures	1/10mm / day
mbal	monthly moisture balance: P/ETP - 1.0	daily avg.
mind	monthly moisture index : P – ETP	1/10mm
swb	annual average site water balance	1/10mm
tave	monthly mean of average temperature (1961-1990)	1/100 deg.C
tmax	monthly mean of maximum temperature (1961-1990)	1/100 deg.C
tmin	monthly mean of minimum temperature (1961-1990)	1/100 deg.C
sfroy	annual average # of frost days during growing season	#day

Window size for climatic variables

Climatic variables were also estimated at varying scale. Therefore, we considered using different window sizes to evaluate the influence of neighbouring values on the correlation between genetic markers and variables. To compute different window sizes, we used the “Simple Filter tool” in SAGA GIS. It computes a raster in which each pixel is the average of neighbouring pixels including the reference pixel itself. For the computation we used the following parameters:

Search mode: Circle; Filter: Smooth; Radius: 3,5,9,17,33 pixels.

3.3 Selection of variables

The variables were retrieved at all sampling location using the tool “Add grid values to points” in SAGA GIS. This function extracts for each point the pixel value directly under it.

Usage of large sets of environmental variables inevitably leads to redundancy in the analysis. In fact, it is common in landscape ecology and landscape genetics to perform a PCA on environmental data to extract the main axes of environmental variation and use PCA axes as environmental variables. On the other hand, one may argue that this procedure is done at the expense of interpretation, since PCA axes often represent several variables of interest. We considered that the number of variables in univariate models should remain high to represent a variety of variables as wide as possible. Therefore, we opted for a computation procedure based on correlations, as used by Stucki (2014). In order to keep a maximum of variables but avoid redundancy, we defined an threshold of 0.9 (in terms of absolute value) for univariate models. This procedure is random in the sense that it randomly selects a variable between the two that show a high correlation (thus higher than +0.9 or lower than -0.9). Correlations between variables were computed using Spearman's rank correlation coefficient in R using cor.test {stats}.

The loop is the following:

1. Threshold of 0.9
2. Search for the highest absolute correlation
3. Random selection of one of the 2 variables concerned
4. Loop until no pairwise-correlations are higher than the threshold

We also kept track of deleted variables and of their correlation with selected variables, in order to facilitate the ecological interpretation later on.

For multivariate cases however, a threshold of 0.9 is not adapted. It results in high multicollinearity leading to biased and unstable regression parameters (Stucki 2014). Multicollinearity can be estimated by calculating variance inflations factors (VIF), which depend on the determination coefficient (R^2) of the linear regression between predictors. VIF is calculated with the following equation and its usual maximum tolerated value is 5 (Dobson & Barnett 2008).

$$VIF = \frac{1}{(1 - R^2)}$$

Equation 3.1 Variance Inflation Factor

Therefore, in a bivariate model, lowering the pairwise correlation threshold to 0.8 corresponds to a maximum VIF of 2.8.

3.4 Genetic data and spatial structure

In this section, we explain the methods used to evaluate the spatial distribution of sampling locations, the methods used to produce the descriptive statistics of the genetic data and population structure of the three case studies.

3.4.1 Spatial distribution of samples

Sampling strategies are diverse in landscape genetics and are typically redefined for each particular study. Some studies focus on populations while others tend to evaluate continuous species responses to environmental gradients. Therefore, the geographic spread of sampling locations might influence the results and it is thus important to assess it.

We estimated geographic clustering of samples in QGIS (QGIS Development Team, 2014. QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org). We used the nearest neighbour index, which expresses the ratio of the observed distance between points and the expected distance between points. The expected distance is an average distance between neighbours in a hypothetical random distribution.

Equation 3.2 Average Nearest Neighbour analysis

$$ANN = \frac{\overline{D}_{obs}}{\overline{D}_{est}} \quad where \quad \overline{D}_{obs} = \frac{\sum_{i=1}^{n} d_i}{N} \quad and \quad \overline{D}_{est} = \frac{1}{2\sqrt{\frac{N}{S}}}$$

Where N is the number of events, S is the surface and d_i is the distance between two points.

This clustering index varies between zero and two (0 for completely clustered, 1 for random and 2 for completely dispersed). Its significance can be estimated with a Z score and informs on whether to reject the null hypothesis of a random distribution or not. For convenience we also converted the Z score into a p-value.

In addition, knowing the distance between samples is essential to estimate the relevance of multi-scale variables. For example, distance between samples can inform us about the minimum resolution we should use. It is also valuable in defining neighbourhood sizes for spatial autocorrelation (3.5.3). We thus computed and provided histograms of the pairwise distances, as well as the shortest distances between individuals.

3.4.2 Descriptive statistics of genetic data

We computed the descriptive statistics of genetic data for each dataset in order to: 1) assess the quality of the genetic data and 2) evaluate gene flow and genetic diversity.

P. major and sheep & goats case studies were analysed with PLINK (Purcell *et al.* 2007) as the underlying genetic data were provided in a corresponding format. The following statistics were computed:

- Missingness per individual
- Missingness per locus
- Minor Allele Frequency
- Heterozygosity
- Inbreeding coefficient

Amplified Fragment Length Polymorphisms (AFLPs) were used in the *B. laevigata* case study and Single Nucleotide Polymorphisms (SNPs) in the two others. AFLP are dominant markers, where

the presence of a band could be a homozygote or a heterozygote individual. Therefore, we could not evaluate heterozygosity using the inbreeding coefficient (Fis) for AFLP data.

For the *B. laevigata* case study, we used SPAGeDi (Hardy & Vekemans 2002) to estimate the mean relationship coefficient between samples, a measure of isolation by distance. The relationship coefficient is assessed for each genetic marker, from which a mean value is computed at each distance interval. To assess the significance of the coefficients, SPAGeDi performs permutations between individuals.

3.4.3 Population structure

Individuals of a dataset are all related with a certain level and some can form distinct groups, or populations. Assessing the population structure within a dataset informs us on the presence of clusters of related individuals and is likely to provide hints on the presence of geographic barriers to gene flow. In addition, as already mentioned in the introduction, population-genetic-based methods require individuals to be assigned to populations in order to evaluate Fst. Two software were used to assess population structure.

Admixture

Admixture (Alexander *et al.* 2009) estimates allelic frequencies of SNPs in ancestral populations as well as the proportion of individual genomes coming from these ancestral populations. Admixture uses bi-allelic SNPs data to assign each individual to a population. It requires the user to define the number of expected clusters (K) and can be launched over several Ks to estimate its most likely value using a cross validation method. For this cross-validation, genotypes are partitioned in a certain number of groups that are masked one by one. At each turn, membership coefficients and allelic frequencies are computed from the visible genotypes. Values of masked genotypes are then predicted from the membership coefficients and allelic frequencies of ancestral populations. Cross validation (CV) error measures the average deviation of predicted genotypes from hidden values and real genotypes. Because this task is performed for each K, the minimal value of CV indicates the most probable value of K. Admixture has the advantage of being much faster than Structure (Pritchard *et al.* 2000) as it does not use Bayesian computation. It is thus appropriate for the *Sheep & Goats* case study.

Structure

Admixture is not appropriate for AFLP data and polyploid species. In addition, it is not recommended to use it when the dataset is small. Therefore, for the *B. laevigata* case study, using AFLP data, and the *P. major* case study, with 464 SNPs only, we used the program Structure.

Structure 2.4 (Pritchard *et al.* 2000) is a model based clustering method capable of identifying distinct genetic populations and of probabilistically assigning individuals to populations. We decided not to use sampling location as prior population information as we expect the relief of both

case studies to create a complex pattern of membership between individuals that does not correspond to geographic relatedness. We chose the model with admixture for the ancestry model, selected a burning length of 20 000 and a number of simulations of 100 000, as recommended in (Pritchard *et al.* 2007). The number of clusters was set between K=2 and K=7 and 20 iterations were computed for each K. Secondly, we used Structure Harvester (Earl & vonHoldt 2012) to find the most likely number of clusters using the log probability of the data and the ΔK statistic (Evanno *et al.* 2005). Afterwards, we used CLUMPP (Jakobsson & Rosenberg 2007) on the most likely K, in order to calculate the mean cluster membership coefficients across the 20 replicates of cluster analysis (Greedy algorithm, 1000 random input orders).

3.5 Identification of signals of selection

This section describes the methods that were used to detect signatures of selection, the differences between them and the reasoning behind the selection of a particular method for each case study.

3.5.1 Correlative methods to detect outlier loci

Correlative approaches in landscape genetics estimate the strength of a relationship between genetic markers and environmental variables. Compared with population genetics approaches, they have the advantage of identifying environmental variables responsible for selection as well as of obtaining a statistically representative number of individuals per landscape type and not per population (Joost *et al.* 2007). We used two correlative approaches to detect signatures of selection.

Samβada

Samβada (Stucki 2014) is an individual-based method performing logistic regressions. It is a variant of linear regression in which the binary genetic marker is either present or absent and correlates with a quantitative environmental variable. Therefore, it provides the probability of occurrence of a genotype for each individual in function of environmental parameters (Joost 2006). The aim of Samβada was to improve SAM, the spatial analysis method (Joost et al. 2008) by increasing its computational power over large datasets on the one hand and by providing multivariate models on the other hand. Because the method is simple, it has the advantage of being fast and can efficiently process millions of models. Finally, it also provides spatial statistics that are helpful for the interpretation of significant results regarding spatial autocorrelation (see section 3.5.3).

Equation 3.3 Logistic regression. The natural logarithm of odds (or logit) is assumed to be linearly related to x, the independent variable.

$$logit(p) = \beta_0 + \beta_1 x$$

Samβada performs a univariate logistic regression for each combination of genotype and environmental variable (Equation 3.3). The significance of each model is assessed by a comparison between the values predicted by the model and observed values. It calculates the parameters of an estimated function that best fits observed values by maximizing the probability of obtaining the observed set of data (Hosmer & Lemeshow 2000). Therefore, the maximum likelihood function is expressing the probability that observed data are a function of unknown parameters.

In Samβada, two tests of significance are performed, likelihood ratio G (Equation 3.4) and Wald test (Equation 3.5). A model is considered significant if both G and Wald tests are significant. In addition, a value of Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and several pseudoR² are calculated for each model (Stucki 2014).

Equation 3.4 Equation of the likelihood ratio G

$$G = -2ln\frac{L}{L'}$$

L is the likelihood of the initial model (with a constant only) and L' the likelihood of the new model including the examined variable. If added parameters are equal to zero, this statistic follows a chi-square distribution with a number of degrees of freedom equal to the number of added parameters (Joost *et al.* 2007).

Equation 3.5 Equation of the Wald test of significance

$$w = \frac{\hat{\beta}_i}{\sigma(\hat{\beta}_i)}$$

The Wald test is obtained by comparing the maximum likelihood estimate of the $\hat{\beta}_i$ parameter with the estimate of its standard error.

Multivariate models

Samβada can also produce multivariate models to assess the effect of several environmental variables on genetic variation. First, Samβada computes univariate models for each marker per environmental variable. In a second analysis, it considers all possible combinations of variables and compares for each of them the G, Wald, AIC and BIC scores with the univariate scores. If any of

the combination models better predicts the frequency of a genetic marker than the univariate model only, it is considered significant and recorded in the results file. Models are then sorted according to their Wald score in a table.

The advantage of multivariate models is that population structure can be included as a co-explanatory variable through a membership coefficient and thus can estimate if the prediction of a locus is better explained by a multivariate model (one or several environmental variables + population structure) or by the membership coefficient alone (Stucki 2014).

Bonferroni correction and false discovery rate.

Because Samβada performs multiple simultaneous tests, a correction of the significance level must be applied to avoid excessive false positives. Bonferroni correction for multiple tests was applied to select the significant models. This correction consists in dividing the defined level of significance α by the number of tests performed (Shaffer 1995).

However, with datasets of increasing size, Bonferroni correction becomes too conservative and one might consider computing a false discovery rate (FDR) instead (Benjamini & Hochberg 1995). FDR is the ratio between the number of false discoveries and the total number of models considered significant (Figure 3.12). FDR implemented in Samβada is based on Storey & Tibshirani (2003) and calculates a q-value for each model based on the G scores. Consequently, by fixing a significance threshold α, we obtained a group of significant models in which a proportion α are false positives. For example, a threshold of 10% (FDR=0.1) implies that 10% of the significant models are false discoveries.

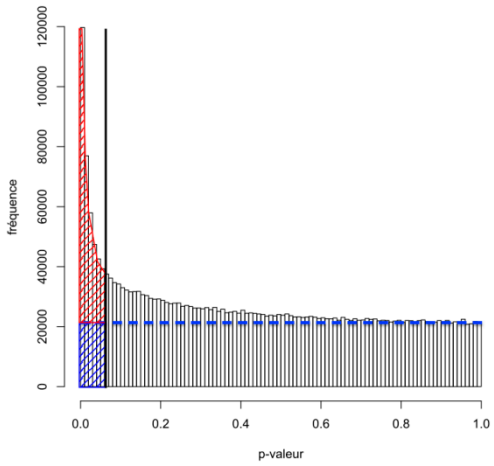


Figure 3.12 Estimation of the false discovery rate. Tests for which the null hypothesis is true are uniformly distributed from 0 to 1, which are estimated by the blue line and is adjusted on the basis of the frequency of tests having p-values close to 1. When a threshold of significance is defined (vertical black line), the proportion of false positives among the significant models is estimated by the ration between the false discoveries (blue surface) and all the discoveries (blue and red surfaces). From Stucki (2014)

RecodePLINK

RecodePLINK is used for recoding of PLINK's .ped and .map files to Samβada's format. Because Samβada's format requires binary data, each bi-allelic SNP is recode as three genotypes (e.g. AG, GG and GG).

LFMM

We used another correlative approach named Latent Factor Mixed Models (LFMM; Frichot *et al.* 2013). It differs from Samβada in the ability to assess the influence of the population structure on correlations. In fact, LFMM uses a hierarchical Bayesian mixed model based on a variant of principal component analysis in which residual population structure is introduced via unobserved factors. It thus detects signals of adaptation at the same time as it infers the background level of population structure. The regression model is a linear mixed-model that contains a genotypic component and a matrix of genetic variability, which, in turn, is not explained by the environment and adjusts its parameters in a Bayesian context.

LFMM only performs univariate models and assess their significance with a z score, on the basis of which a p-value is calculated. Like Samβada, it is possible to apply the FDR method from Storey & Tibshirani (2003) if the distribution of p-values permits it (see Figure 3.12).

LFMM considers bi-allelic genetic markers (such as SNPs) encoded as 0, 1 and 2. The number of simulations of the Gibbs sampler was kept at default value provided in the manual (number of iterations: 10000; burnin: 5000). The user also has to define the number of latent factors K, which can be estimated using the Tracy-Widom theory that provides the number of significant axes of the PCA on genetic data (Patterson et al., 2006). For each dataset, two values of K were tested as recommended.

LFMM was designed to use SNP data only and cannot be applied to AFLP markers, such as in the *B. laevigata* case study. In addition, we were not able to make it work on the SNPs of the *P. major* case study.

3.5.2 Population genetics approaches

BAYESCAN (Foll & Gaggiotti 2008) is based on a Bayesian method and uses differences in allele frequencies between populations in order to identify candidate loci. In particular it decomposes locus–population Fst coefficients into a population-specific component (beta) shared by all loci and a locus-specific component (alpha) shared by all the populations using a logistic regression. A given locus is assumed to be under natural selection when a locus-specific component alpha significantly different from 0 is necessary to explain the observed pattern of diversity. Consequently, for each locus, the posterior probabilities of two alternative models are estimated, including or not the alpha component. Posterior probabilities directly allow for the control of the expected

proportion of false positives among outlier markers using a FDR. BAYESCAN has proven to be robust to a wide range of demographic scenarios and can use very small sample sizes with no particular risk of bias.

We used the settings suggested by Foll & Gaggiotti (2008) with an exception for the length of the pilot run, where we used 10,000 instead of 5000, with a total length of the chain of 100,000 iterations.

Importantly, BAYESCAN cannot be applied to the *Sheep & Goats* case study since we did not identify distinct populations.

3.5.3 Spatial autocorrelation

Spatial autocorrelation occurs when geographically close objects are more related than distant objects (Tobler 1970). Spatial autocorrelation is thus expected to take place in many situations, be it for environmental variables or for genetic data. However, this phenomenon also implies that characteristics of close objects can be partially predicted by their neighbours, thus refuting independence between samples, which constitute the basis of many standard statistic tests (Dobson & Barnett 2008). Therefore, measuring spatial autocorrelation is necessary if one aims to assess the independence between samples as well as the influence of the geographical or environmental component on the spatial structure of samples.

For this purpose, we used the Moran's I measurement of spatial autocorrelation (Moran 1950) for both genetic data (frequency of markers) and environmental variables. Moran's index varies between [-1 1] and indicates outlier values when negative, clustering when positive and randomness when close to 0.

Equation 3.6 Moran's I global autocorrelation coefficient

$$I = \frac{N \sum_i \sum_j W_{i,j}(X_i - \bar{X})(X_j - \bar{X})}{\left(\sum_i \sum_j W_{i,j}\right) \sum_i (X_i - \bar{X})^2}$$

Where N is the number of observation units, $W_{i,j}$ the spatial weight, X_i the value of the variable at location i, X_j the value of the variable at location j and $\bar{X}$ the mean of the variable.

The calculation of Moran's I measurement starts with the definition of a weighting scheme, or neighbourhood, as the value of each point is compared with a weighted average of its neighbours. Two types of weighting schemes exist. The first is named "fixed kernel" and is based on a defined distance, thus involving a variable number of points (e.g. spatial lags); the second is name "adaptive kernel" on a defined number of points to consider, thus involving a variable distance (e.g. nearest neighbours).

We used Samβada to measure Moran's I rather than any GIS, as Samβada is able to process Moran's I for each marker and each variable included in the analysis. Samβada proposes three

weighting schemes using fixed kernels (mobile window, Gaussian kernel, bi-square kernel) and one adaptive kernel (nearest neighbour). Among the proposed weighting schemes, we opted for a nearest neighbour weighting scheme for all case studies as in Stucki (2014).

Equation 3.7 Moran's I pseudo p-value

$$Pvalue = \frac{(Random\ permutations > I) + 1}{Random\ permutations + 1}$$

Where I is the observed Moran's I value.

In addition, to assess spatial autocorrelation over larger distances, we decided to take into account several values of K nearest neighbour, starting from 10 to 100 by steps of 10 for the *Sheep & Goats* case study; and from 20 to 200 by steps of 20 for the *B. laevigata* and *P. major* case studies. We then plotted obtained Moran's I to produce a correlogram. The significance of each measurement was assessed using permutations. We opted for 999 permutations to obtain a significance of three decimals (0.001).

Moran's I is a global measure of spatial autocorrelation but it does not show where autocorrelation is strong and/or significant. Anselin (1995) developed a local index of spatial association (LISA) to solve this problem. LISA informs us on whether a point is significantly correlated with its neighbours and, if so, if the autocorrelation is positive or negative. In our case, LISA allows us to visualize whether the frequency of a genotype is correlated with the frequency in its neighbours and whether the autocorrelation is a local and a global phenomenon. The computation of LISA slightly differs from the computation of Moran's I, however, the sum of LISA indices is proportional to the global autocorrelation measurement (Anselin 1995).

Equation 3.8 Local Index of Spatial Association

$$I_i = \left[\frac{Z_i}{S^2}\right] \sum_{j=1}^{n} w_{ij}\, z_j\ , j \neq i$$

Where Z_i is the deviation from the mean, S the standard deviation of the dataset, w_{ij} the weight.

LISA's significance is assessed by permutations where the local value is fixed and the values of its neighbours are randomly picked from the entire dataset (N-1, since the point of interest is fixed).

The parameters and weighting scheme we chose are the same like for the global Moran's I. However, the significance level α is set to 0.01 instead of 0.05, which is considered to be a more appropriate cut-off value (Anselin 1995).

3.6 Visualisation of the results

The case studies described use large datasets with many variables processed by several methods of detection, resulting in many different outputs that cannot only be analysed in tables. We wanted to separate results by locus or chromosome, by variable, by resolution, and visualise maps of candidate loci to understand their spatial distribution and assess the robustness of their detection.

Therefore, a clear visualisation framework was necessary to make it possible to compare the results obtained through all the methods described above. We decided to use first and foremost graphical outputs for all the data. Importantly, we scripted the production of these graphics.

For each method, an overview of the results including for example histograms of scores and bar plot per variable (e.g. Figure 6.11 p127). These graphics significantly facilitated the examination of the results. Tables themselves were modified from the original output to include descriptive statistics, such as minor allele frequencies, spatial autocorrelation scores as well as a "detection score" produced by other methods (e.g. Table 4.6 p84).

In addition, in the sheep & goats case study, we compared results between methods by plotting their score versus their position on the chromosomes. Doing so, we were able to evaluate the presence of clusters of detected SNPs in the genome and to track for their presence in the results produced by each method (e.g. Figure 6.13 p130).

In addition to the overviews per method, we focused on the representation of significant associations between genetic markers and environmental variables. We subdivided the graphs into different compartments in order to include the map of the locus, its local spatial autocorrelation indices, the correlated environmental variable, as well as the evolution of the scores with resolutions. In addition, we provided a table containing important descriptive statistics (an example is provided in Figure 3.13 and described in the next sub-chapter).

Finally, we produced scatterplots showing the relationship between results produced by the different methods. For example comparing Samβada's G score to LFMM's p-value, to Moran's I etc. In each of these graphs, we classified the genetic markers by colour, representing either those commonly detected or those detected by one method only.

We used Matlab to load and treat results from each outlier detection method as well as from the spatial analysis. We performed the treatment of the results for each case study in a similar way, and used Matlab to produce the corresponding graphs.

3.6.1 Visualisation of significant associations between genetic markers and environmental variables

In Figure 3.13, a map showing the genetic marker analysed is provided in graph (A), close to its correlated environmental variable in graph (B). This allows us to visualise the level of clustering of the marker and its correlation with the environmental variable. In the cases of SNPs, the title of the graph (A) indicates which genotype is involved in the association model. In addition, the fre-

quencies of each genotype are provided in the legend. Similarly in graph (B), a legend is added to display the minimum, average, and maximum values of the variable. Units of the environmental variables can be found in sections 3.1.4 and 3.2. In cases of multi-scale variables, only the variable with the best association with the genetic marker according to the G score is shown.

Graph (C) illustrates the local indices of spatial association (LISA) performed for the genetic marker with an adaptive weighting scheme of 20 neighbours. The legend informs us on the different categories of spatial autocorrelation: positive autocorrelation HH (High-High) and LL (Low-Low), negative autocorrelation HL (High-Low) and LH (Low-High), and non-significant autocorrelation (NS). Point size of the four significant categories of LISA are bigger than non-significant to highlight them. For global autocorrelation, graph (D) is a correlogram that displays the evolution of Moran's I for all genetic markers of the dataset at different distances or number of neighbours. Here, all loci of the datasets are coloured in grey and the genetic marker of interest in red. It allows us to evaluate how strong is spatial autocorrelation for this genetic marker, how it evolves with distance and how it behaves compared with other markers.

Then, the evolution of the significance of the model with decreasing resolution (or increasing window size) can be observed in graph (E). An horizontal blue line indicates the threshold of significance specified in each case study.

Finally, general information on the model is provided in graph (F), including the best G score (among the different resolutions), the Wald score of this model, its q-value, the Moran's I of the marker and its p-value, the Moran's I of the variable and its p-value, the best scores in other methods and their corresponding variable.

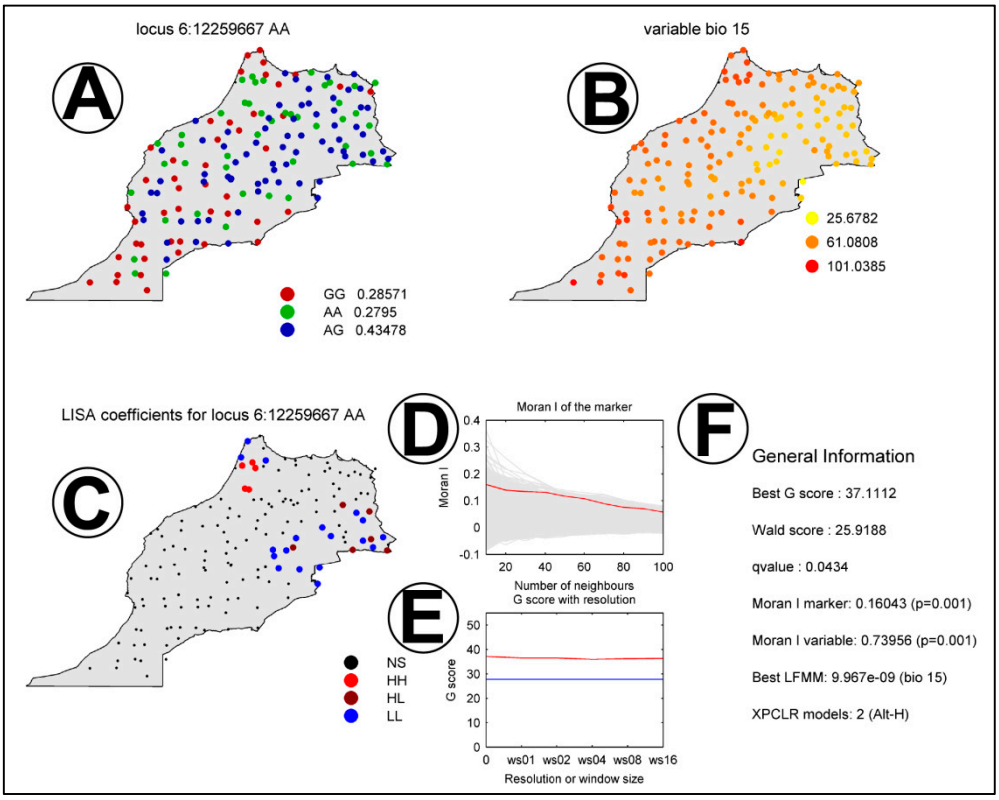
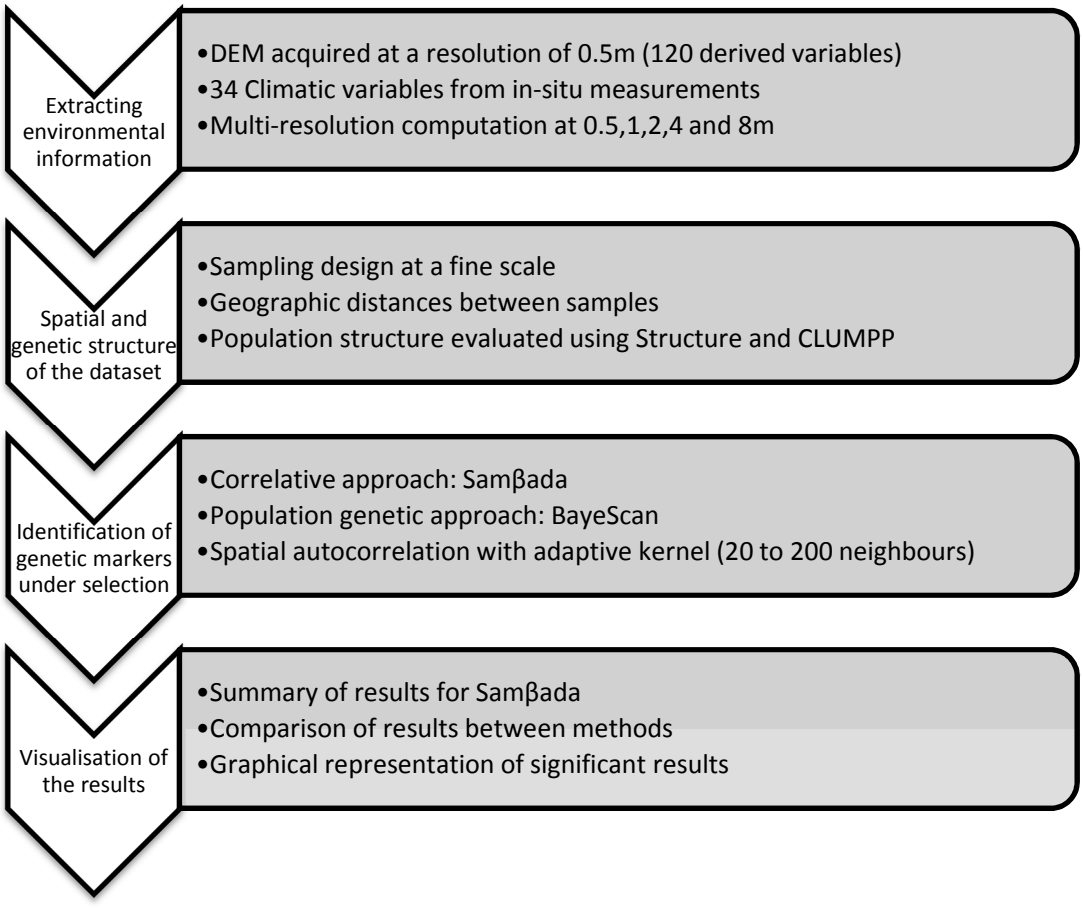


*Figure 3.13 Visualisation of the significant association between genotype 6 12259667 AA and Precipitation Seasonality (bio15) in **goats** detected by Samβada. Explanations on the different graphs can be found on p. 65.*

Chapter 4 *Biscutella Laevigata* in Les Rochers-de-Naye

Observing adaptation of alpine plants at a local scale requires fine scale environmental data. However, while it is often expected that a higher precision should bring more accurate results, it should not be forgotten that a high amount of details might blur the output signal.

As part of the project Velux Stiftung (Project 705), 361 *B. laevigata* were sampled along the ridge of "Les Rochers-de-Naye" (Western Swiss Alps). Our purpose was to assess on one hand whether fine scale topography obtained from a very high resolution DEM can model ecologically relevant features such as temperature or humidity, to settle their usefulness at a local scale, and on the other hand to find out if DEM-derived proxies of environmental features can detect signatures of selections. The strength of this study is to take advantage of very high resolution to evaluate the scale dependency of microhabitat modelling and of signatures of selection at a local scale.



B. laevigata is a widespread autotetraploid Brassicaceae, which occur in small patches in warm and dry areas (Parisod & Bonvin 2008). It is a perennial self-incompatible alpine plant and generalist Diptera and Lepidoptera achieve pollen dispersal while seeds disperse through wind and gravity. It survived the glacial ages as a diploid (Manton 1937) and is thought to have recolonized the alps starting from multiple refugia, with the peripheral alps representing the historical core of the lineage (Parisod & Besnard 2007). The study zone is situated at « les Rochers-de-Naye » (N46°26‘00‘‘ E6°58‘50‘‘) where *B. laevigata* forms a natural hybrid zone between closely related lineages. Parisod & Christin (2008) and Parisod & Joost (2010) showed that individuals consistently presented similar genotypes in habitats of contrasting solar radiation.

To investigate its adaptation to a local mountainous environment, we acquired a high resolution DEM to model micro-habitat conditions and used temperature loggers to study the variability of climatic conditions. The first part of the study demonstrates the usefulness of DEMs as a surrogate to crucial climatic variables and highlights the ability of DEM variables to model micro-habitat conditions as those encountered by plants (Leempoel *et al.*; Körner 2003). This research was submitted to Methods in Ecology and Evolution and can be found in Appendix II. The second part explores the population structure and exposes the correlations between local environmental data and genetic variation. In addition, a multi-scale approach was applied on DEMs in order to evaluate the scale dependency of these correlations.

4.1 Environmental Data

We computed variables from the VHR DEM (0.5m spatial resolution) detailed in section 3.1.2. These variables (see Table 3.2) were computed at resolutions of 0.5, 1, 2, 4 and 8m. Unlike the other case studies, we decided not to use 6 but only 5 different resolutions since we found 16m to be too coarse for this case. In fact, the delineation of the ridge at a coarse resolution might not be accurate enough to model the habitat of the plants and could introduce an important bias.

4.1.1 Remote sensing data

In addition to climatic variables detailed below, we extracted two additional variables from remote sensing data. The first one is an aerial Infrared image obtained from Swisstopo (Swissimage © 2013 swisstopo, JD100064), with a resolution of 0.5m. The second one is an orthophoto obtained in winter with the same drone as described in section 3.1.2. We hoped to identify uncovered areas and estimate sun reflectance of the snow by measuring the intensity of light on the image. Values of each band (Red, Green, and Blue) were extracted at each sampling location and summed into one total intensity variable.

4.1.2 Ecological relevance of VHR DEM variables

The local scale of this case study requires high-resolution climatic variables. Indeed, we mentioned that the eco-climatic variables from Zimmermann & Kienast (1999) were irrelevant due to

their coarse resolution and high correlation with altitude (Leempoel & Joost 2012). Therefore, we decided to install a series of temperature and humidity loggers along the ridge. The purpose of this analysis is twofold. First, it allowed us to evaluate the role of topography on local climatic variability, by performing multivariate regression models between climatic measurements and VHR DEM-derived variables. Secondly, temperature measures could be directly used in association models with the frequency of genetic markers to account for micro-habitat conditions in possible signatures of local adaptation.

Environmental information was measured with 60 uncovered temperature loggers placed at the centre of each plant sampling plot (Figure 4.4) and 20 additional uncovered temperature loggers installed at random locations, outside of these plots, along the ridge to measure direct air temperature (Figure 4.1 A). Furthermore, 25 temperature and humidity covered loggers were placed close to 1 uncovered logger over 3 to measure ambient temperature and humidity.

Following this design, uncovered I Button loggers (1922L) from Maxim Integrated (http://www.maximintegrated.com/) were placed 15cm above the ground to estimate direct air temperature (DT) as perceived by the plant, while covered temperature and humidity loggers (I Button 1923) measured ambient temperature (AT) and humidity (HU) at 15cm above the ground (see Figure 4.2 for an example). These loggers were covered with a white shield pierced with several holes to avoid stagnant air. Loggers were set to record information with a frequency of 30 minutes during 126 days, from June 15, 2013 to October 18, 2013, with an accuracy level of 0.5 degrees C° and 5% for humidity. The devices were used to produce i) direct air temperature (DT), ii) ambient temperature (AT) and iii) air humidity (HU) variables. The 126 days mentioned above were grouped in 9 periods of 14 days (P1: June 15 to 28; P2: June 29 to July 12; P3: July 13 to 26; P4: July 27 to August 9; P5: August 10 to 23; P6: August 24 to September 6; P7: September 7 to 20; P8: September 21 to October 4; P9: October 5 to 18).

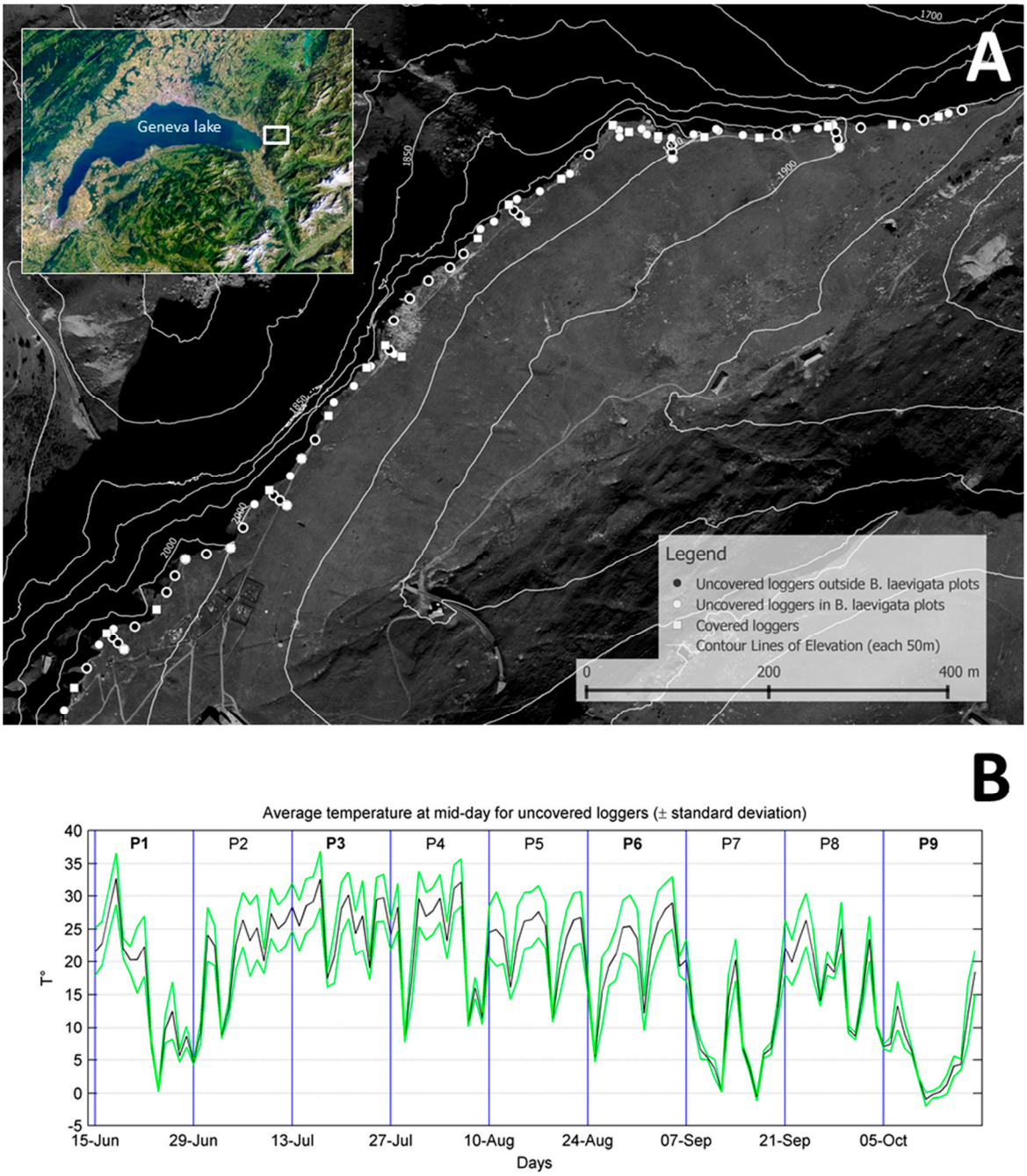


Figure 4.1 (A) Study zone and sampling locations of loggers on the ridge of Les Rochers-de-Naye in the Swiss Western Alps. Loggers were placed at and between Biscutella laevigata locations (not shown here, see Figure 4.4). Uncovered and covered loggers were used to measure direct air temperature and ambient temperature respectively. (Background image with 50 m isoelevation lines: Swissimage © 2013 swisstopo (JD100064)). (B) Mean daily direct air temperature and standard deviation (in green) from the 15 June to the 18 October 2013, measured with uncovered loggers. Vertical lines delimit the defined periods. The periods retained for following analyses are displayed in bold.

Figure 4.2 Illustration of loggers placed on the ridge of Les Rochers-de-Naye. Loggers measuring temperature and humidity were covered with a white shield while loggers measuring direct air temperature were uncovered (situated just above the cover logger on the picture).

The following descriptive statistics were computed for direct air temperature (DT), ambient temperature (AT) and humidity (HU) during each period: minimum (MIN), maximum (MAX), mean (MEA), standard deviation (SD), median (MED), mean value at 1am (M1A), mean value at 1pm (M1P), mean daily range (MDR).

In addition, soil moisture was measured by Geiser (2014) at 201 representative sampling locations. The soil volumetric water content was evaluated once with a FieldScout TDR 300 Soil Moisture Meter (Spectrum Technologies, Aurora, USA, http://www.specmeters.com/). According to le Roux *et al.* (2012), soil moisture values are highly correlated among distinct sampling events and therefore only one measure was carried out (more than 24 hours after previous rainfall) to get a reliable percent soil moisture value (MSM).

To perform multivariate regression models between DEM-derived variables and climatic variables measured in the field, we used the original DEM (resolution of 50cm) described in section 3.1.2 gradually generalized to 1, 2 and 4 meters using the Gaussian Pyramid (section 3.1.3). On the basis of each of these DEMs, we computed DEM-derived variables (see section 3.1.4 p49). In the cases of Direct air Temperature (DT), and Soil Moisture Measurements (MSM), we had a set of eight DEM-derived variables, and five in the case of Ambient Temperature (AT) and Ambient Humidity (HU). We used a Stepwise Generalized Linear Models (SGLM) (Nelder & Wedderburn 1972) with a Gaussian family and controlled the addition or removal of a term based on the Akaike Information Criterion (AIC). After model completion, co-linearity between variables was controlled using Variance Inflation Factors (VIF; Montgomery & Peck 1982), based on the threshold >3; (Zuur *et al.* 2009). Models with variables showing VIF>3 were processed again, excluding the inflating variables. Adjusted R^2 ((N-1)/(N-k-1) where N = number of observations and k = number of predictors) were calculated for each model.

Instead of GLMs, Generalized linear mixed models (GLMM) (Breslow & Clayton 1993; Bolker *et al.* 2009) were performed on the dataset of soil moisture to take into account the possible effect of spatial autocorrelation. These variables were collected in plots and the merging by plot was thus considered as the random effect variable in the multivariate GLMMs. GLMMs were performed with the lme4 R package (Bates & Maechler 2009). As the package does not support step procedure, we used the resulting DEM-derived variables from SGLMs procedures as fixed effects in the GLMMs. GLMMs were not performed on air temperature and humidity data since only one logger at most was located in each plot.

Results

The distribution of average direct air temperature over the whole sampling period provides a global view on climatic conditions during summer 2013 (Figure 4.1 B). The average for the whole period was 12.1C° and we focused here on four among the nine periods of 14 days representative of contrasted weather conditions at such altitude. P1 is representative of early spring, with the beginning of the growing season and containing a cold episode, whereas P3 and P6 are characterized by warm averages and a high standard deviation, representing early and late summer conditions, respectively. Finally, P9 is representative of late fall conditions and contains a snowy episode.

Together with altitude, terrain wetness index (twi), vector ruggedness measure (vrm), eastness (eas) and slope (slo) are the DEM-derived variables that best explained the variance of measured environmental variables. Hereunder, the VHR DEM-derived variables showing the best goodness-of-fit to explain the variability of measured environmental variables and ecological factors, depending on different spatial resolutions and periods of time, are presented.

Direct air temperature (DT)

Among all DT models, twi was the most frequently significant DEM-derived variable (47% of the models). It was positively correlated with measured variables related to high temperatures (M1P, MAX, MDR) and negatively correlated with those related to cold temperatures (M1A, MIN, MEA) (see Table 4.1 and Appendix III). Other DEM-derived variables, such as slope, eastness and ddg, were less frequently significant. Altitude is also frequently significant (55% of the models), but mainly with measured variables related to cold temperatures (M1A, MED, MEA, MIN).

Significance of DEM-derived variables varied considerably with spatial resolution, whereas it remains constant to a large extent at all resolutions for elevation. Although significance was lower when computed at 0.5 or 1m than at coarser resolutions for twi (Appendix III), adjusted R^2 (aR^2) were usually highest in models at 0.5 or 2m resolution and almost systematically lower at 4m. Noticeably, aR^2 are higher for all measured variables (except for mean range) during periods P1 and P9, which correspond to the two coldest periods among the four analysed.

Table 4.1 Summary of the results of multivariate generalized linear models sorted by adjusted R^2(aR^2) in decreasing order for DIRECT AIR TEMPERATURE (DT), measured with uncovered loggers at 15 cm above the ground. First column is the abbreviation of the model shown, with different calculated variables and time periods. The second column tells at which resolution (Res) the highest aR^2 was found. Coefficients for each variable show when the variable is significant and its significance is expressed with "" where p-values <0.001 correspond to ***, <0.01: **, <0.05: *. All models at all resolutions can be found in Appendix III. Abbreviations are the following. Measured variables: minimum (MIN), maximum (MAX), mean (MEA), median (MED), mean temperature at 1am (M1A), mean temperature at 1pm (M1P), mean daily range (MDR). Time periods: P1=15 to 28 June, P3=13 to 26 July, P6=24 August to 06 September, P9=05 to 18 October. DEM-derived variables : Altitude (alt), Terrain Wetness Index (twi), Vector Ruggedness Measure (vrm), Eastness (eas), Slope (slo), Horizontal Curvature (hcu), Vertical Curvature (vcu), Downslope Distance Gradient (ddg)*

Model	Res	aR^2	alt	twi	vrm	eas	slo	hcu	vcu	ddg
DT-M1A-P9	0.5	0.69	-0.71^{***}	0.17^{*}	-0.21^{*}					
DT-MIN-P9	2	0.50					0.28^{**}			
DT-M1A-P6	1	0.46	-0.49^{***}	-0.81^{***}		0.25^{**}		-0.20^{*}		
DT-MED-P3	2	0.37	-0.40^{***}	-0.57^{***}						
DT-MEA-P6	2	0.32	-0.35^{**}	-0.80^{***}			0.41^{**}			-0.45^{*}
DT-MDR-P3	0.5	0.22	0.25^{*}	0.47^{***}		-0.41^{***}				
DT-MDR-P1	2	0.19	-0.25^{*}		-0.38^{***}					
DT-MIN-P1	0.5	0.13	-0.37^{**}							

Ambient temperature (AT)

Significant correlation between DEM-derived variables and AT are much less frequent (49% of the models) than for previously presented DT models (91%; Appendix III). However, relevant predictors are the same like DT models, except that horizontal curvature (hcu) is significant at a 2 m resolution (Table 4.2). Like DT models, twi is positively correlated with measured variables related to high temperatures, and negatively correlated with cold temperatures. The goodness-of-fit of altitude is high in all models and is involved in the models with the highest R^2, particularly during the snow episode (P9).

Table 4.2 Summary of the results of multivariate generalized linear models sorted by adjusted R^2(aR^2) in decreasing order for AMBIENT TEMPERATURE (AT), measured with uncovered loggers 15 cm above the ground. First column is the abbreviation of the model show, with different measured variables and time periods. The second column tells at which resolution (Res) the highest aR^2 was observed. Coefficients of each variable are showed when the variable is significant in a model and its significance is expressed with "" where p-values <0.001 correspond to ***, <0.01: **, <0.05: *. All models at all resolutions can be found in Appendix III. Abbreviations as in Table 4.1.*

Model	Res	aR^2	alt	twi	eas	slo	hcu
AT-MED-P9	0.5	0.89	-0.94^{***}	-0.35^{**}			
AT-MED-P6	4	0.80	-0.74^{***}	-0.44^{**}			
AT-MDR-P3	2	0.49	0.43^{*}		0.52^{**}		-0.69^{***}
AT-MAX-P6	2	0.43				0.48^{*}	-0.44^{*}
AT-M1A-P3	2	0.40	-0.74^{***}				0.48^{*}
AT-MIN-P1	2	0.38	-0.81^{***}	0.87^{**}	-0.75^{**}		0.55^{*}
AT-MDR-P6	1	0.31				0.58^{*}	
AT-MDR-P9	0.5	0.31				0.58^{*}	

Ambient humidity (HU)

Among the 112 HU models computed, only 35 (40%) were significant (Appendix III), contrasting with models for DT (90%) and AT (70%). Such a low rate of significant models was related to the rare significance of altitude and of DEM-derived variables such as eastness, slo and hcu in HU models (5% of them). On the other hand, twi was the DEM-derived variable with most frequently and highly significant models (37%). It was significant for all categories of measured variables and all periods analysed, except during the snowy episode (P9). Like DT models, resolution influences twi significance and models have an aR^2 optimum at 1 or 2m (Table 4.3).

Table 4.3 Summary of the results for multivariate generalized linear models sorted by adjusted R^2 (aR^2) in decreasing order for AMBIENT HUMIDITY (HU), measured with uncovered loggers 15 cm above ground. First column is the abbreviation of the model showed, with different measured variables and time periods. The second column tells at which resolution (Res) the highest aR^2 was found. Coefficients of each variable are showed when the variable is significant and its significance is expressed with "" where p-values <0.001 correspond to ***, <0.01: **, <0.05: *. All models at all resolutions can be found in Appendix III. Abbreviations as in Table 4.1.*

Model	Res	aR^2	alt	twi	eas	slo	hcu
HU-M1A-P6	1	0.76		0.82^{***}		0.48^{**}	0.54^{**}
HU-MDR-P1	2	0.48		-0.75^{***}	0.42^{*}		
HU-MED-P3	2	0.47		0.70^{**}			
HU-M1P-P6	0.5	0.38		0.55^{*}		-0.53^{*}	
HU-M1P-P1	2	0.28		0.59^{**}			
HU-MDR-P6	0.5	0.27		-0.63^{*}		0.51^{*}	
HU-M1P-P9	1	0.23		-0.47^{*}			
HU-MDR-P3	1	0.19		-0.76^{*}			

To assess the importance of the time-period for the three categories of environmental variables (DT, AT, HU), we computed models between DEM-derived variables and measured variables over the entire fieldwork season (i.e. 15 June to 18 October) (Appendix III). Although the same predictors are significant for roughly the same measured variables, our results show that periods with cloud cover (P1) or snow cover (P9) contrasted with those of sunshine (P3, P6) and that this contrast could be explained by the higher goodness-of-fit of topography in the latter case (weaker significance of altitude, stronger significance of eas, slo, twi). In addition, the use of several measured variables is justified in order to distinguish different ecological conditions, as recommended by (Ashcroft *et al.* 2011; Vercauteren *et al.* 2012).

Soil moisture

In soil moisture models, vector ruggedness measure (vrm) was the only DEM-derived variable that showed a significant contribution across resolutions (Table 4.4). However, its contribution was dependent on resolution, as models were less and less significant with coarser resolutions. Given that altitude showed a stable contribution though scales, the highest aR^2 was obtained at 0.5m resolution due to the highest goodness-of-fit of vrm at that resolution.

Table 4.4 Summary of multivariate GLMMs on one-time measurements of SOIL MOISTURE sorted by adjusted R^2 (aR^2). Coefficients of each variable are showed when the variable is significant and its significance is expressed with "" where p-values <0.001 correspond to ***, <0.01: **, <0.05: *. Abbreviations as in Table 4.1.*

Res	aR^2	alt	twi	vrm	eas	slo	hcu	vcu	ddg
0.5	0.46	-0.26^{**}		-0.43^{***}					
1	0.43	-0.45^{***}		-0.19^{**}					
2	0.41	-0.46^{***}		-0.20^{*}					
4	0.35	-0.44^{***}				-0.23^{**}			

4.1.3 Extraction of climatic variables at sampling locations

We retrieved DT measurements for each plant by extracting temperature data from their closest logger. In the cases where temperature data were missing, we extracted data from the second closest logger, up to the third closest. In cases where DT measurements showed missing data, we deleted variables for which more than 10% of individuals had missing data. In total, 34 variables from loggers were used in the following analysis.

4.1.4 Variables selected

A total of 160 variables were considered for associations with genetic markers. The selection procedure ended up with 66 uncorrelated variables for a maximum correlation threshold of 0.9. Among them, 12 (out of 34) are temperature measurements from loggers and 50 (out of 120) were DEM-derived variables. However, we noticed that most of the DEM variables computed

were not correlated with any other selected variable for at least one resolution, except for altitude (correlated with coordinates) and two solar radiation variables, diffuse and direct insolation, correlated with total insolation. Therefore, we decided to keep all DEM-derived variables in the dataset, expect those mentioned above, since we wanted to calculate their correlation with genetic markers at multiple resolutions. The remaining variables are coordinates, population structure (see below), Infrared and Snow reflection.

4.2 Spatial and genetic structure of the dataset

4.2.1 Sampling design

Precise coordinates are essential for local scale studies as it is crucial to estimate accurately the positions of plants in the corresponding pixels of environmental rasters. All sampling points and loggers were thus geo-referenced with a differential GPS unit offering a horizontal accuracy of ~2-3cm and a vertical accuracy of ~3-4cm (TOPCON-HIPer Pro, http://www.topcon.com.sg/survey/hiperpro.html).

Sampling locations were selected by Geiser (2014) along the ridge. She followed a random cluster sampling, guided by the population density of *B. laevigata* and guaranteeing that all data points are located within pixels representing 0.5x0.5m in the field, corresponding to the pixel size of the DEM. A total of 361 individuals of the focal species were sampled in 60 4x4m areas with at least five individuals per area. When less than five plants were reported, a new area was selected at a random distance ranging from 0 to 25m (see transects and resulting distribution in Figure 4.4).

4.2.2 Spatial structure

B. laevigata show a highly clustered spatial distribution, with a nearest neighbour index of 0.021 (p-value < 0.0001). The average observed distance between sampling locations is of 0.44m distributed at 1.2km maximum (Figure 4.3). The spatial distribution of *B. laevigata* reflects its sampling but was also coherent with its natural distribution along the ridge.

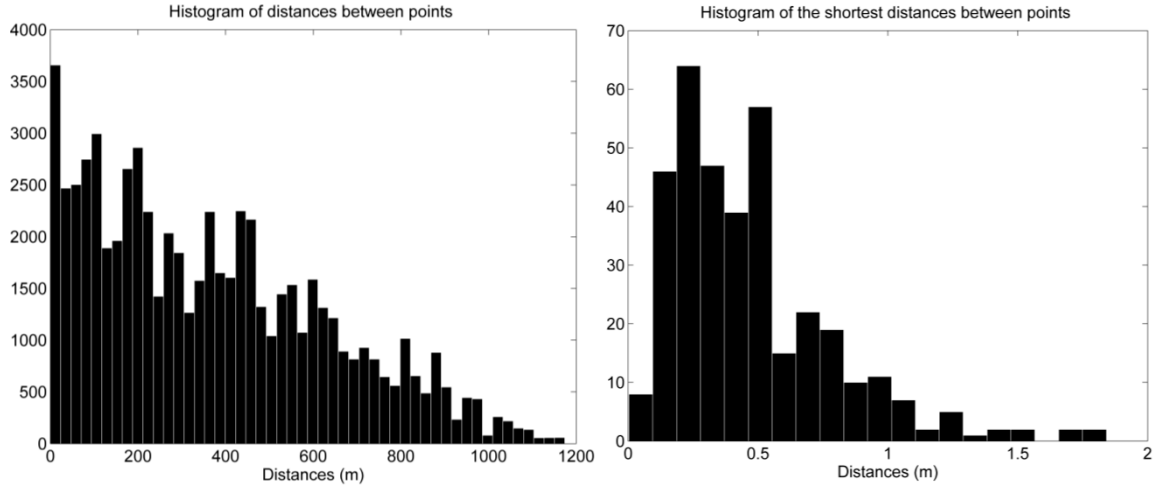


*Figure 4.3 Histograms of pairwise distances between samples (left) and shortest distances between samples (right) for **B. laevigata.** Samples were collected continuously along the ridge. Therefore, the histogram of distances between samples shows a linear decrease as the sampling is linear along the ridge. In addition, the closest neighbour of each sample is often situated at less than 1m because plants were sampled in plots.*

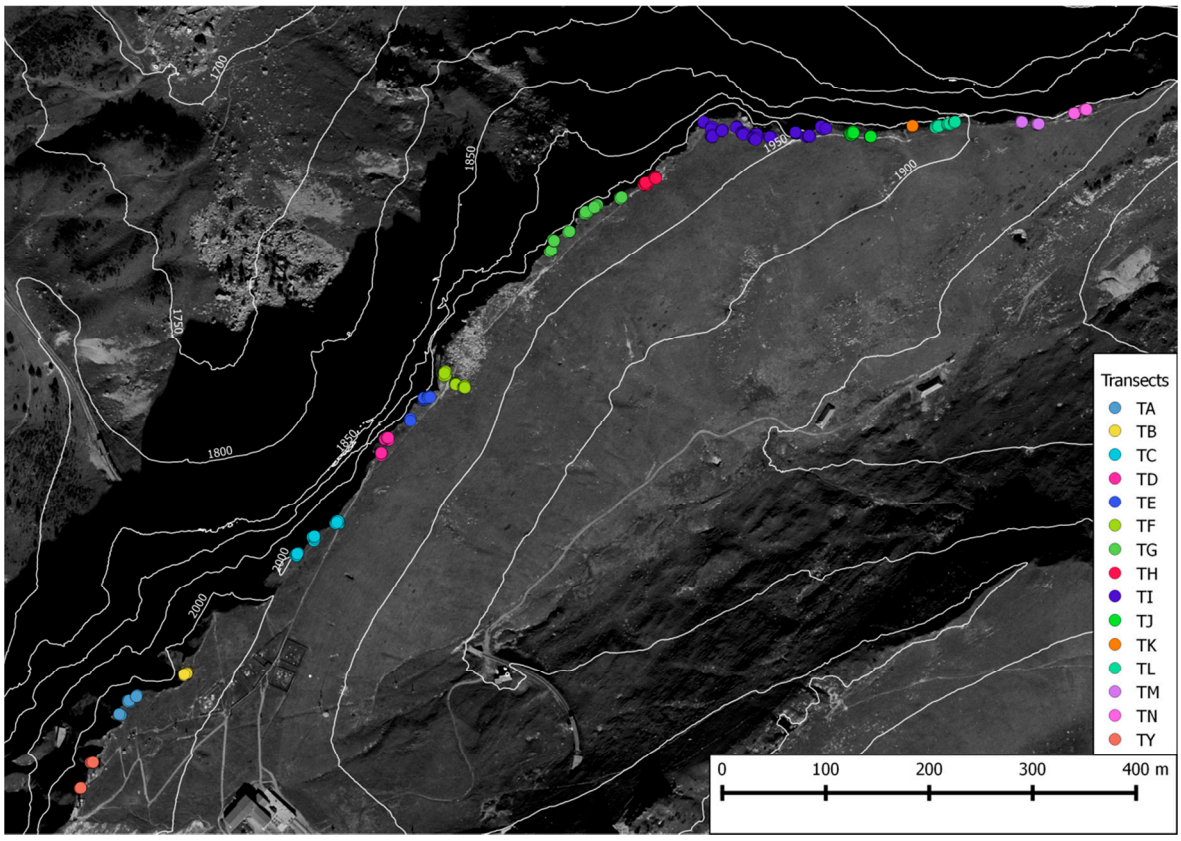


*Figure 4.4 Sampling locations of **B. laevigata** and transects. Sampling locations were selected along the ridge following a random cluster sampling guided by the population density of the plant. Transect F is the only exception with an orthogonal direction from the ridge. A total of 361 points representing individuals were sampled in 60 4x4m plots with at least five individuals per plot.*

4.2.3 Genetic data and population structure

Genetic data were developed by collaborators Céline Geiser and Christian Parisod from the University of Neuchatel using amplified fragment length polymorphisms (AFLPs). They first tested a set of 38 AFLP primer combinations and retained the six best regarding polymorphism and reproducibility (MCAG/EATC, EAGG/MCGG, MCAG/EAAT, EACT/MCAC, MCGA/EATA and MCGG/EATA). To evaluate the error rate, they replicated 15% of the individuals and obtained an error rate of 2.93% (Geiser 2014).

Among the 266 AFLP markers obtained, 233 were polymorphic (frequency of minor variant >0.05). The following histogram shows the distribution of the minor allele frequency, which is skewed towards low frequencies. Only these markers are used in subsequent analysis.

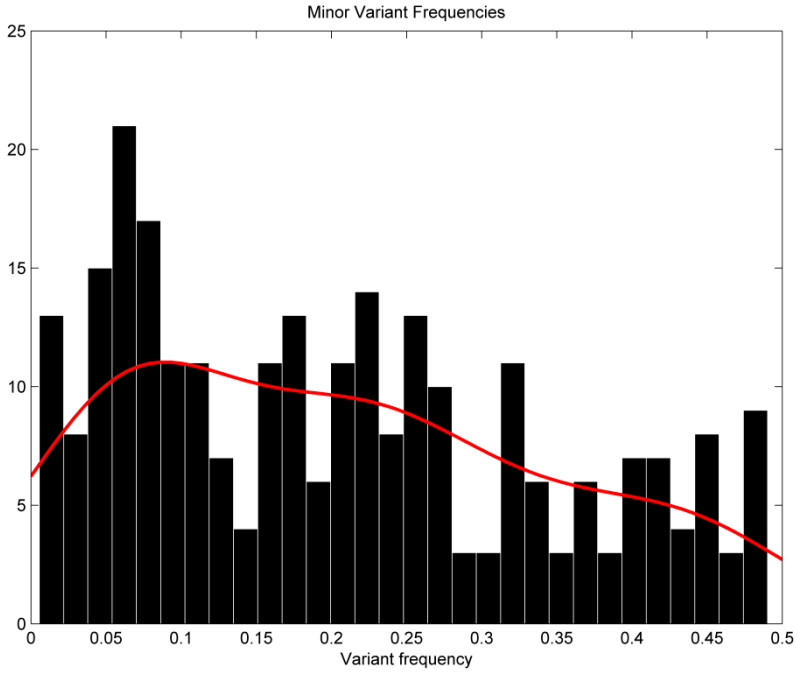


*Figure 4.5 Minor variant frequency for **B. laevigata** AFLP dataset. A kernel window curve (in red) was added to facilitate the visualization of the distribution. The distribution is skewed towards low frequencies. Only markers with a polymorphism >0.05 were used in subsequent analysis*

Because AFLP are dominant markers, we could not gain information on the heterozygosity of the dataset. Therefore, other analyses are often performed on AFLP, such as pairwise relationship coefficients computed in SPAGeDI. The pairwise relationship coefficient was computed at 20 distance intervals and the significance was assessed with 9999 permutations of individuals' values (described in section 3.4.2). Figure 4.6 illustrates the fine-scale genetic structure of *B. laevigata* and shows that nearby individuals were genetically related until the fourth distance class. Pairwise relationship is strong within the 233 markers used, and decreases rapidly with distance.

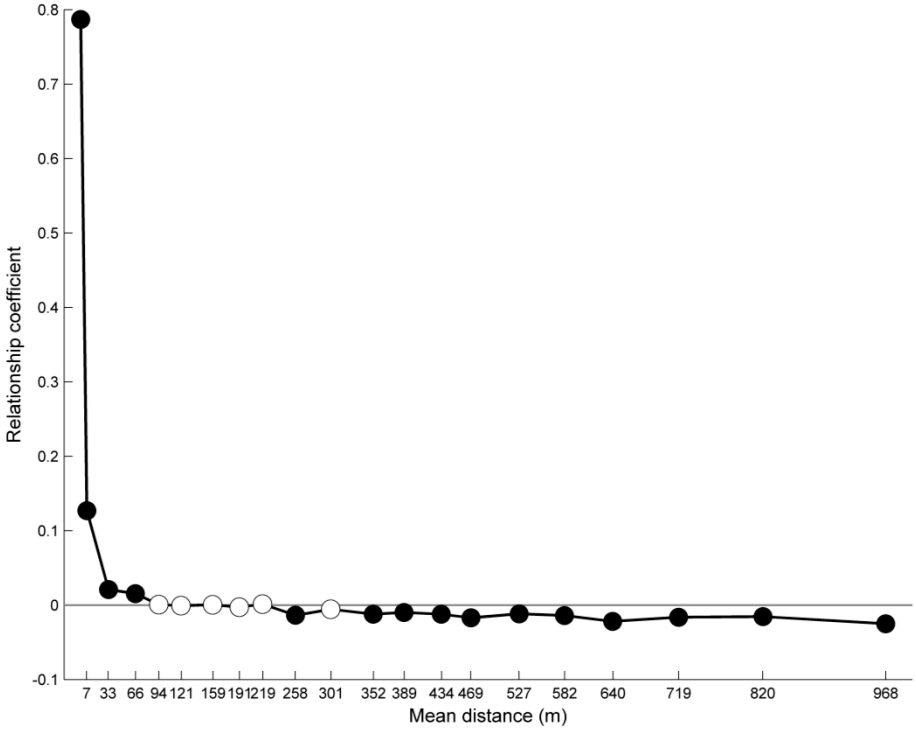


*Figure 4.6 Pairwise relationship coefficient for dominant markers in **B. laevigata**, assuming an inbreeding coefficient of 0.5.Pairwise relationships are calculated for 20 intervals of distances and are shown in black when significant (p-value < 0.05/20) and in white when not significant. The graph shows a sharp decrease of pairwise relationship at short distances.*

We used Structure (section 3.4.3) to assess population structure, using a model with admixture and adding a row of recessive alleles for AFLP data, as described in the manual. We set a burnin of 20 000 and 100 000 simulations and each K was run 20 times. Afterwards, we uploaded the results in Structure harvester, which indicated two distinct clusters using the Evanno method (Table 4.5). These results are consistent with Geiser (2014), who also identified two distinct cluster using K-means clustering, with a similar spatial distribution of the two groups. Finally, CLUMPP was used to obtain membership coefficients of each individual.

*Table 4.5 Results of Structure Harvester for **B. laevigata***. Twenty iterations were performed for each K (K=1:6) in Structure and evaluated in Structure Harvester. The names of the columns designate the *mean likelihood LnP(K) and the variance per value of K;* the rate of change of the likelihood distribution Ln'(K); the a*bsolute values of the second order rate of change of the likelihood distribution* |Ln''(K)|; the *Delta K.*

K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
1	20	-41483	0.3			
2	20	-40861	4.8	621.6	190.9	39.16
3	20	-40430	14.3	430.6	269.1	18.71
4	20	-40269	522.4	161.5	405.2	0.77
5	20	-39702	66.3	566.7	240.3	3.62
6	20	-39376	55.1	326.4		

On the map showing the clusters (Figure 4.7), we can observe the clear geographic distinction between Group 1 and 2. These two groups are separated by an area were no *B. laevigata* are growing and correspond to a stony area. We used the delimited populations A and B (populations definition based on both genetic and geographic structure) to perform the population method BayeScan.

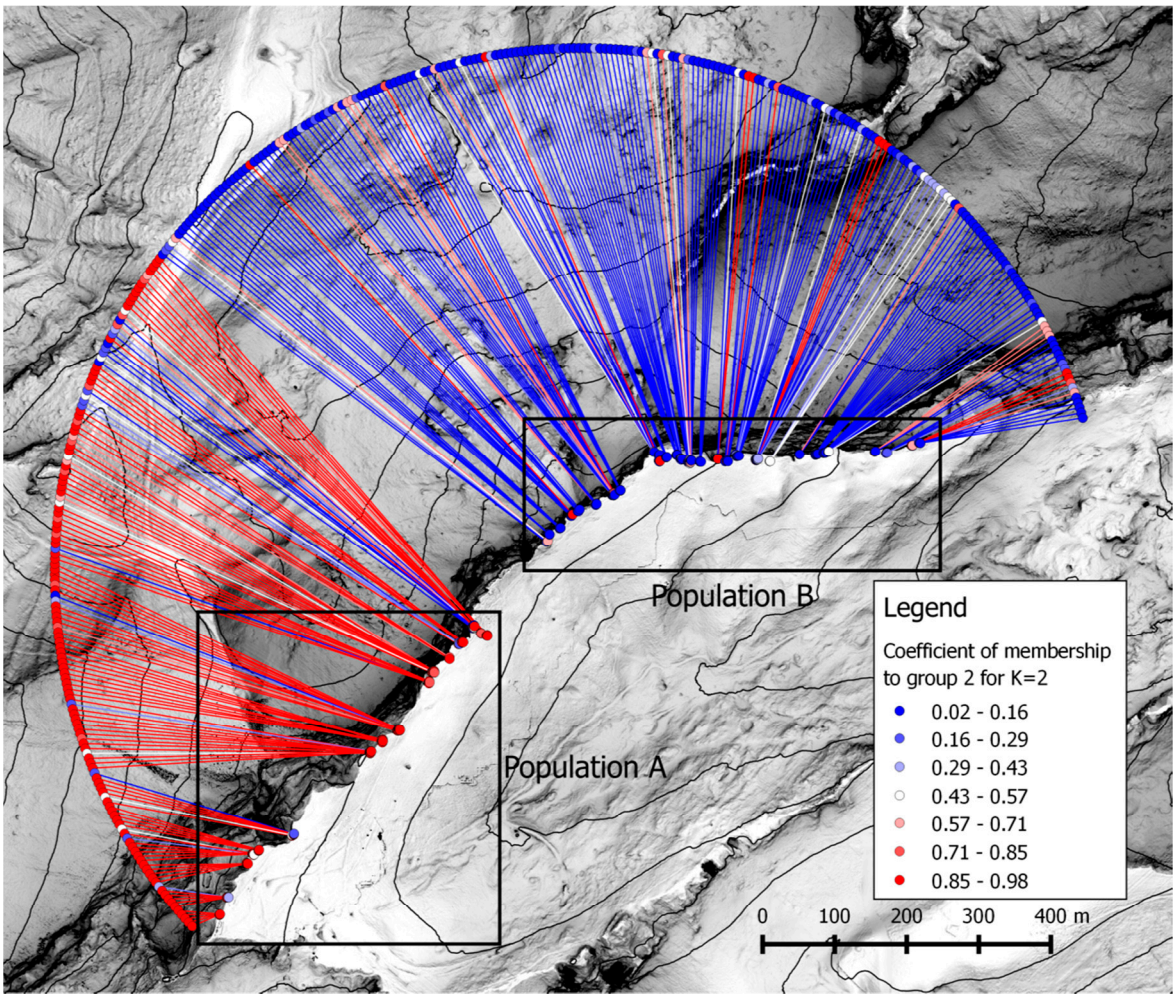


*Figure 4.7 Map showing the membership coefficient [0-1] to the second group identified for **B. laevigata** in case of K=2. Two populations were identified in Structure and the corresponding 20 iterations were aggregated with Clumpp. Starting from this result, we decided to define two populations (A and B) for further use with population genetics methods. A partial circle was added to facilitate the visualization of the 361 individuals.*

4.3 Identification of genetic markers under selection

4.3.1 Samβada

A total of 32154 models were computed in Samβada. The significance threshold is based on a Bonferroni correction (0.01/number of models computed) and corresponded to a minimum G and Wald scores of 29.3 and 26.4 respectively. At this level, Samβada detected 34 significant associations involving 21 genetic markers. Among these models, 18 involve membership coefficient of population structure, latitude and longitude. The remaining associations involve seven DEM-derived variables and four temperature variables (Table 4.6) with moderate G and Wald scores (mean of 38 and 29 respectively).

Four genetic markers are associated with temperature variables, and height with DEM variables. Genetic markers are most frequently correlated to primary terrain attributes of curvatures (Hcu, Vcu, and Cu), followed by Orientation (structure tensor variable) and single occurrences with ruggedness (VRM), protection index (MPI) and sky view factor (SVF). There is no significant association with solar radiation or hydrology variables. We can notice the sensitivity of models to spatial resolution, as most markers are significantly associated with these variables at one or rarely two spatial resolutions, usually of 4 and 8 meters. On the other hand, significant markers correlated with temperature variables are mostly associated to mean temperature (mea) or mean temperature a 1pm (m1p) measured from late June to mid-July. One marker (C1V428) is detected by both types of environmental variables but its association with SVF was biased because of the presence of an outlier (see Appendix II.a). We show in Figure 4.14 its association with mean temperature between the 13 to 26 of July.

Seven markers are associated to latitude, longitude or membership coefficient of population structure only and 3 with these variables and environmental variables. Marker C2N142 for example, was associated with orientation variable (structure tensor) but also with latitude, longitude and membership coefficient (Appendix II.a). Similarly, marker C1N272 (Figure 4.16) is associated to the minimum temperature (05 to 18 October) and with latitude and longitude.

Multivariate models

Before computing bivariate models, we first had to assess again the correlations between variables in order to limit collinearity (section 3.3 p55). We used the same script as the one defined in section 3.3 but with a maximum correlation threshold of 0.8. The resulting dataset contained 59 variables (including latitude, longitude and coefficient of membership). Afterwards, we set Samβada parameters to output only significant models with improved explanatory power compared to univariate models (see section 3.5.1 p59).

Only one association model came out as significant, for marker C2V404 with longitude and altitude (4m). However, this model should not be retained since these two variables are highly correlated (-0.91). In fact, our expectations were to evaluate whether genetic markers associated with geographic or population structure would be better explained by including an additional environ-

mental variable. Therefore, we had to add coordinates in the dataset, despite their eventual correlations with other variables.

*Table 4.6 Results for univariate multi-resolution Samβada results on **B. laevigata**. The table shows one model per line with the following columns: the AFLP marker, the associated variable, the resolution of the best G score for the models involving a DEM variable, the best G and Wald scores among all resolutions, the AIC of the model at the best resolution, the minor allele frequency of the marker (MAF), the Q value and Fst of the marker in BayeScan results, the Moran's I and its p-value for the marker with a neighbourhood of 20 individuals, the Moran's I of the variables with a neighbourhood of 20 individuals. Models are ranked according to their Wald score.*

Marker	Variable	Best Resolution	Best G score	Best Wald score	AIC	MAF	Bayescan Qvalue	Bayescan Fst	# of significant resolutions	Moran for marker	Moran P-value for marker	Moran for variable	Moran P-value for variable
C1N109	Hcu_08	8	39.58	34.69	338.6	0.21	0.902	0.034	1	0.139	0.001	0.316	0.001
C1V342	VRM_04	4	36.94	33.09	278.6	0.16	0.898	0.032	1	0.215	0.001	0.448	0.001
C1N109	Orientation	8	41.28	32.43	336.9	0.21	0.902	0.034	2	0.139	0.001	0.953	0.001
C1N256	M1pJun1528		34.97	30.99	309.4	0.18	0.907	0.033		0.176	0.001	0.763	0.001
C1V138	Hcu_08	8	30.92	30.83	149.6	0.07	0.895	0.033	1	0.116	0.001	0.316	0.001
C1N81	meaJul1326		42.94	29.55	392.8	0.29	0.910	0.034		0.134	0.001	0.422	0.001
C1V428	meaJun2912		43.43	28.96	212.0	0.11	0.854	0.036		0.271	0.001	0.641	0.001
C1V200	Cu_08	8	34.11	28.57	447.2	0.37	0.869	0.032	1	0.120	0.001	0.562	0.001
C1V428	SVF_01	1	36.73	27.93	218.7	0.11	0.854	0.036	2	0.271	0.001	0.136	0.001
C1V485	M1pJun1528		37.17	27.36	253.6	0.14	0.904	0.034		0.129	0.001	0.763	0.001
C1V206	meaJul1326		39.80	27.29	235.7	0.12	0.898	0.032		0.131	0.001	0.422	0.001
C1V200	Vcu_08	8	30.42	27.07	450.8	0.37	0.869	0.032	1	0.120	0.001	0.583	0.001
C1V344	MPI_02	2	31.71	27.06	427.1	0.32	0.911	0.033	1	0.113	0.001	0.273	0.001
C2N142	Orientation	4	32.58	26.75	314.8	0.18	0.841	0.038	4	0.219	0.001	0.807	0.001
C1N272	minOct0518		62.27	26.46	328.6	0.23	0.900	0.035		0.119	0.001	0.533	0.001

4.3.2 BayeScan

For BayeScan, we used the populations defined in section 4.2.3 and converted the Structure file to BayeScan file format using PGDspider (Lischer & Excoffier 2012). Because we could not compute inbreeding coefficient (Fis) from AFLP data, we decided to set a higher bound of 0.5 for this parameter, as proposed in BayeScan manual. This way, BayeScan can move freely the value of Fis within its prior range in order to incorporate the uncertainty on this parameter.

BayeScan does not detect any markers under selection (see Appendix II.a). The Q-values ranged from 0.9 to 0.78, far from showing any decisive evidence of selection.

4.3.3 Comparison between methods

Comparisons between methods have a limited utility in this case study since BayeScan did not detect any markers under selection. However, we are also looking for correlations between methods and to distinguish significant markers from neutral ones. However, Figure 4.8 shows that there is no correlation between BayeScan's Q-value and Samβada's Wald score. The same observation is done between BayeScan's Q-value and Samβada's G score (not shown). Similarly, Figure 4.9 shows no correlation between BayeScan's Q-value and Samβada's AIC. There is thus no difference of BayeScan Q-value between significant and non-significant markers identified in Samβada.

On Figure 4.10, we can notice that all detected markers by Samβada show a higher spatial autocorrelation compared to neutral loci. SA of detected markers decreases after spatial lags of 20 and 40 neighbours and stabilizes around the mean value taken by most of neutral markers.

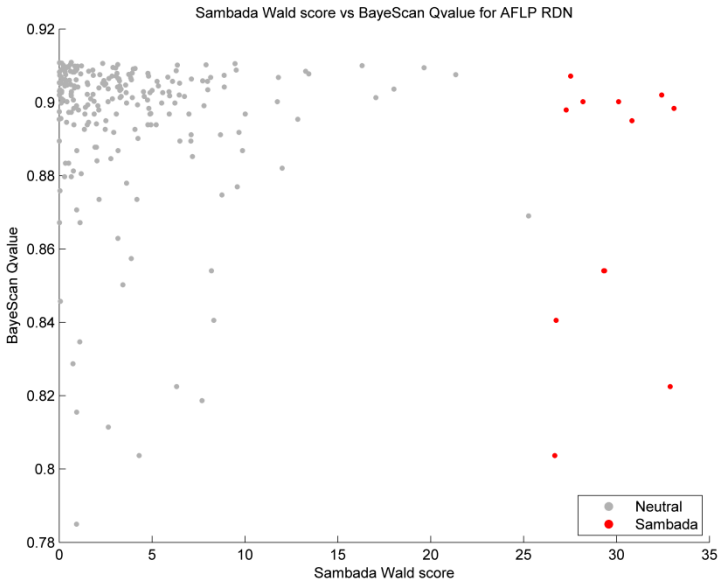


*Figure 4.8 Scatter plot showing BayeScan's Q-value against Samβada's G score for **B. laevigata**. Markers possibly under selection by environmental variables are displayed in red. The figure shows that there is no correlation between these independent methods.*

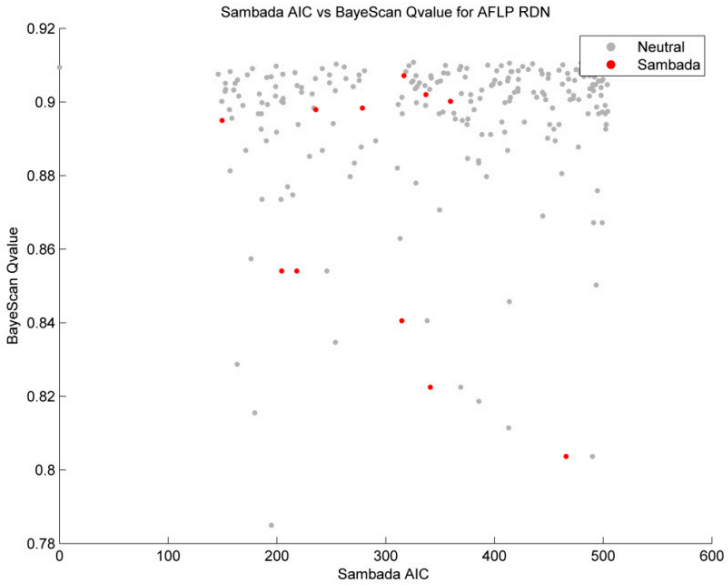


*Figure 4.9 Scatter plot showing BayeScan's Q-value against Sambada's AIC for **B. laevigata**. Markers possibly under selection by environmental variables are displayed in red. The plot shows that there is no correlation between these independent methods*

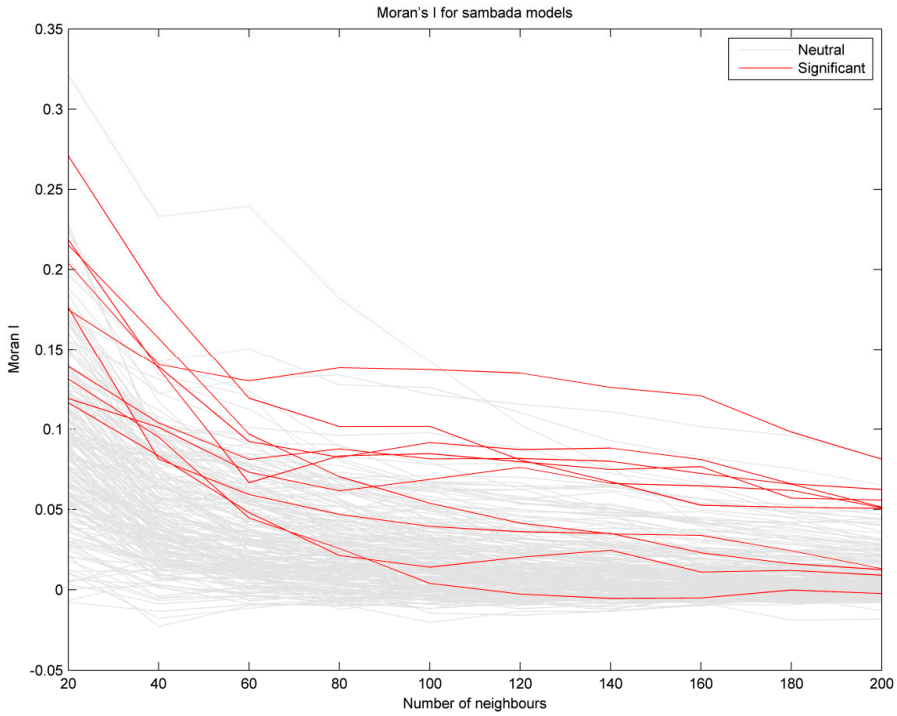


*Figure 4.10 Moran's I Correlogram for **B. laevigata** markers using increasing spatial lags from 20 to 200 neighbours. Neutral markers are shown in grey and markers possibly under selection by environmental variables in red (see Table 4.6 for the detected markers). Significant genetic markers in Sambada are amongst the most spatially autocorrelated*

4.3.4 Visualisation of significant associations

Spatial analysis of significant models allows us to visualise the correlation between genetic markers and environmental variables as well as to analyse spatial autocorrelation patterns. The main purpose of comparing these graphs is to find common trends and differences between the significant models. Details on the different parts of these graphs can be found in section 3.6 (p65).

Figure 4.11 to Figure 4.16 illustrate six of the significant Samβada models (other significant models are provided in Appendix II.a). They were chosen because they represent the most significant associations involving DEM-derived variables (Figure 4.11 to Figure 4.13) or temperature measurements (Figure 4.14 to Figure 4.16).

The first observation we can make is that DEM-derived variables involved in significant models do not have the same response to change of spatial resolution. In fact, some show a sharp increase of Samβada's scores at 8m (Figure 4.11 for example) while others show a local optima at 4m (Figure 4.13). On the other hand, it is important to observe that DEM-derived variables show clustered values at coarse resolution and a high spatial autocorrelation. While this was inevitable for loggers variables (because the same logger measurements were attributed to several samples), it is less obvious for DEM-derived variables. Indeed, by degrading very high resolution at multiple scales, close points were more likely to be located on the same pixel at coarser resolutions. This is well illustrated in Figure 4.11, where the G and Wald score are low except at 8m. In Figure 4.13, however, we can see that VRM was close to significance at 2m before showing a peak at 4m. Similarly, minor alleles of most significant markers are highly spatially clustered, as illustrated by their maps and by their Moran's I correlograms. However, local indices of spatial autocorrelation (LISA) rarely indicate the presence of local clusters.

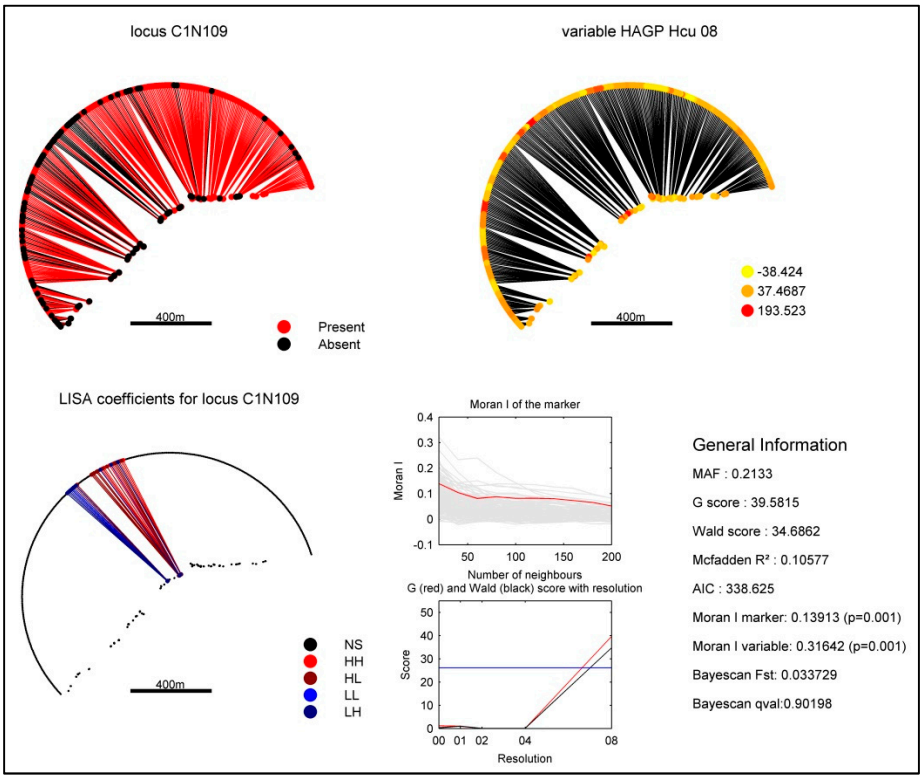


Figure 4.11 Visualisation of the main results for the model involving marker C1N109 and variable Curvature at 8m.

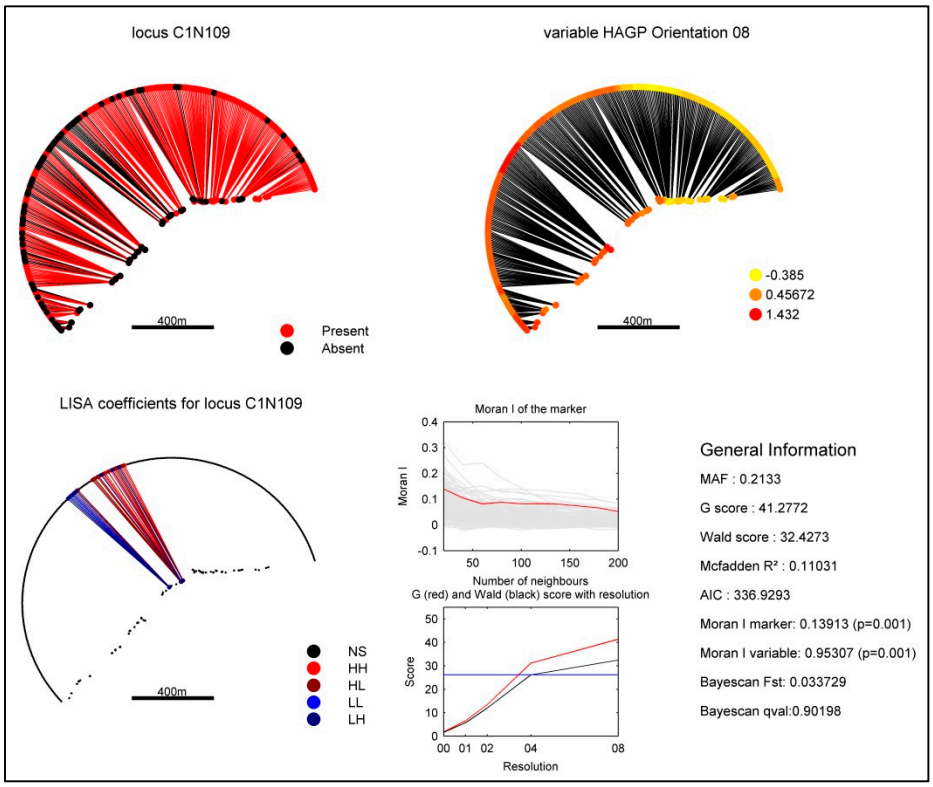


Figure 4.12 Visualisation of the main results for the model involving marker C1N109 and variable Orientation at 8m.

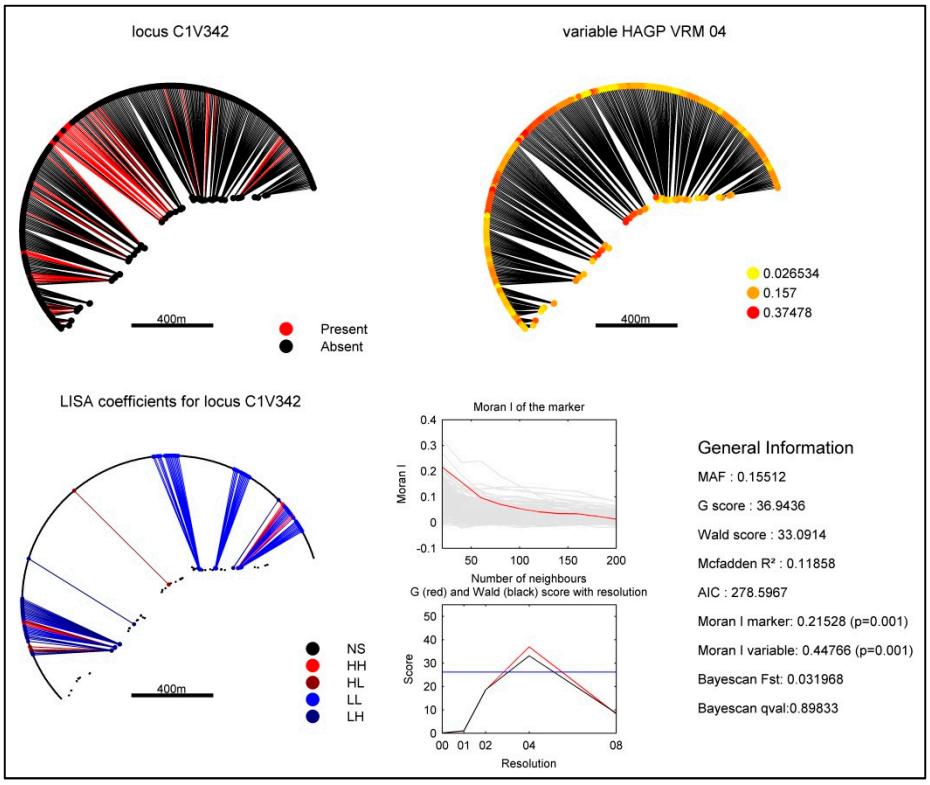


Figure 4.13 Visualisation of the main results for the model involving marker C1V342 and the Vector Ruggedness Measure at 4m.

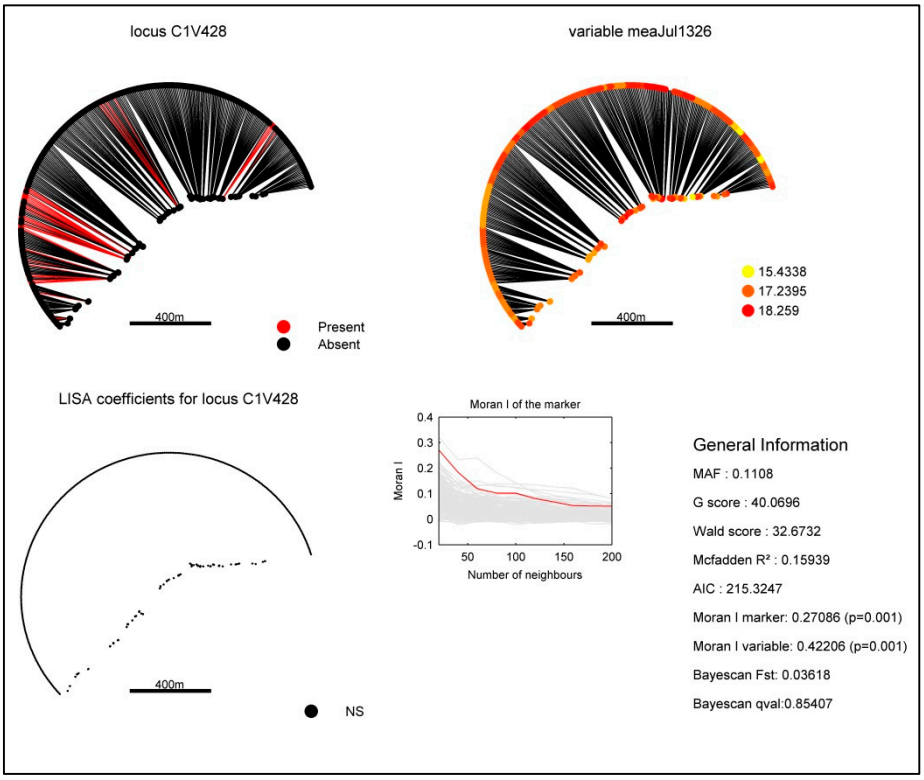


Figure 4.14 Visualisation of the main results for the model involving marker C1V428 and mean temperature between the 13 and 26 of July.

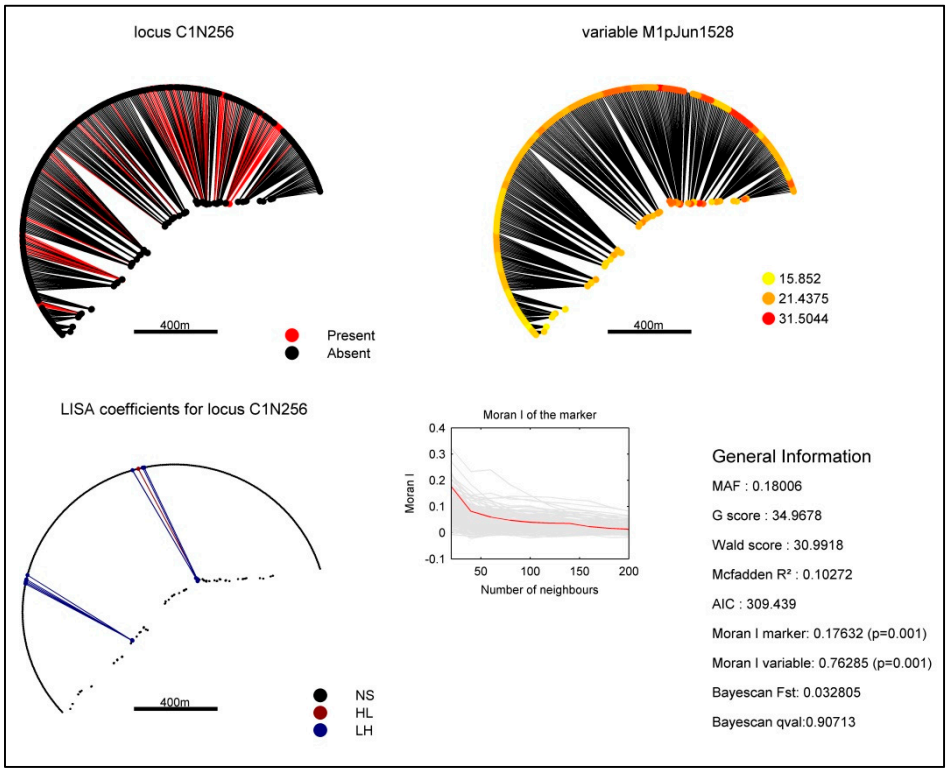


Figure 4.15 Visualisation of the main results for the model involving marker C1N256 and mean temperature at 1pm between the 15 and 28 of June.

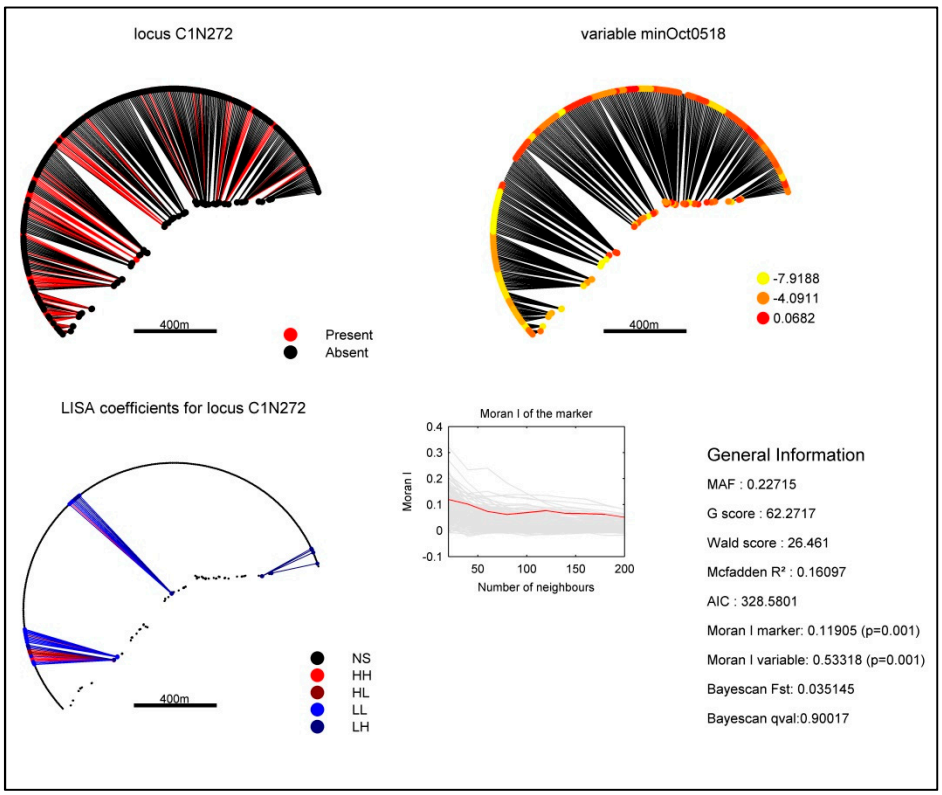


Figure 4.16 Visualisation of the main results for the model involving marker C1N272 and minimum temperature between the 05 and 18 of October.

4.4 Discussion

Relevance of VHR DEM-derived variables to model microclimatic variability

The apparently homogeneous ridge on which *B. laevigata* was sampled, showing a constant slope and slight changes in orientation, turned out to be highly heterogeneous at a high resolution. Prior work on ecotypes of *Biscutella laevigata* (Parisod & Christin 2008) suggested a mosaic distribution of subalpine and alpine habitats, and the use of VHR DEM-derived variables here brought clear evidence of topographic control on micro-climatic patterns.

Our climatic models consistently reported decreasing adjusted R^2 at the coarsest spatial resolution (4m), supporting the hypothesis that VHR provides better predictions of climatic variables in heterogeneous areas such as mountains. However, our models did not generally converge towards a clear optimal resolution and reveal that the most suitable resolution depends on the type of DEM-derived variable considered. This is particularly well illustrated by vrm, showing the highest significance at 0.5m and highlighting that soil characteristics are best grasped when initially computed with as much details as possible, whereas hydrology variables, such as twi, reach optima at different resolutions (Böhner & Selige 2006; Buchanan *et al.* 2013).

Our results further bring advantages of using a large panel of DEM-derived variables. Terrain wetness index (twi) showed the highest explanatory power among the DEM-derived variables here tested, highlighting a relevant proxy for dryness across the studied landscape (Appendix III). In addition, models including more variables such as eastness and slope best predicted temperature, probably because these primary attributes have a high influence on radiation and wind exposure (Wilson & Gallant 2000; McVicar *et al.* 2007; Appendix S5). For instance, in our specific study area, twi partially accounted for the distance to the ridge as well as for the protection from wind, which could further contribute to temperature and humidity variability. In fact, distance to ridge and twi were moderately correlated at very high resolutions (i.e. 0.6 at 0.5m and 0.7 at 1m) and dropped to 0.3 at coarser resolutions. Although such correlations are inevitable and likely blur interpretations, our models showed that most of the significant contribution of twi were obtained at 0.5 and 2m, when the correlation between twi and distance to ridge were not the strongest. This highlights the relevance of multi-scale analysis.

Among other overlooked DEM-derived variables in the literature, vector ruggedness measure (vrm) appeared as the most important predictor of soil moisture (MSM), suggesting that vrm at such high resolution is a suitable proxy for the distribution of stony soils along the ridge and thus for soils with different porosities. Accordingly, the negative coefficients observed here support the hypothesis that high roughness highlight stony soils implying low soil moisture, whereas low roughness reflects developed soils retaining higher moisture. This vrm variable, measuring vector dispersion across the central pixel rather than being a derivative of slope, represents a much better proxy than related proxies such as Terrain Ruggedness Index (Appendix S3&4), as previously stressed by Sappington *et al.* (2007).

Although our results show a significant contribution of micro-topography to model micro-habitat, unmeasured factors may play a major role. For instance, it is generally admitted that high elevation and exposed sites are more likely to be coupled with free air environment as compared with

low elevation sites that are protected (Pepin & Seidel 2005). However, we observed 5°C difference in ranges for AT and up to 8°C for DT. Such important temperature variability over short distances cannot only be due to large scale effects and support our evidences for a micro-topographic control (Fridley 2009). In addition, VHR DEM-derived variables in our models highlighted the lower relevance of elevation as compared with studies at regional or continental scale. Despite a correlation of -0.99 reported between temperature and elevation across Switzerland (Zimmermann & Kienast 1999), we showed here that the 0.5°C decrease per 100m elevation increase do not hold at a local scale. Therefore, the important variability of temperature observed here is likely valid in various mountainous areas, even when microhabitats variability is only partially distinguished from large scale factors. Our results thus confirm that proxies other than elevation can - and in fact probably better - account for temperature variability in as mountainous areas.

On top of micro-climatic factors, meso-climatic ones might affect climatic variables in the study area. For instance, varying wind patterns and cloud cover across the studied ridge could impact on the variability of local climates. The results obtained here for micro-topography are however not disqualified by meso-climatic patterns. In contrast to common cloudiness on the highest part of the study area early and late during the growth season, the contribution of DEM-derived variables appeared consistently significant at different time periods, demonstrating a substantial effect of micro-topography. In addition, several DEM variables such as protection index, sky view factor or ruggedness might constitute surrogates for protection from wind at a micro-climatic level. Noticeably, temperatures measured during the snow episode provide an indirect measure of snow cover, as loggers situated under the snow during this period (P9) did not show a daily cycle of temperature at sampling locations. Therefore, modelling of snow cover heterogeneity could be improved by combining topographic variables (Gottfried *et al.* 1998; Randin *et al.* 2009) with the daily cycles of loggers. Our results thus highlight the role of micro-topographic effects and the need to consider different measured variables and temporal variability at a scale pertinent for plants, as previously reported by Körner (2003) and Scherrer & Körner (2011).

Detecting signature of natural selection with multi-scale DEMs and climatic variables

We studied the adaptation of *B. laevigata* at a local scale with two goals: first, to characterize gene flow and to find signatures of selection at a local scale. Second, to assess the relevance of VHR DEM-derived variables and logger measurements to detect these signatures.

Regarding the likelihood of observing selection patterns, it should be noted that there is no strong evidence of selection. In fact, significance of associations in Samβada is moderate and BayeScan did not detect any markers under selection, despite the high amount of individuals. This lack of common ground suggests that Samβada's detections probably contain a large proportion of false positives and an additional population genetics approach, such as FDIST or Mcheza (Antao & Beaumont 2011), would have provided a valuable comparison to BayeScan results.

One explanation for the absence of local adaptation could be the limited gene flow. Figure 4.6 indeed shows a steep decline in pairwise relationships between individuals over 30m and high-

lights that, beyond the patches in which several individuals were sampled, little dispersal is expected. In addition, this observation corroborates the results obtained by Parisod & Christin (2008), who also observed a strong decline in pairwise relationship at short distances for the same population, with another set of AFLP markers. Furthermore, the different distribution maps of genetic markers we propose in section 4.3.4 all show a strong clustering of genetic variants in patches and a limited spread in surrounding patches, indicating a strong isolation-by-distance. Finally, a clear population structure was evidenced due to a geographic barrier (Figure 4.7), which consists in a large and steep stony area where soil cover is poor. This area is also barely covered by vegetation and thus provides a likely explanation to the limited gene flow between the two populations identified. In addition, the steep terrain conditions must render the dispersion of seeds difficult as *B. laevigata* seeds disperses by wind and gravity (Parisod & Christin 2008). However, Geiser (2014) identified instead a pre-zygotic barrier, suggesting that this is probably a major force shaping the genetic structure in this population. A possible role of a post-zygotic barrier was not excluded either. Geiser (2014) also found that peaks of open flowers were not shifted in time between both populations but that very few flowers of both populations co-occur at the same period. However, this observation was made for only one of the two growing seasons evaluated.

Correlations between genetic markers and *in-situ* measurement of climatic variability most often occurred with the mean temperature (either with the mean of all measurements during a range of days or with the mean of measurements at 1pm during a range of days). We show that spatial analysis of these significant models provided important information regarding spatial distribution patterns of genetic markers. In fact, markers associated to temperature variables revealed two spatial patterns in genetic data. One is a cluster of presence/absence in the highest elevation area (southwest), where temperature is slightly lower than average (Figure 4.14 and Figure 4.16 and other examples in Appendix II.a). The second pattern is a cluster located on the north-east of the ridge, where orientation is changing (facing south) and where temperatures are usually higher (Figure 4.15 and other examples in Appendix II.a). Both patterns were expected after fieldwork and analysis of loggers' data as we noticed that the change of orientation on the ridge implies an increase in temperature and we could expect it to be important for plants habitat. The other pattern of cold temperature close to the summit (south-west) was also evidenced in our data and in the field, as this part of the ridge has a higher elevation and is more often exposed to wind and cloud cover. These results show that *in-situ* measurements of climatic variables are relevant at such scale.

DEM-derived variables involved in significant models are related to morphometry (vrm, curvature, mpi), which are proxies to the variability of stony soils, snow cover and soil humidity. On the other hand, no markers are associated to insolation or aspect, which were our main expectations in a mountainous area. These variables may not show enough variability at sampling locations. In fact, solar radiation variables were highly correlated with slope and aspect, meaning that shade from local relief does not influence solar radiation in the study area. Furthermore, we knew from the regressions models between climatic and DEM variables that solar radiation variables do not contribute to temperature variability.

Regarding resolution, its impact is strong on correlations and highest scores are observed at coarsest resolutions in the range proposed (i.e. 0.5 to 8m). It highlights once again that relevant

topographic features can only be detected with a multi-scale approach, and that starting from a very high resolution is essential here to delimit the ridge correctly. One interesting example is illustrated in Figure 4.13 where marker C1V342 is correlated with vrm at a resolution of 4 meters, resolution at which vrm is moderately clustered, thus limiting a potential bias due to pseudo-replicas. In addition, the distribution of the variants is more widespread than in other models and partially depends on the proximity of the geographic barrier, which corresponds well to the observation that this area is unusually covered by rocks and that vrm could be a proxy for this geomorphological feature. In addition, other significant genetic markers show a similar pattern and are associated with curvature, a known proxy for stony soils or soil humidity (Appendix II.a). Spatial representation also allow us to disqualify models, such as marker C1V428 with variable svf (see Appendix II.a). In this case, the unique low cluster of values of svf biases the regression as these samples are situated just on the other side of the ridge where the view distance is limited to the north.

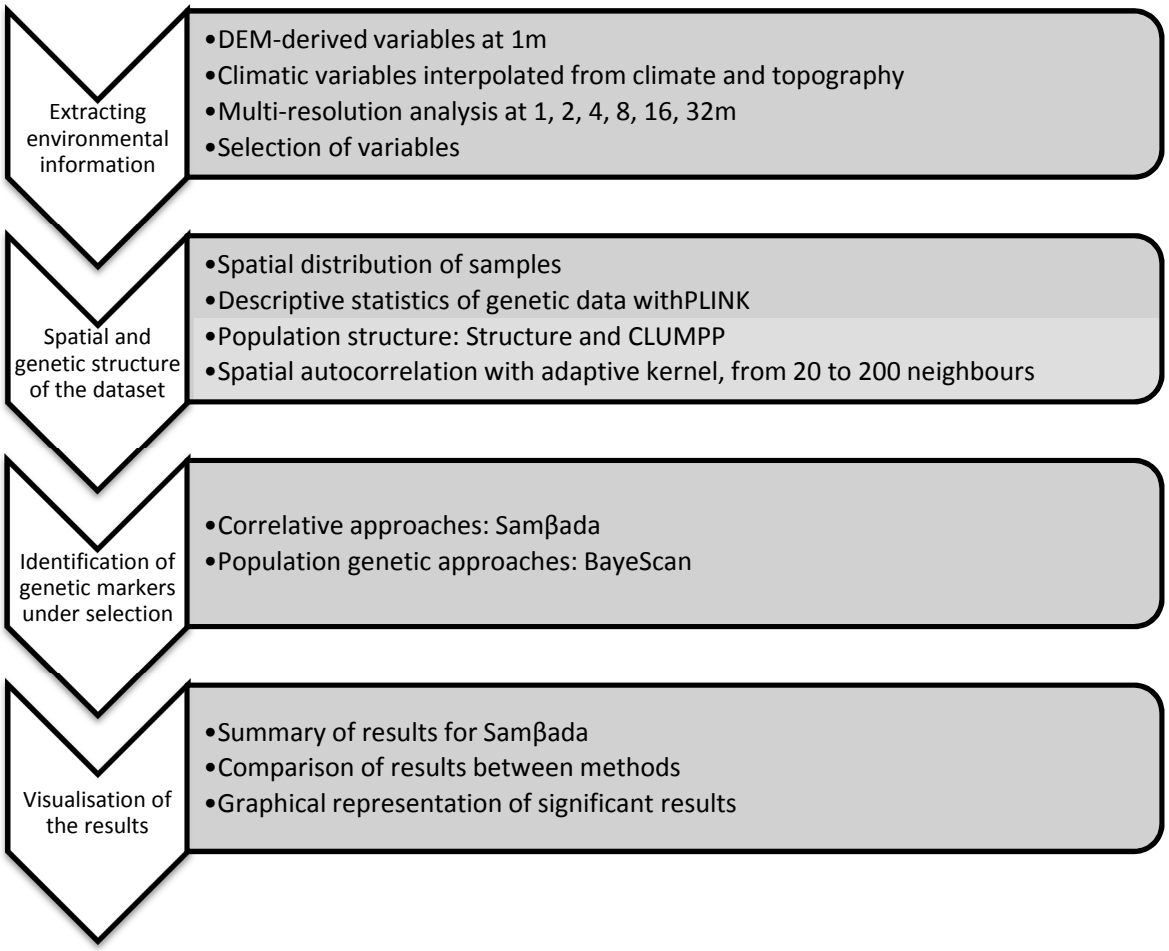
B. laevigata has been studied several times over the past decade (Parisod & Besnard 2007; Parisod & Bonvin 2008; Parisod & Christin 2008). The publication of Parisod & Joost (2010) is of particular interest to compare with our results, as they evaluated signatures of selection in two populations of *B. laevigata*, including the one at “Les Rochers-de-Naye”. In contrary to our results, they found that BayeScan detected more markers under selection than a correlative approach. In addition, they identified several common markers, one of them being highly significant in both methods. Identified markers were mainly associated with altitude and degree-days, which were both spatially autocorrelated. However, both variables had a resolution of 25m, which we considered to be too coarse to delineate the ridge and could be a source of spurious correlations due to pseudo-replication.

To conclude this chapter, we consider that the important variability of the temperature observed on this site should be typically observed in many mountainous areas where microhabitats variability can be distinguished from large-scale factors. On the other hand, we found weak evidences of adaptation of *B. laevigata* but show that a clear population structure can be identified at a local scale in mountainous areas, probably because of limited dispersal and of the presence of a geographic barrier. Although Samβada models are moderately significant, we found that environmental heterogeneity measured with VHR DEMs and in *in-situ* measurement of climate at such a local scale is substantial and could have a major influence on gene flow and adaptation.

Chapter 5 *Plantago Major* in Geneva

Green spaces and biodiversity in general have an important direct and indirect economic value. However, the recent fragmentation and anthropogenic pressure on these habitats is increasing and modifies the connectivity and the way species adapt to urban environment. Therefore, the analysis of their genetic population structure, gene flow and local adaptation is primordial in a perspective of conservation (Cushman *et al.* 2006). In addition, despite being well studied in natural habitats, fragmentation is rarely assessed in urban environments where it is known to be faster (Di Giulio *et al.* 2009).

As part of the Urbangene project on Geneva urban biodiversity, leaves from 479 *P. major* plant individuals were sampled along five transects departing from the city centre. Our purpose was to assess whether climate, topography or urban environment are important to understand both its population structure and signals of local adaptation to the environment.



Plantago major is an abundant and widely distributed synanthropic plant (species that live near, and benefit from, an association with humans). *P. major* grows in a wide variety of habitat and often occurs in human disturbed areas such as in lawns or along roadsides and sidewalks. It often grows in compacted or disturbed soils and thus naturally grows in urban environments, making it important for soil rehabilitation. In Geneva, *P. major* is widely distributed and is thus a relevant species to measure fragmentation and anthropogenic pressure.

Our purpose was to assess whether climate, topography or urban characteristics are important to understand both its population structure and local adaptation to the environment. A sampling design of five transects were chosen along the right shore of the lake, in direction of Geneva airport, along the Rhone river, in direction of the Salève hill and in direction of Anemasse (Figure 5.2). Along these transects, the plants face different environmental conditions influenced by 1) topography, with several small mountains surrounding the city; 2) climate, with the mountain massifs of the French Alps and the Jura creating local variability of climatic conditions; and 3) urban density.

5.1 Environmental Data

Climatic variables at *P. major* sampling locations where recovered from the Swiss eco-climatic dataset (Table 3.4). Although these variables show a large variability over the territory of Switzerland, it was not the case at the scale of Geneva. In fact, 12 variables were discarded because they showed little variability at the study site and were more qualitative than quantitative. On the variables selected, different window sizes were applied to extract the variables at 6 different distances around each sampling point (3x3, 5x5, 9x9, 17x17, 33x33 pixels), resulting in a dataset of 72 climate variables (12x6).

The DEM was obtained from the Geneva canton, at a resolution of 1m. However, it was not possible to produce DEM-derived variables at such spatial resolution because the model contains 105 million pixels. Therefore, we reduced the resolution to 2m before obtaining DEMs at 5 generalized resolutions (4, 8, 16, 32, 64m). Twenty-three DEM-derived variables were computed for each resolution, resulting in a dataset of 138 DEM-derived variables (see Table 3.2). Structure tensor variables, however, could not be computed at 2m due to systematic failure of the software and were thus processed at 4m only.

5.1.1 Variables selected

In order to avoid multi-collinearity, 87 variables were kept after filtering with a threshold of 0.9 (see section 3.3). Eventually, only four variables were coming from the eco-climatic dataset and 80 from the DEM. Like in *B. laevigata* case study, most DEM variables showed a weak correlation with any other variable for at least one resolution, except Total insolation, which was deleted due to its high correlation with Direct insolation. The environmental dataset thus contained 156 variables to which we added latitude and longitude as well as the population structure variable (see section 5.2.2).

5.2 Spatial and genetic structure of the dataset

5.2.1 Spatial structure

All sampling locations were geo-referenced with a standard GPS device during fieldwork (precision <7m). Leave tissue from 479 *P. major* plant individuals were sampled along the five transects by Ivo Widmer (EPFL, LaSIG). The nearest neighbour index described in section 3.4.1 showed significant clustering (0.37, p-value < 0.00001), due to the sampling design in transects. This index was computed on 464 individuals instead of the 479 sampled, due to insufficient quality of genetic data (see section 5.2.2). Regarding the pairwise geographic distances, a few samples were close to each other (<25m) but the average shortest distance is about a hundred meters, with a maximum below 600m. The maximal distance between two points is less than 2.5km.

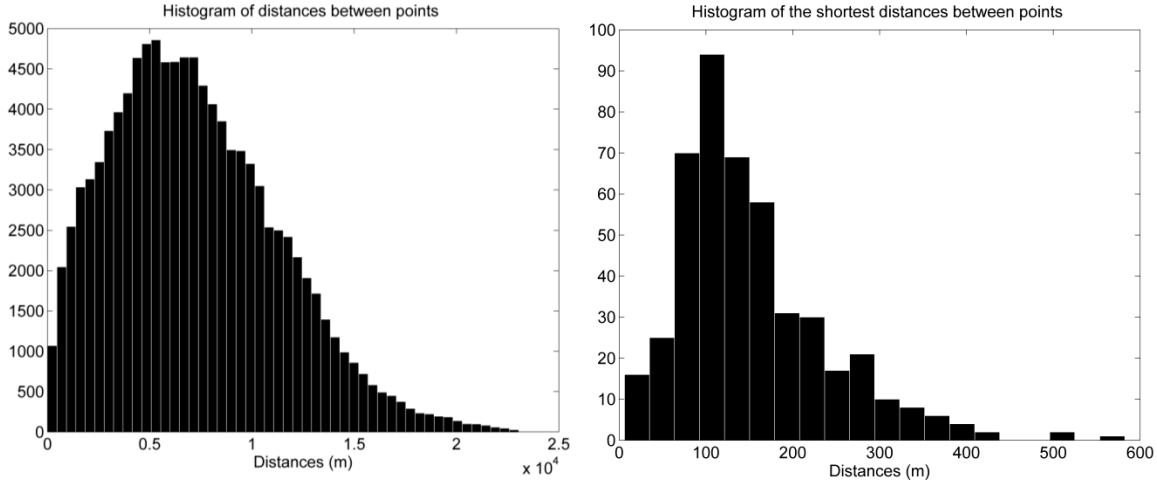


*Figure 5.1 Histogram of pairwise distances between samples (left) and shortest distance between samples (right) of **P. major**. Sampling design is made of 5 transects along which plants were sampled at ≈100m intervals.*

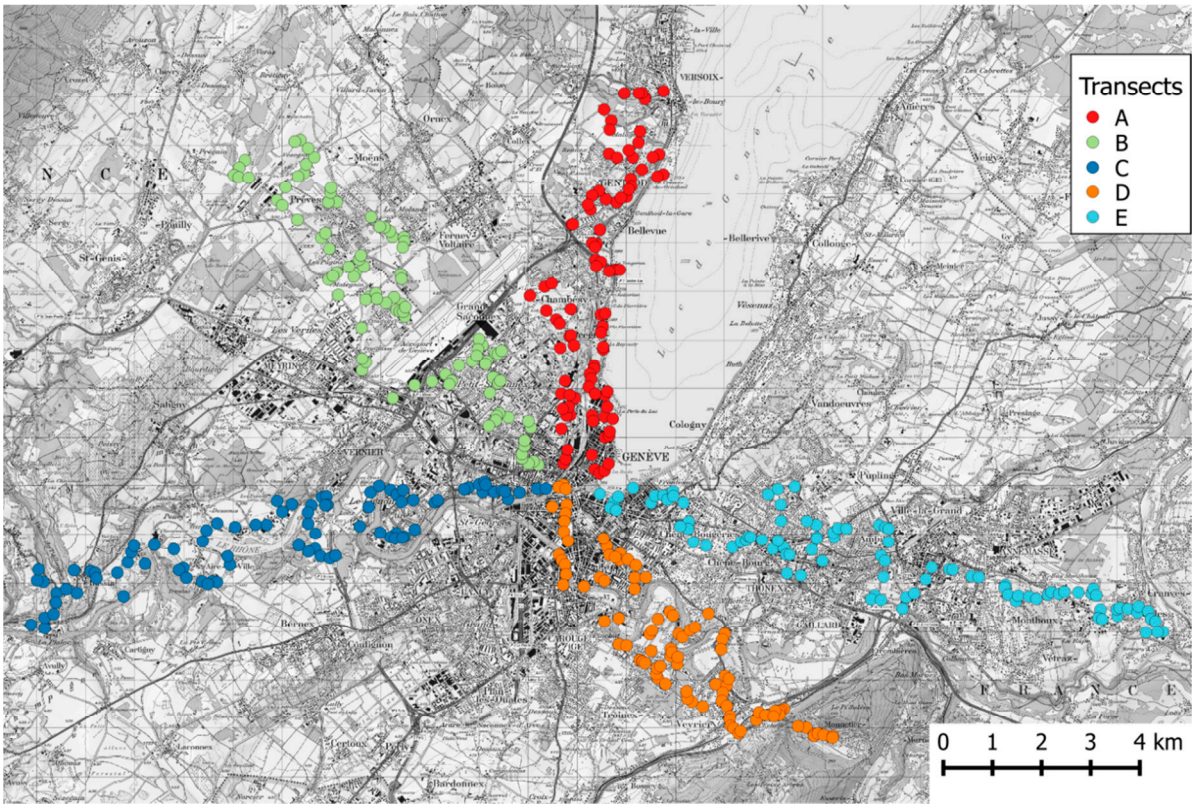


*Figure 5.2 Sampling locations of **P. major** along transects. The sampling strategy was defined to cover five transects departing from Geneva city centre, in order to assess the impact of urbanization and other environmental variables on gene flow and on the spatial distribution of genetic diversity. The present figure shows the coordinates of 479 sampled locations, among which 464 were used.*

5.2.2 Genetic data and population structure

Samples were sequenced by LGC Genomics, a private company based in Berlin, Germany. DNA was extracted from leaves using sbeadex maxi plant kits (LGC Genomics) and restriction-site associated DNA sequencing (RADseq) was performed using the Illumina HiSeq 2000 platform (Illumina). Due to budget limitations, 192 amplified libraries were pooled per lane (3 lanes), thus reducing coverage. The raw genetic dataset contained 20420 SNPs and was pre-filtered for a minimum coverage of six. However, the dataset contained missing data and we performed a series of filtering tasks using VCFtools (Danecek *et al.* 2011) and PLINK . First, we filtered the VCF file in order to keep loci with at most 40% missingness, resulting in a dataset of 5110 SNPs. Then, we computed statistics such as depth of coverage and minor allele frequency (MAF) to check for the quality of the dataset. Afterwards, we converted the VCF file to a PLINK PED file. On the latter one, we filtered all SNPs showing a minor allele frequency below 5% as well as those with a missingness-per-site above 20%. Finally, these filtering steps resulted in a dataset of 464 bi-allelic SNP markers that was used for further analyses. The code for computing these steps can be found in 0.

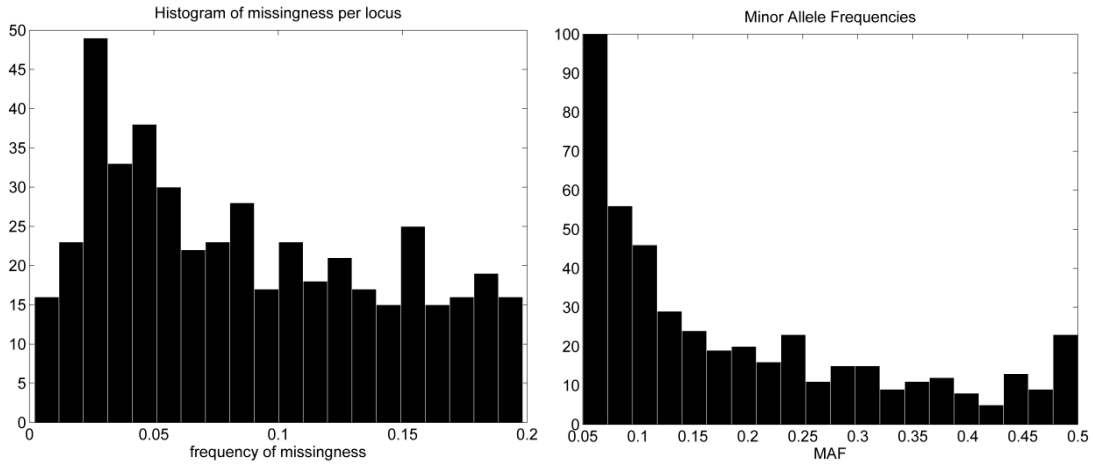


*Figure 5.3 Histogram of missingness-per-site (left) and of minor allele frequency (right) for the 479 individuals of **P. major**. Missingness-per-site was high and we decided to filter loci with more than 20% of the data missing. The MAF histogram shows a high occurrence of low frequencies.*

Figure 5.3 shows the missingness-per-site and the minor allele frequency (MAF) statistic for the 464 SNPs. Missingness-per-site was cut at 20% and there is not particular observation to be made on the first graph. Regarding MAF, we can see that most of the sites show a low minor allele frequency, which did not substantially increase before 0.5. In Figure 5.4, we can observe that, despite the previous filtering operations, missingness-per-individual was still high for a few individuals. Therefore, we deleted individuals with a missingness above 50%, resulting in a dataset of 464 individuals.

The histogram of inbreeding coefficient (Fis; Figure 5.5) shows a strong variability in the dataset with a considerable proportion of individuals experiencing negative Fis values. In addition, we can see in Figure 5.6 that the distribution of negative Fis values seem to be spatially clustered, with transect B showing most of the negative values. We adress this point in the discussion.

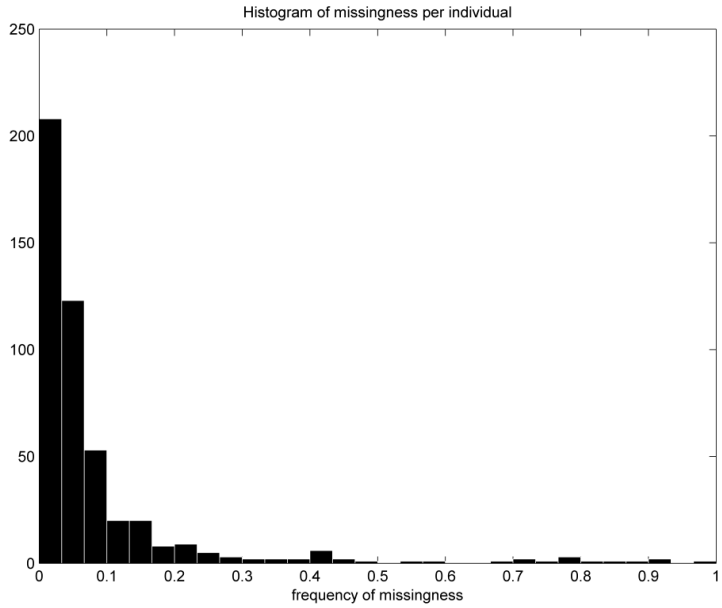


*Figure 5.4 Histogram of missingness-per-individual (left) for the 479 individuals of **P. major**. Missingness-per-individual decreases sharply and only a few samples have more than 50% of the data missing.*

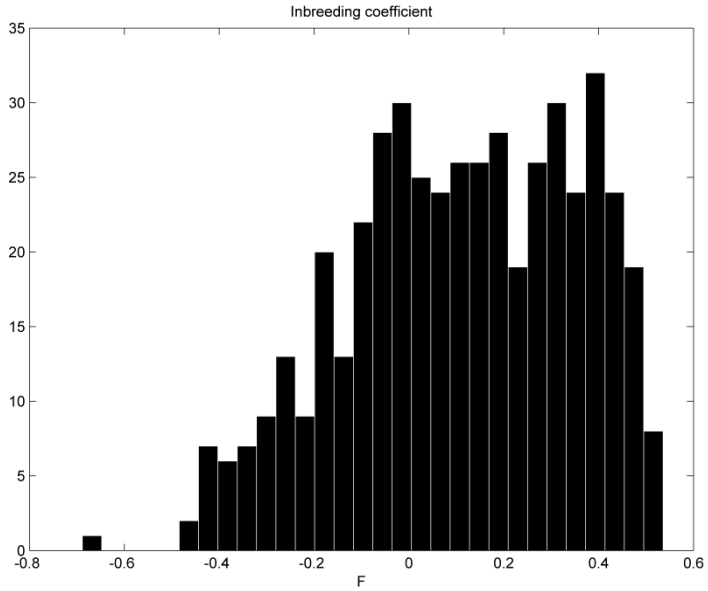


*Figure 5.5 Histograms of inbreeding coefficient (Fis) per individual for **P. major**. Fis of a large part of the samples are negative due to an excess of heterozygotes. This could indicate the presence of a polyploid sub-species.*

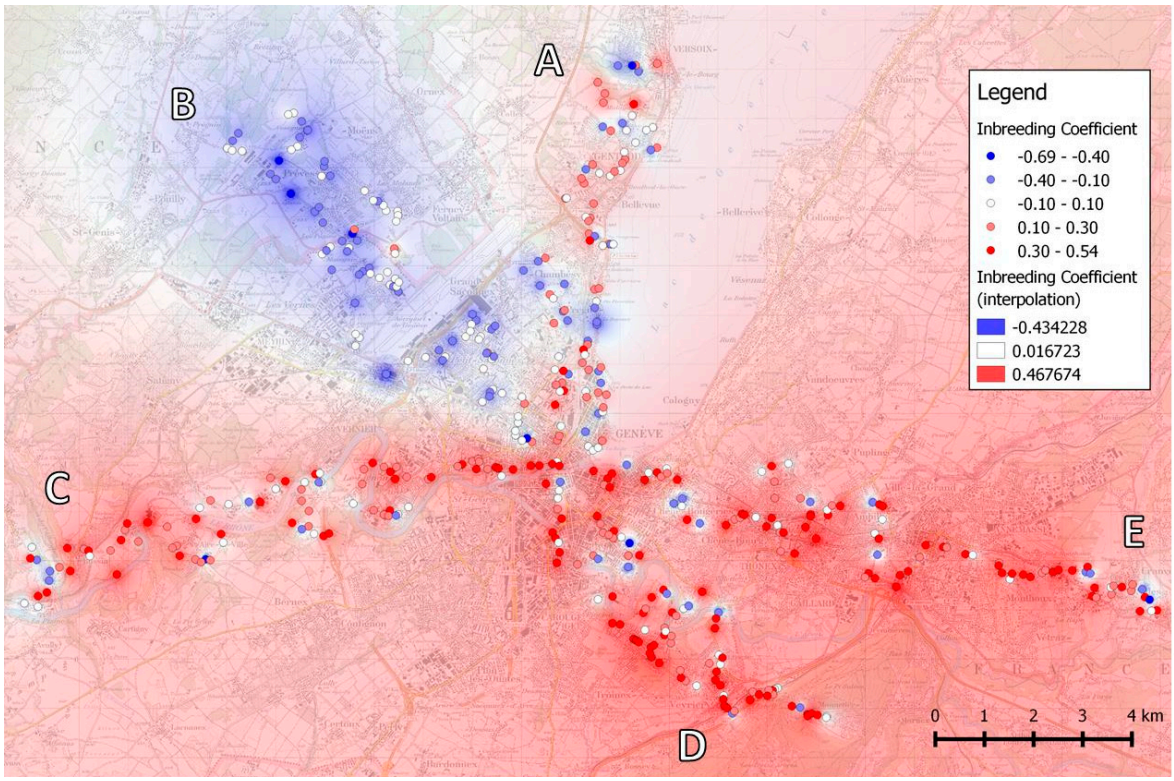


*Figure 5.6 Map of inbreeding coefficient (Fis) per individual for **P. major.** An Inverse Distance Weighted (IDW) interpolation was performed on the Fis coefficients to improve its visualisation. Negative Fis samples are mostly located in transect B, which could indicate the presence of a polyploid sub-species in this area.*

We assessed population structure using the program Structure (see section 3.4.3) and used the Evanno method in Structure Harvester to indicate the most likely number of genetic clusters. In Structure, we used the admixture model with a burnin of 20 000 and 100 000 simulations. Each K was run 20 times. The Delta K indicates that the most likely value is K=3 (Table 5.1). Afterwards, CLUMPP was used to aggregate the twenty runs performed. Finally, individuals were assigned to populations based on the maximum membership coefficient found for one of the three populations. Figure 5.7 illustrates the population structure of the dataset and allows us to distinguish population 1 from 3, while population 2 is composed of a few individuals only. However, Fst divergence is weak between population 1 and 3 (0.068), but higher between these two and population 2 (0.159 and 0.180). Although the spatial segregation of genetic populations is not clear-cut, it seems that in the southern transects (C,D,E) the majority of individuals are assigned to the genetic populations 1 while individuals from populations 3 are more clustered in in the northern transects (A, B). However, the distribution is difficult to interpret due to the admixed transect C (Rhone river) and the presence of individuals from population 1 up to the north of transect A. On the other hand, individuals from population 3 were barely present in transects D and E. Membership coefficients to population 1 only were used further on in Samβada models as population 2 was small and population 1 was largely negatively correlated to population 3.

*Table 5.1 Structure output results for **P. major***. Twenty iterations were performed for each K (K=1:6) in Structure and evaluated in Structure Harvester. The names of the columns designate the *mean likelihood LnP(K) and the variance per value of K;* the rate of change of the likelihood distribution Ln'(K); the a*bsolute values of the second order rate of change of the likelihood distribution* |Ln''(K)|; the *Delta K, showing a most likely value of K=3.*

K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
1	20	-173999	1.9			
2	20	-169021	504.3	4977.2	1446.3	2.86
3	20	-165491	12.2	3530.9	2233.1	182.02
4	20	-164193	162.4	1297.7	285.9	1.76
5	20	-163181	258.8	1011.8	388.1	1.49
6	20	-162557	1946.3	623.6		

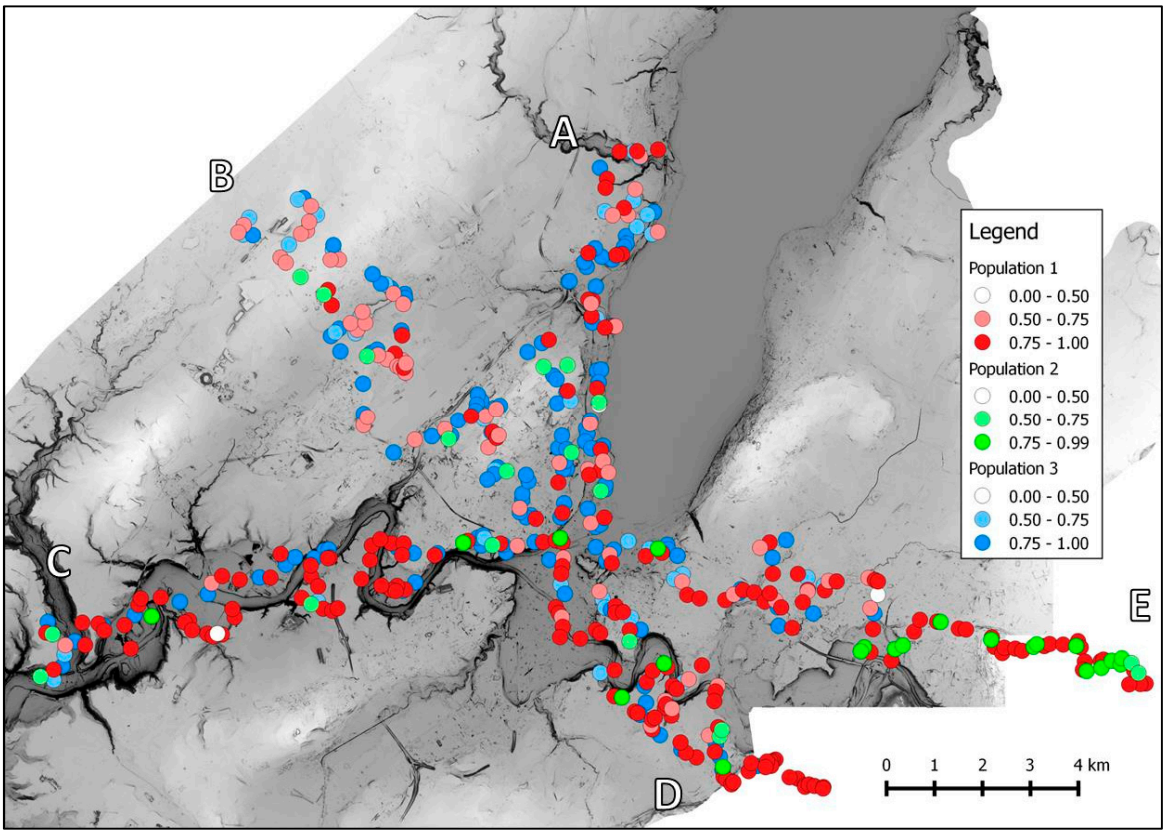


*Figure 5.7 Genetic population structure as calculated by Structure and CLUMPP for three genetic clusters for **P. major**. Individuals were assigned to populations based on their maximum membership coefficient to one of the three populations. Background raster is the Sky View Factor (2m spatial resolution).The two main populations (1 and 3) are not clearly distinct but seem to segregate along a north/south axis. Population 2 is essentially located on transect E.*

5.3 Identification of genetic markers under selection

5.3.1 Samβada

The PLINK file was converted to Samβada's format using RecodePLINK. Each SNP was thus recoded as three genotypes (see section 3.5.1).

A total of 265 872 models were performed in Samβada for 159 variables and 1392 genotypes. The significance threshold is based on a Bonferroni correction (0.01/number of models computed), corresponding to a minimum G and Wald score of 34.44 and 30.25 respectively.

With this threshold, 300 models are considered significant. Among them, a large part involve membership coefficient (population structure) with 45 models, latitude (50 models) and longitude (2 models). Other significant models involve genotypes associated with precipitation in December (prec12) at different window sizes (194 models) and seasonality of climate (gamst) with 9 models, also at different window sizes (both variables are described in Table 3.4). It must be noted that window size does not influence the associations, consequently most genotypes associated with these climatic variables are significant at every scale. On the other hand, we should mention that DEM-derived variables are never involved in significant models.

When considering genotypes associated with climatic variable (regardless of the resolution), we obtain 34 significant models, involving 34 genotypes (Table 5.2). Two of them are detected by both prec12 and gamts, which are both kept in the procedure of selection of variables (correlation of -0.82). However, most genotypes associated with environmental variables are also associated with latitude or/and membership coefficient. In these cases, G and Wald scores are generally higher with the latter ones compared to prec12. However, it was not systematically the case and 11 genotypes (8 SNPs) are detected by precipitation only.

Multivariate models

Multivariate models were computed on 49 variables including latitude, longitude and membership coefficient (population structure). We had to reduce the environmental dataset in order to avoid multicollinearity and variance inflation, and thus used the procedure of selection described in section 3.3, with a threshold of 0.8. Eleven bivariate models are considered significant, mostly involving average temperature in December (tave12) and altitude at 2m resolution. However, genotypes detected in bivariate models were already detected in univariate models, either with precipitation or latitude (see section 3.5.1 for details on the bivariate method). It is thus surprising to note that these genotypes are associated with two "new" variables in bivariate models and not with their significant parent plus another variable. Only one genotype has a significant parent and is this time associated with latitude and membership coefficient.

5.3.2 BayeScan

For BayeScan analyses, we used the three populations identified with Structure (Figure 5.7). Because the genetic population structure is weak, we decided to keep only individuals with a membership coefficient of 0.75 or more, resulting in the deletion of 125 individuals (see Figure 5.7). The PED file was converted to a BayeScan file format using PGDspider (Lischer & Excoffier 2012). BayeScan parameters used are described in section 3.5.2.

BayeScan detected six SNPs with decisive evidence for selection (FDR threshold = 0.1; Figure 5.8). Two of the loci most likely to be under selection were also detected by Samβada with variables prec12, but also with latitude and membership coefficient. Another one, MC01915554:13, is only associated with membership coefficient in Samβada, while the remaining three, the least significant of them, are not detected by Samβada. We also observe that most insignificant loci had a similar Fst.

*Table 5.2 Significant Sambada models for **P. major**. The table shows one model per line with the following columns: the genotype, the associated variable, the resolution of the best Wald score, the best G and Wald scores among all resolutions, the AIC of the model at the best resolution, the frequency of the genotype, the missingness of the SNP, the Q value and Fst of the SNP in BayeScan results, the Moran's I and its p-value for the marker with a neighbourhood of 20 individuals, the Moran's I of the variables with the same neighbourhood. Finally, an "X" is present if the genotype was also significant with either geographic or population structure variables. Models are ranked according to their Wald score*

Genotype	Variable	Best moving window size	Highest G score	Highest Wald score	AIC	Frequency	Site missingness	BayeScan Qvalue	BayeScan Fst	Moran for marker	Moran p-value for marker	Moran for variable	Detection by latitude, longitude or population structure
MC03993697:21_CC	G_prec12	1	93.6	67.3	470.5	0.51	0.11	5.93E-05	0.09	0.37	0.001	0.88	X
MC03993697:21_CG	G_prec12	1	80.1	59.3	475.1	0.35	0.11	5.93E-05	0.09	0.31	0.001	0.88	X
MC06309558:73_CT	G_prec12	16	70.6	50.9	467.1	0.27	0.02	0.80173	0.02	0.24	0.001	0.88	X
MC06309558:73_CC	G_prec12	16	70.6	50.9	467.1	0.71	0.02	0.80173	0.02	0.24	0.001	0.88	X
MC06929001:75_GG	G_prec12	2	50.6	42.6	548.5	0.42	0.07	0.77161	0.02	0.20	0.001	0.88	X
MC06929001:75_CG	G_prec12	2	50.6	42.6	548.5	0.52	0.07	0.77161	0.02	0.20	0.001	0.88	X
MC00827814:20_AC	G_prec12	original	47.7	42.1	428.6	0.25	0.16	0.75435	0.02	0.20	0.001	0.88	
MC06681055:73_CC	G_prec12	original	49.4	42.1	481.2	0.34	0.16	9.62E-07	0.11	0.44	0.001	0.88	X
MC00247904:97_AC	G_prec12	4	45.1	40.3	537.3	0.34	0.05	NaN	NaN	0.15	0.001	0.88	X
MC06698177:33_AG	G_prec12	original	44.7	40.3	543.8	0.33	0.02	0.79602	0.02	0.18	0.001	0.88	X
MC01643098:103_TT	G_prec12	original	44.1	39.1	483.2	0.56	0.13	0.73718	0.02	0.15	0.001	0.88	
MC06085250:39_AG	G_prec12	16	42.1	38.6	514.5	0.29	0.01	0.68304	0.02	0.13	0.001	0.88	
MC01834138:78_CT	G_prec12	16	41.4	38.5	432.1	0.20	0.00	0.6451	0.02	0.13	0.001	0.88	X
MC01651581:47_AT	G_prec12	original	43.4	38.4	478.9	0.30	0.14	0.59404	0.02	0.17	0.001	0.88	X
MC01585839:73_CT	G_prec12	2	48.3	37.6	466.9	0.25	0.02	0.74508	0.02	0.22	0.001	0.88	X
MC01585839:73_CC	G_prec12	2	48.3	37.6	466.9	0.73	0.02	0.74508	0.02	0.22	0.001	0.88	X
MC06784148:25_GT	G_prec12	2	39.5	36.5	454.3	0.23	0.05	0.74828	0.02	0.16	0.001	0.88	
MC01643098:103_CT	G_prec12	1	39.9	36.4	406.0	0.21	0.13	0.73718	0.02	0.15	0.001	0.88	
MC05667273:31_AG	G_prec12	1	39.6	36.1	393.7	0.20	0.15	0.72188	0.02	0.18	0.001	0.88	
MC03690815:99_AT	G_prec12	original	39.3	35.9	273.8	0.11	0.04	0.76275	0.02	0.16	0.001	0.88	X
MC03690815:99_AA	G_prec12	original	39.3	35.9	273.8	0.86	0.04	0.76275	0.02	0.16	0.001	0.88	X
MC01301563:54_GT	G_prec12	original	39.8	35.8	476.8	0.29	0.13	0.68597	0.02	0.11	0.001	0.88	X
MC01883648:43_GG	G_prec12	2	43.2	35.5	532.4	0.31	0.00	0.79423	0.02	0.13	0.001	0.88	X
MC05778243:40_AA	G_prec12	original	41.2	34.8	563.1	0.35	0.00	NaN	NaN	0.21	0.001	0.88	X
MC01276423:101_CT	G_prec12	16	36.9	34.8	424.2	0.20	0.01	0.70939	0.02	0.13	0.001	0.88	X
MC03993697:21_CC	G_gamst	16	132.3	34.0	431.8	0.51	0.11	5.93E-05	0.09	0.37	0.001	0.97	X
MC00185267:124_AG	G_prec12	original	36.2	33.6	538.1	0.31	0.01	0.6529	0.02	0.06	0.001	0.88	
MC06784148:25_TT	G_prec12	original	36.3	33.2	560.4	0.58	0.05	0.74828	0.02	0.11	0.001	0.88	
MC03895169:134_CT	G_prec12	original	35.8	33.2	552.7	0.33	0.01	0.79632	0.02	0.17	0.001	0.88	X
MC05667273:31_AA	G_prec12	1	36.4	33.1	482.1	0.55	0.15	0.72188	0.02	0.18	0.001	0.88	
MC02223852:142_CG	G_prec12	16	35.7	32.9	523.1	0.32	0.07	0.75911	0.02	0.10	0.001	0.88	X
MC02192085:71_AA	G_prec12	original	40.6	31.1	402.4	0.19	0.06	0.80005	0.02	0.14	0.001	0.88	X
MC06681055:73_AC	G_prec12	16	35.3	30.6	507.9	0.39	0.16	9.62E-07	0.11	0.22	0.001	0.88	X
MC03993697:21_CG	G_gamst	16	121.8	30.6	433.4	0.35	0.11	5.93E-05	0.09	0.31	0.001	0.97	X

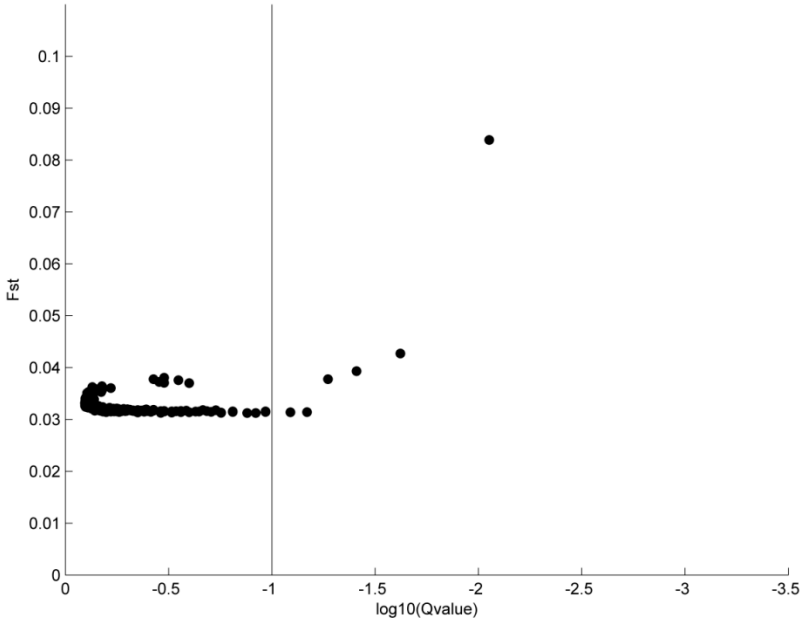


*Figure 5.8 BayeScan results for **P. major**. The FDR threshold was set to 0.1 and corresponds to a log(PO) of 1. Six SNPs show a decisive evidence of selection. Most SNPs have a low and similar Fst.*

5.3.3 Comparison between methods

We mention that we tried to use LFMM, but did not manage to make it work, even though we tried with several environmental variables, different levels of missingness per site and per individual or deletion of individuals without variable values.

The spatial autocorrelation of detected genotypes in Samβada is higher than that for the neutral markers (Figure 5.9). We note that the two SNP markers that are detected by both methods (BayeScan and associated with prec12 in Sambada) show the highest values of Moran's I (two genotypes for each of these SNPs are significant in Samβada).

Figure 5.10 shows BayeScan results versus the Samβada Wald score. Two out of six BayeScan detections are also detected by Samβada. However, the most significant Wald scores are amongst BayeScan neutral loci.

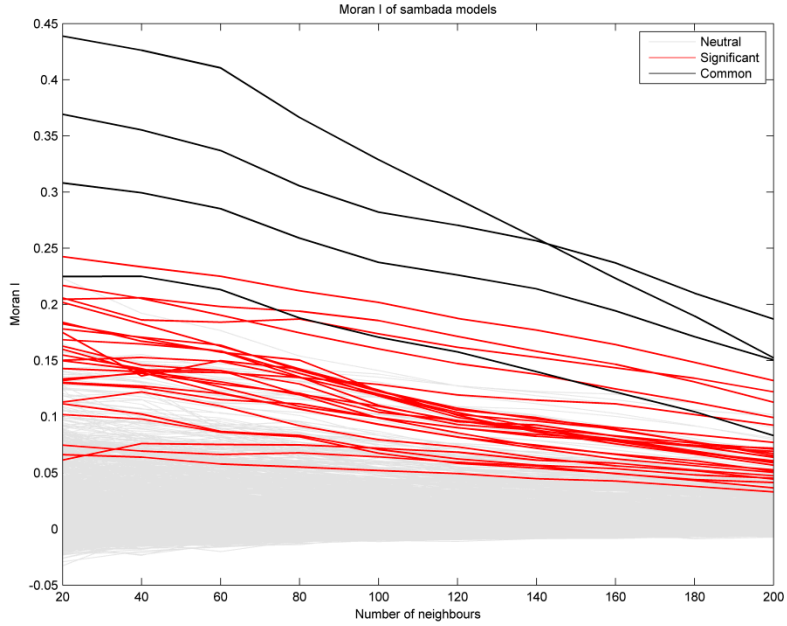


*Figure 5.9 Moran's I Correlogram for **P. major** genotypes. Neutral genotypes are in grey, genotypes detected by Samβada's and involving environmental variables are shown in red. Commonly detected genotypes are in black. The figure shows that significant loci are more spatially autocorrelated than neutral loci. In addition, commonly detected loci show a higher SA than all other loci.*

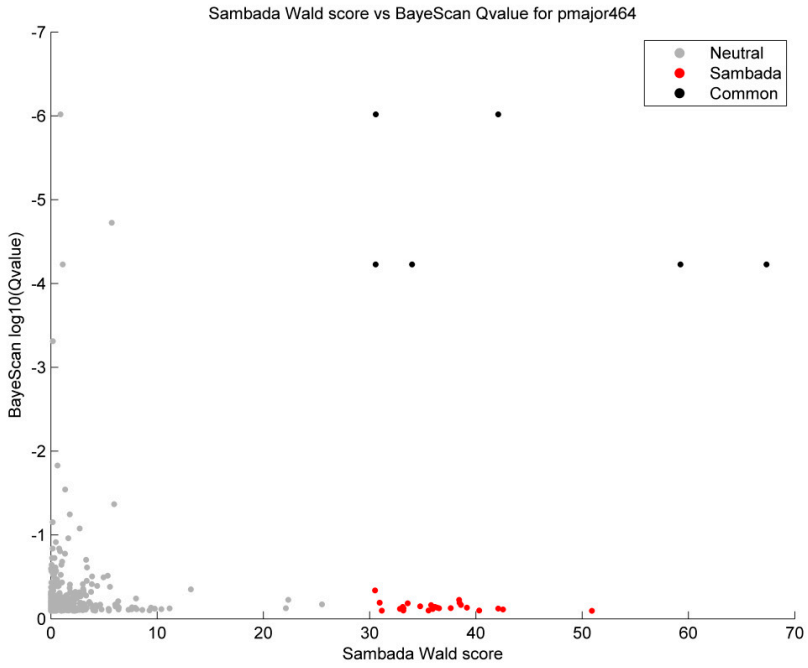


*Figure 5.10 Scatter plot of BayeScan Q-value against Samβada G score for **P. major**. Genotypes associated with environmental variables in Samβada are in red and commonly detected genotypes are in black. The figure shows that there is no correlation between the scores of these two independent methods. However, commonly detected loci are amongst the most significant in BayeScan.*

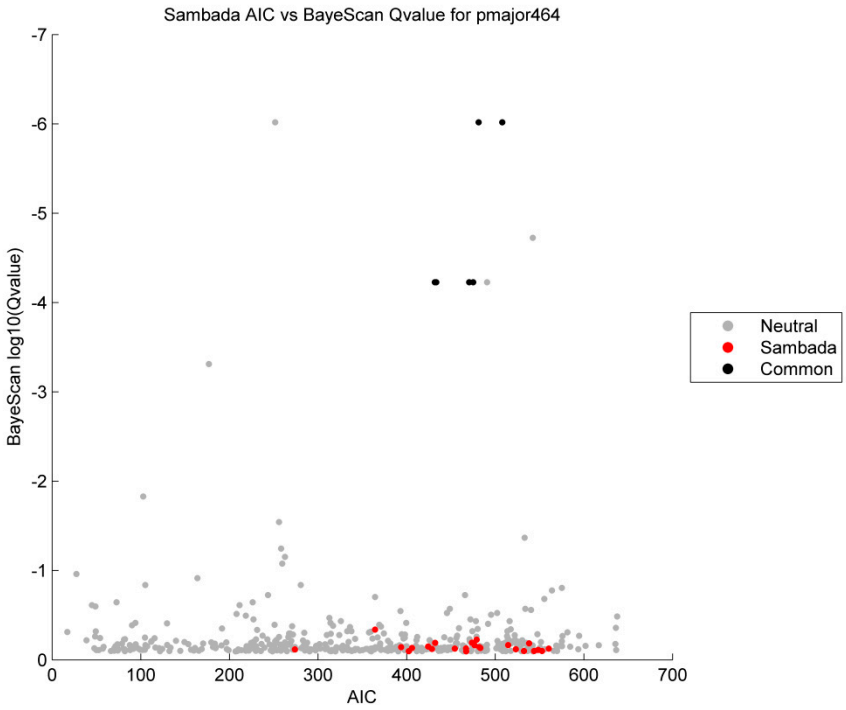


*Figure 5.11 Scatter plot of BayeScan Q-value against Samβada AIC for **P. major**. Genotypes associated with environmental variables in Samβada are in red and commonly detected genotypes are in black. The figure shows that there is no correlation between the scores of these two independent methods. However, commonly detected loci are amongst the most significant in BayeScan.*

5.3.4 Visualisation of significant associations

The main purpose of comparing these graphs of significant associations is to find common trends or differences between them. Details on the different parts of these graphs can be found in section 3.6 (p65).

Figure 5.12 to Figure 5.14 illustrate three examples of genotypes only associated with the amount of precipitation in December (prec12). They show an uneven spatial distribution along an axis that goes through transects B and C (NW-SE direction), with an evident clustering. Indeed, for each of them, LISA coefficients are significant in this transect and not elsewhere. The map of the variable shows a precipitation gradient with decreasing winter precipitation in SE direction. Winter precipitation is highest in transect B and highlights important differences in environmental conditions in the study region.

Figure 5.15 and 5.16 illustrate two SNPs commonly detected by Samβada and BayeScan. Both are amongst the most significant models for both methods and are also detected in Samβada by latitude and membership coefficient. They both show a clear-cut distribution between north and south. These SNPs also show the two highest values in Moran's I correlograms (see also Figure 5.9).

In general, all detected variables show a strong global spatial autocorrelation and both variables and markers show their strongest Moran's I in the first spatial lag of 20 neighbours (Figure 5.12 to 5.16 and appendix II.b). In addition, LISA coefficients were often significant and well identify the presence of clusters. They show a general trend of being significantly positive in the two northern transects, A and B. Most of the other significant models display a similar trend like as in figure 5.15 and 5.16 shown below (see Appendix II.b).

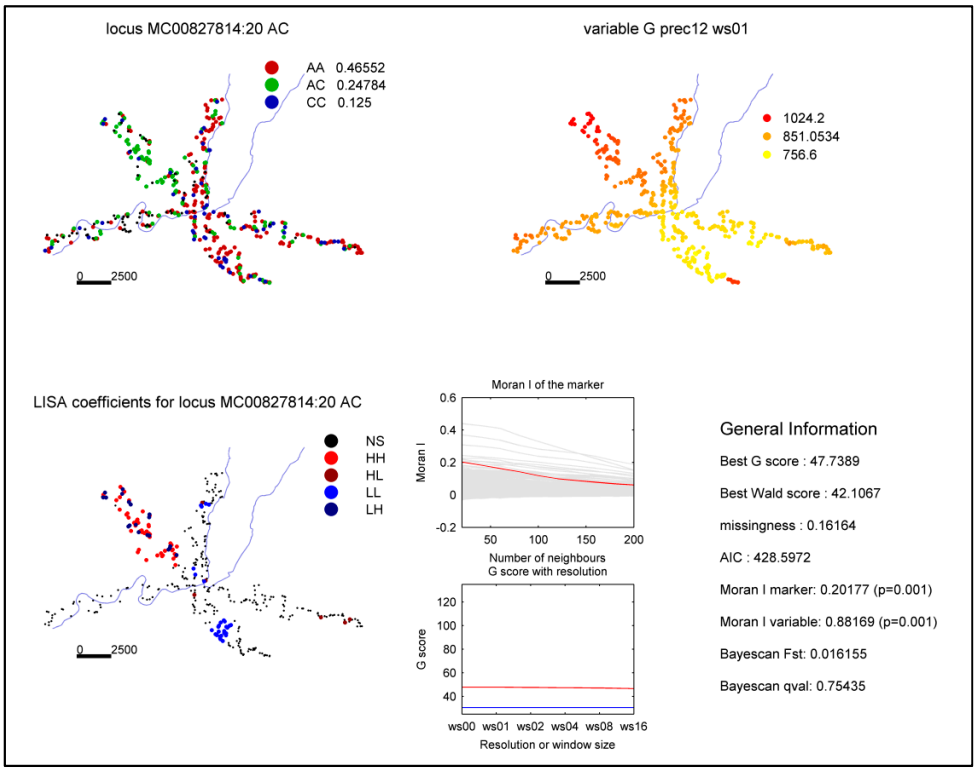


Figure 5.12 Visualisation of the main results for the model involving genotype MC00827814:20_AC and precipitation in December. This genotype was only detected by Samβada and only associated with prec12.

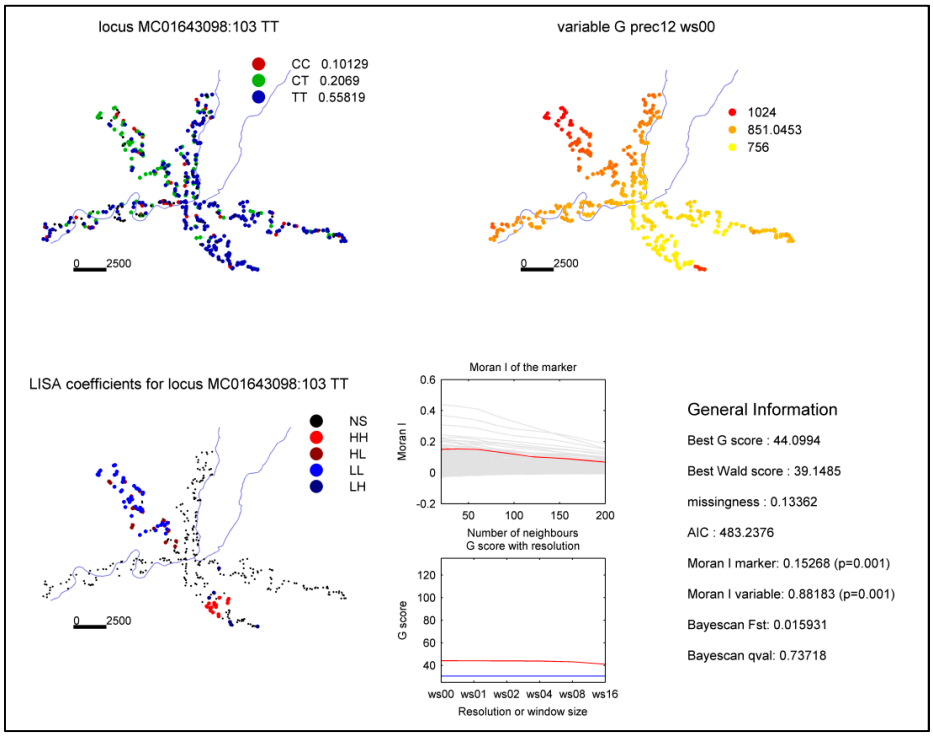


Figure 5.13 Visualisation of the main results for the model involving genotype MC01643098:103_TT and precipitation in December. This genotype was only detected by Samβada and only associated with prec12.

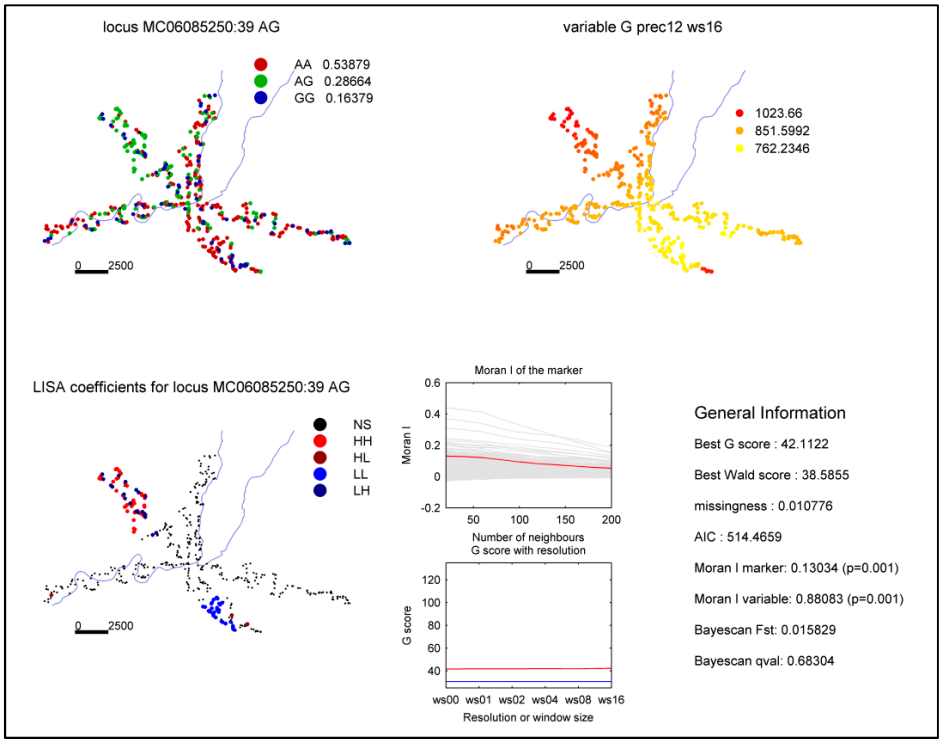


Figure 5.14 Visualisation of the main results for the model involving genotype _AG and precipitation in December. This genotype was only detected by Samβada and only associated with prec12.

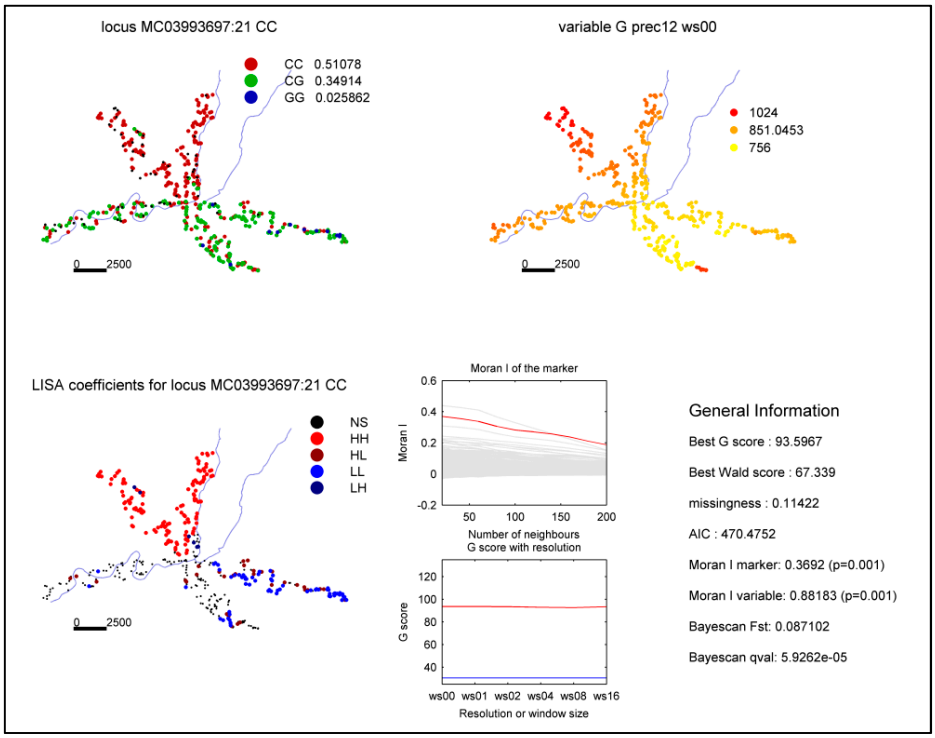


Figure 5.15 Visualisation of the main results for the model involving genotype MC03993697:21_CC and precipitation in December. This genotype is the most significant with prec12 in Samβada and was also detected with latitude, membership coefficient and gamts. It is also one of the significant SNPs in BayeScan.

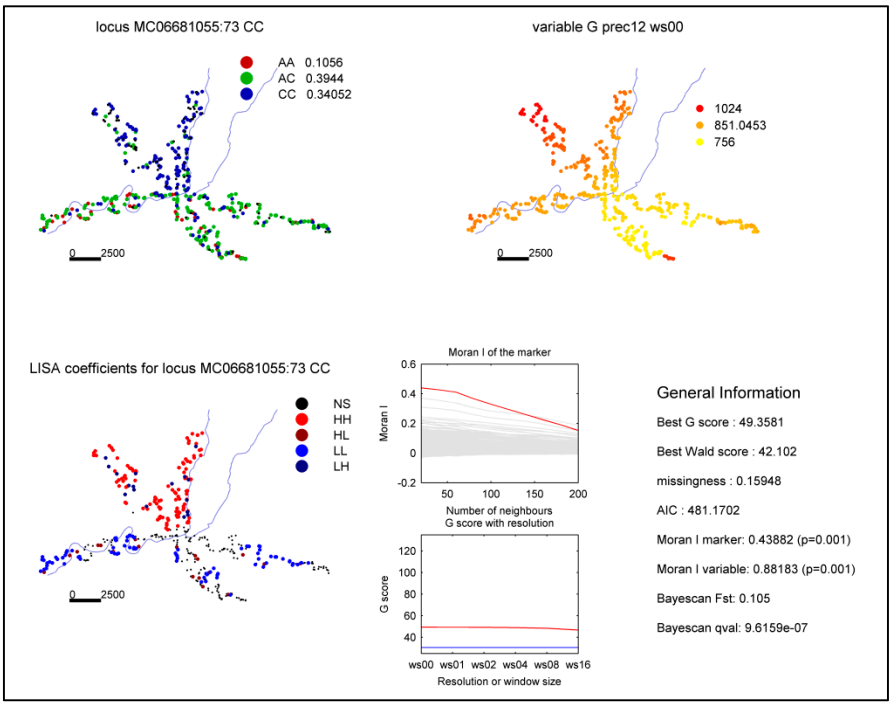


Figure 5.16 Visualisation of the main results for the model involving genotypeMC06681055:73 _CC and precipitation in December. This genotype is the most significant with prec12 in Samβada and was also detected with latitude, membership coefficient and gamts. It is also one of the significant SNPs in BayeScan.

5.4 Discussion

Investigations regarding *P. major* around Geneva are part of a larger project aiming to assess the impact of urbanization on population structure and adaptation of several species. Particularly, within this study we focused on the impact of topography and climate at a regional scale on a plant submitted to a potentially high anthropogenic pressure.

We encountered several difficulties to handle *P. major* genetic dataset. The genetic dataset for P. major consisted of 464 SNP markers, which is low compared with recent papers that assessed population structure with SNPs (Namroud *et al.* 2008; Lamaze *et al.* 2012; Hübner *et al.* 2012; Moore *et al.* 2014) but should be enough to assess it, particularly since we have a large sample size (Morin *et al.* 2009). However, genetic differentiation between the two main genetic clusters was low. Figure 5.6 shows that many individuals in transect B but also in other transects show negative inbreeding. We also recalculated the Fis without transect B, still encountering negative Fis. Nevertheless, the SNP markers are statistically reliable since only SNPs that occur in more than 80% of individuals with a minor allele frequency of 5% are included in this dataset and we found several trends and spatial patterns that are worth discussing further.

The population structure is moderate and shows a distinct distribution of population 1 and 2 on one side of Geneva and of the Rhone River, and of population 3 on the other side. However, Fst between the two main populations is low and does not evidence a clear separation. Nevertheless, this boundary is supported by the sharp spatial structure of 39 SNPs correlated with latitude and of 2 SNPs with longitude. The spatial distribution of these SNPs shows a clear-cut distribution along the barrier constituted by the river, the city and the lake (see Figure 5.15, 5.16 and Appendix II.b). This could mean that both major populations are quite isolated and experience restricted migration to and from the global distribution of *P. major*. In fact, small isolated populations are likely to experiment a higher level of inbreeding and a reduction of genetic diversity through genetic drift.

Precipitation in December (prec12) and continentality (gamst) are the only two environmental variables involved in potentially adaptive genetic variation among the study region, indicating to be selection factors for *P. major*. However, prec12 and latitude largely detected the same genotypes, despite these variables are only moderately correlated between each other (0.79). It is thus impossible to disentangle the influence of these two variables in our dataset. In addition, multivariate models do not better explain the spatial distribution of these SNPs, showing that there is no combined effect of these two variables. In addition, their score in Samβada are higher with latitude than with precipitation, suggesting that latitudinal structure is more likely to explain the spatial distribution of these SNPs. However, since most detected loci respond to latitude, membership coefficient and prec12, we can hypothesize that the population structure is influenced by an adaptation to precipitation. A few additional SNPs were correlated to precipitation only and are illustrated in figure 5.12 to 5.14. These models have a different spatial pattern than most of the other detections. In fact, they show a strong spatial autocorrelation in transect B, highlighting a remarkable clustering that could be due to precipitation.

If we suppose that most significant models are true positives and that Samβada has a low risk of missing adaptive loci, precipitation is the only relevant variables to explain local adaptation of *P.*

major in the Geneva cross-border area. In fact, Eco-climatic variables are largely redundant at this regional scale and show higher pairwise correlations than the values given for the entire country (Guisan & Zimmermann 2000). In addition, using different window sizes does not influence the correlations of these interpolated variables with genetic markers, despite the fact that they take topography into account. On the other hand, DEM-derived variables were irrelevant in this study. The effect of topographic variability might not be sufficient in the flat area of Geneva to influence plant habitat, or might simply be negligible compared with anthropogenic pressure, also considered in this project. Indeed, we will investigate the role of urban-related variables on the connectivity between populations and on the adaptation of the species. For example, these variables will include the type of landcover and the percentage of green spaces as these variables can impact the connectivity between populations. We will also consider the height of the buildings, which could be derived from the surface model, as it strongly affects wind that is essential for the spread of seeds.

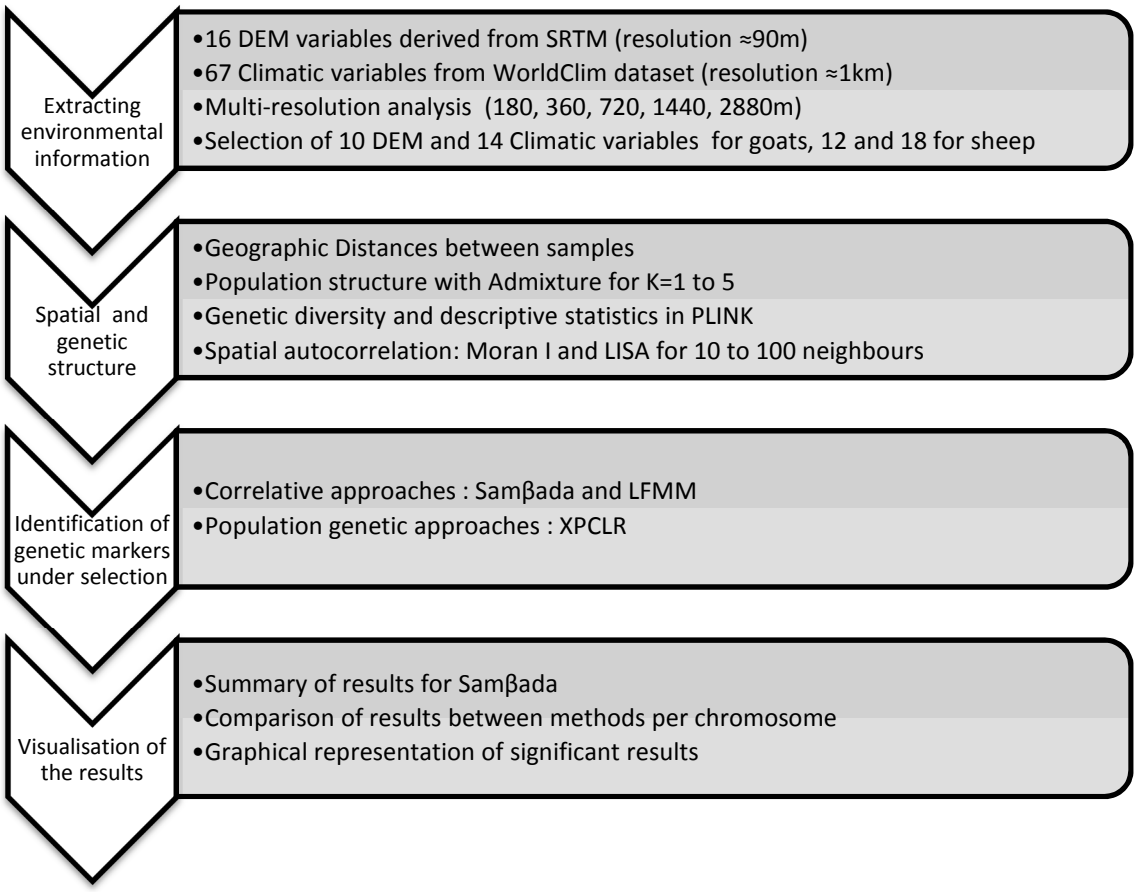
On the other hand, BayeScan identified 6 SNPs under selection. Three of them are also detected by Samβada with prec12, but also by latitude. If we consider that BayeScan is a powerful method with little risk of detecting false positives (De Mita *et al.* 2013), these common detections comfort us in the hypothesis that precipitation, is a relevant selection factor in this case, despite its confusing detections with latitude. However, the weak population genetic structure most probably makes it more difficult to detect signatures of selection with the BayeScan method. In fact, BayeScan uses the entire genetic dataset to evaluate the neutral background and is detections might thus be biased in this case. In addition, BayeScan's best performances take place when more genetic data are used to estimate properly the neutral variability and when a larger number of populations is considered (De Mita *et al.* 2013). With 464 SNPs, BayeScan may not have enough data to model neutral variability, ending up in identifying loci that may be neutral contributors to the moderate population structure observed around Geneva. We also note that, with BayeScan alone, we would not have detected the peculiar distribution in transect B.

To conclude, Samβada and BayeScan both detected substantial and common signatures of selection due to precipitation. However, their high correlation with latitude, corresponding to a geographic barrier (lake, river, and city) does not allow us to conclude for a strong evidence of selection.

Chapter 6 Sheep & goats in Morocco

Domesticated species are submitted to both human and environmental pressure, leading to locally adapted populations. However, local breeds are nowadays threatened by industrial breeds with better production values but with little evaluation of their capacity to adapt to new habitats. In this case study, we evaluate the local adaptation of sheep (*Ovis Aries L.*) and goats (*Capra Hircus L.)* in Morocco by assessing breeds uniqueness and their adaptive genetic resources. Therefore, identifying both neutral and adaptive variations may highlight the necessity of keeping these resources.

For this purpose, sheep and goats were uniformly sampled across a region showing marked variation in environmental conditions between mountains, plains and deserts. Because the study area is large and that the mobility of these species is considerable, we expected that climatic variables would play a more important role than topography in the detection of signatures of selection. In addition, with two species sampled in a similar way, we may identify similar genomic regions associated with the same environmental variables.



This case study is part of the project NEXTGEN (http://nextGen.epfl.ch/), which is the first project aiming at a comparative analysis of whole genome sequencing (WGS) data and high-density sequencing for sheep, goats and cattle. The purpose of the project is to study the genomic basis of adaptation and resistance to parasites to raise awareness on the preservation of genetic resources. Indeed, future breeding programmes will be designed to exploit whole genome data in livestock populations but should also seek for durable strategies by maintaining genetic diversity. In fact, industrial breeding has become more widespread and exert an increasing pressure on traditional breeds (Taberlet *et al.* 2008). However, direct anthropic selection is relatively modest on these Moroccan populations and until recently, it was difficult to distinguish well-defined breeds. Furthermore, livestock farming could be endangered on the long term by the extinction of the local well-adapted breeds and, therefore, an evaluation of wild ancestors as reservoirs of genetic diversity is crucial. In addition, Benjelloun *et al.* (2015) previously assessed the genetic diversity of three phenotypically distinct indigenous breed, showing high genetic diversity.

The project NEXTGEN is also the first project to use WGS in a landscape genomics context and opens new perspective in the detection of candidate SNPs and genes under selective pressure. In fact, it also aims to provide the necessary tools for the exploitation of next generation sequencing in conservation genetics and farm animal practices.

6.1 Environmental Data

6.1.1 Environmental variables

Climatic variables were recovered from WorldClim dataset (Table 3.3). To estimate these climatic variables at different scales, we extracted their values with different window sizes of 3x3, 5x5, 9x9, 17x17, 33x33 pixels, corresponding to distances from ≈3 to ≈33km. It should be noted that weather stations used to interpolate the climatic variables are not disseminated uniformly across the territory of Morocco, where more stations can be found along the coast than in the mountain range and further east (see Figure 2.4).

DEM-derived variables were computed from the SRTM DEM (see section 3.1.1). Its initial projection system is WGS84 but SAGA GIS requires projected DEMs in a metric system and we decided to use the Merchich / Nord Maroc projection system (EPSG: 26191). We estimated that this system is more adapted than the Sud Maroc projection because the majority of samples are situated in the northern half of the country and because the territory is large enough to be covered by two UTM zones, thus making the selection difficult. The DEM was projected with a resolution of 90m. The list of variables computed can be found in Table 3.2. Multi-scale DEM variables were computed using the Gaussian Pyramid described in section 3.1.3 at resolutions of 180, 360, 640, 1024 and 2048 meters. Variables from structure tensors could not be produced at this original resolution, and were computed at a resolution of 180m. They were thus not included in the Multi-scale analysis, as we could not compute them at each resolution.

6.1.2 Variables selected

Despite sheep & goats sampling design is the same and covers the entire territory of Morocco, sampling locations are different (Figure 6.1 and Figure 6.2). Consequently, environmental variability can be different between both samples sets and thus, we may not obtain the same set of variables when selecting uncorrelated variables (see section 3.3). Therefore, we performed pairwise correlations between all 81 variables (67 climatic variables and 14 DEM-derived variables) for both datasets independently. By doing so, we expect to cover a maximum of the environmental space for both samples sets while deleting redundant variables. The list of variables deleted was kept as they may help interpretation of adaptation (Appendix II.c). Latitude and longitude variables were also included in the analysis to account for spatial structure. However, no membership coefficient was added to the dataset since population structure was weak (see below).

The number of variables selected for sheep is a bit larger than for goats, 31 for sheep compared with 27 in goats out of 81 (see Appendix II.c). We found a high redundancy between WorldClim temperature variables and precipitation. Most of the other bioclimatic variables were also highly correlated with temperature and precipitation variables. On the other hand, most of DEM-derived variables were kept.

6.2 Spatial and genetic structure of the dataset

The sampling strategy we applied was constrained by several criteria. Among the 412 farms for goats and 432 for sheep, we had to select only 164 individuals for each species due to budget limitations. In fact, it is currently impossible to fund whole genome sequencing of ≈3000 individuals and we had to choose sequenced individuals wisely. The most important criterion was to optimize the selection in order to have the widest possible range of environmental conditions among our samples. The second criterion was to take the geographic space into account. Indeed, we wanted to maximise the spread of individuals over the covered area in order to ensure a spatial representativeness of all the regions of Morocco. Finally, we tried to take enough individuals per breed in order to maximise the chances of finding genetic differences among them.

However, traditional random sampling cannot take these criteria into account. Therefore, we opted for a sampling scheme that maximised environmental information and spatial spread by applying a regular grid over the study area in which each cell contains between 1 and 18 individuals distributed among several farms (Figure 6.1 and Figure 6.2).

In order to choose samples as different as possible, Stucki (2014) first performed a principal component analysis (PCA) on the 117 variables extracted from the Climatic Research Unit (CRU) dataset (New *et al.* 2002). The PCA allows maximising the ecological distance between the farms (separately for sheep and goats). Afterwards, she performed an ascending hierarchical classification on the first 7 PCA-axes (96 % of the variance) to regroup farms according to the ecological distances between them. Using the Ward criteria, we reduced the number of classes from 432 and 413 for sheep & goats respectively to the desired 164 classes, corresponding to the 164 individuals retained (Escoffier & Pages 2008).

After regrouping classes, individuals among classes were randomly selected but since we wanted to guarantee spatial representativeness as well, 50 different random samplings were performed. Among these 50 samplings, we chose the one with the maximal index of spatial distribution, which was the sum of distances between each farm and its nearest neighbour. Finally, we found that all breeds were sufficiently well represented, as the number of individuals per breed did not change much from one random sample to another (Figure 6.1 and 6.2).

Finally, a couple of individuals were not sequenced, thus reducing the datasets to 161 individuals each.

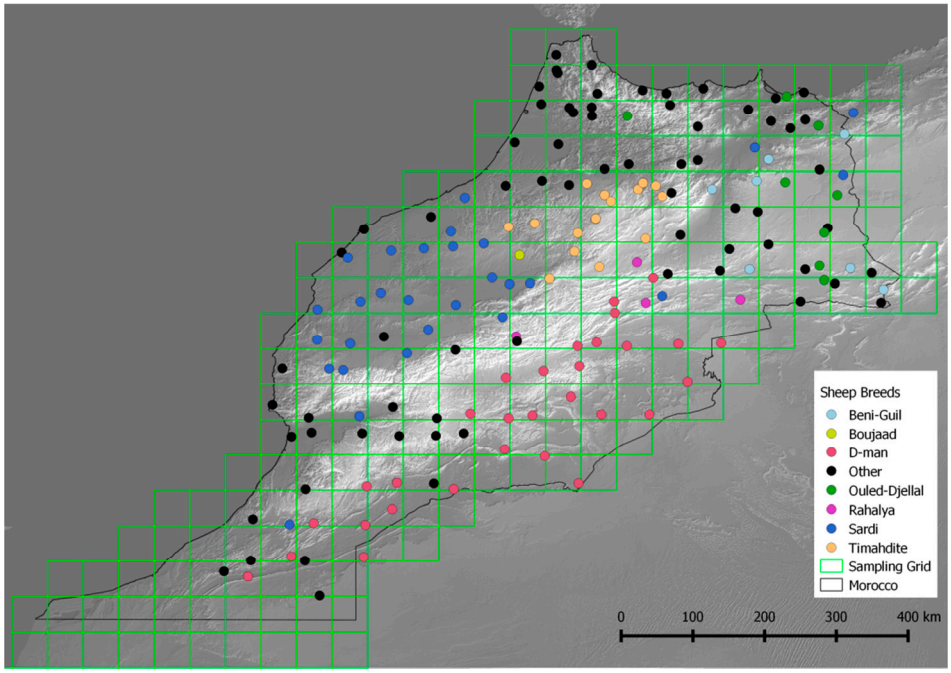


*Figure 6.1 Spatial distribution of sampled **sheep** in Morocco. Illustration of the sampling strategy using a regular grid distribution. The purpose is to guarantee a representative spatial and environmental distribution as well as breed diversity. In total, 161 individuals from 6 breeds and non-breed individuals were sequenced.*

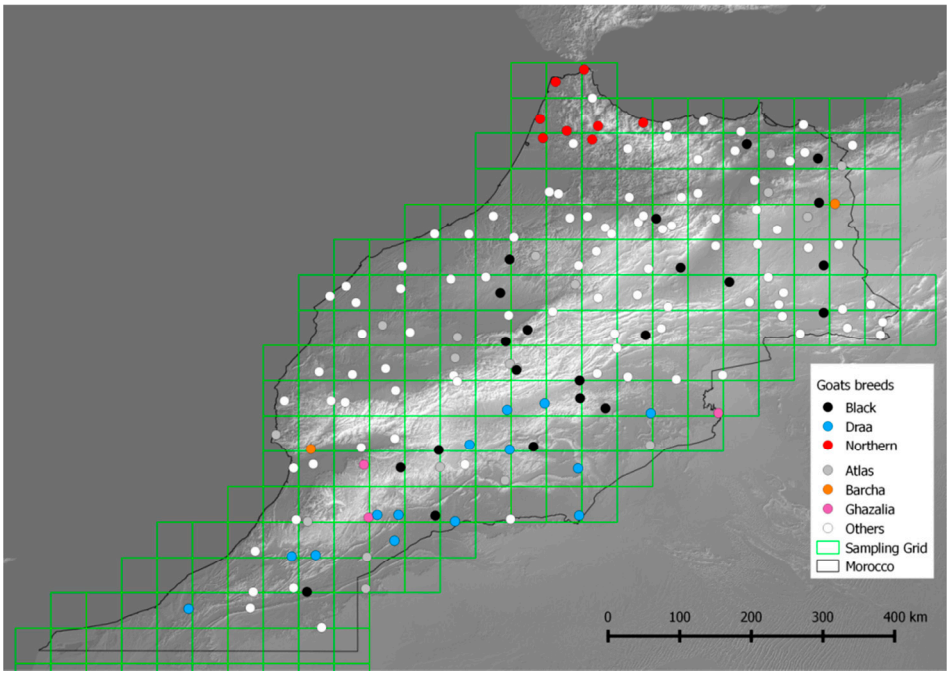


*Figure 6.2 Spatial distribution of sampled **goats** in Morocco. Illustration of the sampling strategy using a regular grid distribution. The purpose is to guarantee a representative spatial and environmental distribution as well as breed diversity. In total, 161 individuals from 5 breeds and non-breed individuals were sequenced.*

6.2.1 Spatial structure

Both datasets benefited from a random stratified sampling over the territory of Morocco. We used the average nearest neighbour measurement to evaluate spatial clustering (3.4.1, p56) and obtained for sheep a value of 0.95 (P-value = 0.2, non-significant) and for goats 0.89 (P-value = 0.001, significant clustering). Both are closer to 1 (random distribution) than for the previous case studies, showing that the method of selection of samples was appropriate to minimize spatial clustering and optimize environmental variability. Therefore, the probability of encountering samples with the same habitat is low.

Histograms of pairwise distance show that maximal distances between two points do not exceed 1200km (Figure 6.3 top). Shortest distances histograms are close to a normal distribution with a mean around 30km and a maximum lower than 100km.

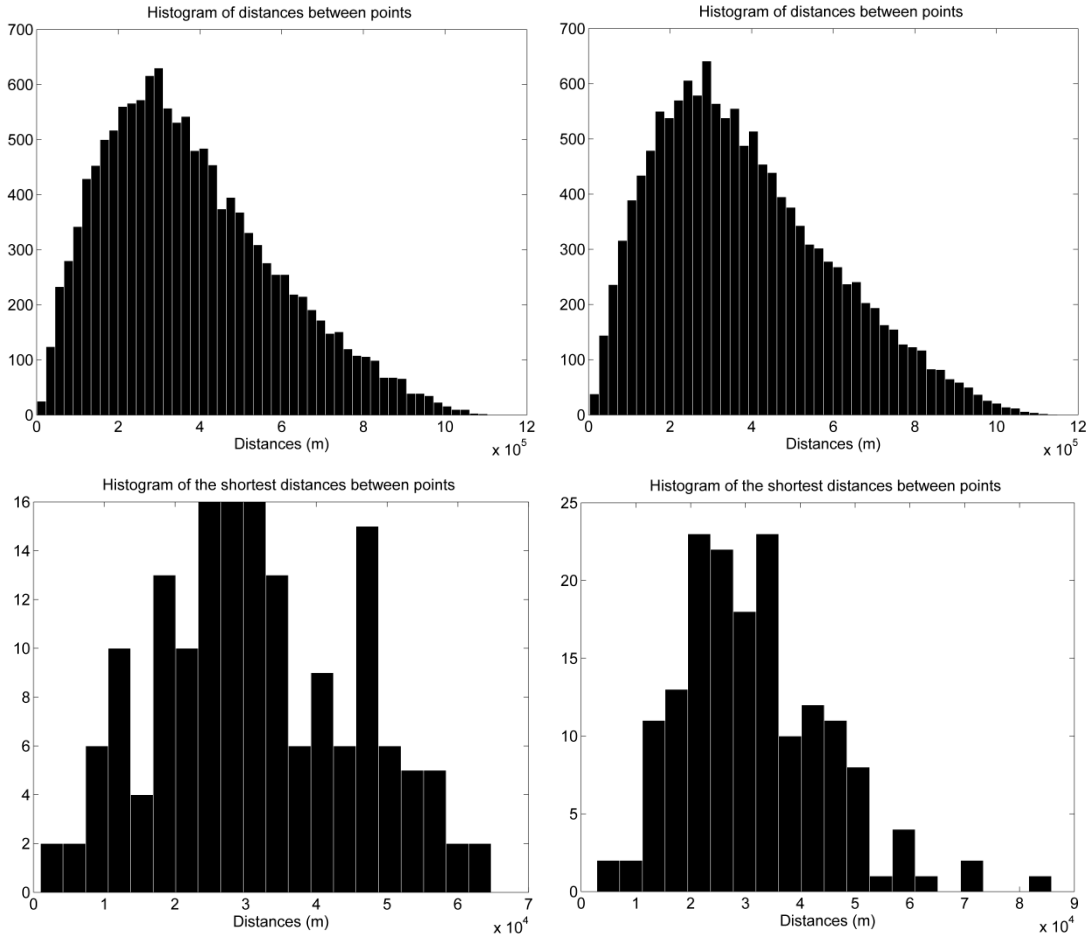


*Figure 6.3 Histograms of pairwise distances between samples (top) and shortest distances between samples (bottom) for **sheep** (left) and **goats** (right). Both histograms of distances between samples are close to a Poisson distribution, which is expected when individuals are randomly sampled. These histograms also show that the spatial distribution of samples is similar for both species.*

6.2.2 Genetic data and population structure

Sheep and goats genetic datasets were obtained with whole genome sequencing at 12X coverage. A detailed description of their production and filtering for goats can be found in Benjelloun *et al.* (2015). Briefly, DNA extraction was done using Puregene Tissue Kit from Qiagen. Samples were then sent to the Génoscope (Centre national de séquençage, Paris) where whole genome sequences were obtained using Illumina Hiseq2000. Afterwards, paired-end reads were mapped to the goat reference genome (CHIR v1.0, GenBank assembly GCA_00317765.1) and sheep oar3.1 reference genome, respectively. Variant calling was done using three different algorithms: Samtools mpileup (Li *et al.* 2009), GATCK UnifiedGenotyper (McKenna *et al.* 2010) and Freebayes (Garrison & Marth 2012). Original SNP datasets contained ≈39 million and ≈32 million variants for sheep and goat respectively. These datasets were first filtered to remove sites with more than two allelic variants and indels. Afterwards, PLINK was used to filter these datasets according to three criteria: missingness-per-individual of maximum 5%, missingness-per-site of maximum 10%

and a minor allele frequency (MAF) of minimum 5%. Finally, SNPs in linkage disequilibrium were suppressed. Linkage disequilibrium occurs when alleles are non-randomly associated, which is the case with combinations of alleles at very close positions. SNPs in linkage disequilibrium are thus redundant in correlative analysis. This filtering was performed in PLINK with windows of 50 SNPs and steps of 10 SNPs with a r^2 of 0.2. The final datasets contained 1.7 and 1.8 million SNPs for sheep and goats respectively. We note that all genotypes were renamed AA, AG and GG to facilitate recoding in RecodePLINK (see section 3.5.1).

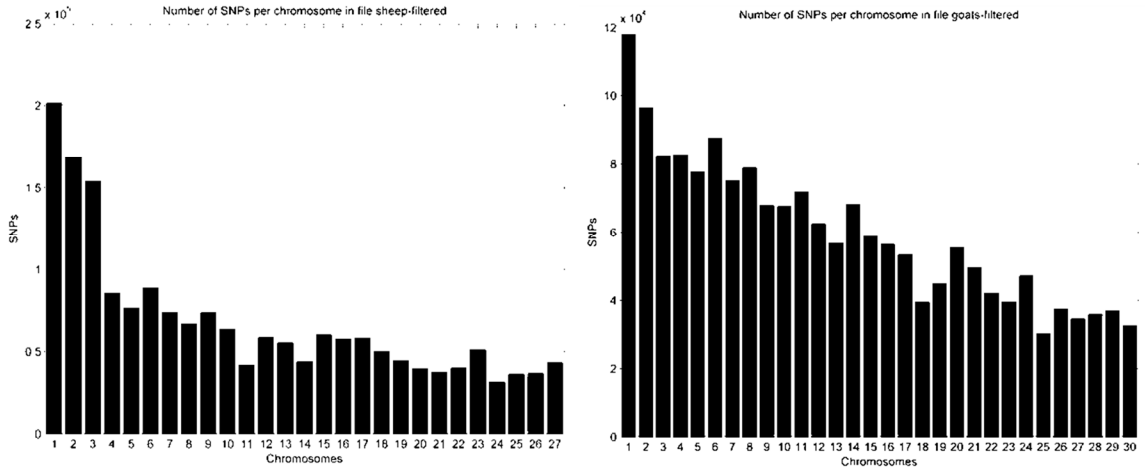


*Figure 6.4 Number of SNPs per chromosome after LD filtering for **sheep** (left) and **goats** (right). The number of SNPs per chromosome is decreasing with the reference number of the chromosome as they are arranged by size.*

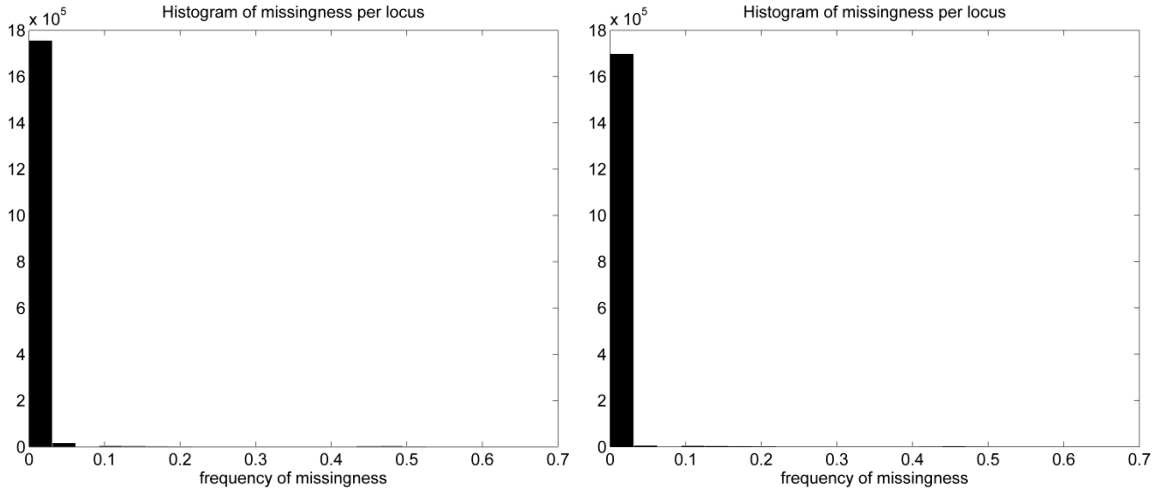


*Figure 6.5 Histograms of missingness-per-site for **sheep** (left) and **goats** (right). Missing data are rare thanks to whole genome sequencing data and the sufficient coverage of each individual.*

In Figure 6.4, the number of SNPs per chromosome is decreasing with the reference number of the chromosome; the latter being usually defined by the decreasing size of the chromosomes. In sheep for example, the first three chromosomes are much larger than the others.

Figure 6.5 shows that missingness-per-site was low and, while there are a few outliers SNPs for both species, it is evident that the data are of high quality and missingness rarely exceeds 5% thanks to WGS. In addition, missingness-per-individual does not exceed 2%.

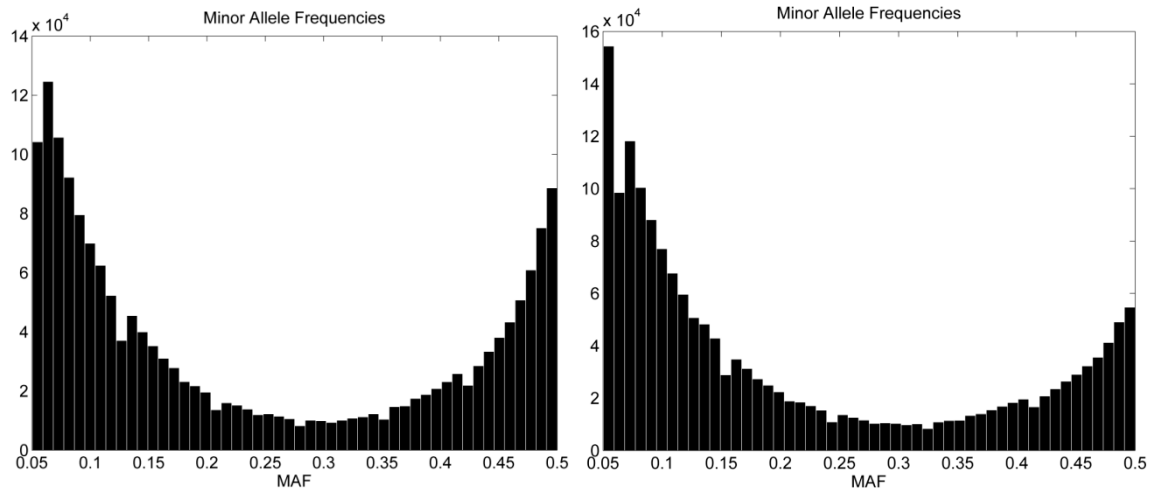


*Figure 6.6 Histograms of minor allele frequencies (MAF) for the 1.7 and 1.8 million SNPs in filtered datasets of **sheep** (left) and **goats** (right). Both species show a higher frequency of minor alleles at both end of the spectrum. In the case of sheep, the amount of high frequencies is higher than for goats.*

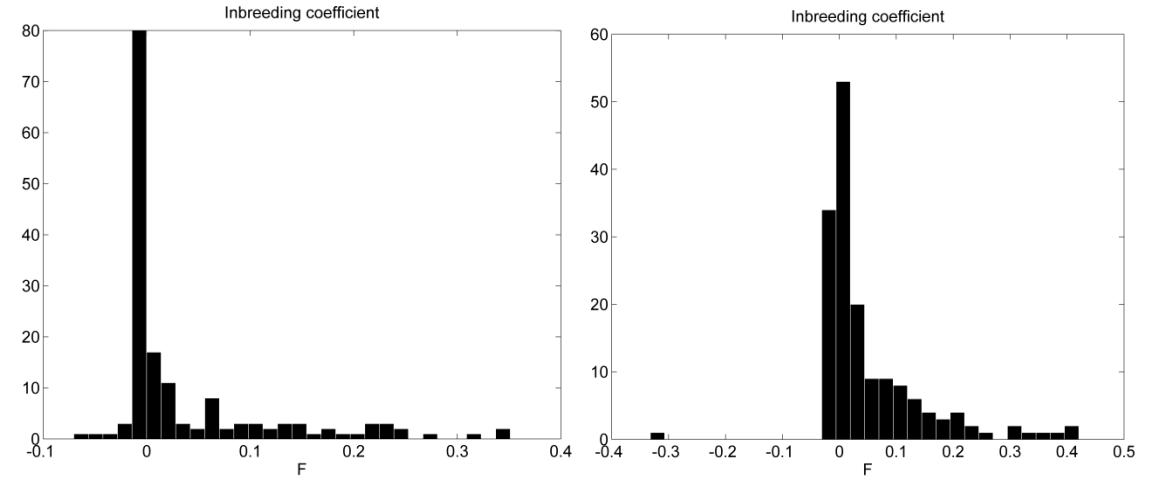


*Figure 6.7 Histograms of inbreeding coefficients per individual for **sheep** (left) and **goats** (right). Few individuals showing an excess of heterozygotes are found in both species. Most individuals have an inbreeding coefficient close to zero but the distribution is larger in the case of goats.*

Histograms of minor allele frequency (Figure 6.6) are similar between species. We can note that sheep have a higher occurence of minor alleles with a frequency close to 0.5.

Histograms of inbreeding coefficients show low values for both datasets (Figure 6.7). A visual inspection of these coefficients on a map (not shown) did not show any spatial patterns for either the lower or higher Fis outliers.

Population structure was assessed with admixture for values of K between 1 and 5. Both species showed a weak population structure with increasing cross-validation error, indicating K=1 as the most likely number of populations. In the case of K=2, we found a Fst of 0.054 for sheep and 0.048 for goats, indicating, again little differences among individuals and we concluded that distinct populations could not be defined for both datasets. Nevertheless, we decided to illustrate the membership coefficients in the case of K=2 on maps. Figure 6.8 shows that most of sheep have an admixed membership to both populations. However, two small clusters can be distinguished. The first is situated between the sea and the high-Atlas mountains, and the second one is located southeast of the mountains, close to the desert. For goats (Figure 6.9), most of individuals belong to the first cluster, ranging over the entire territory, and only a few individuals are attributed to the second cluster along the coast.

*Table 6.1 Cross-validation error from Admixture results for **sheep and goats** from K=1 to K=5. Convergence is assessed by studying the Log-likelihood and cross validation error. These results show that K=1 is the most likely number of clusters in both species.*

	Sheep			Goats		
K	Iterations	Log likelihood	CV error	Iterations	Log likelihood	CV error
1	1	-272728869.4	0.533	1	-254103443.4	0.520
2	76	-271396236.5	0.548	69	-252618371.7	0.531
3	166	-270116337.9	0.565	89	-251305468	0.552
4	563	-268868329.8	0.604	74	-250168974.9	0.576
5	129	-267744118	0.614	108	-248936278.1	0.600

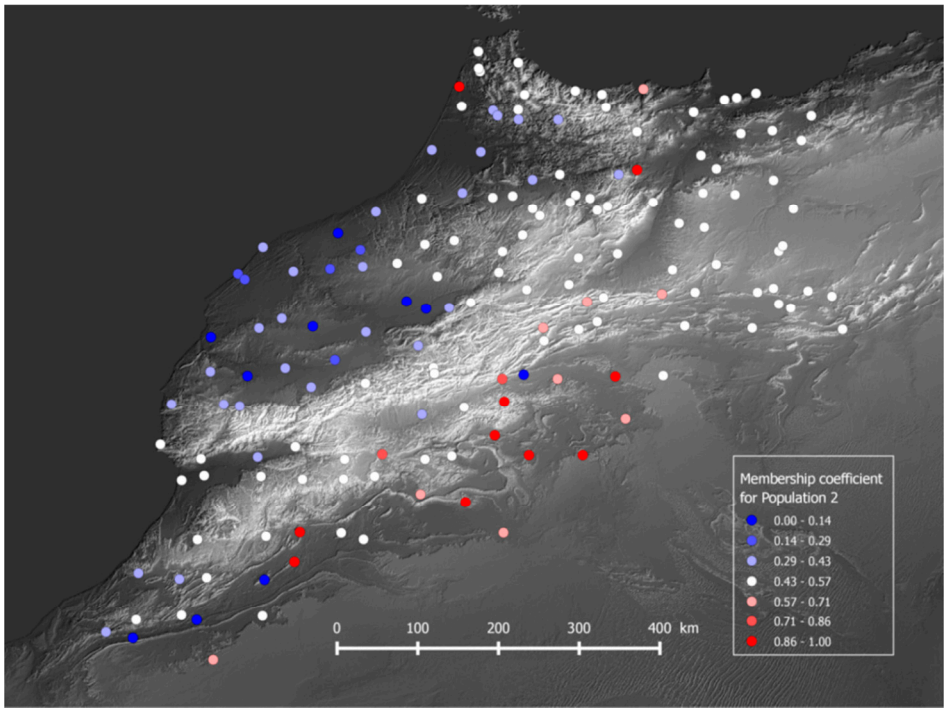


*Figure 6.8 Map of membership coefficient to population 2 for **sheep** in the case of K=2.*

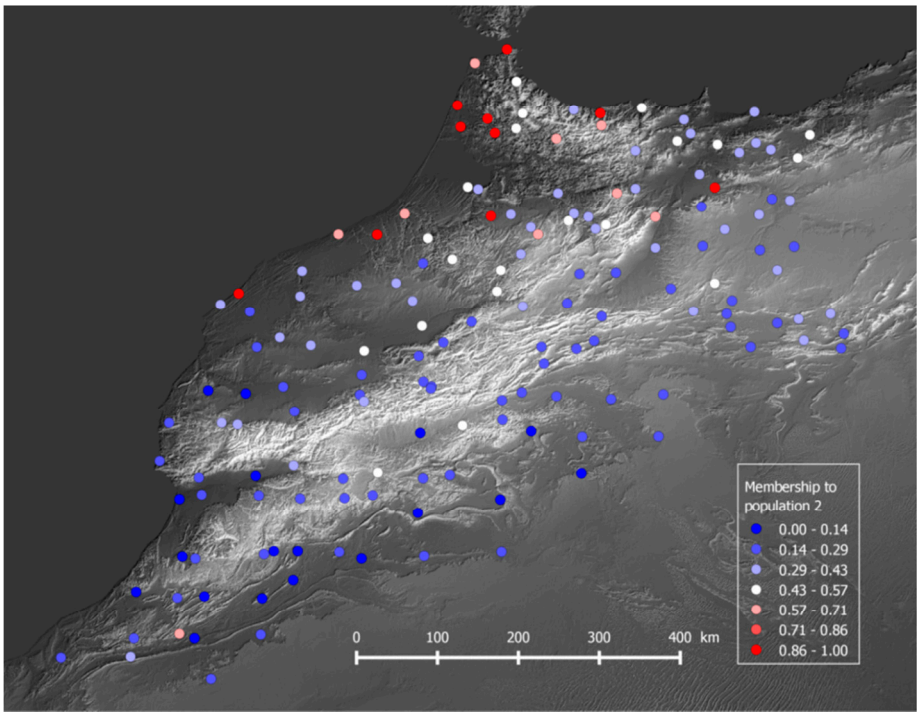


*Figure 6.9 Map of membership coefficient to population 2 for **goats** in the case of K=2.*

6.3 Identification of genetic markers under selection in sheep

Due to datasets sizes, we decided to first perform univariate models on original variables in order to compute associations within a reasonable time and facilitate the handling of output files. In fact, after recoding SNPs as three genotypes (see recodePLINK in section 3.5.1), genetic datasets were constituted of 5.1 and 5.3 million genotypes for sheep and goats respectively. Therefore, multi-resolution analysis were performed afterwards on a subset of SNPs with all selected variables, as described in next section.

We note that the weak population structure observed does not allow us to apply population genetics approaches to the datasets, or to use population structure variables to perform multivariate analysis in Samβada. Another method, using artificially defined populations, is presented in section 6.3.2.

6.3.1 Samβada results

Original resolution models

A False discovery threshold of 0.2 was applied on Samβada's results to define the significance level (see section 3.5.1 p59). With this threshold, Samβada identifies only 40 significant models for sheep (Table 6.2). However, we note that a FDR threshold of 0.1 would have been sufficient to detect most of these associations (28 out of 40), as some are showing low Q-values (minimum of 0.0001).

Genotypes in significant models are recurrently associated with precipitation (prec4, prec8, prec9, bio14, bio15), accounting for 25 amongst the most significant associations (Figure 6.10). A few genotypes are associated to maximum temperatures. Longitude and latitude, however, are rarely involved in significant models as only 4 genotypes are associated with longitude, although they are amongst the most significant models. On the other hand, only three genotypes are associated with DEM-derived variables.

Occurrences by chromosome highlights an uneven distribution of genotypes involved in significant models (Figure 6.10). Chromosome 23 is the most represented and its significant genotypes are mostly correlated to precipitation and maximum temperature in April (Figure 6.14). The next most remarkable chromosome is chromosome 19, including all 4 genotypes correlated with longitude. Among them, one is detected by longitude only while the others are also correlated with precipitation variables (prec8, prec9, bio14).

*Table 6.2 Summary of significant models (FDR=0.2), genotypes and variables at the original resolution with Samβada in **Sheep***

	# of Significant models	# of models computed
Number of models	40	166 545 473
Number of genotypes	26	1 842 823 x3
Number of variables	14	32

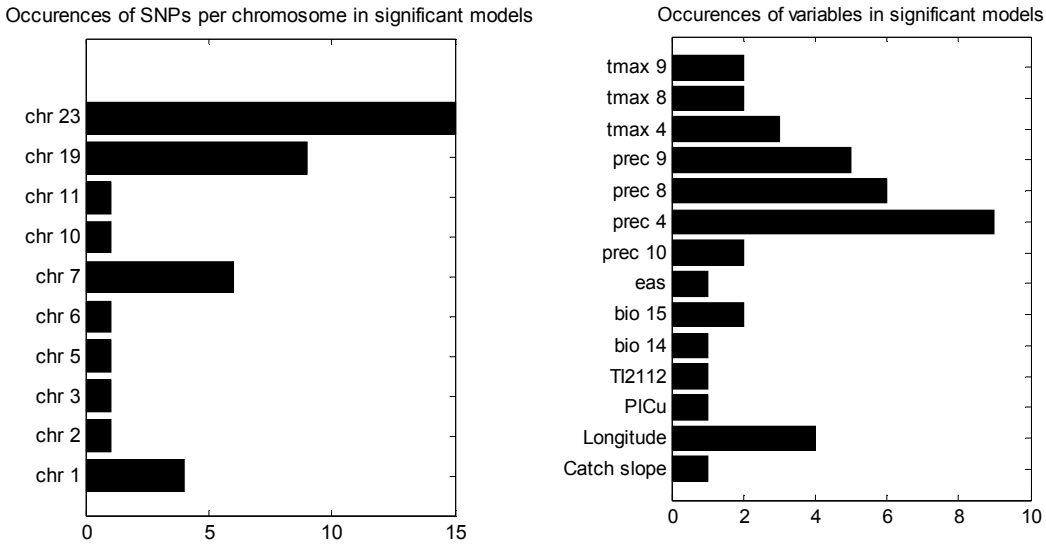


*Figure 6.10 Distribution of the frequency of environmental variables at their original resolution and of SNPs per chromosome involved in significant models from Samβada for **sheep**.*

Models involving multi-scale variables

As mentioned above, computing multi-scale models on the entire genetic dataset cannot be achieved in a reasonable time and we decided to select a subset of loci that were associated with variables at their original resolution with a Q-value lower than 0.4. With this threshold, 412 models were kept, comprising 243 SNPs.

Samβada was thus run again on this set of 243 SNPs (729 genotypes) and 120 variables (28 variables x 5 resolutions). In that case, a model is kept if at least one resolution showed a Q-value lower than the FDR threshold selected (0.2). To focus on multi-scale environmental associations, the following results exclude latitude and longitude. Orientation and Coherency were not considered either, as they could not be computed at the highest resolution.

In these models involving multi-resolution variables, 11 additional genotypes are detected with a FDR threshold of 0.2 (Table 6.3). However, with a FDR threshold of 0.1, there is no change with respect to the original resolution models. Most additional SNPs detected are thus showing weak

significance. Table 6.5 (p133) shows the significant associations between genotypes and multi-scale variables for sheep.

These newly detected genotypes are mostly correlated with precipitation or DEM variables (VRM, TI2112). Regarding occurrences per chromosome, chromosome 7 shows the double of significant models compared to original resolution, all of these genotypes being related to precipitation variables and showing high significance with window sizes of 5x5 and 9x9 (Figure 6.11).

We note that without a membership coefficient and considering the limited number of models involving latitude and longitude, we did not consider relevant to perform multivariate models.

*Table 6.3 Summary of significant models from multi-scale analysis with Samβada in **Sheep***

	# of significant models	# of models computed
Number of models	53	126 315
Number of genotypes	33	243 x 3
Number of variables	16	28

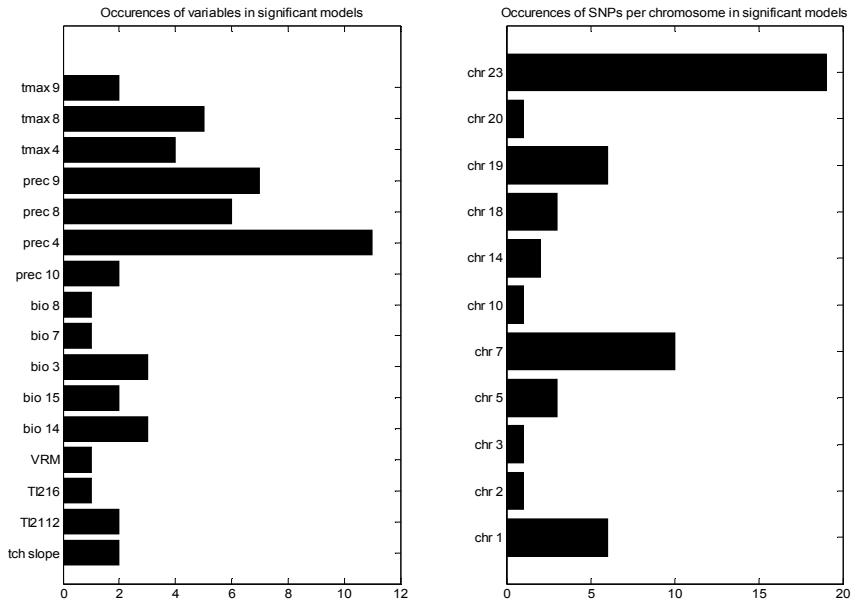


*Figure 6.11 Distribution of the frequency of multi-scale environmental variables and of SNPs per chromosome involved in significant models from Samβada for **sheep**.*

6.3.2 Comparison between methods

LFMM was also applied for each species (section 3.5.1). However, because computation time for several millions of markers on each variable would have been too long to run, it was only applied to environmental variables that either showed up frequently in Samβada's significant models or

were involved in highly significant models. We chose the following variables: prec 9, prec 4, prec 8, tmax 4, and bio 15. We computed LFMM models with 1, 2 and 3 latent factors. Since the number of latent factors did not influence substantially the SNPs detected, we decided to discuss those encountered with K=1 only. It must be noted that the distribution of p-values did not allow us to apply FDR to LFMM results. Therefore, we opted for a Bonferroni correction (0.05/number of models computed).

A total of 32 significant models are identified by LFMM (Table 6.6). Detected SNPs are significantly associated with all variables but prec 9, prec 8 and tmax 4 are more frequent (Figure 6.12). Seven SNPs are detected twice, mostly with prec 8 and prec 9. For LFMM results, there is no particular observation regarding the distribution of SNPs per chromosome.

*Table 6.4 Summary of models of association between SNPs and a subset of variables (prec 9, prec 4, prec 8, tmax 4, bio 15) with LFMM in **sheep***

	# of significant models	**# of models computed**
Number of models	32	8 024 225
Number of SNPs	25	1 604 845
Number of variables	5	5

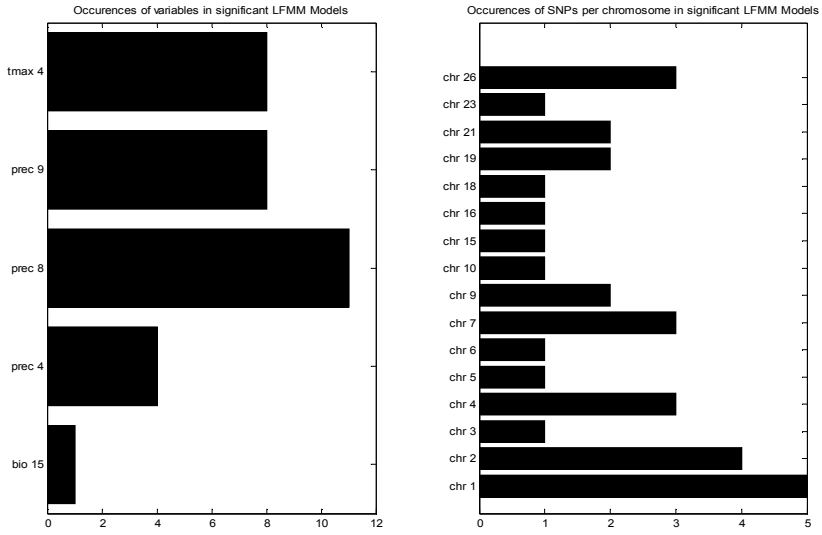


*Figure 6.12 Distribution of the frequency of a subset of environmental variables and of SNPs per chromosome involved in significant models from LFMM for **sheep**.*

In addition to Samβada and LFMM, collaborators Badr Benjelloun and François Pompanon from the Laboratoire d'Ecologie Appliquée (LECA; Université Joseph Fourier, Grenoble) applied XP-CLR to detect signatures of selection (Chen *et al.* 2010). XP-CLR is a genome scan approach that attempt to identify selective sweeps between two populations. Because XP-CLR is a population-based method, it was decided to define populations based on environmental criteria, due to the weak population structure identified in both species. Therefore, 2 groups of 20 individuals were

extracted at both ends of the four environmental gradients (i.e. altitude, bio7, bio15, prec4). For example, the first population is constituted of 20 goats encountered at the lowest altitude values and the second population is constituted of 20 goats encountered at the highest altitude values. The major drawback is that it does not necessarily guarantee a random spatial spread of samples and thus might generate correlations between environmental variables higher than what we observed for the whole case study. We decided to compare directly the significant XP-CLR scores to those calculated by Samβada and by LFMM, looking at their positions on chromosomes. In XP-CLR, a moving window is centred on a position with a distance defined at 2500 base pairs on each side. Therefore, correspondences with significant SNPs detected by LFMM and Samβada were searched at a distance of 2500 base pairs on each side of SNPs positions.

XP-CLR detected many more regions than Samβada or LFMM. Its results are shown in graphs per chromosome (Figure 6.14 to Figure 6.15 and in Appendix II.c). XP-CLR detected 8 loci in common with Samβada but none with LFMM. Each of the four variables tested in XP-CLR were regularly involved in significant detections.

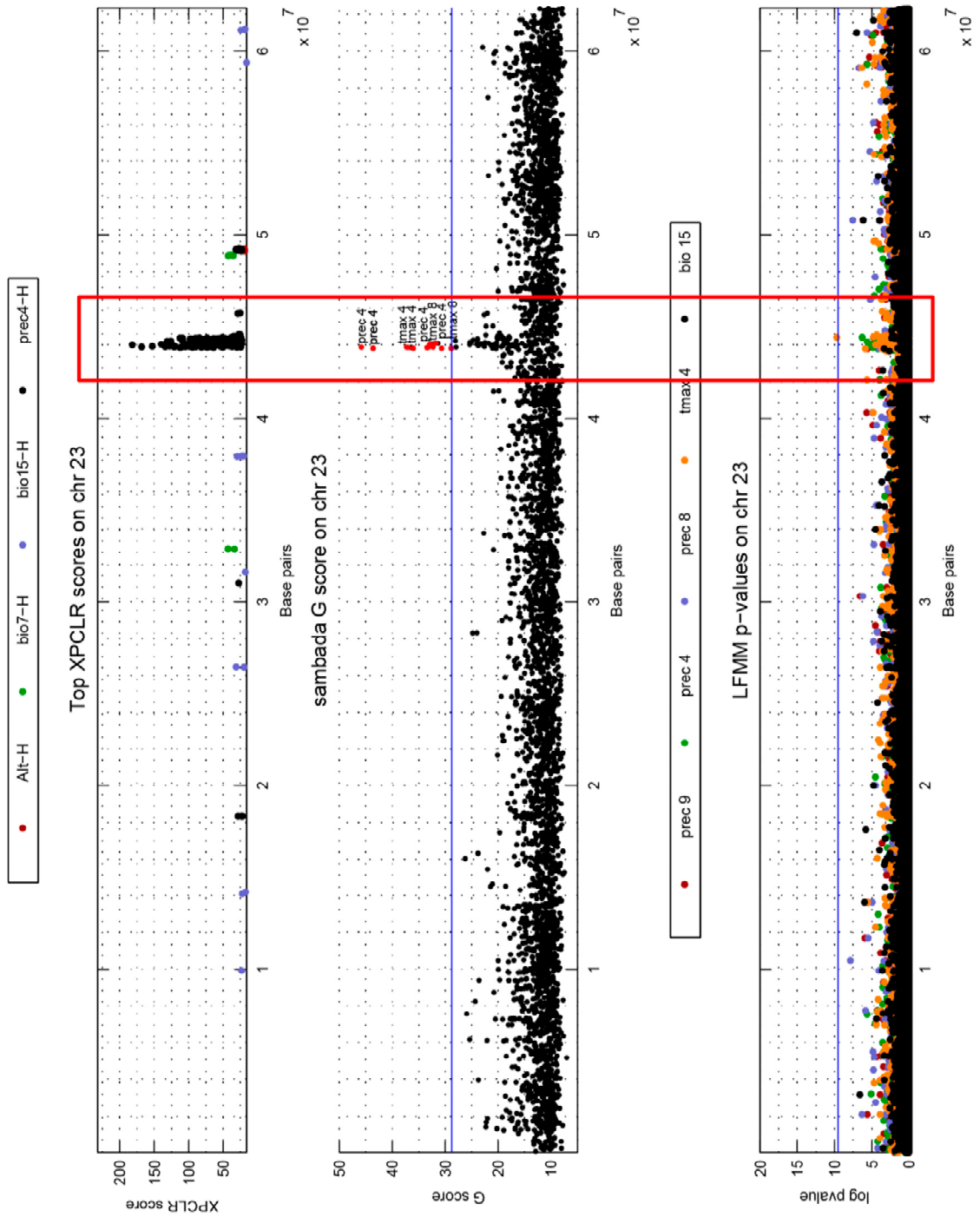


*Figure 6.13 Comparison between significant XP-CLR, Sambada and LFMM results according to their position on chromosome 23 for **sheep**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Sambada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance. The peak is highlighted with a red frame.*

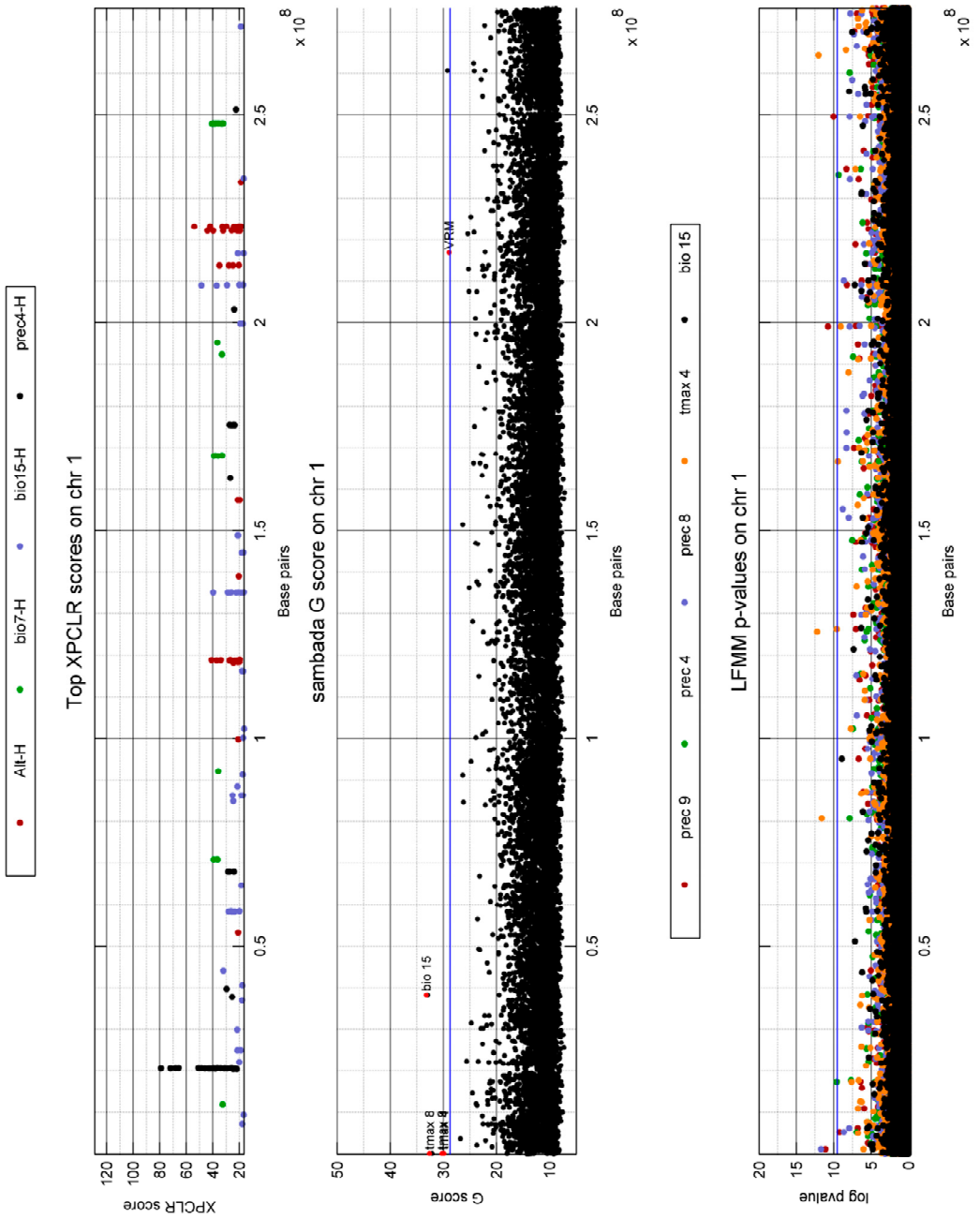


*Figure 6.14 Comparison between significant XP-CLR, Sambada and LFMM results according to their position on chromosome 1 for **sheep**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Sambada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

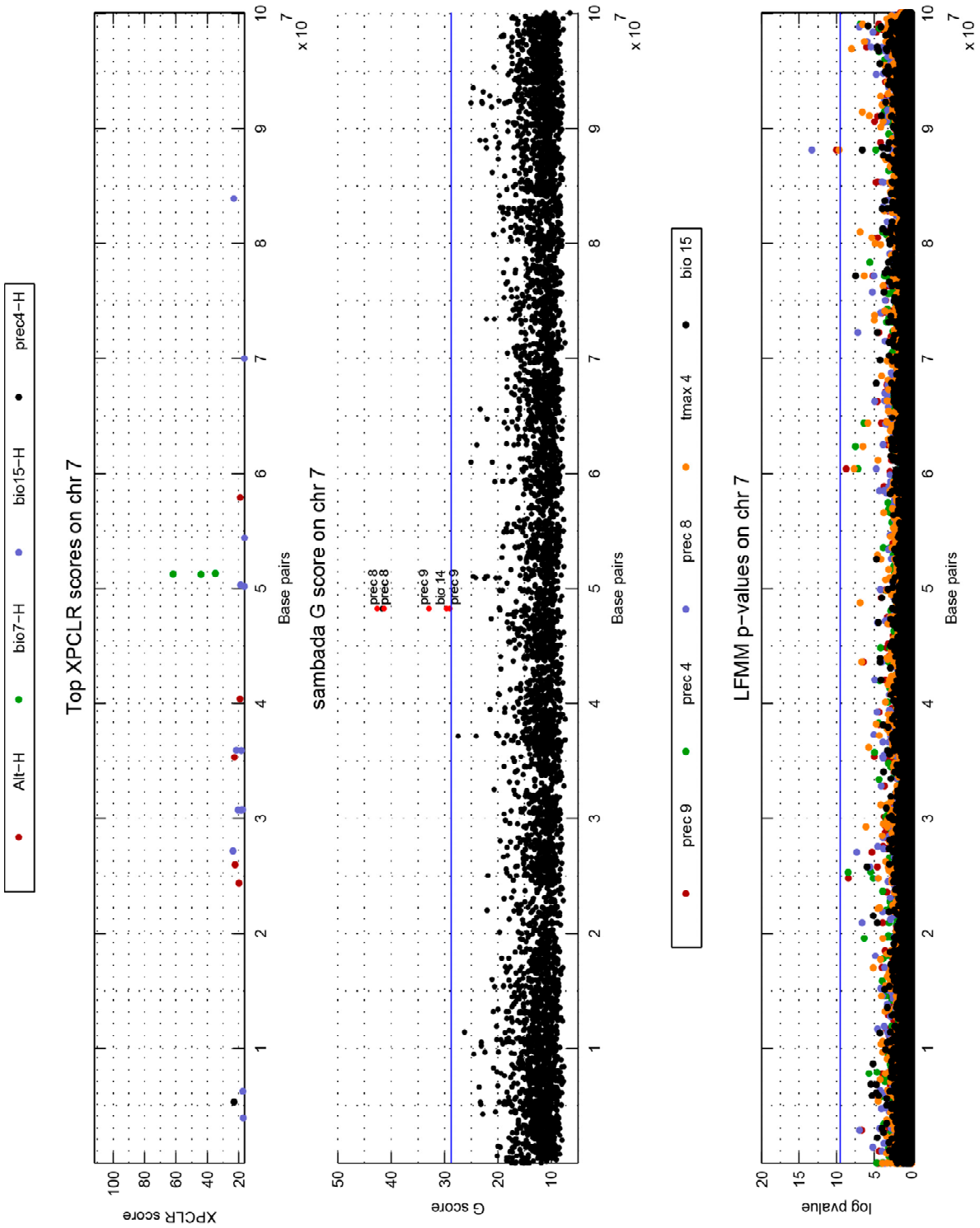


*Figure 6.15 Comparison between significant XP-CLR, Sambada and LFMM results according to their position on chromosome 7 for **sheep**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Sambada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

Chromosome 23 shows the largest number of significant SNPs detected by Samβada (Figure 6.13). Comparing positions allowed us to notice the presence of a peak of significant SNPs in each method, even though it is not significant in LFMM. The peak is also clear in XP-CLR and corresponds to an association with the environmental variable prec 4. This is the only peak detected in sheep with multiple methods.

SNPs detected by LFMM are mostly located on chromosome 1 and we can see that many other SNPs are close to being significant in this method (Figure 6.13). However, none of them is detected by either Samβada or XP-CLR.

Finally, some of the most significant genotypes detected in Samβada are located on chromosome 7 (Figure 6.15). Here as well, there is no common detection between Samβada and the two other methods.

*Table 6.5 Samβada's significant results involving multi-resolution variables in **sheep**. The table shows one model per line including the genotype, the associated environmental variable, the resolution of the variable at which the highest G score was found (Best resolution), the highest G and Wald scores of the model involving the variable at the best resolution, the frequency of the genotype, the most significant LFMM p-value of the corresponding SNP and its associated variable, the highest XP-CLR score and its corresponding variable, the Moran's I of the genotype obtained with a neighbourhood of 20 individuals. Models are ranked according to their G score. SNPs identified in the peak on chr 23 are in bold.*

	Genotypes	Variable	Best resolution	Highest G	Highest Wald	Frequency	Highest LFMM	LFMM variable	Highest XP-CLR	XP-CLR variable	Moran
Fig 6.18	19:2170224_AA	prec_9	ws01	51.69	28.96	0.269	5.89E-09	prec 9			0.31
Fig 6.19	**23:43874160_GG**	**prec_4**	**initial**	**45.84**	**28.87**	**0.219**	**0.00013**	**tmax 4**	**117.83**	**bio7-H**	**0.37**
Fig 6.20	**23:43794976_GG**	**prec_4**	**initial**	**43.61**	**28.68**	**0.238**	**9.09E-05**	**prec 4**			**0.31**
Fig 6.21	**23:43812782_GG**	**prec_4**	**initial**	**43.61**	**28.68**	**0.238**	**8.16E-05**	**prec 4**			**0.31**
	7:48256781_AG	prec_8	ws08	42.61	20.06	0.106	0.010115	prec 8			0.11
	7:48262822_AG	prec_8	ws08	42.61	20.06	0.106	0.00925	prec 8			0.11
	19:2167574_AA	prec_9	initial	42.32	27.13	0.294	3.22E-08	prec 9			0.31
Fig 6.24	7:48256781_GG	prec_8	ws04	41.33	22.10	0.119	0.010115	prec 8			0.17
	7:48262822_GG	prec_8	ws04	41.33	22.10	0.119	0.00925	prec 8			0.17
	19:2167574_AA	prec_8	ws02	37.93	23.44	0.294	3.22E-08	prec 9			0.31
	23:43874160_GG	**tmax_4**	**initial**	**37.31**	**27.14**	**0.219**	**0.00013**	**tmax 4**	**117.83**	**bio7-H**	**0.37**
	19:2170224_AA	prec_8	ws08	37.11	22.07	0.269	5.89E-09	prec 9			0.31
Fig 6.22	**23:43861704_GG**	**prec_4**	**ws08**	**37.02**	**27.67**	**0.406**	**3.04E-06**	**prec 4**	**73.95**	**Alt-H**	**0.22**
	23:43794976_GG	**tmax_4**	**initial**	**36.04**	**26.52**	**0.238**	**9.09E-05**	**prec 4**			**0.31**
	23:43812782_GG	**tmax_4**	**initial**	**36.04**	**26.52**	**0.238**	**8.16E-05**	**prec 4**			**0.31**
	23:43847594_GG	**prec_4**	**ws01**	**33.47**	**25.79**	**0.463**	**3.63E-06**	**prec 4**	**32.90**	**Alt-H**	**0.15**
	1:38304177_AG	bio_15	ws08	33.21	18.04	0.106	0.052543	bio 15			0.12
	1:38304177_GG	bio_15	ws08	33.21	18.04	0.106	0.052543	bio 15			0.12
	23:43874160_GG	**tmax_8**	**initial**	**33.12**	**24.50**	**0.219**	**0.00013**	**tmax 4**	**117.83**	**bio7-H**	**0.37**
	7:48256781_AG	prec_9	ws08	32.91	20.66	0.106	0.010115	prec 8			0.11
	7:48262822_AG	prec_9	ws08	32.91	20.66	0.106	0.00925	prec 8			0.11
Fig 6.23	**23:44038684_AA**	**prec_4**	**ws01**	**32.87**	**22.67**	**0.163**	**2.01E-05**	**prec 4**	**105.87**	**bio7-H**	**0.14**
	1:190582_AA	tmax_8	ws08	32.58	18.13	0.063	0.002206	tmax 4			0.11
	23:43874160_AG	**prec_4**	**ws16**	**32.30**	**21.50**	**0.163**	**0.00013**	**tmax 4**	**117.83**	**bio7-H**	**0.26**
	23:44095648_AA	**prec_4**	**ws01**	**31.90**	**21.74**	**0.144**	**4.42E-05**	**prec 4**			**0.19**
	23:44084253_AA	**prec_4**	**ws01**	**31.70**	**21.59**	**0.144**	**1.85E-05**	**prec 4**			**0.19**

Fig 6.25	20:50510912_GG	Catch_slope	0180	31.60	15.53	0.206	0.034827	prec 9	29.29	Alt-H	0.24
	3:211734411_AA	tmax_9	initial	30.88	23.49	0.244	0.002991	tmax 4			0.07
	19:2162818_GG	prec_9	ws01	30.84	22.71	0.325	6.66E-06	prec 9			0.28
	23:43794976_GG	**prec_10**	**initial**	**30.68**	**20.35**	**0.238**	**9.09E-05**	**prec 4**			**0.31**
	23:43812782_GG	**prec_10**	**initial**	**30.68**	**20.35**	**0.238**	**8.16E-05**	**prec 4**			**0.31**
	1:190582_AA	tmax_9	ws04	30.26	17.28	0.063	0.002206	tmax 4			0.11
	18:60885624_AG	bio_8	ws04	30.10	19.95	0.125	0.123731	tmax 4			0.01
	19:2170224_AA	bio_14	initial	30.00	16.49	0.269	5.89E-09	prec 9			0.31
	1:190582_AA	tmax_4	ws04	29.92	18.56	0.063	0.002206	tmax 4			0.11
	5:70648057_GG	TI2112	initial	29.91	20.38	0.138	0.324822	prec 8			0.04
	10:13537408_GG	Catch_slop	initial	29.86	23.56	0.3	0.384299	prec 4			0.01
	18:11745070_AG	bio_3	ws04	29.65	22.34	0.156	0.214845	bio 15			0.12
	7:48256781_AG	bio_14	ws04	29.62	22.59	0.106	0.010115	prec 8			0.11
	7:48262822_AG	bio_14	ws04	29.62	22.59	0.106	0.00925	prec 8			0.11
	18:11745070_GG	bio_3	ws04	29.56	22.37	0.163	0.214845	bio 15			0.11
	14:875672_AG	tmax_8	initial	29.40	21.88	0.475	0.026939	prec 4			0.13
	5:60766984_AG	TI216	initial	29.37	22.59	0.444	0.218402	tmax 4			0.07
	5:70648057_AG	TI2112	initial	29.32	20.02	0.119	0.324822	prec 8			0.04
	14:875672_AG	bio_7	ws16	29.09	23.16	0.475	0.026939	prec 4			0.13
	7:48256781_GG	prec_9	ws08	29.04	20.37	0.119	0.010115	prec 8			0.17
	7:48262822_GG	prec_9	ws08	29.04	20.37	0.119	0.00925	prec 8			0.17
	2:66041147_AG	bio_3	ws16	29.03	22.69	0.244	0.030539	prec 9			0.11
	1:216819947_AG	VRM	initial	28.93	16.38	0.188	0.200267	tmax 4			0.05
	23:43794976_AG	**prec_4**	**ws16**	**28.87**	**20.46**	**0.175**	**9.09E-05**	**prec 4**			**0.18**
	23:43812782_AG	**prec_4**	**ws16**	**28.87**	**20.46**	**0.175**	**8.16E-05**	**prec 4**			**0.18**
	23:43794976_GG	**tmax_8**	**initial**	**28.84**	**22.41**	**0.238**	**9.09E-05**	**prec 4**			**0.31**
	23:43812782_GG	**tmax_8**	**initial**	**28.84**	**22.41**	**0.238**	**8.16E-05**	**prec 4**			**0.31**

*Table 6.6 LFMM's significant results for univariate models involving a subset of environmental variables in **sheep**. The table shows one model per line including the locus, the associated variable, its z score and p-value, the minor allele frequency, the mean and maximum G score between the three genotypes in Samβada for the same environmental variable, the maximum Moran's' I between the three genotypes and the distance at which it was found. Models are ranked according to their P-value.*

	SNPs	Variable	Zscore	P-value	MAF	Mean G score	Max G score	Max Moran	Distance of max Moran
Fig 6.26	7:88146918	prec_8	8.2671	5.13E-14	0.242	4.039	7.739	0.029	5
	4:12303967	prec_8	8.0433	1.89E-13	0.243	0.727	1.434	0.016	1
Fig 6.28	1:125752841	tmax_4	7.4421	5.86E-12	0.053	0.409	0.852	0.014	3
Fig 6.27	16:16446171	prec_8	7.4242	6.48E-12	0.013	0.657	1.066	0.039	1
	21:5710355	tmax_4	7.4048	7.23E-12	0.296	5.926	13.317	0.012	5
	2:212487548	prec_9	7.3296	1.10E-11	0.227	1.418	3.357	0.033	1
	1:264377461	tmax_4	7.3047	1.26E-11	0.230	0.277	0.731	-0.004	3
	26:29680812	prec_8	7.2689	1.54E-11	0.039	3.173	9.238	0.099	1
	5:95006908	prec_4	7.2318	1.90E-11	0.105	2.532	4.037	0.042	2
	26:38634496	tmax_4	7.1867	2.43E-11	0.235	1.828	4.371	-0.001	2
	1:1189051	prec_9	7.1732	2.62E-11	0.123	1.658	3.309	0.008	1
	6:7438056	tmax_4	7.1573	2.86E-11	0.270	0.145	0.337	0.053	1
	1:80764418	tmax_4	7.1391	3.16E-11	0.000	0.008	0.016	-0.008	10
	15:8048267	prec_9	7.1301	3.32E-11	0.224	0.494	0.881	0.003	2
	1:1189051	prec_8	7.1276	3.37E-11	0.123	0.387	1.005	0.008	1
	10:47538166	prec_8	7.0627	4.81E-11	0.065	0.279	0.622	0.009	5
	2:104456155	prec_9	6.9853	7.35E-11	0.125	0.275	0.696	0.022	1

9:9714798	prec_4	6.9784	7.63E-11	0.170	0.011	0.025	0.023	1
4:12303967	prec_9	6.9656	8.18E-11	0.243	0.233	0.524	0.016	1
7:88146918	prec_9	6.9311	9.86E-11	0.242	3.053	5.952	0.029	5
19:43221906	prec_8	6.9118	1.09E-10	0.255	3.577	8.397	0.012	2
2:222236016	prec_8	6.9053	1.13E-10	0.258	1.568	3.587	0.008	8
21:4412095	prec_4	6.8466	1.56E-10	0.263	1.948	4.329	0.042	4
3:180621248	prec_8	6.8374	1.64E-10	0.085	0.198	0.605	0.091	1
19:43221906	prec_9	6.8131	1.87E-10	0.255	1.481	3.823	0.012	2
4:32900303	prec_8	6.8125	1.87E-10	0.156	2.726	5.894	0.081	3
23:44417365	tmax_4	-6.795	2.06E-10	0.239	8.333	17.276	0.028	4
7:88146918	tmax_4	-6.788	2.14E-10	0.242	3.354	7.440	0.029	5
26:29680812	bio_15	-6.774	2.30E-10	0.039	2.180	5.608	0.099	1
18:57803509	prec_9	6.7696	2.36E-10	0.059	3.560	6.226	0.102	2
2:212487548	prec_8	6.7495	2.63E-10	0.227	1.585	3.595	0.033	1
9:83776120	prec_4	-6.731	2.90E-10	0.191	6.362	11.478	0.081	1

The analysis of the relationship between LFMM's p-values and Samβada's G score confirmed that there is no relationship between their results. SNPs detected by both methods are clearly distinct (Figure 6.16). We can also note that those detected by both Samβada and XP-CLR are not necessarily showing a high G score in Samβada or a low p-value in LFMM.

Figure 6.17 shows the Moran's I correlograms for 729 genotypes including the significant genotypes in Samβada (33 genotypes) and LFMM (25 SNPs, highest Moran's I among the three genotypes of each SNP was plotted). The difference between both methods is clear: Samβada identifies mostly genotypes with a strong spatial autocorrelation, while LFMM detects loci with weak or null spatial autocorrelation.

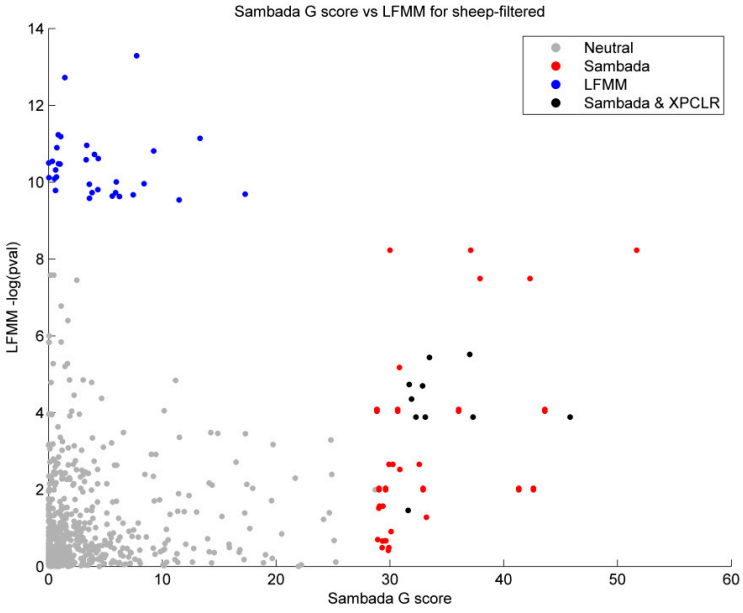


*Figure 6.16 Scatterplot of Samβada's G score against LFMM's p-value in **sheep**. Neutral loci are shown in grey, those detected by Samβada in red, those by Samβada and XP-CLR in black and those by LFMM in blue. Loci detected by LFMM are clearly distinct from those detected by Samβada. There is no distinction possible between genotypes detected by Samβada and XP-CLR.*

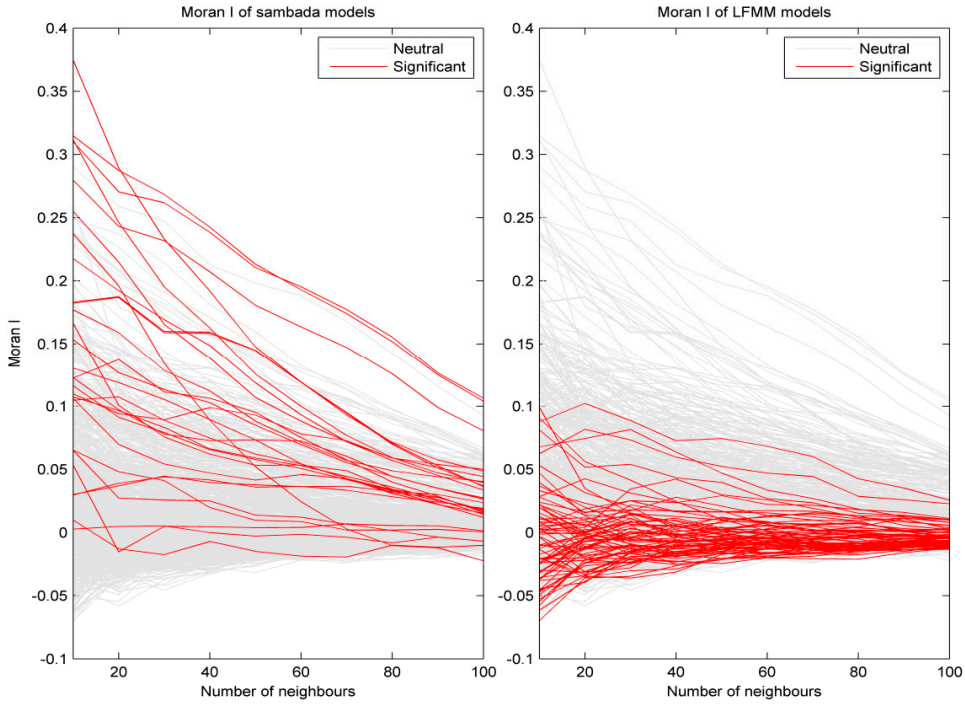


*Figure 6.17 Moran's I correlograms for **sheep**. Neutral genotypes are shown in grey and genotypes associated with environmental variables, either by Samβada (left) or LFMM (right) are in red.*

6.3.3 Visualisation of significant associations

Figures 6.18 to 6.28 allow us to compare the spatial distribution of genotypes significantly associated with environmental variables, their spatial autocorrelation, the influence of spatial resolution (or moving-window in the case of climatic variable) and other information like best models detected by LFMM and XP-CLR for the corresponding SNP. Details on the different parts of these graphs can be found in section 3.6 (p65).

The six examples shown in figures 6.18 to 6.24 are representative of the most significant models processed in Samβada. The first notable observation is the high level of spatial autocorrelation of the genotypes involved in significant models, which is clearly illustrated on the map showing the spatial distribution of SNPs and on the Moran's I autocorrelogram. In addition, LISA coefficients show one or two locally significant clusters of positive autocorrelation. These examples also illustrate the high spatial autocorrelation of climatic variables. This remark, however, is not valid for DEM variables such as for the catchment slope variable (see for instance Figure 6.25). Regarding the impact of multi-scale variables, there is a limited influence of window size on climatic variables but a strong influence of resolution on DEMs variables.

On the other, close SNPs in the peak on chromosome 23 display similar patterns of spatial distribution (Figure 6.18 to Figure 6.23), although the dataset was corrected to take linkage disequilibrium into account.

Afterwards, we show some of the most significant LFMM models (Figure 6.26 to Figure 6.28). Spatial patterns of distributions of significant SNPs are less autocorrelated than in Samβada's results and, thus, their association with variables is less straightforward. In fact, Moran's I is low for all of them and does not change much with distance. This is also true locally, as few points have significant LISA values.

Significant models identified by Samβada

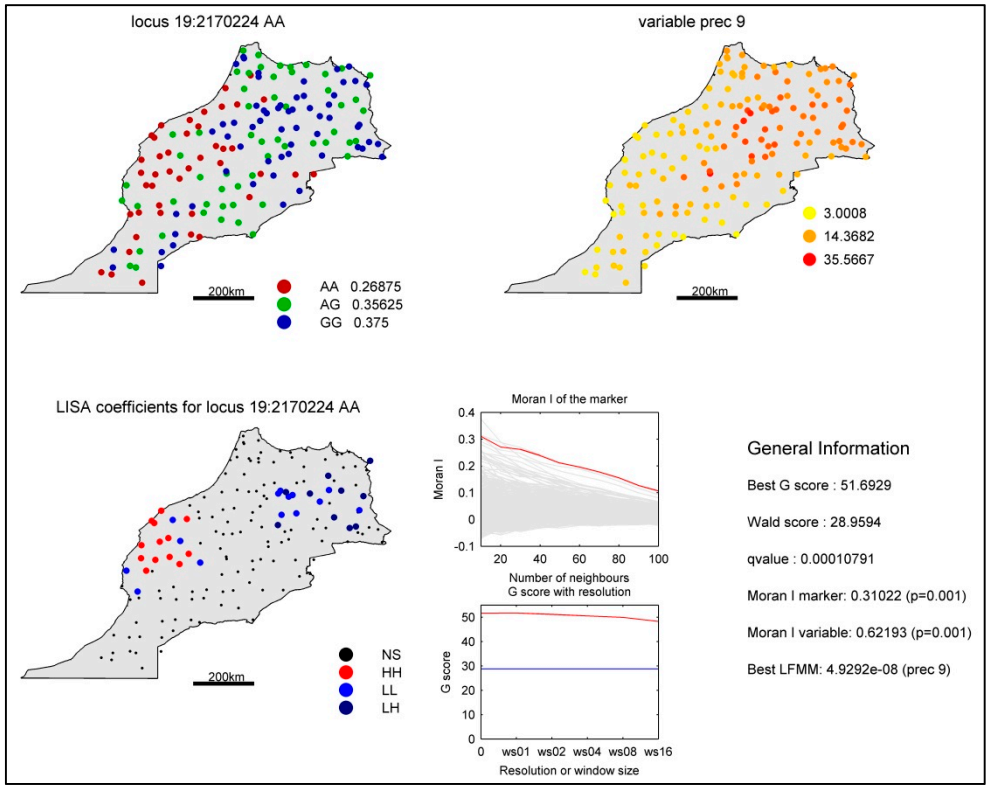


*Figure 6.18 Visualisation of the significant association between genotype 19:_AA and precipitation in September in **sheep** detected by Samβada.*

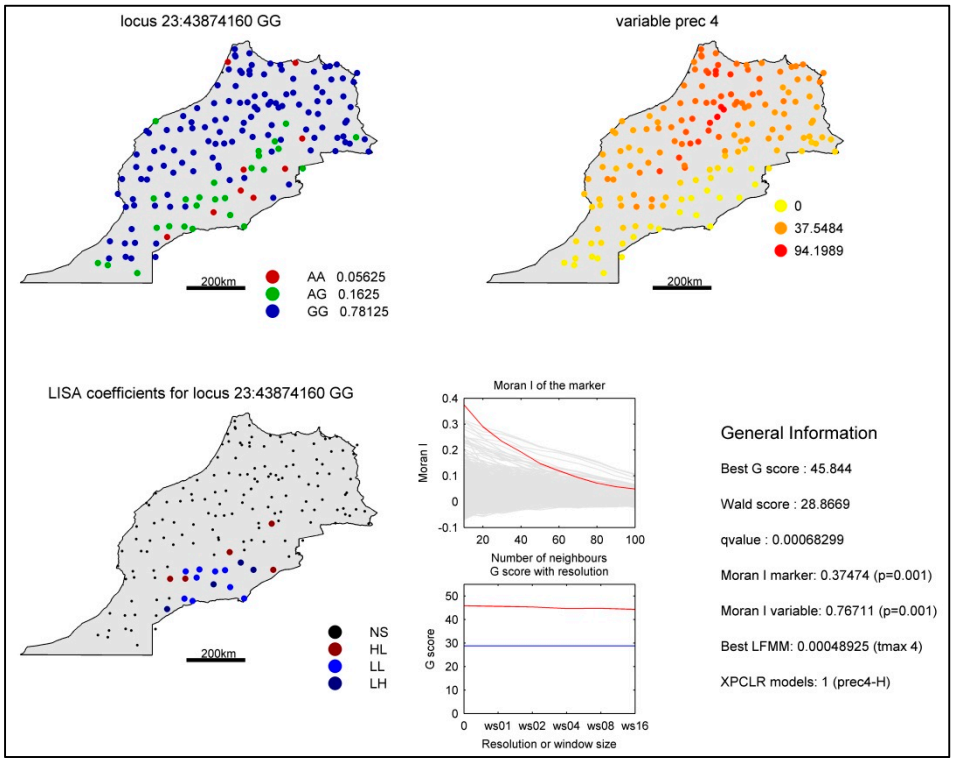


*Figure 6.19 Visualisation of the significant association between genotype 23:43874160_GG and precipitation in April in **sheep** detected by SamBada*

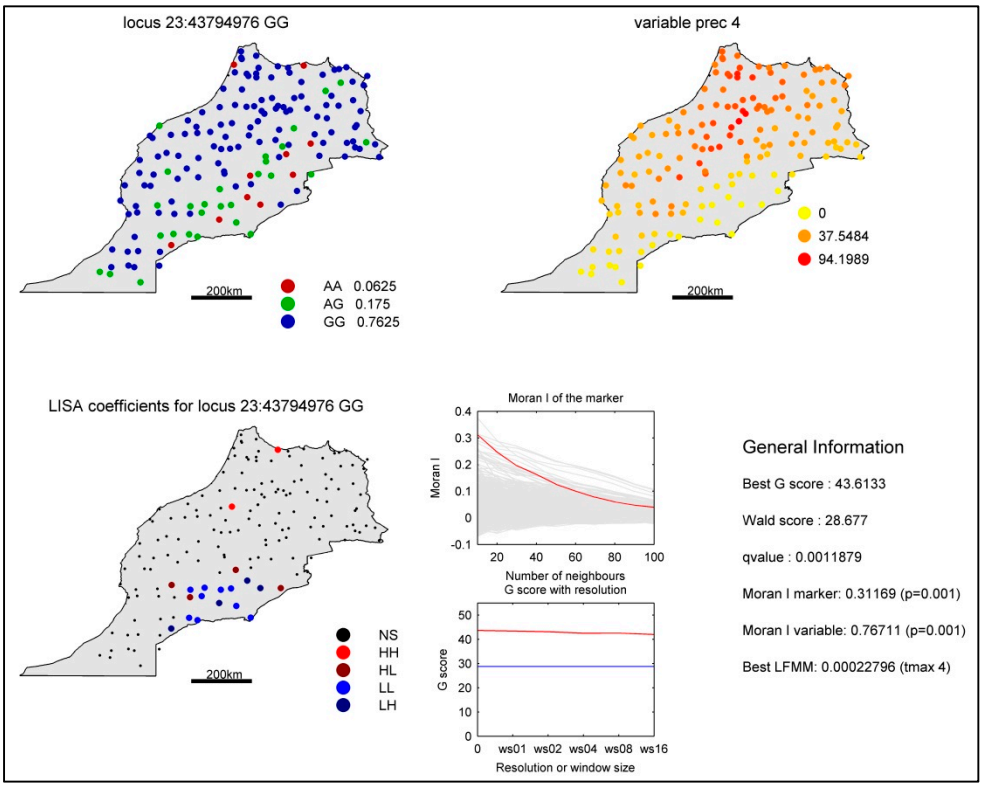


*Figure 6.20 Visualisation of the significant association between genotype 23:43794976_GG and precipitation in April in **sheep** detected by SamBada*

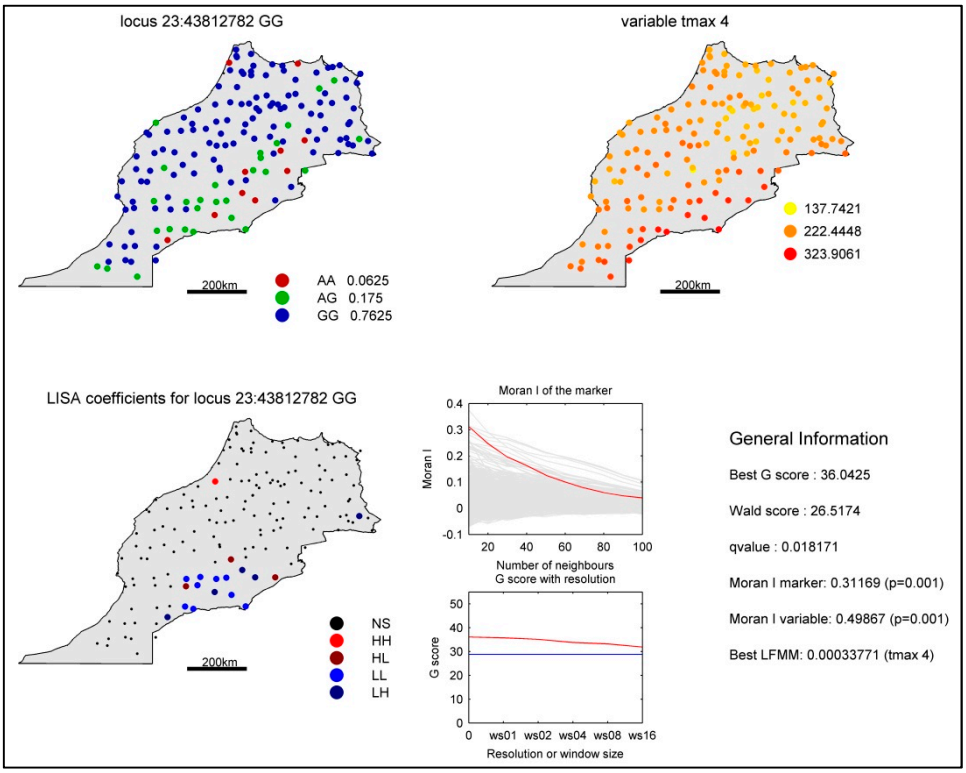


*Figure 6.21 Visualisation of the significant association between genotype 23:43823782_GG and maximal temperature in April in **sheep** detected by Samβada*

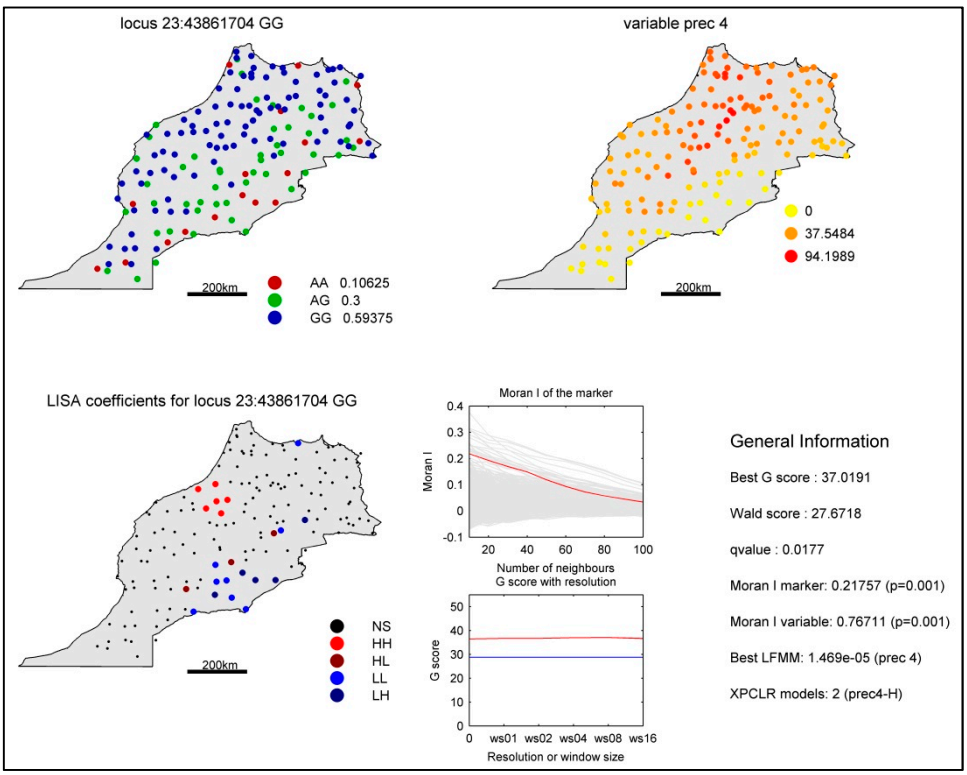


*Figure 6.22 Visualisation of the significant association between genotype 23:43861704_GG and precipitation in April in **sheep** detected by Samβada*

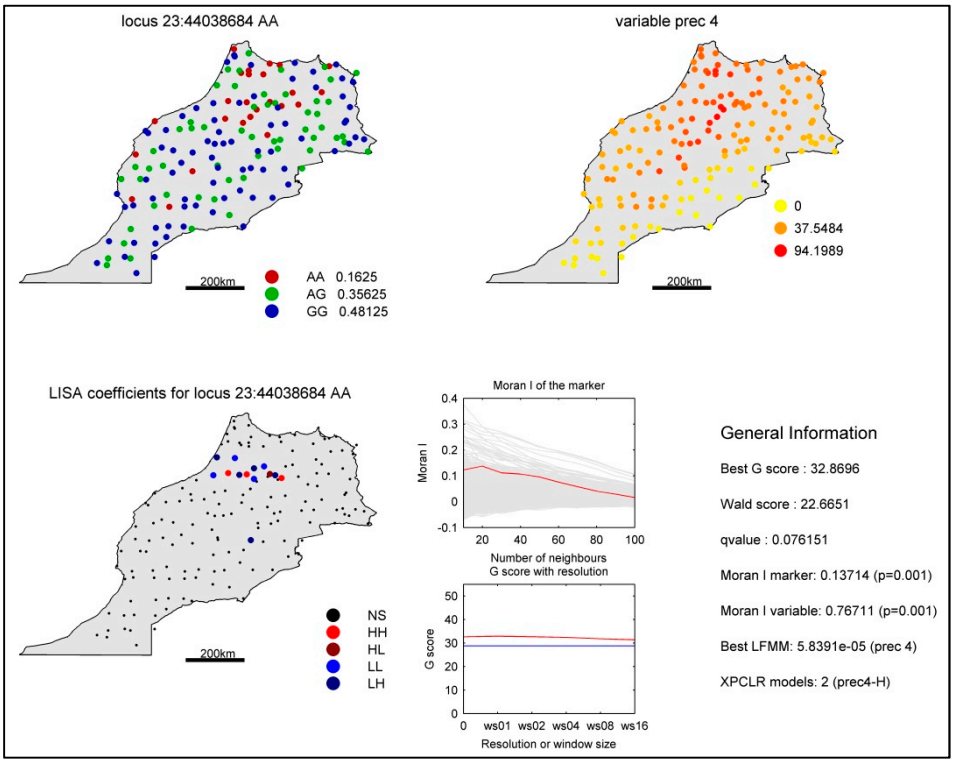


*Figure 6.23 Visualisation of the significant association between genotype 23:44038684_AA and precipitation in April in **sheep** detected by Samβada*

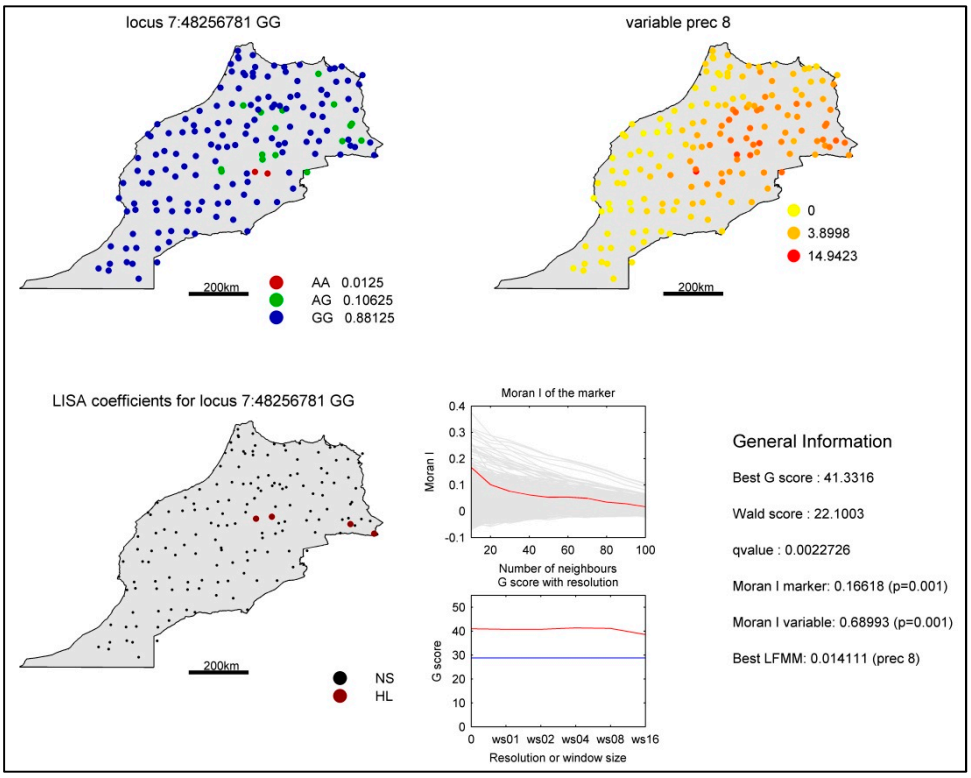


*Figure 6.24 Visualisation of the significant association between genotype 7:48256781_GG and precipitation in August in **sheep** detected by Samβada*

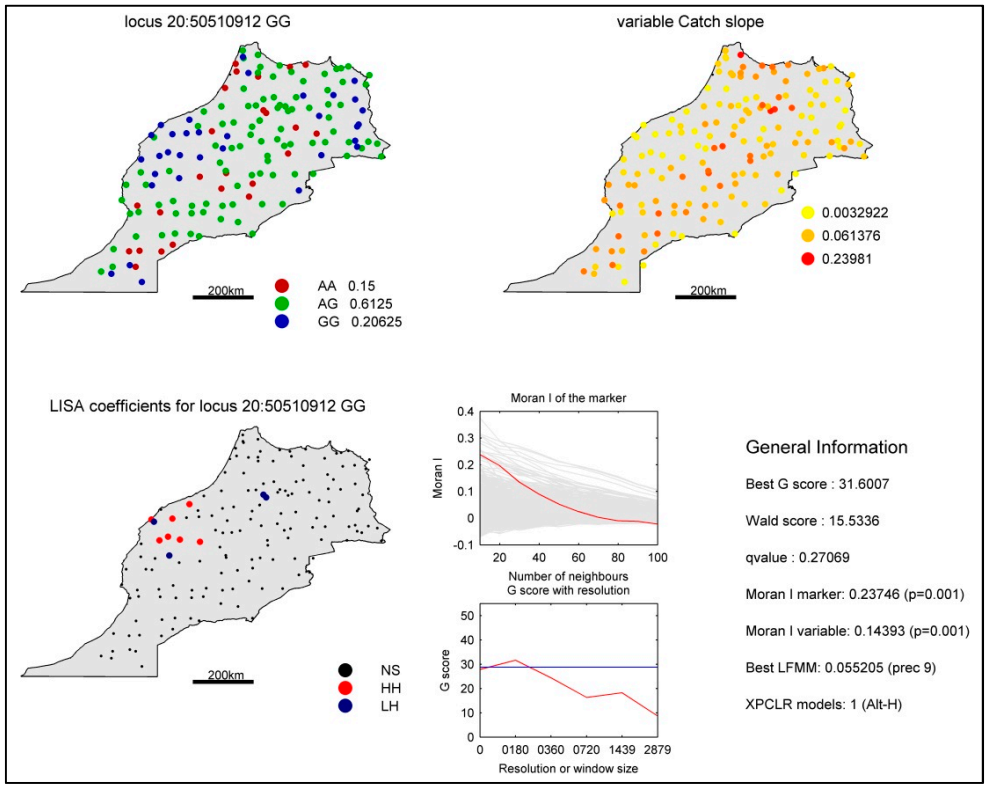


*Figure 6.25 Visualisation of the significant association between genotype 20:50510912_GG and Catchment slope (spatial resolution 180m) in **sheep** detected by Samβada.*

Significant models identified by LFMM

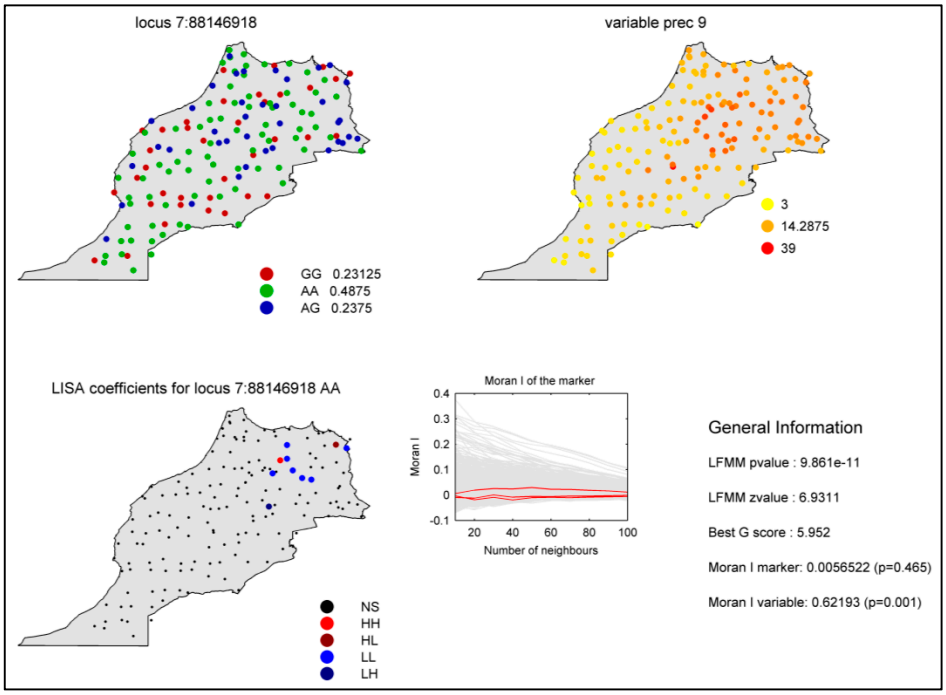


*Figure 6.26 Visualisation of the significant association between SNP 7:88146918 and precipitation in September in **sheep** detected by LFMM.*

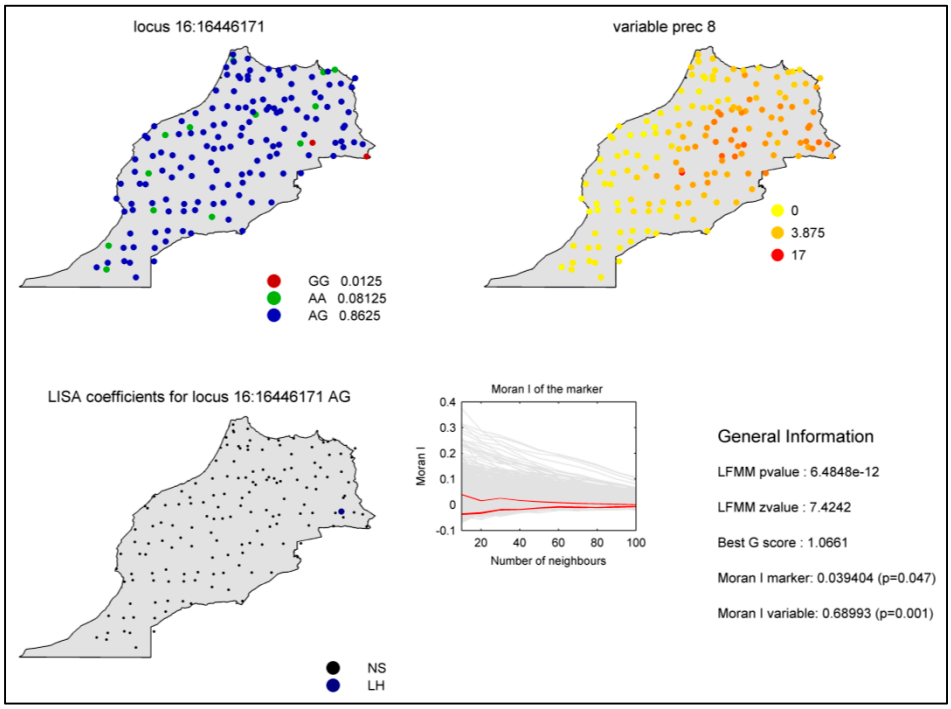


*Figure 6.27 Visualisation of the significant association between SNP 16:16446171 and precipitation in August in **sheep** detected by LFMM.*

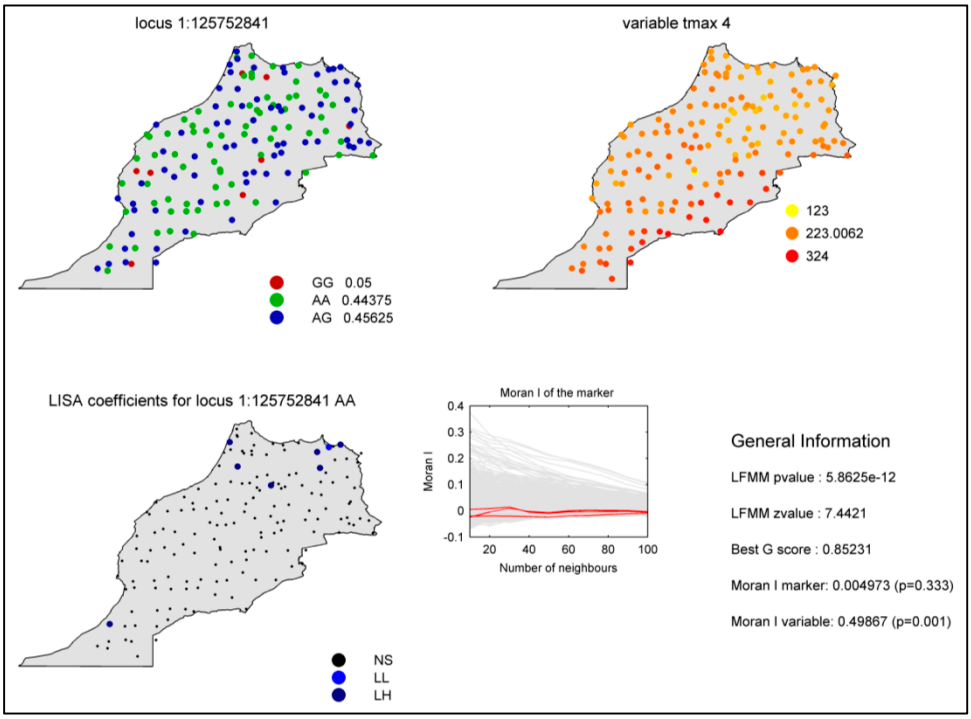


*Figure 6.28 Visualisation of the significant association between SNP 1:125752841 and maximum temperature in April in **sheep** detected by LFMM.*

6.4 Identification of genetic markers under selection in goats

6.4.1 Samβada results

Original resolution models

For goats, a total of 82 genotypes are identified by Samβada with weak Q-values (minimum of 0.04). Contrary to sheep, the amount of detections depend heavily on the false discovery threshold (see Table 6.7). Indeed, while a threshold of 0.2 identifies 95 significant models, a threshold of 0.1 identifies 9 significant models only, involving 8 SNPs associated with 3 variables (latitude, bio 15 and bio 7).

*Table 6.7 Summary of significant models (FDR=0.2), genotypes and variables at the original resolution with Samβada in **goats***

	# of significant models	# of models computed
Number of models	95	142 997 233
Number of genotypes	82	1 789 702 x3
Number of variables	20	28

Genotypes associated to latitude are by far the most represented in significant models. Indeed, among the 82 genotypes detected, 34 are significantly associated with latitude only and 4 with longitude only, supporting the weak latitudinal population structure identified in Figure 6.9 (p124). The other genotypes are significantly associated with environmental variables but with different occurrences per variables (Figure 6.29). Bio 15 (Precipitation Seasonality) is the most frequent variables involved in significant models and counting among the most significant models. Genotypes correlated with Bio 7 are the second most frequent (Temperature Annual Range) followed by altitude (SRTM). Two genotypes are associated with both bio 7 and tmax 7.

Occurrences of significant genotypes per chromosome is also uneven (Figure 6.29). Chromosome 6 shows the highest number of significant genotypes, which are mostly correlated with precipitation variables (bio 15, bio 14, prec8), altitude and latitude. On the other hand, chromosome 4 has 12 genotypes out of 17 correlated with latitude, the others with precipitation variables (bio 15, prec 1).

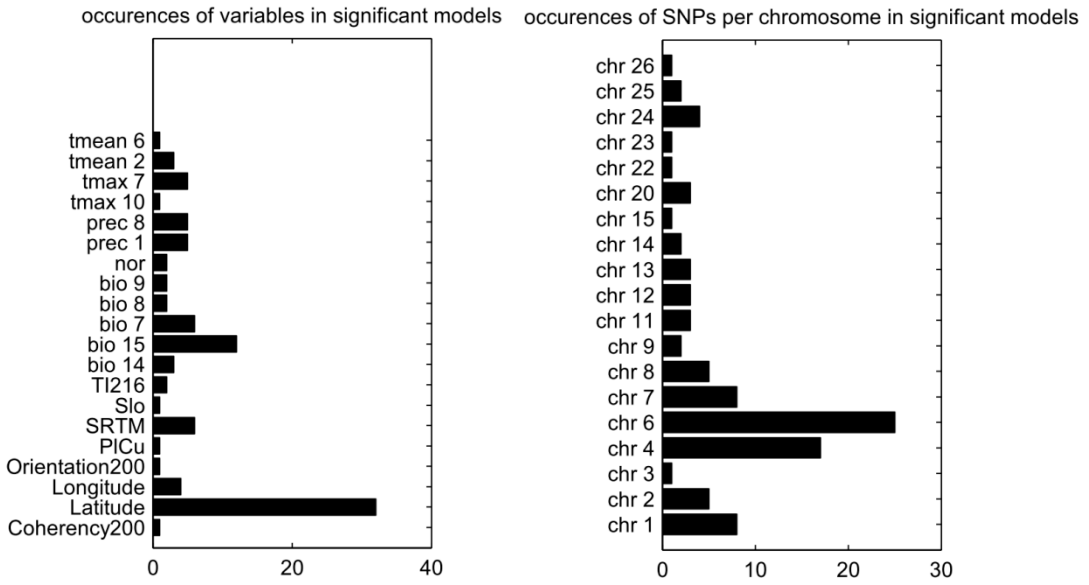


*Figure 6.29 Distribution of the frequency of environmental variables at their original resolution and of SNPs per chromosome involved in significant models from Samβada for **goats**.*

Multi-resolution results

To perform multi-resolution, we selected loci from models with a Q-value lower than 0.4 as explained in section 6.3.1 (p125). With this threshold, we kept 1878 models comprising 1341 SNPs.

Samβada processed this set of 1341 SNPs versus 120 variables (24 variables x 5 resolutions). A model is considered to be significant if at least one resolution shows a Q-value lower than the FDR threshold selected (0.2). To focus on multi-scale environmental associations, the following results exclude latitude and longitude. Orientation and Coherency were not considered either, as they could not be computed at all resolutions.

A total of 22 genotypes were newly detected (Table 6.8), mostly associated with DEM variables altitude and total insolation in June (Ti216). However, if we lower the significance threshold to 0.1, genotypes detected are the same than at original resolution and are still associated with variables bio 15 and bio 7. The major change in occurrences per chromosome takes place on chromosomes 4 (Figure 6.30). In fact, most of its significant SNPs were correlated to latitude, which we did not consider in the multi-resolution analysis. On the other hand, chromosome 6 has three more significant SNPs thanks to associations with altitude (SRTM) and bio 14.

Regarding the impact of multi-scale variables, window sizes of climatic variables had little influence on score of models. One exception is bio 14 (Precipitation of Driest Month) with four more detections (Figure 6.30), usually showing best results with moving window sizes of 9x9 and 17x17. Regarding DEM-derived variables, however, the number of significant genotypes associated with them increased substantially with multi-scale variables, in particular with TI216 (+4 associated genotypes) and for altitude (+5). It is interesting to note that most models involving TI216, Nor

and Slo show their best score at a resolution of 180m. These SNPs are located in seven different chromosomes. On the other hand, there is no optimal resolution for altitude.

We note that without a membership coefficient and considering the limited number of models involving latitude and longitude, we did not consider relevant to perform multivariate models.

*Table 6.8 Summary of significant models from multi-scale analysis with Samβada in **goats***

	Significant models	**Total**
Number of models	81	500 689
Number of genotypes	64	1341 x 3
Number of variables	19	25

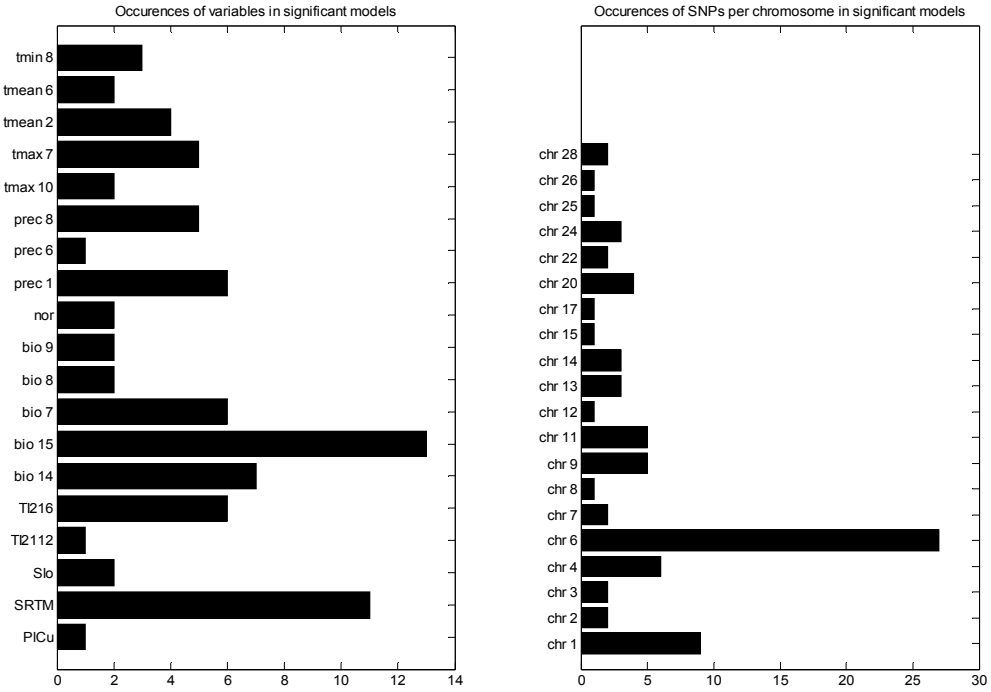


*Figure 6.30 Distribution of the frequency of multi-scale environmental variables and of SNPs per chromosome involved in significant models from Samβada for **goats**.*

6.4.2 Comparison between methods

To apply LFMM to the goats' dataset, we selected variable that were either involved in the most significant models in Samβada or were frequently involved in these models. We chose the following variables: bio15, bio 7, prec 7, TI216 (180m), SRTM (1440m).

One and two latent factors were tested for goats, but we did not observe major differences in the significant associations. With K=1, LFMM identifies a few additional significant SNPs compared to

K=2 and includes all those detected with K=2. Therefore, we chose to discuss on the results for K=1 only. It must be noted that the distribution of p-values did not allow us to apply FDR on LFMM results. Therefore, we opted for a Bonferroni correction (0.05/number of models computed).

A total of 45 significant models are detected by LFMM (Table 6.11). All variables significantly detected at least one SNP, but their frequency of occurrence differs from Samβada results. Indeed, genotypes associated with precipitation in January (prec 1) are by far the most frequent in LFMM models and those with bio 15 the least frequent (Figure 6.31). Two SNPs are detected by two different variables, one with bio15 and SRTM, the other with bio7 and Ti216. We can also observe that the distribution per chromosome is different from Samβada's results. Here, no chromosome has much more detected SNPs than another.

*Table 6.9 Summary of models of association between SNPs and a subset of variables (bio15, bio 7, prec 7, TI216 180m, SRTM 1440m) with LFMM in **goats***

	# of significant models	# of models computed
Number of models	45	1.7 million x 5
Number of SNPs	43	1.7 million
Number of variables	5	5

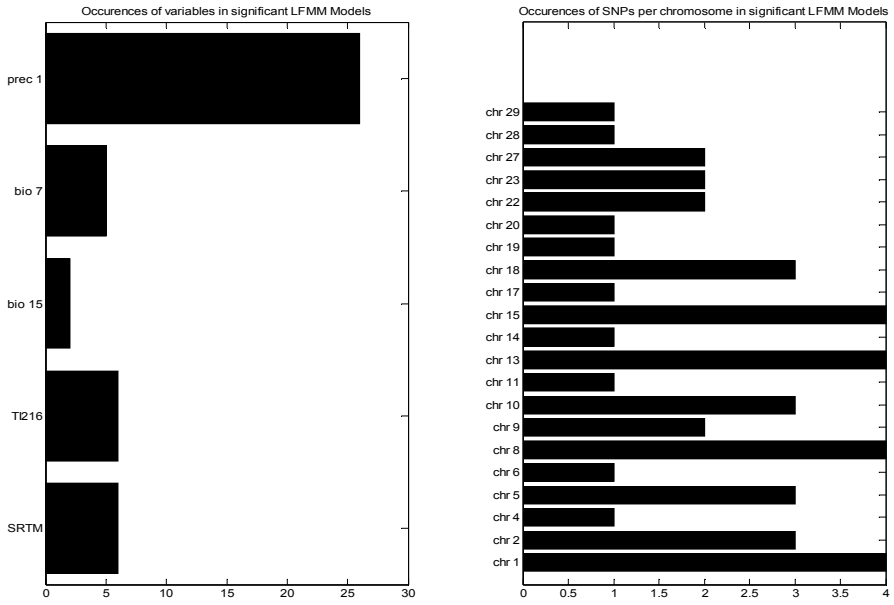


*Figure 6.31 Distribution of the frequency of selected environmental variables and of SNPs per chromosome involved in significant models from LFMM for **goats**.*

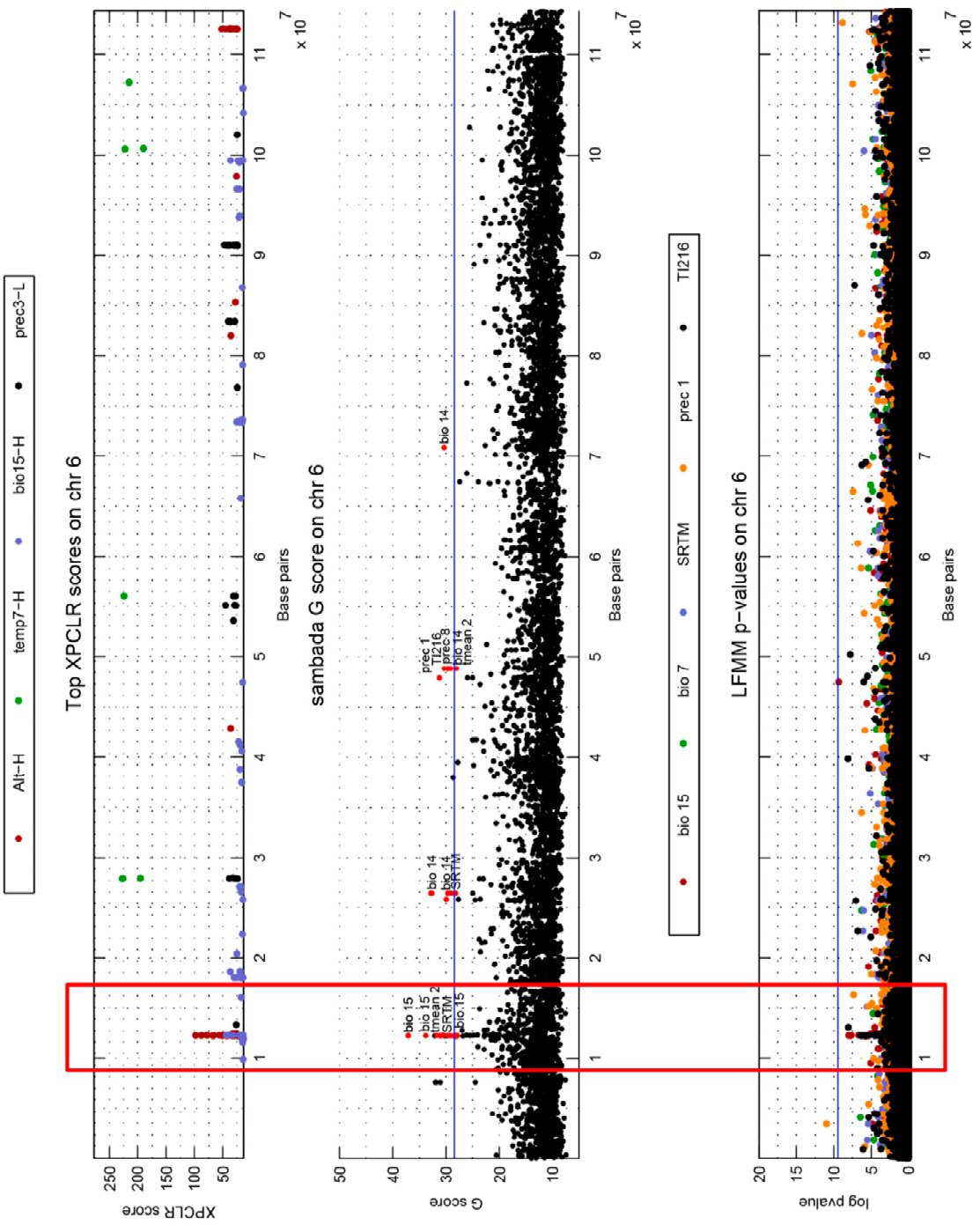


*Figure 6.32 Comparison between significant XP-CLR, Samβada and LFMM results according to their position on chromosome 6 for **goats**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Samβada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance. The peak is highlighted with a red frame.*

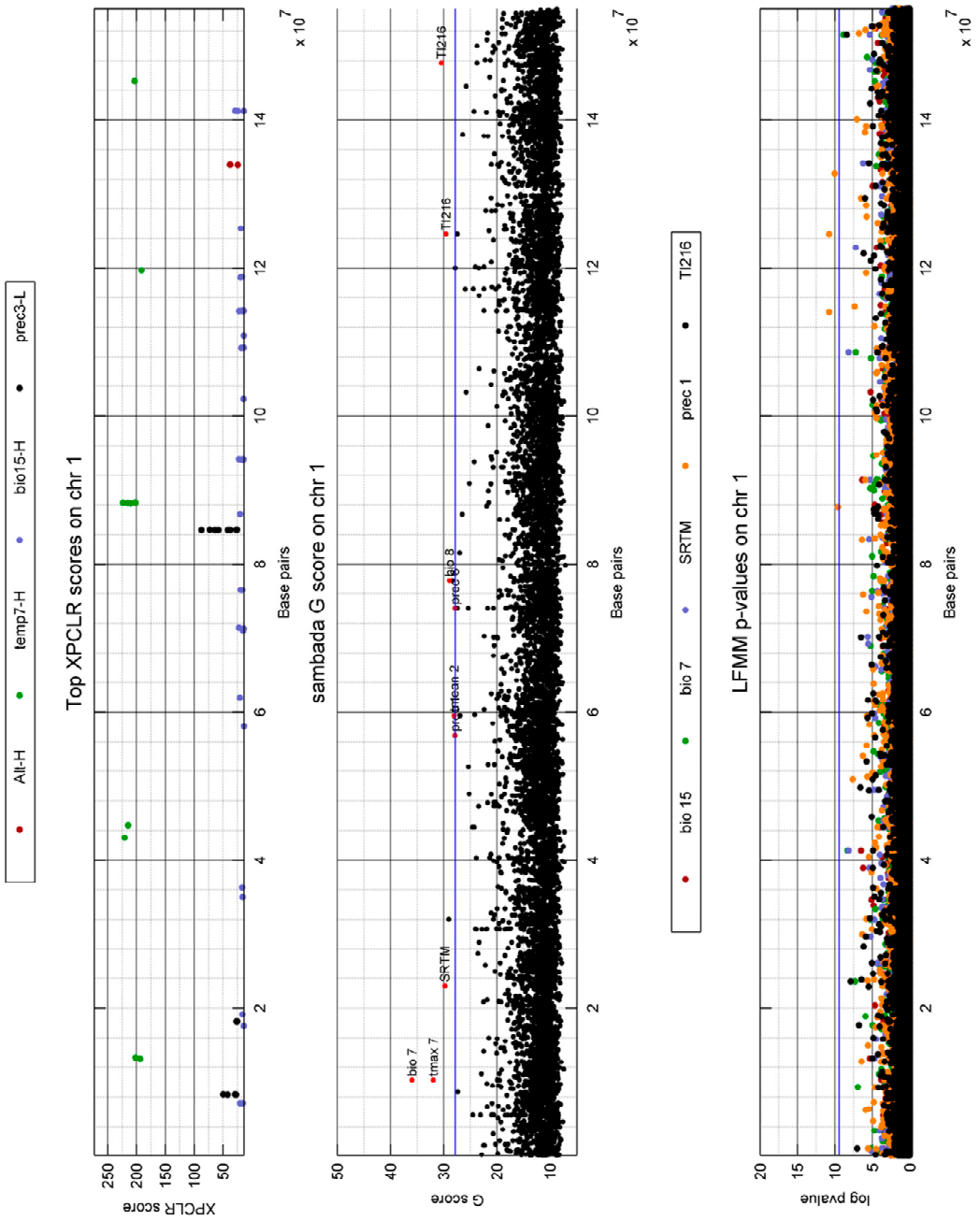


*Figure 6.33 Comparison between significant XP-CLR, Sambada and LFMM results according to their position on chromosome 1 for **goats**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Sambada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

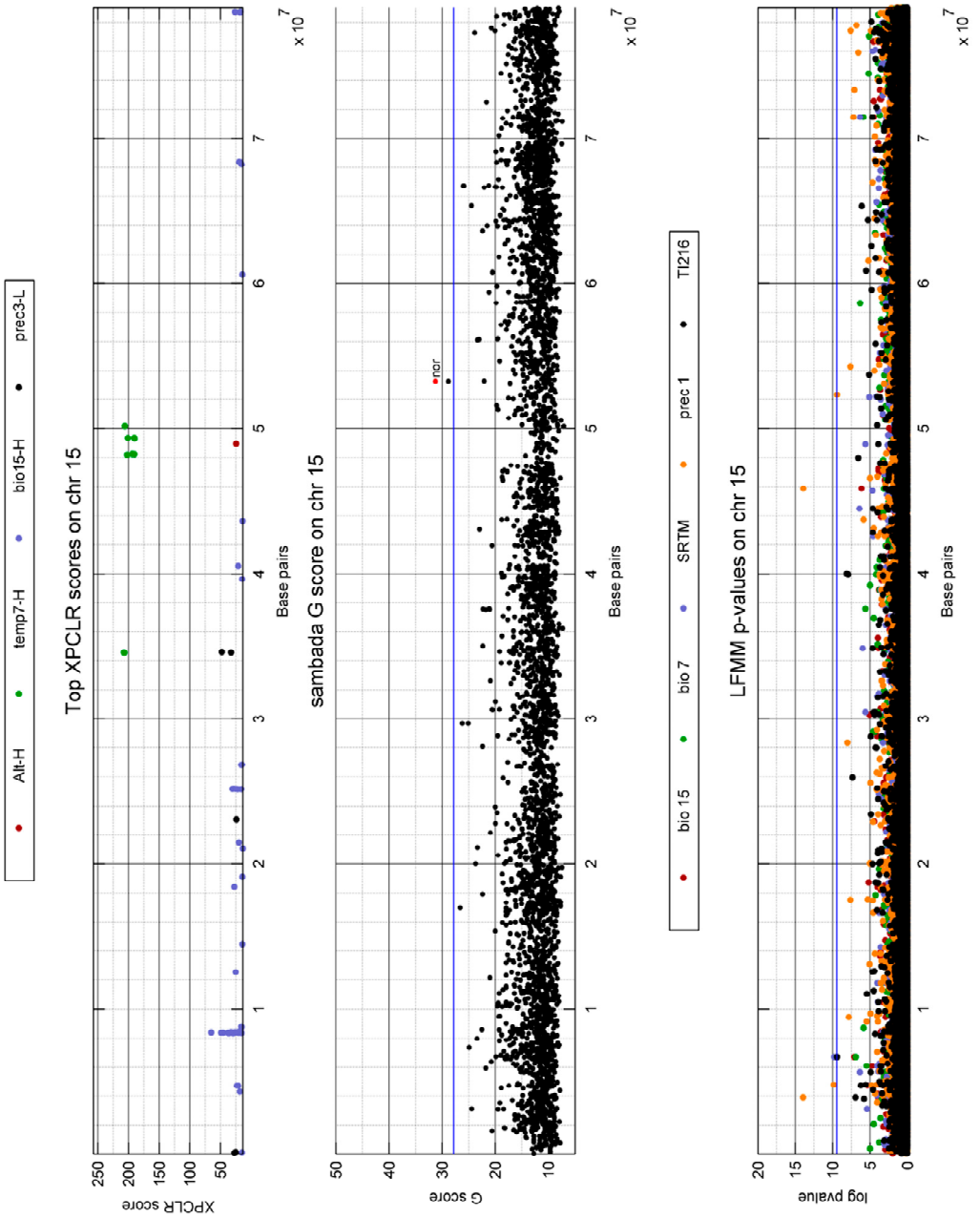


*Figure 6.34 Comparison between significant XP-CLR, Samβada and LFMM results according to their position on chromosome 15 for **goats**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Samβada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

SNPs on chromosome 6 are the most frequent in Samβada and are mainly associated with bio15 and altitude (Figure 6.32). Comparing the results according to their positions allows us to detect a peak on chromosome 6 in each method, even though it is not significant in LFMM. The peak in XP-CLR is clear when using populations based on altitude and bio 15. The latter one is also the most frequent variable in Samβada.

Chromosome 1, involving the second highest number of significant genotypes in Samβada, shows a different pattern (Figure 6.33). Here, Samβada's significant genotypes are scattered and do not correspond to high scores shown by the other methods. In fact, none of Samβada significant results correspond to low p-values in LFMM or to significant scores in XP-CLR. Chromosome 1 also shows a high amount of significant SNPs in LFMM but they do not match with the other methods.

Similarly, two SNPs from chromosome 15 (Figure 6.34) are involved in the most significant models in LFMM (both with prec 1) but we do not detect any match with the other methods. Other graphs per chromosome are provided in Appendix II.c.

At this point, a few general observations can be made. First, XP-CLR detects much more significant regions than both Samβada and LFMM. The only matches are found with Samβada in the peak on chromosome 6 (12 SNPs matching). On the other hand, there is no match between LFMM and Samβada or between LFMM and XP-CLR. Although LFMM detected the peak, its results are more different from Samβada and XP-CLR than these two against each other. In Table 6.10, we show the significant results from Samβada and we can observe that most of their LFMM counterparts identify the same variable as Samβada, at least those with low p-values.

*Table 6.10 Samβada's significant results involving multi-resolution variables in **goats**. The table shows one model per line including the genotype, the associated environmental variable, the resolution of the variable at which the highest G score was found (Best resolution), the highest G and Wald scores of the model involving the variable at the best resolution, the frequency of the genotype, the most significant LFMM p-value of the corresponding SNP and its associated variable, the highest XP-CLR score and its corresponding variable, the Moran's I of the genotype obtained with a neighbourhood of 20 individuals.. Models are ranked according to their G score. SNPs identified in the peak on chr 6 are in bold.*

	Genotypes	Variable	Best resolution	Highest G	Highest Wald	Frequency	Highest LFMM	LFMM variable	Highest XP-CLR	XP-CLR variable	Moran
Fig 6.37	**6:12259667_AA**	**bio_15**	**initial**	**37.11**	**25.92**	**0.286**	**9.97E-09**	**bio 15**	**98.04**	**temp7-H**	**0.16**
Fig 6.40	**6:12254244_AA**	**bio_15**	**initial**	**37.11**	**25.92**	**0.286**	**1.64E-08**	**bio 15**	**87.61**	**temp7-H**	**0.16**
Fig 6.41	4:95035251_AG	bio_15	ws01	36.46	22.23	0.143	0.05025	bio 15			0.16
	24:19436980_GG	bio_15	ws16	36.26	26.44	0.236					0.17
	1:10309616_AA	bio_7	initial	35.94	12.48	0.056	0.00388	bio 7			0.11
Fig 6.39	**6:12242353_GG**	**bio_15**	**ws16**	**33.82**	**25.57**	**0.381**	**2.11E-08**	**bio 15**	**66.69**	**temp7-H**	**0.15**
	13:74456761_AA	bio_7	ws08	33.52	19.37	0.093	0.00482	bio 7			0.08
Fig 6.42	3:104936887_GG	TI2112	0360	32.97	23.34	0.369	1.42E-01	bio 7			0.11
	6:26455416_GG	**SRTM**	**0720**	**32.85**	**19.90**	**0.161**	**0.0132**	**SRTM**			**0.05**
	6:26455416_GG	**bio_14**	**ws16**	**32.72**	**13.87**	**0.161**	**0.0132**	**SRTM**			**0.05**
	4:25093566_GG	prec_1	initial	32.52	22.47	0.118	2.91E-03	prec 1			0.27
	11:15823825_AG	bio_7	ws08	32.37	24.86	0.211	0.01227	bio 7			0.12

	1:10309616_AA	tmax_7	initial	31.94	13.25	0.056	3.88E-03	bio 7			0.11
	6:12254244_GG	**bio_15**	**ws08**	**31.71**	**25.26**	**0.435**	**1.64E-08**	**bio 15**	**87.61**	**temp7-H**	**0.13**
	15:53255703_GG	nor	0180	31.26	23.74	0.230	1.02E-01	bio 7			0.02
	6:47914533_AA	prec_1	initial	31.25	15.85	0.118	0.00191	prec 1			0.23
	22:42794456_GG	SRTM	1439	31.20	20.89	0.199	0.00659	SRTM			0.07
	6:12254244_GG	**SRTM**	**2879**	**31.15**	**24.77**	**0.435**	**1.64E-08**	**bio 15**	**87.61**	**temp7-H**	**0.13**
Fig 6.38	**6:12218302_GG**	**SRTM**	**2879**	**31.08**	**24.68**	**0.388**	**1.18E-06**	**bio 15**	**23.6**	**Alt-H**	**0.15**
	4:51418120_GG	prec_1	ws16	31.06	21.99	0.112	0.00894	prec 1			0.33
	26:3818155_GG	bio_8	ws01	30.78	21.40	0.106					0.07
	6:12254244_GG	**prec_8**	**initial**	**30.75**	**25.24**	**0.435**	**1.64E-08**	**bio 15**	**87.61**	**temp7-H**	**0.13**
	4:95035251_AG	prec_8	ws04	30.68	22.09	0.143	0.05025	bio 15			0.16
	20:4481114_GG	bio_7	ws04	30.65	19.58	0.099	0.03437	bio 7			0.09
	6:12276649_AA	**bio_15**	**initial**	**30.51**	**22.74**	**0.280**	**2.08E-07**	**bio 15**			**0.15**
	1:147737581_GG	TI216	0180	30.41	23.31	0.481	2.98E-05	bio 7			0.14
	6:48842226_GG	TI216	0180	30.35	23.83	0.342	0.00294	Ti216			0.10
	6:70862842_GG	bio_14	initial	30.34	18.16	0.296	0.03976	bio 15			0.09
	22:41805444_AA	Slo	0180	30.31	7.72	0.082	2.14E-02	prec 1			-0.01
	6:25849772_AG	SRTM	initial	29.94	15.88	0.112	0.05414	bio 7			0.07
	12:17515212_AG	bio_9	initial	29.86	18.54	0.344	4.64E-01	SRTM			0.06
	1:23002164_AA	SRTM	1439	29.78	16.11	0.120	2.96E-04	bio 7			0.12
	11:15823825_GG	bio_7	ws16	29.69	23.57	0.224	1.23E-02	bio 7			0.10
	6:26455416_GG	tmean_2	ws02	29.68	19.63	0.161	0.0132	SRTM			0.05
	13:78939678_AG	PlCu	initial	29.68	13.96	0.118	0.69111	bio 15			0.03
	6:48842226_GG	prec_8	initial	29.65	24.70	0.342	0.00294	Ti216			0.10
	9:38675107_GG	TI216	0180	29.58	23.26	0.422	1.76E-05	bio 7			0.21
	9:38668915_GG	TI216	0180	29.58	23.26	0.422	1.45E-05	bio 7			0.21
	24:28807953_AA	bio_15	initial	29.55	18.61	0.125					0.14
	1:124570528_AG	TI216	initial	29.55	19.38	0.199	5.64E-02	Ti216			0.04
	8:46850695_AG	Slo	initial	29.51	16.49	0.415	0.34179	Ti216			0.00
	6:12259667_GG	**bio_15**	**ws08**	**29.46**	**23.84**	**0.435**	**9.97E-09**	**bio 15**	**98.04**	**temp7-H**	**0.10**
	14:26405742_GG	bio_9	ws16	29.44	22.08	0.323	2.36E-01	SRTM			0.03
	25:6830009_AG	tmean_2	initial	29.30	23.66	0.491	0.08164	Ti216			0.05
	9:55810891_GG	bio_14	ws16	29.19	24.23	0.261	0.00815	Ti216			0.12
	6:12218302_AA	**bio_15**	**ws16**	**29.15**	**22.31**	**0.319**	**1.18E-06**	**bio 15**	**23.6**	**Alt-H**	**0.14**
	6:48842226_GG	bio_14	ws08	29.09	23.48	0.342	2.94E-03	Ti216			0.10
	14:45215445_GG	prec_1	ws16	29.05	21.41	0.112	0.00732	prec 1			0.20
	6:26455416_AG	bio_14	ws16	29.04	12.05	0.137	0.0132	SRTM			0.05
	28:40372626_GG	SRTM	1439	28.98	22.37	0.199	0.01514	SRTM			0.04
	1:77790195_AG	bio_8	ws02	28.91	21.67	0.472	4.87E-01	bio 7			0.09
	20:28161553_GG	tmax_10	ws01	28.89	22.12	0.335	5.50E-05	bio 15			0.15
	11:25042928_AG	tmin_8	ws16	28.85	18.15	0.100	0.2361	SRTM			0.10
	11:25030523_AG	tmin_8	ws16	28.70	18.28	0.106	5.90E-02	SRTM			0.10
	14:1335043_GG	prec_1	ws16	28.69	20.39	0.242	0.00412	prec 1			0.13
	24:19436980_AG	bio_15	ws16	28.69	22.10	0.217					0.11
	20:25965154_AG	nor	0180	28.61	23.36	0.344	0.0568	bio 7			0.02
	2:133961081_GG	tmax_7	initial	28.53	21.88	0.425	0.04215	bio 7			0.12
	6:12207826_GG	**SRTM**	**2879**	**28.53**	**23.15**	**0.379**	**1.02E-06**	**SRTM**			**0.15**
	2:133961081_GG	tmean_6	initial	28.49	21.20	0.425	4.21E-02	bio 7			0.12
	4:95035251_AG	bio_14	ws04	28.44	23.28	0.143	0.05025	bio 15			0.16
	6:12254244_GG	**bio_14**	**ws08**	**28.44**	**22.02**	**0.435**	**1.64E-08**	**bio 15**	**87.61**	**temp7-H**	**0.13**
	4:95035251_GG	bio_15	ws01	28.44	20.24	0.155	0.05025	bio 15			0.13
	13:74456761_AA	tmax_7	initial	28.43	16.53	0.093	4.82E-03	bio 7			0.08
	9:14309947_GG	bio_7	ws16	28.43	18.95	0.099	0.04395	bio 7			0.08
	7:79661490_AA	TI216	initial	28.43	17.35	0.099	6.35E-04	Ti216			0.01
	9:38857907_GG	tmax_7	initial	28.36	22.08	0.354	1.05E-03	bio 7			0.16
	6:26455416_AG	SRTM	0720	28.20	17.10	0.137	1.32E-02	SRTM			0.05

	20:4481114_GG	tmax_7	initial	28.18	16.76	0.099	3.44E-02	bio 7			0.09
	6:12187316_GG	**bio_15**	**initial**	**28.08**	**20.10**	**0.193**	**7.61E-03**	**bio 15**			**0.23**
	1:59516318_GG	tmean_2	ws08	28.04	21.67	0.329	4.29E-03	SRTM			0.10
	6:48842226_GG	tmean_2	initial	28.03	22.94	0.342	2.94E-03	Ti216			0.10
	6:12259667_GG	**prec_8**	**initial**	**28.02**	**23.43**	**0.435**	**9.97E-09**	**bio 15**	**98.04**	**temp7-H**	**0.10**
	7:54532252_AA	prec_8	initial	28.01	15.79	0.168	1.49E-02	SRTM			0.12
	3:36013665_AG	tmean_6	ws08	27.99	17.18	0.106	1.91E-01	prec 1			0.00
	1:56842826_AA	prec_1	ws02	27.92	20.50	0.063	1.74E-05	prec 1			0.36
	1:74038394_AA	prec_6	ws01	27.91	12.32	0.081	0.00794	bio 15			0.13
	11:25021331_AG	tmin_8	ws16	27.90	17.86	0.099	0.3131	SRTM			0.08
	6:12259667_GG	**SRTM**	**2879**	**27.88**	**22.73**	**0.435**	**9.97E-09**	**bio 15**	**98.04**	**temp7-H**	**0.10**
	17:4322848_GG	tmax_10	ws04	27.83	21.65	0.313	0.05285	SRTM			0.06
	28:40372626_AG	SRTM	1439	27.82	21.66	0.193	0.01514	SRTM			0.03

*Table 6.11 LFMM's significant results for univariate models involving a subset of environmental variables in **goats**. The table shows one model per line including the locus, the associated variable, its z score and p-value, the minor allele frequency, the mean and maximum G score between the three genotypes in Samβada for the same environmental variable, the maximum Moran's' I between the three genotypes and the distance at which it was found. Models are ranked according to their P-value.*

	SNPs	Variable	Zscore	P value	MAF	Mean G score	Max G score	Max Moran	Distance of max Moran
Fig 6.43	4:5923331	prec_1	9.18677	2.03E-16	0.15484	1.21892	2.127	0.039	1
Fig 6.45	17:2263702	TI216	-9.14991	2.54E-16	0.08497	0.22179	0.7185	0.019	1
Fig 6.44	15:3909036	prec_1	8.51769	1.13E-14	0.10191	2.02485	3.1125	0.042	4
	15:45878555	prec_1	8.50812	1.20E-14	0.25806	2.22142	3.8869	0.008	1
	10:29730546	prec_1	8.14834	9.99E-14	0.09615	4.55115	7.4974	0.019	6
	27:42551769	prec_1	8.00561	2.29E-13	0.09615	0.55146	1.4409	0.024	1
	13:43643932	prec_1	7.85768	5.39E-13	0.01948	4.0131	6.929	0.060	1
	9:88900192	bio_7	-7.72437	1.16E-12	0.0719	0.10504	0.2	0.069	3
	19:16896720	TI216	-7.64165	1.86E-12	0.06494	0.13327	0.3641	0.023	1
	11:97481069	prec_1	7.57495	2.71E-12	0.01911	2.54324	3.7031	0.019	3
	8:1761930	prec_1	7.55728	3.00E-12	0.23077	2.4394	5.7013	0.053	1
	10:80063740	prec_1	7.45367	5.38E-12	0.15385	1.09282	2.7444	0.023	2
	6:3446921	prec_1	7.33412	1.05E-11	0.13725	0.42352	0.8642	0.004	2
	5:71106895	SRTM	7.30185	1.26E-11	0.05195	0.62947	1.2034	0.013	1
	5:89094139	prec_1	7.29644	1.30E-11	0.24026	1.66023	2.9687	0.004	5
	2:5663077	SRTM	7.25431	1.64E-11	0.21019	7.37206	15.611	0.053	2
	1:114098068	prec_1	7.25062	1.67E-11	0.08861	0.85468	1.5454	0.010	2
	1:124546289	prec_1	7.2492	1.69E-11	0.23529	2.48821	5.0381	0.027	1
	2:4657541	bio_15	7.23758	1.80E-11	0.24183	1.5616	2.6398	0.012	1
	10:33822403	SRTM	7.18854	2.36E-11	0.20915	1.55	3.7999	0.022	4
	13:34211339	prec_1	7.15316	2.87E-11	0.04487	0.82829	1.4438	0.009	3
	23:24525923	TI216	-7.13568	3.16E-11	0.2013	8.48158	18.956	0.111	1
	2:58995772	bio_7	7.1291	3.28E-11	0.13725	1.88712	4.5233	0.045	3
	13:4414264	prec_1	7.10113	3.83E-11	0.28205	0.79441	1.8274	0.023	3

18:59196467	prec_1	7.05767	4.86E-11	0.2129	2.26856	4.473	0.021	1
22:35260639	TI216	-7.0555	4.92E-11	0.32692	0.25217	0.8965	-0.005	8
18:9500578	prec_1	7.02113	5.94E-11	0.2013	1.77365	3.1243	0.065	1
13:48099783	prec_1	7.02103	5.94E-11	0.12903	0.98941	1.6708	0.076	1
28:5541929	prec_1	6.98978	7.05E-11	0.16993	0.39813	0.844	-0.009	10
27:36189456	SRTM	-6.93667	9.41E-11	0.0719	2.24656	3.9621	0.017	3
1:132775653	prec_1	6.93386	9.56E-11	0	0.81898	0.9129	0.030	1
14:12686959	bio_7	-6.88416	1.25E-10	0.30065	1.68659	2.8483	0.027	2
15:4780144	prec_1	6.86518	1.39E-10	0.25	1.05422	2.0146	0.009	1
8:35490226	bio_15	-6.83662	1.62E-10	0.01307	0.89036	2.3857	0.002	3
15:6707714	SRTM	6.83371	1.64E-10	0.03922	0.09075	0.2526	0.001	3
9:69563755	prec_1	6.83314	1.65E-10	0.20915	2.17063	3.41	0.046	1
22:55984785	prec_1	6.8297	1.68E-10	0.22581	0.05967	0.13	0.005	4
18:41102518	prec_1	6.81835	1.79E-10	0.26144	0.03289	0.0731	-0.001	4
8:35490226	SRTM	6.80688	1.90E-10	0.01307	1.55512	3.2111	0.002	3
29:1957547	prec_1	6.80002	1.97E-10	0.07643	0.35467	0.9722	0.012	2
8:110668946	TI216	6.763	2.41E-10	0.25	3.61157	9.3634	0.021	1
20:65900817	TI216	-6.75505	2.51E-10	0	0.0983	0.3816	-0.007	7
1:87669244	prec_1	6.74881	2.60E-10	0.20645	0.70403	1.5134	0.012	7
23:24525923	bio_7	-6.73395	2.81E-10	0.2013	3.91396	7.4263	0.111	1
5:94548328	bio_7	-6.72966	2.88E-10	0.08387	1.03963	2.3933	0.037	1

Figure 6.35 shows that there is no correlation between Samβada's G score and LFMM's p-value. Significant SNPs detected in both methods are clearly distinct. However, we note that SNPs detected by Samβada and XP-CLR have a higher significance in LFMM than those detected by Samβada only.

Regarding spatial autocorrelation, most of the significant genotypes that correlated with climatic variables show a moderate Moran I. This is not the case for genotypes associated with DEM variables, in which spatial autocorrelation is lower. On the other hand, SNPs detected in LFMM systematically show low Moran's I (Figure 6.36).

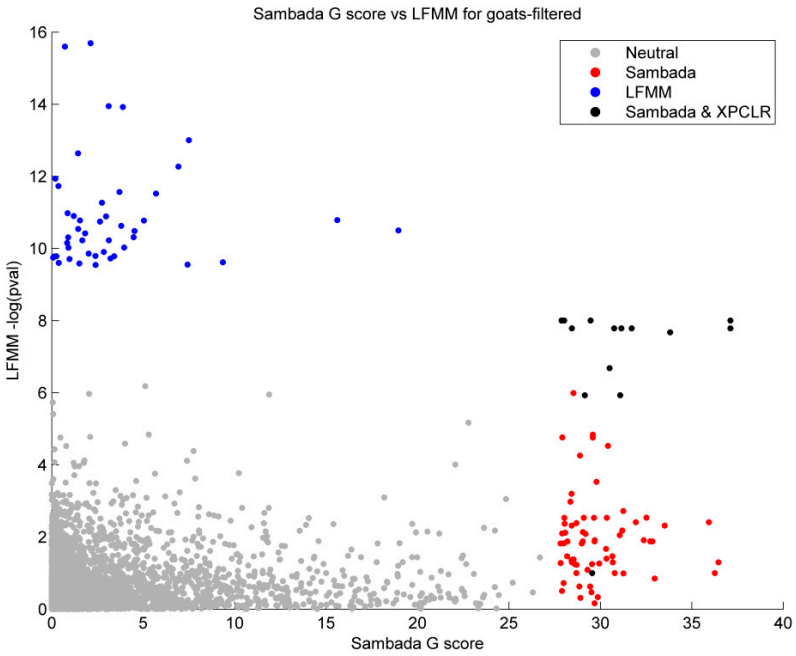


*Figure 6.35 Scatterplot of Sambada's G score against LFMM's p-value in **goats**. Neutral loci are shown in grey, those detected by Sambada in red, those by Sambada and XP-CLR in black and those by LFMM in blue.*

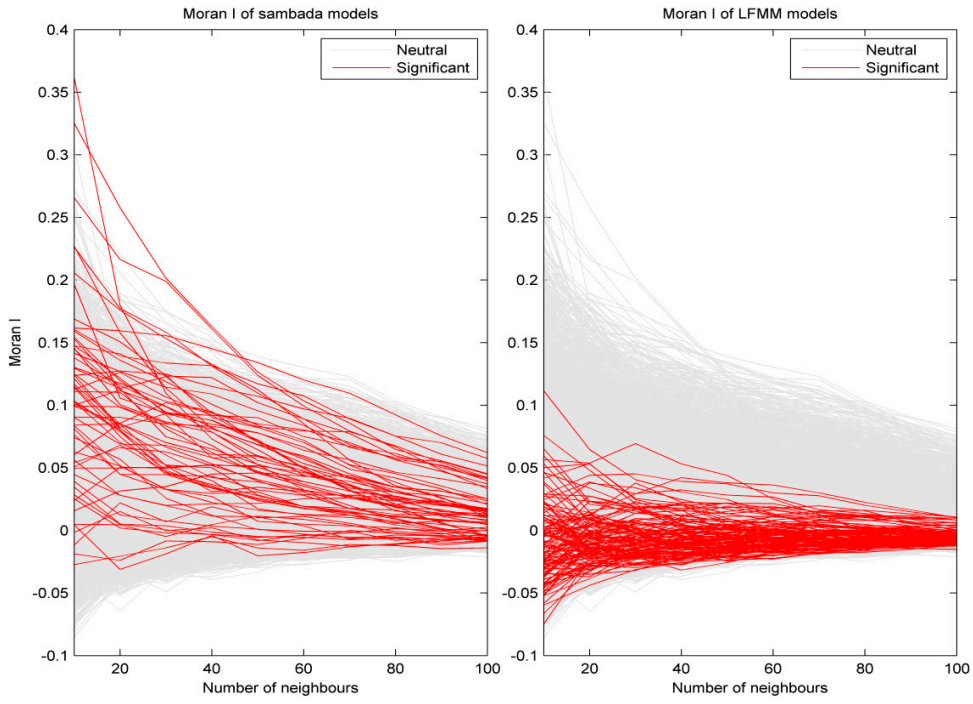


*Figure 6.36 Moran's I correlograms for **goats**. Neutral genotypes are shown in grey and genotypes associated with environmental variables, either by Sambada (left) or LFMM (right) are in red.*

6.4.3 Visualisation of significant associations

Details on the different parts of these graphs can be found in section 3.6 (p65).

Figure 6.37 to Figure 6.42 are representative of some of the most significant models in Samβada. We first note that most models involving climatic variables do not evidence an influence of window size on Samβada's scores. In addition, models that became significant by increasing window size remain poorly significant. For genotypes associated with DEM-derived variables, however, multi-scale variables always plays a major role in the final scores but we cannot identify a general trend. This is well illustrated in Figure 6.42, where the model involving the original resolution is not significant but those involving the variable at 180 and 360m are significant.

From Figure 6.37 to Figure 6.40, we show several SNPs identified in the peak on chromosome 6 and associated with bio15. They all show a similar pattern of spatial distribution. For example, genotype 6 12259667 AA is involved in the most significant model in Samβada (Figure 6.37). Its Moran's I is high compared with other loci but the LISA coefficients are rarely significant.

Genotype 4 95035251 AG is also one of the most significant genotype in Samβada's models and is associated with bio 15 (Figure 6.41). Here as well, LISA coefficient are not significant even though Moran's I is high. However, we can visually identify a cluster of AG genotypes in the northeast, which well fits the low values of bio 15.

Figure 6.42 shows the most significant genotype associated with a DEM-derived variables. This SNP is only associated with total insolation in December, with a best score at a resolution of 360m. We notice that different spatial resolutions do not generate sharp differences in Samβada's scores.

Significant models identified by Samβada

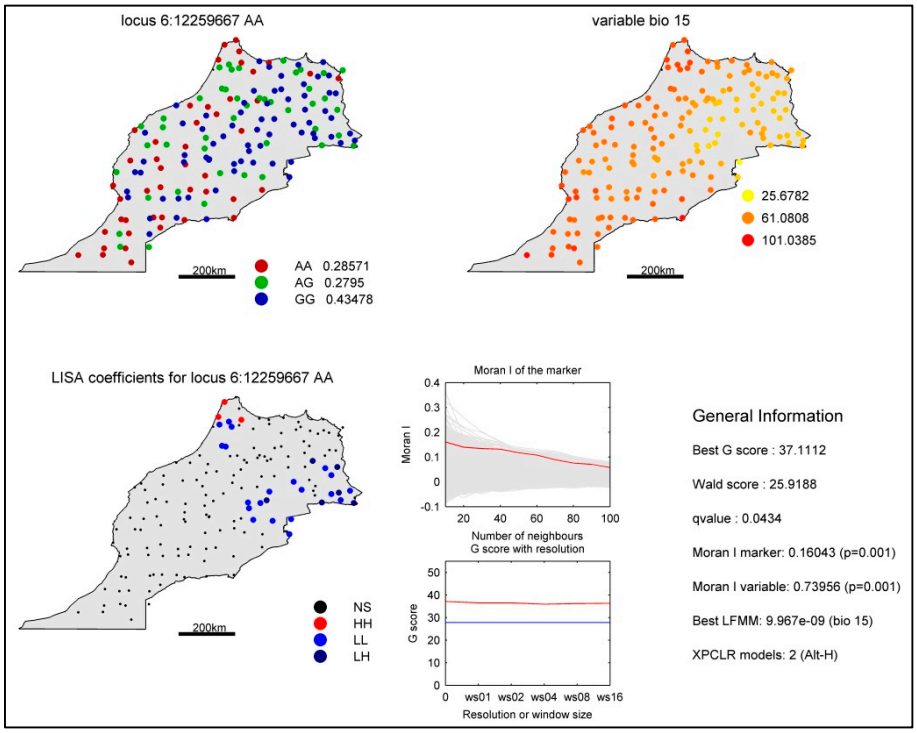


*Figure 6.37 Visualisation of the significant association between genotype 6 12259667 AA and Precipitation Seasonality (bio15) in **goats** detected by Samβada.*

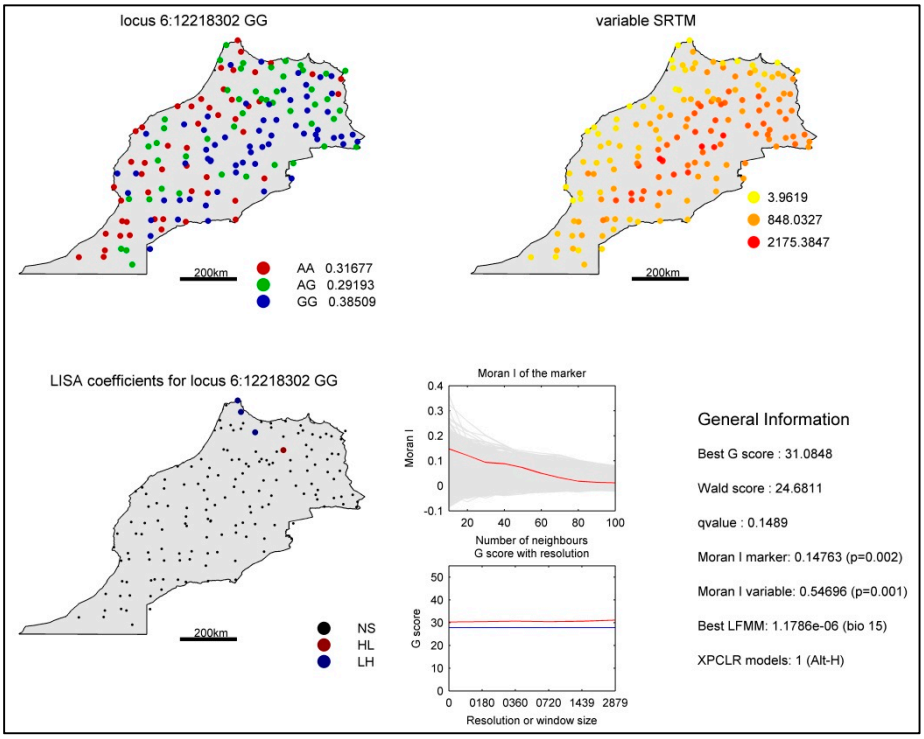


*Figure 6.38 Visualisation of the significant association between genotype 6 GG and altitude (SRTM) in **goats** detected by Samβada.*

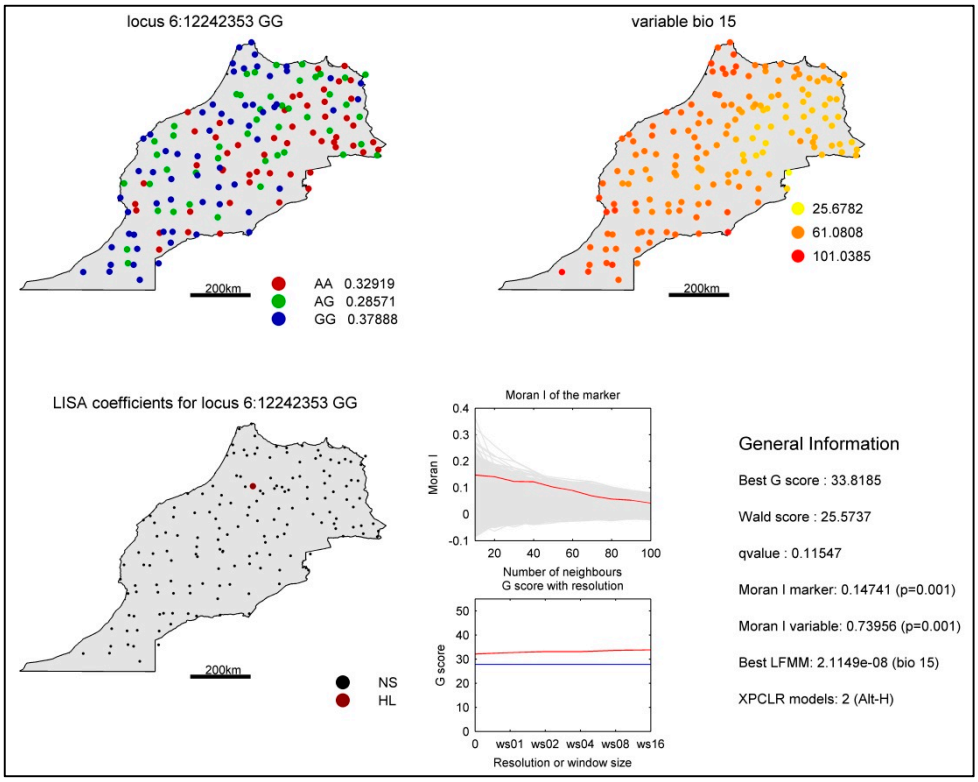


*Figure 6.39 Visualisation of the significant association between genotype 6 GG and Precipitation Seasonality (bio15) in **goats** detected by SamBada.*

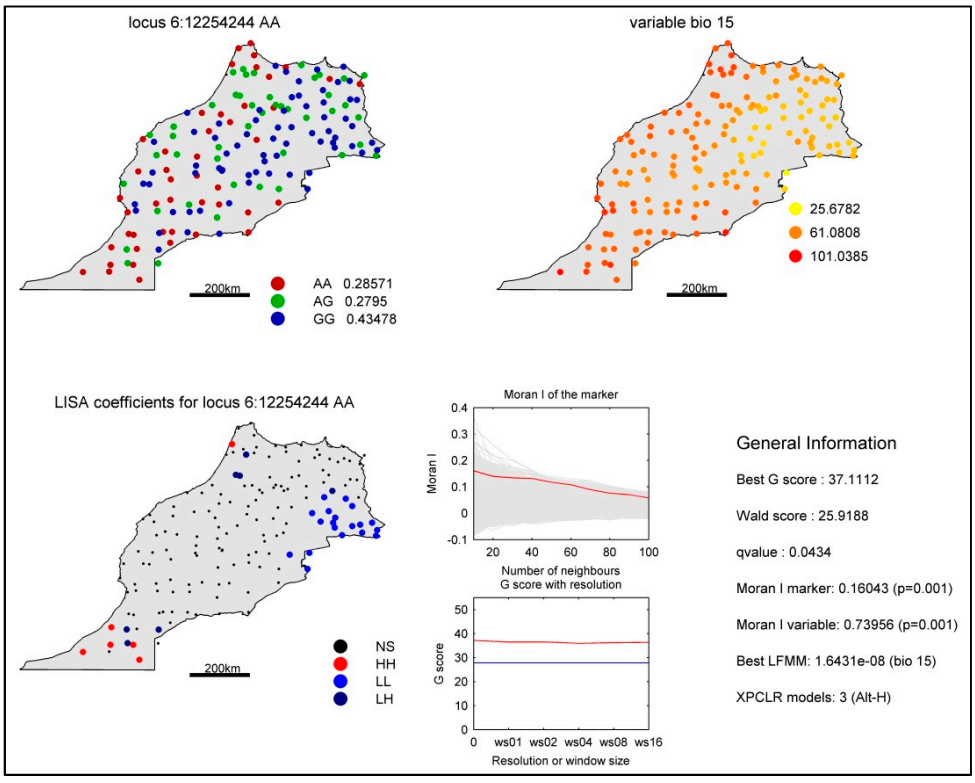


*Figure 6.40 Visualisation of the significant association between genotype 6 12254244 AA and Precipitation Seasonality (bio15) in **goats** detected by SamBada.*

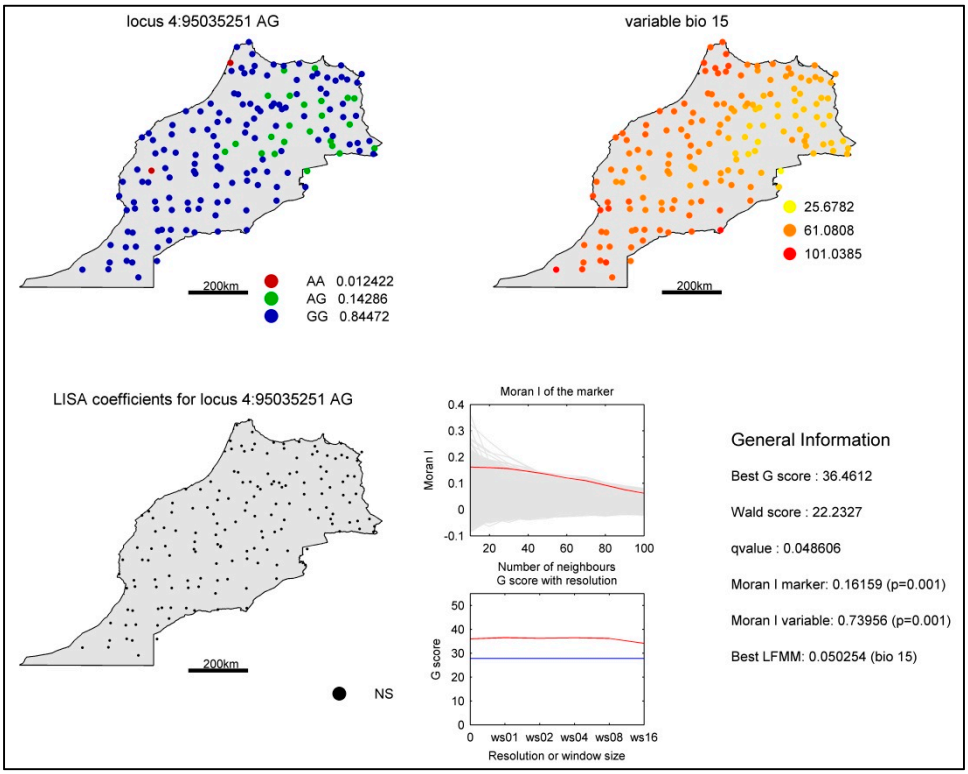


*Figure 6.41 Visualisation of the significant association between genotype 4 95035251 AG and Precipitation Seasonality (bio15) in **goats** detected by Sam*β*ada.*

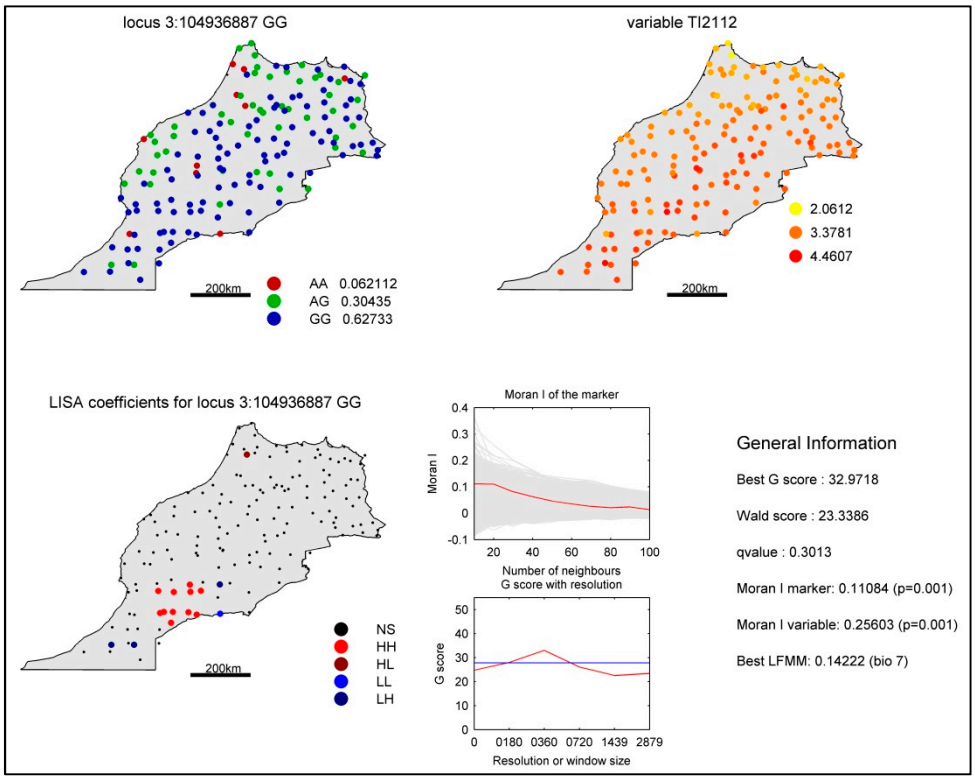


*Figure 6.42 Visualisation of the significant association between genotype 3:104936887 GG and Total Insolation on the 21 of December (Ti2112) in **goats** detected by Sam*β*ada.*

Significant models identified by LFMM

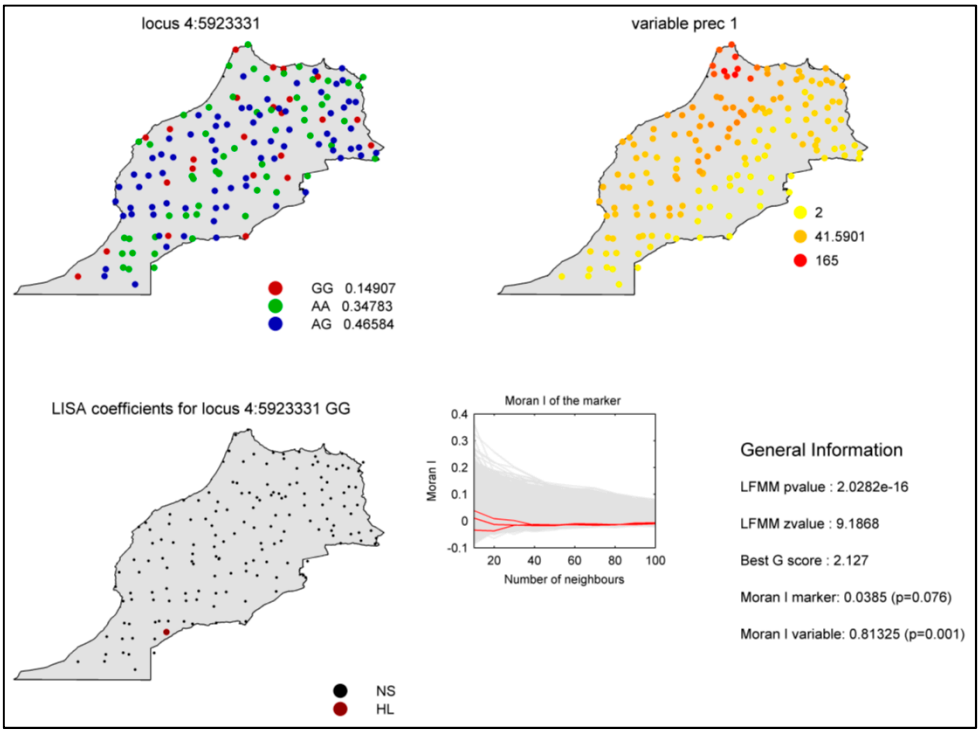


*Figure 6.43 Visualisation of the significant association between SNP 4:5923331 and precipitation in January in **goats** detected by LFMM.*

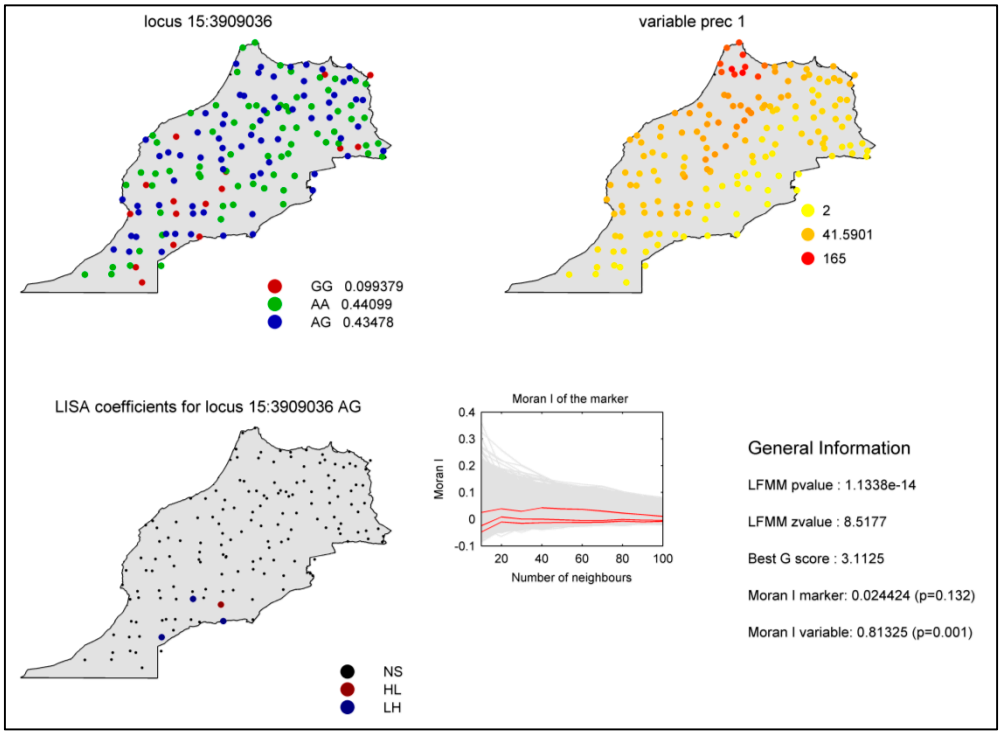


*Figure 6.44 Visualisation of the significant association between SNP 15:3909036 and precipitation in January in **goats** detected by LFMM.*

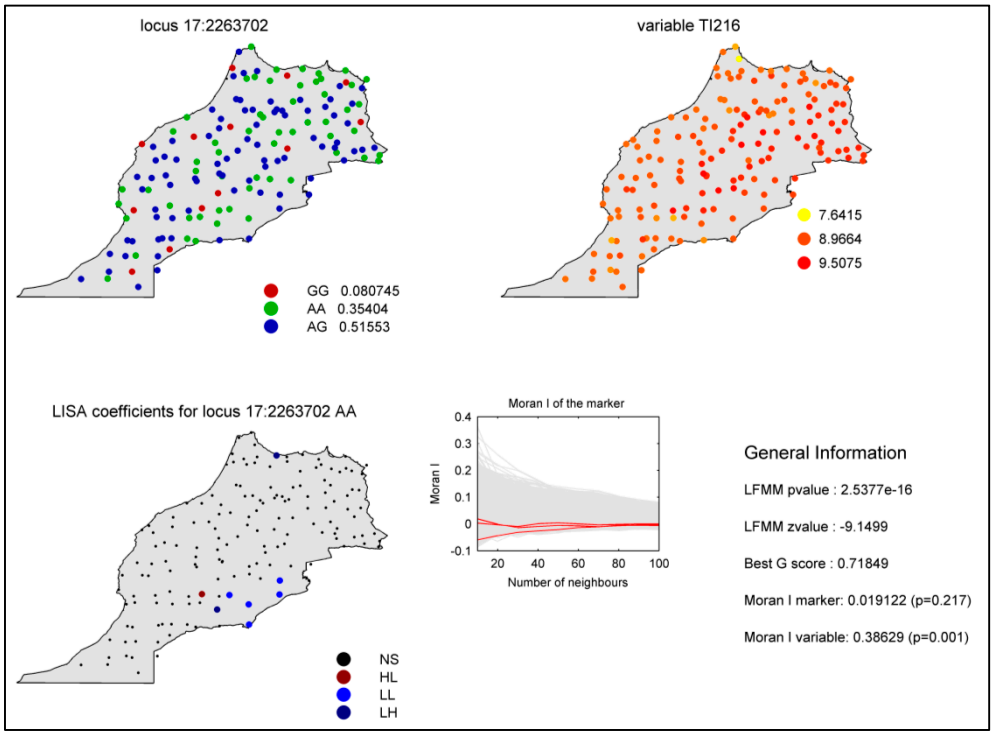


*Figure 6.45 Visualisation of the significant association between SNP 17:2263702 and Total insolation on the 21 of June (Ti216) in **goats** detected by LFMM.*

6.5 Discussion

In this chapter, we analysed the correlation between genetic markers in sheep and goats versus DEM-derived as well as climatic variables to detect consistent and robust signatures of selection. In addition, with the help of WGS data, we expected to identify annotated genes, or highlight unknown genes, that would support some of these signals of selection.

Despite the high amount of genotypic variation covering their entire genome, both sheep and goats show a limited amount of significant signatures of adaptation and weak scores in Samβada & LFMM. From these results, it is obvious that adaptation traits, if present, are discreet in these domesticated mammals.

Nevertheless, we found one remarkable result in each species with candidate regions in the form of “peaks” of selection obtained on chromosome 23 in sheep (Figure 6.14) and on chromosome 6 in goats (Figure 6.32). These peaks are clearly visible in each method, and represent a remarkable signature of selection. Despite SNPs in these peaks were not significant in LFMM, the peaks are clearly visible in this method and, thus, increase the robustness of these signatures. Furthermore, the same variables were involved in the detected models in each method (bio 15 for goats, prec 4 and tmax 4 for sheep) (see Table 6.5 and Table 6.10). Furthermore, the seven SNPs detected by Samβada in this peak on chromosome 6 of goats show a similar spatial distribution and a spatial

autocorrelation above 0.1 (see Figure 6.19 and Figure 6.37 for representative examples). The spatial distribution of bio 15, prec 8 and SRTM are similar, explaining why these variables are sometimes significant for the same genotypes. In sheep also, eight significant SNPs identified by Samβada in the peak on chromosome 23 have a similar spatial distribution (Figure 6.19) and a spatial autocorrelation above 0.2.

We decided to investigate these two peaks in more details and to look for proximal genes. We used the genome assembly of NCBI genetic databases, accessed at the following addresses:

- for sheep http://www.ensembl.org/Ovis_aries/Info/Index
- for goats http://www.ncbi.nlm.nih.gov/genome?term=capra%20hircus

In the peak identified on chromosome 23 in sheep (Figure 6.46), we identified five genes (LDLRAD4, FAM210, RNMT, MC1R and ACTHR) that were corresponding to the detections in Samβada and XP-CLR. Among the five SNPs identified by Samβada, three are located in the coding sequences of genes LDLRAD4, FAM210 and RNMT. Detections by XP-CLR, however, were located in or around genes RMNT, MC1R and ACTHR.

For these genes, we investigated their function and whether a causal relationship can be found with the associated environmental variable. We note that so far we haven't investigated the genomic regions in which other significant SNPs were detected. In addition, we will further compare genetic sequences of significant SNPs associated with the same variable between the two species. In fact, if two species have similar signatures of selection due to the same environmental variables, it increases the robustness of these signatures.

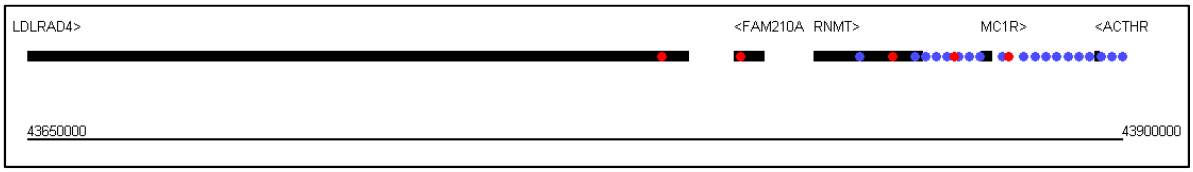


*Figure 6.46 Zoom on the peak encountered for **sheep** on chromosome 23 between positions 43 650 000 and 43 900 000. Black lines are showing the different genes in this window, red dots are the significant Samβada results and blue dots are the significant XP-CLR results. Arrows (< >) indicate if the gene is on the forward or reverse strand.*

The gene MC1R is the melanocortin-1 receptor. It binds to a class of pituitary peptide hormones known as the melanocortins located on the plasma membrane of melanocytes, which produce the pigment melanin (Benjelloun *et al.* 2015). MC1R is one of the key proteins involved in regulating mammalian coat and hair colour and has been investigated in many mammals (Fontanesi *et al.* 2009). As a result of its central situation in the peak identified and its importance in mammalian studies, we consider MC1R as a robust candidate for selection in Moroccan goats. In addition, significant SNPs detected in MC1R are associated to precipitation and maximum temperature in

April and the spatial distribution of these SNPs shows the prevalence of certain genotypes at the edge of the Sahara desert, where temperature is particularly high and almost no precipitation occur. MC1R polymorphism could thus express a different phenotypes of coat or hair colour due to desert conditions.

RNMT is an RNA (guanine-7-) methyltransferase enzyme with a methyltransferase domain and an N-terminal domain whose function is unclear (Aregger & Cowling 2013). It is conserved in mammals, but not required for cap methyltransferase activity. The RNMT N-terminal domain is required for transcript expression, translation and cell proliferation. We could not find a link between this function and the patterns of selection observed.

On the peak identified on chromosome 6 in goats (Figure 6.32 p147), incomplete genome annotation of goats assembly prevents us from identifying the gene detected (D4S2842) (Figure 6.47). Only one position, significant in XP-CLR, was identified in the gene. However, it is interesting to note that this peak was also identified as a signature of selection in the Black population in Benjelloun *et al.* (2015).

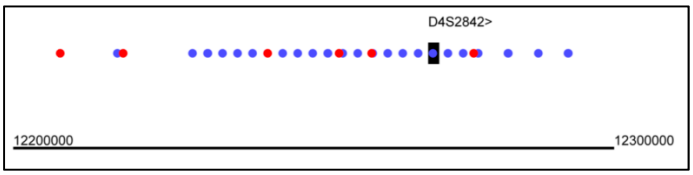


*Figure 6.47 Zoom on the peak encountered for **goats** on chromosome 6 between positions 12 200 000 12 300 000. Black line is showing the gene in this window, red dots are the significant Samβada results and blue dots are the significant XP-CLR results. Arrows (< >) indicate if the gene is on the forward or reverse strand.*

In both peaks, SNPs were mostly associated to precipitation variables. Furthermore, top candidate SNPs outside these peaks are mainly correlated with precipitation variables in both species and for both LFMM and Samβada. It demonstrates that precipitation in Morocco, which vary substantially from coast to desert, are likely to be the main driver of local adaptation to environment in both species. This could corroborate to the results from Benjelloun *et al.* (2015), who found several candidate genes giving insights on a possible adaptation to panting. On the other hand, we found that SNPs are rarely associated to DEM-derived variables and show weak significance with these variables. It indicates that, as expected, topographic variables are not relevant at such scale for a mobile species.

Multi-scale variables increased the amount of detected SNPs but the newly detected SNPs are moderately significant. Regarding climatic variables, window size has little effect on the significance of models. For DEM-derived variables, however, spatial resolution had a substantial influence on the goodness-of-fit but there is no general trend for a decrease or an increase of significance throughout the continuous representation of scales.

In the population structure of sheep (Figure 6.8), we can see that the strongest membership coefficients partially correspond to Sardi and D-man breeds (see the map of breeds in Figure 6.1). We note that there is a link with the peak on chromosome 6. In fact, there is a moderate spatial correspondence between the local cluster observed in all SNPs identified in this peak (Figure 6.19 to Figure 6.23) and the membership coefficient to population 2.

For goats, strongest coefficients to the second population correspond to the northern breed (Figure 6.2 and Figure 6.9). Using a subset of this dataset, Benjelloun *et al.* (2015) previously assessed the genetic diversity of three phenotypically distinct indigenous populations: the Black, Draa and Northern breeds (see Figure 6.2 p119 for their spatial distribution). They could not identify either a clear population structure between these 3 breeds using either mitochondrial haplotypes of WGS variants. They attribute this lack of population structure to uncontrolled breeding strategies in which extensive breeding systems favour high gene flow. We mention also that they found a very high genetic diversity (24 million variants) and a low linkage disequilibrium in between the three populations. In comparison the human 1000 genome project identified ≈15 million SNPs.

Whole genome sequence data also allowed us to apply a false discovery rate (FDR) on the p-values obtained in Samβada, which is by default implemented with a Bonferroni correction for multiple tests. Here we compared a threshold at 0.1 and 0.2, meaning that we assume 10 or 20% of false discoveries in the results. Both species should be distinguished at this point. In fact, for goats, a threshold of 0.2 increased considerably the number of detections while for sheep, most models with this threshold already showed q-values below 0.1. However, multi-scale variables do not improve the number of detected SNPs below the 0.1 threshold. In addition, Wald scores are low, often lower than the accepted threshold for the two other case studies, in which much less associations were computed. Furthermore, compared to the two previous case studies, Moran correlograms of detected loci are more spread in the neutral background and choosing a more conservative threshold would have led to a clearer distinction between neutral and selected loci. Therefore, a FDR of 0.1 might be a better option in these cases, particularly in sheep.

On the other hand, Samβada scores are low compared with those obtained by Stucki (2014) in Ugandan cattle. There are two major differences with these results. First, the population structure is weak or non-existing in our case while five populations were identified in Uganda. Second, we only had 160 individuals, compared with 804 in Uganda. These two parameters partially explain why our scores are so low and why we do not obtain common detections between LFMM and Samβada. In fact, both types of approaches depend largely on the number of individuals (Lotterhos & Whitlock 2015). However, it is surprising to notice the complete absence of common detections between the Samβada and LFMM, contrasting with the results from Stucki (2014) who identified many common SNPs. Here, LFMM identified SNPs at very different genomic positions. In addition, there is no visual evidence highlighting correlation when comparing the map of the SNP detected by LFMM and the map of its associated variable. The weak population structure and LFMM's loss of power under IBD models (Lotterhos & Whitlock 2015) might explain why it opposes to other methods so strongly in this case study. In fact, LFMM was developed to cope with the

effects of populations' structure, not the absence of it. We also note that we tried values of 1, 2 and 3 for K without causing any major change in the SNPs detected. Therefore, Samβada seem to be more appropriate than LFMM in this case.

To conclude this chapter, we found weak evidences of adaptation in sheep and goats but identified, in both cases, peaks of signatures of selection in their genome. Further investigation of these peaks indicated polymorphism in and nearby genes that were, for most of them, related to precipitation patterns. These results demonstrate the relevance of genome screening and point the way ahead for further functional research.

Chapter 7 General discussion and conclusion

In this thesis, we assessed the local adaptation to environment based on three case studies. To do so, it is important to define how local a study should be and how local we do expect adaptation to operate. However, defining the scale of a study (i.e. the grain and extent to be used) raises important questions regarding the relevance of environmental variables, particularly their spatial resolution. Therefore, our main goal was to evaluate how fine the resolution of environmental variables should be to detect local adaptation and how sensitive are detections of signatures of selection to the spatial resolution of environmental variables.

To tackle this problematic, we proposed a multi-scale landscape genomics framework to identify signatures of selection. Particularly, we focused on the relevance of Very High Resolution Digital Elevation Models (VHR DEMs) and the application of a multi-resolution analysis to detect adaptation at different scales. We explored three case studies that differ between each other on their scale (or extent), species, topography, genetic data and sampling scheme. To increase the robustness of these detections, we compared several methods of outlier detection.

Relevance of DEM-derived variables in a landscape genomic context

Our results validate two essential concerns regarding DEMs: i) multi-scale approaches are valuable when facing topographic heterogeneity, and ii) investigating a large diversity of DEM-derived variables is crucial in order to evaluate all topographic aspects that might influence climatic variability.

At a local scale, we were able to show that DEM-derived variables can be used as relevant surrogates for environmental variables and to better understand relationships with local topography. Indeed, physiological activity and adaptation of plants are affected by temperature, humidity and soil characteristics (Körner 2003; Böhner & Selige 2006; Manel *et al.* 2012b) and we found that signatures of selections related to topographic heterogeneity were substantial at the local scale of the *B. laevigata* case study. However, the relevance of DEM-derived variables at wider extents was null in a flat urban environment at a regional scale and weak for mobile mammal species at a large scale. These results show that DEM variables are relevant in a landscape genomics context and provide proxies to essential ecological conditions assessed in many geomorphological studies (Wilson & Gallant 2000). In addition, DEM variables widen the diversity of environmental variables available to offer a maximum of potential pressures of selection.

Furthermore, DEM-derived variables are easy-to-compute proxies for environmental features, are cheap, and involve limited fieldwork but good knowledge of Geographic Information Systems. DEM-derived variables should thus be widely used as proxies of environmental features in ecology and evolution (Kozak *et al.* 2008). In addition, open source GIS alternatives (e.g. SAGA GIS, Quantum GIS and GRASS) provide algorithms to process a variety of secondary terrain attributes. By scripting the computation of variables in RSAGA, we facilitated investigations of a large spectrum of variables at different scales. Finally, several DEMs with a global cover are freely available and the upcoming release of the TANDEM-X model with a resolution of ≈10m and a high accuracy holds great promises for future use of multi-scale DEM-derived variables (Krieger *et al.* 2007). As regards very high resolution, LIDAR represents the best DEM acquisition technology at the moment, providing great precision and high resolution across hardly accessible terrains, but still expensive (Xiaoye Liu 2008). Although they do not show the same level of precision like LIDAR, stereo-photogrammetry from Unmanned Aerial Vehicles (UAV) constitutes a less powerful but suitable and cheaper alternative subject to intense research (Leempoel & Joost 2012).

Despite these advantages, DEMs remain underexploited in landscape genetics. Although they are often used for altitude, primary attributes or solar radiation, usage of more complex variables like morphometric or hydrology indices is underestimated in the literature and we recommend to go beyond the traditional use of DEMs (Dobrowski 2011). Indeed, the present models demonstrated that DEM-derived variables, such as slope, eastness, vrm or twi, provide suitable or complementary surrogates to *in situ* measurements for characterization of plant habitat. We voluntarily assessed a very large panel of DEM variables, sometimes at the expense of redundancy, to make a point regarding variables diversity. However assessing the collinearity between variables is thus crucial, not only to avoid redundancy but also to improve ecological interpretation, particularly for multivariate models.

Furthermore, most DEM-derived variables used in this thesis were shown in the literature to be surrogates for many relevant ecological features (Wilson & Gallant 2000; Böhner & Selige 2006). However, while we demonstrated that DEM-derived variables can approximate climatic variables at a local scale, this relationship may not be valid in other places or with different spatial resolution. Indeed, it is important to keep in mind that relationships between properties of interest and terrain attributes cannot be transferred from one area to another with different characteristics (Wilson & Gallant 2000). While this does not question the relevance of topographic variables in a landscape genetics context, the ecological or geomorphological interpretation of DEM variables found in the literature only partially improves our understanding of the correlation between genetic markers and the DEM variable simply because we do not know if the ecological interpretation is valid in our cases. On the other hand, DEM variables obtained from structure tensors that we tested in these case studies also lack relationship with environmental features. In fact, these variables were first applied by Kalbermatten (2010) on a landslide to help visualising the different structures in the landscape, but he did not assess the relationship between these variables and ecological or geomorphological processes. Therefore, we cannot interpret the associations between genetic markers and structure tensor variables. For example, we expect that energy and coherency are proxies for terrain ruggedness, but they were not correlated with other terrain ruggedness variables. On the other hand, we found that orientation from structure tensors is less

noisy than aspect when compared at the same resolution. Nevertheless, the fact we detected significant associations with these variables is a strong argument favouring further investigation.

The spatial resolution of environmental data

By addressing to what extent a finer resolution provides ecologically relevant predictions, results from *B. laevigata* case study shed new light on these scale issues. While it is often expected that a higher precision should bring results that are more accurate, we demonstrated that the highest resolution available is rarely contributing to the most significant models. In addition, associations' strength heavily depended on the resolution of the variable. In fact, variations in the goodness-of-fit due to spatial resolutions indicates that multi-scale approaches should systematically be considered to model micro-climatic variables and association with genetic markers. We argue that using DEMs at their original grid resolution, without consideration of scale representativeness, likely leads to an underestimated role of topographic features in ecological models. Indeed, a too fine resolution may hold an excess of details and generate too much noise, while too coarse resolution would only show generalized properties of the landscape and lose explanatory power (Cavazzi *et al.* 2013). This issue is rarely addressed in the landscape genetics literature and most studies using DEMs at their original resolution often ended up with a minor contribution of topography (Zimmermann & Kienast 1999; Manel *et al.* 2010b; Vercauteren *et al.* 2012; Patsiou *et al.* 2014).

The main difficulty with multi-scale variables is to interpret the detected spatial resolution. For example, the most recurrent resolution of DEM variables involved in significant models in goat was 180m, but only a couple of multi-scale models involving DEM variables were detected. Usually, each DEM-derived variable is scale dependent in a different way: for example curvatures or hydrology variables are strongly dependent on the resolution (Wilson & Gallant 2000). We believe that using different resolutions allows calibrating the DEM variables and spurring them to fit climatic features, rather than seeking one optimal resolution. The difficulty, however, is to find an ecological feature that may correspond to the DEM-derived variable at that resolution, which is what we achieved in the first part of *B. laevigata* case study (Leempoel *et al. accepted*). In fact, the high sensitivity of terrain analysis to DEMs' spatial resolution questions even more the interpretability of significant associations. Indeed, we remind that relationships between terrain attributes and environmental parameters are only valid at the scale for which they have been derived (Wilson & Gallant 2000).

Another important point regarding spatial resolution is the presence of pseudo-replicates when the resolution is not appropriate for the sampling design. Pseudo-replication occurs when the number of samples are treated inappropriately as independent replicates. It means that observations may not be independent if the observations are correlated in space. For example, the sampling design of *B. laevigata* (Figure 4.4), imposed by its peculiar spatial distribution, contains pseudo-replicates, and thus increases the probability to abusively take into account recurrent high frequencies for a given marker due in reality to the fact that close individuals are genetically related (spatial autocorrelation due to demography). When the spatial resolution of the DEM is coarse, these close samples will retrieve their DEM variables from the same pixels, thus inflating

autocorrelation due to spurious correlations. Therefore, GLM estimates are inflated and standard errors are likely to be too small between closely-related individuals. This questions the detection of *B. laevigata* models with DEM-derived variables since these variables were mostly significant at their coarsest resolutions. In this case, methods such as General Estimating Equations (GEE; Hanley *et al.* 2003; Hardin & Hilbe 2003; Poncet *et al.* 2010) or generalized linear mixed models (GLMM; Liang & Zeger 1986; Bolker *et al.* 2009), that take into account pseudo-replicates might have been more appropriate. In these methods, the sampling design is integrated as supplementary information in order to account for autocorrelation and to estimate correctly the correlation matrix between individuals. In the sheep & goats case study, however, pseudo-replicates were avoided by selecting only one individual per farm (section 6.2), guaranteeing the spatial spread of sampling locations.

On the other hand, we found that window size of climatic variables did not influence substantially associations with genotypes, for both *P. major* and sheep & goats case studies, even when these variables considered topography. This is due to the interpolation process itself, which results in smoothing the variable between points (weather stations). The limited number of weather stations, located in and around the sampling area, is not sufficient to model local climatic variability at these extents.

Detection of signatures of selection

Many methods exist to detect footprints of selection. They differ mainly on the different principles or theory they are based on and thus separate in two categories: correlative approaches and population genetic approaches. They also differ by their sensitivity to demographic effects, inclusion of population structure or of spatial autocorrelation. In fact, population genetics methods of detection of selection are well established and have been tested against many demographic scenarios (Beaumont & Nichols 1996). On the other hand, correlative approaches often show a high rate of false positives but are more powerful as they can identify environmental pressures of selection (De Mita *et al.* 2013). Common detections should thus increase the robustness of signatures of selection by separating true from false positives and by identifying the environmental actor responsible for selection. However, only a few comparisons between methods have been published using simulations (Pérez-Figueroa *et al.* 2010; De Mita *et al.* 2013 and references therein). In these papers, BayeScan is shown to have a low rate of false positives and a good power of detection, unless selection pressure is low. On the other hand, correlation based methods were sufficiently powerful to detect true positives even when selection pressure was low and regardless of the demographic scenario. Finally, we note that in Lotterhos & Whitlock (2015) both types of methods failed to identify loci under weak selection, which probably constitute the majority of loci.

In the three case studies, comparison between methods was difficult. In fact, not all methods could be applied to each case study, due to the lack of population structure or to systematic failure of software. In fact, population genetics approaches require assignment of individuals to populations, which we could not achieve for sheep & goats in Morocco. Even if we had found a significant population structure, we doubt that methods like BayeScan could have processed WGS da-

tasets in a reasonable time. Regarding the systematic failure in LFMM in P. major case study, we mentioned in section 5.3.3 that we tried to filter the genetic dataset differently or try other parameters, without success. Similarly, applying LFMM on the sheep dataset sometimes resulted in crashes of the software.

Common detections turned out to be rare in all case studies. There is none for *B. laevigata*, despite Samβada identified moderate correlations. Here, BayeScan does not detect any markers under selection between the two populations identified or even when we defined populations based on transects. Using an additional population genetic approach, like Fdist or Mcheza, could have provided an interesting comparison. In *P. major*, we found two common detections between Samβada and BayeScan suggesting that precipitation is a major actor of natural selection. However, precipitation and latitude largely detected the same loci and it is thus impossible to disentangle the effect of both variables. In this case, either BayeScan does not have enough power due to the small genetic dataset and the few populations identified; or the spatial distribution of these common detections is possibly due to precipitation and latitude. In sheep & goats, validation of models depend more on the advantages of having the positions of SNPs rather than on the common detections. In fact, peaks of high scores were observed in each method at the same locations, with one example for each mammal. However, LFMM failed to identify significant SNPs in these peaks and there is no trend between LFMM and GLM models, not even regarding their spatial patterns.

In these three case studies, our philosophy was to focus on a correlative approach (Samβada) and then to compare its results to other methods (BayeScan, LFMM, XP-CLR). We highlight several advantages of GLM as a first approach, corroborating those identified by Stucki (2014). First, Samβada is a good first approach to detect adaptive signals as it does not have any pre-requisite and does not make theoretical assumptions like population genetics methods (Joost *et al.* 2007). On the contrary, LFMM and BayeScan use a theoretical background in population genetics for their models and require parameters to be defined. For example, LFMM requires the number of latent factors and BayeScan requires assignment of individuals to populations. A second advantage of simple correlative approaches is their ability to screen efficiently large genome datasets with hundreds of environmental variables to look for significant associations. This permitted to filter the environmental dataset for other methods such as LFMM, which are much longer to run. We consider this step safe as correlative approaches are known to have a high rate of false positives but little-to-no false negatives (De Mita *et al.* 2013).

However, the main issue in Samβada remains that it cannot distinguish true from false positives. Therefore, we investigated the usefulness of multivariate models as well as spatial autocorrelation measurements to tackle this issue. Regarding multivariate models, our hope was to better explain the distribution of genetic markers by a combination of an environmental variables and a variable related to demographic processes (membership coefficient, latitude, longitude). However, this problem is more complex than initially thought as historical movements of populations are often confronted with landscape patterns (Prunier *et al.* 2013). For example, most of the genotypes associated with precipitation in Samβada for *P. major* are also significantly correlated to latitude and longitude but were not significant in the multivariate case. In these cases, it is thus not possible to know whether the coordinates or the environmental variable is responsible for the

spatial pattern of the locus of interest. In fact, we never encountered significant multivariate models in our case studies, unlike Stucki (2014) with Ugandan cattle. She obtained several significant bivariate models including membership coefficient (population structure) and one environmental variable, which brings us close to the LFMM functioning. A great improvement for Samβada would be to integrate directly variables of population structure and geographic vectors, such as Moran's eigenvector maps (Borcard & Legendre 2002; Manel *et al.* 2010b), to perform multivariate models.

Another more general issue is that these methods have been developed independently and lack a coherent framework (Wagner & Fortin 2013). While we recommend using systematically several approaches (at least one correlative and one differential), the major difficulty is to define a sampling design that could suit both landscape and population genetics purposes (Joost *et al.* 2013). In fact, differential and correlative methods do not apprehend populations or individuals the same way. Therefore, differential methods often show spurious autocorrelation due to spatial clustering of samples data; and correlative approaches at the individual level may end up with limited population structure, thus discarding differential methods. More broadly, both types of methods do not have yet a good understanding of how sampling design affects performances of detection (Lotterhos & Whitlock 2015).

To conclude this section, let us mention that common identification of genetic markers under natural selection was rare and that this lack of common ground prevented us from providing robust signatures of selection. We do not consider correlative approaches as a sufficient tool to detect signatures of selection but as part of a panel of methods, in which it is a valuable initial step. In addition, we recommend identifying systematically correlations between genetic data and environmental variables on one side, and coordinates and population structure on the other side, to estimate whether detected patterns of adaptation are a combination of environmental pressure and demographic processes or if they can be due to environment only.

Questions related to spatial analysis

When looking for signatures of selection due to environmental variables, we expect to find spatial autocorrelation (SA) in detected genetic markers as SA is a natural component of environmental variables (Legendre 1993). Indeed, the SA we are looking for is induced by the environment, but we hope to distinguish it from the one due either to demographic effect (isolation by distance) or to sampling design (spurious SA) (Cushman & Landguth 2010). However, SA induced by patterns of selection is often confounded with demographic effects such as isolation by distance (Vekemans & Hardy 2004). In fact, in population genetics, autocorrelation is most often mentioned as caused by demographic processes and thus, its measurement serves as a descriptive analysis of the spatial structure of genetic data (Sokal *et al.* 1998; Diniz-Filho *et al.* 2009). Therefore, Moran's I correlograms are used in population genetics to describe the complexity of IBD patterns, both in original variable and model's residuals, as SA violates the assumption of independent error of non-spatial linear models. If SA does occur, analysis should be modified to account for it (Wagner & Fortin 2005). To better understand this dilemma, many authors have reviewed SA and tried to find solutions to limit its confounding and spurious effects (Legendre 1993;

Wagner & Fortin 2005; Dale & Fortin 2009; Diniz-Filho *et al.* 2009). However, studies comparing different methods of detection of natural selection only included SA as a point of discussion and rarely measured it (Pérez-Figueroa *et al.* 2010; De Mita *et al.* 2013; Lotterhos & Whitlock 2015). In these cases, SA is mentioned for the spurious correlations it might generate but not for its inevitable presence in signatures of selection.

In all case studies, Samβada's significant models were all strongly spatially autocorrelated, even (more) in common detections with BayeScan. They constituted most of the highest SA in the correlograms, which depicted neutral vs detected loci (Figure 4.10,Figure 5.9, Figure 6.17 and Figure 6.36). In fact, we found high spatial autocorrelation in each significant environmental variable and we thus expected that it would induce SA in detected loci. However, we know that logistic regressions are subject to high false positive rates, partly due to demographic or spurious autocorrelation effects (Fortin *et al.* 2002) that inflates false positive errors, thus overestimating detections of selection (Diniz-Filho *et al.* 2009; De Mita *et al.* 2013). However, Samβada's results contrasts sharply with those from LFMM in sheep & goats, as the SA of genotypes detected by LFMM is close to zero (spatial independence). Therefore, while it makes sense that LFMM results are less autocorrelated because LFMM corrects associations by using population structure, the absence of SA questions the validity of these models. In other words, we consider that SNPs detected by LFMM in sheep & goats might not be valid because they are not spatially autocorrelated.

It is difficult to compare our results to the literature because SA is rarely used to characterize signatures of selection. Nevertheless, Stucki (2014) compared Moran's I and Samβada G score for models she computed on Ugandan cattle. She found a strong correlation between these two variables (Figure 7.1) and distinguished neutral markers (non-significant G score, weak SA) from common detections between Samβada and LFMM (significant but moderate G score, moderate SA) and those only detected by Samβada (high G score, high SA). Common detections were thus showing substantial SA, but genotypes with very high SA were ignored by LFMM. Her results show that LFMM can detect spatially autocorrelated loci and support our statement that LFMM cannot perform well under the absence of strong population structure. In our case, we also found that high Samβada scores often imply high global spatial-autocorrelation but more importantly, we assessed the spatial autocorrelation of all loci and were able to illustrate that selected loci are more spatially autocorrelated than neutral loci.

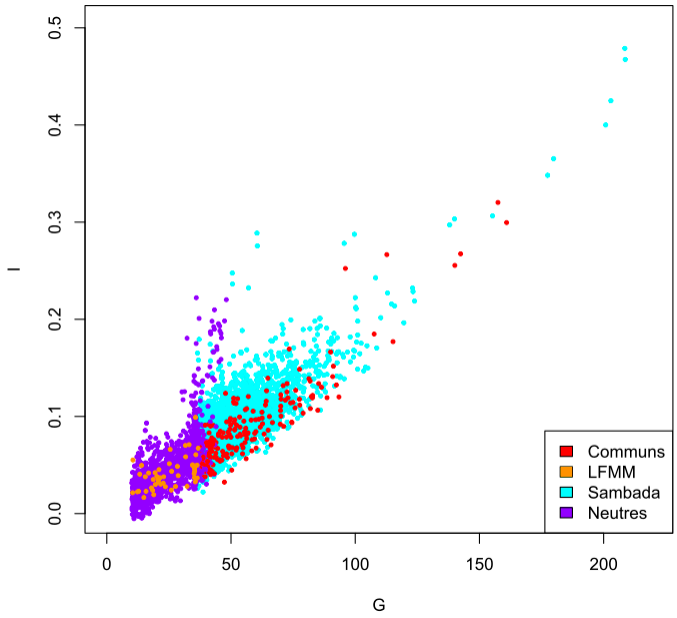


Figure 7.1 Comparison of SNPs detected by Samβada and LFMM. The graph shows the Moran's I in function of the G score of correlative models in Ugandan cattle. Neutral loci are in violet, loci detected by LFMM only are in orange, loci detected by Samβada in blue and loci detected by Samβada and LFMM in red. From (Stucki 2014).

Regarding *B. laevigata* and *P. major*, we cannot eliminate a potential SA due to sampling design. In fact, transects have been pointed out as a source of false SA (Dale & Fortin 2002). On the other hand, sheep & goats case study is based on a grid and on stratified random sampling, thus avoiding clusters. We consider this design as the best for landscape genomics studies, which should seek to reduce biases due to sampling design and focus more on the diversifications of habitat conditions rather than to correspond to the actual distribution of a population.

In addition to global autocorrelation measurements, Samβada can also compute Local Indices of Spatial Association, or LISA (Anselin 1995). Measuring local spatial autocorrelation of significant genetic markers could allow us to identify best candidates of local adaptation by studying the spatial relationship of genetic markers with environmental features. In fact, LISA indices assess whether samples are locally dependent and detect the presence of significant spatial clusters (Stucki 2014). However, LISA measurements in Samβada were disappointing as we found out that fixed kernels, such as Gaussian and Bi-square weighting schemes, did not work. We expected these weighting schemes to provide more significant measurements of local clusters, as the values of neighbour will be weighted by the distance to the point of interest, rather than the binary presence/absence only. We also suggest developing a local Moran index for alleles rather than for genotypes. In fact, alleles at sampling locations vary from zero to two in diploid data and could be more powerful to detect clusters. Finally, we note that LISA is never used in landscape and population genetics and that, at the exception of Samβada, packages to compute LISA coefficient in an automatic way are rare.

To conclude this part, we hope to raise awareness on the issue of SA and recommend to systematically measuring it in landscape genomics. Unfortunately, comprehensive measurements of SA are inexistent in the literature. While our results cannot provide a clear methodology to distinguish true and false positives, we believe that spatial autocorrelation is an important characteristic of signatures of selection when one wants to find correlations with environmental variables, which are naturally autocorrelated. In addition, despite this point is clear in the literature, SA is rarely measured in case studies but only mentioned (Dale & Fortin 2009). Finally, to improve Samβada's detections, we suggest analysing SA in residuals to assess the bias it produces.

Promises of whole genome sequencing data

Whole genome sequencing (WGS) holds great promises in landscape genomics. The main advantage of using WGS data for interpreting signatures of selection is the ability to assess the exact position of the detected SNP within the genome. In conjugation with genome assembly, WGS allows to assess the link between the environmental variables involved in the detection and the function of the gene.

In the sheep & goats case study, we demonstrated that searching for signatures of adaptation in whole genome sequence datasets greatly benefited from usage of a correlative method. First, we identified peaks of selection. Screening the entire genome allowed us to identify places in the genome that are worth considering for deep investigation at the gene level. Second, we found that SNPs detected in these peaks were associated with the same (or similar) environmental variables and displayed similar distribution patterns, reinforcing the probability of detecting selection due to the environment. In addition, the identified peaks of selection were identified in independent method and allowed us to visually target specific regions of the genome, without counting on different thresholds of significance and results filtering. Despite LFMM failed to detect significant loci at the peak position, the peak detected by this method is still clearly visible. In these cases, we can assume that these peaks of high scores are the translation of selective sweeps acting on these genomic regions, shaping the frequency of nearby positions and the spatial distribution of nearby polymorphisms.

When relevant positions are identified, we can benefit from the large online databases (like NCBI) and map the location of nearby genes. Indeed, knowing the genes involved allowed us to search for their functionality and evaluate if it concords with the environmental variable the SNP is associated with. Furthermore, finding a phenotypic variation related to a gene greatly could improve our understanding of adaptation and demonstrates the power of genome screening using correlative approaches. In the other case, when the detected gene has not been described yet, the associated environmental variable informs on the functionality of the gene. Finally, assessing whether the detected SNPs are synonymous or non-synonymous variations could inform further on a potential selection. In fact, only non-synonymous variations can change the amino-acid sequence of the coded protein. If the SNP is synonymous (no change of amino-acid), there is normally no reason to believe a selection pressure is exerted on it.

Another advantage of WGS is the possibility to compare similar genes in different species against the same environmental variables. In fact, we detected in both species SNPs that were associated with the same environmental variables. Therefore, further investigation of genetic sequences around these SNPs could greatly increase the reliability of these signatures of adaptation.

One minor drawback when comparing methods is the large size of the WGS datasets, which require heavy computational power for the analyses and powerful software that could process the results. In our case, we found that a first analysis at one resolution was a good start to filter the data with a loose threshold of significance. Afterwards, the subset of SNPs was analysed again with the multi-scale environmental dataset. Using logistic regression as a first exploratory analysis was here pertinent to facilitate the analysis and interpretation. Still, the recent availability of high performance computers will likely suppress this issue.

Perspectives

The three case studies analysed in this thesis comprise several species at different extents. Such diversity of situations does not allow to accurately assess the role of scale in species adaptation. Instead, to study the operational scale of adaptation, we should study the same species at multiple extents (i.e. by increasing the size of the study area). For example, an alpine plant would be a good candidate for this type of study, as we would expect to find signatures of selection mainly related to topographic features and climatic conditions, respectively at a local and at a large scale.

At the same time, we could take advantage of a model species to better understand the mechanisms of adaptation at different scales. Indeed, by using a model species, we often benefit from a reference genome on which we can align SNPs sequences to identify genomic regions and genes under selection. In addition, we can benefit from the literature on the model species regarding its functional genomics and adaptation to the environment. *Arabis alpina* for example, is an alpine plant often studied in its natural environment and a large part of its genome has been sequenced (Ansell *et al.* 2008; Manel *et al.* 2010b; Poncet *et al.* 2010; Melodelima & Lobréaux 2013; Lobréaux *et al.* 2014). Conversely, one can also imagine a reverse approach where a known gene with well-studied phenotypic variability is analysed along an environmental gradient corresponding to the alleged selection pressure.

We showed that DEM-derived variables are relevant proxies for important ecological features. However, the relationship between DEM and environmental features depends on the spatial resolution of the DEM, on the extent of the case study, and can vary from one location to another. Finding these relationships is therefore a crucial step to interpret the signatures of selections. In the *B. laevigata* case study, signatures of selection were not robust and it would be interesting to apply the same framework to other species in mountainous areas, in order to compare with our results. In addition, the *B. laevigata* case study did not benefit from recent high-throughput genetic data, such as SNPs. We therefore suggest producing a wide range of SNPs from the plants already sampled and reproduce associative models using the extensive environmental database that we have acquired for this thesis.

Defining a sampling design that is suitable for both population and landscape genetics approaches is a complicated step as these distinct approaches have different pre-requisites. In fact, the three case studies differed in the sampling design and we showed that this could cause problems with certain approaches of detection of selection. We discuss two important points concerning the sampling design. First, we emphasize the importance of considering false detections that spatial autocorrelation can generate with specific sampling designs. Indeed, when several individuals are sampled in a few populations, we increase the risk of false positives due to a strong spatial autocorrelation. In addition, a small number of populations does not allow estimating changes in allele frequency along an environmental gradient. Conversely, sampling along an environmental gradient may turn the assignment of individuals to populations more difficult and thus, does not allow combining both types of approaches. Secondly, De Mita *et al.* (2013) propose to favour a high number of populations with few samples rather than a small number of populations with many samples. These authors observed that this type of sampling design improves the performance of logistic regressions and of BayeScan. Such an approach indeed allow on one hand to apply populations genetics approaches such as BayeScan by considering each plot as a population, with no particular risk of bias, and on the other hand correlative approaches. However, more complex logistic regressions must then be used to take into account pseudo-replicas effects (i.e. several samples in the same plot), such as GEE or GLMM.

Finally, we suggest three main strategies found in the literature, with their pros and cons, to increase the robustness of detections in correlative approaches. One is to sample the same population at different time periods, but it is expensive and there should be correlations between generations. A second one is to do common garden experiments, were individuals from different populations along an environmental gradient are interchanged and diagnosed over one or several growing seasons (De Kort *et al.* 2014). However, this could be too slow for selection to be detected. A third one is to replicate a study by sampling different populations of the same species in geographically distinct habitats, which is not always possible (Schwartz & McKelvey 2009). In this case, detecting the same association in similar habitats provides a robust evidence of environmental adaptation.

Conclusion

What can we learn about the adaptation of species to the environment by applying a multi-scale landscape genomic framework?

The framework that we proposed in this thesis highlighted the benefits of a multidisciplinary approach (i.e. GIS, spatial analysis, environmental modelling, population genetics and computer science) for the exploration of genetic and environmental data in evolutionary biology. i) We established the relevance of DEMs in providing proxies to particular micro-habitat conditions for plants, and outlined the large panel of variables that can be computed from DEMs. ii) We contributed to a better understanding of scale in adaptation studies by using multi-scale environmental data, and demonstrated the need to consider the scale representativeness of topographic features. iii) We combined population and landscape genetics methods to detect signatures of selection and showed that both types of approaches are mutually beneficial. iv) We used spatial statis-

tics and demonstrated the importance of spatial autocorrelation in the interpretation of signatures of selection. v) Finally, we improved the investigation and interpretation of signatures of selection by providing appropriate graphical representations combining the different information existing for each significant model. In these graphs, we highlighted the complementarity between spatial analysis, cartography and correlative approaches. Importantly, we developed an automated procedure to produce these numerous graphs.

In the following paragraphs, we summarize the findings of this thesis regarding the research questions that were asked in the objectives section (p. 17).

1. How important is topography in modelling of the micro-habitat conditions encountered by plants and in detecting signatures of selection?

DEM-derived variables are relevant for modelling micro-climatic conditions in mountainous areas and may constitute good proxies to environmental factors involved in the adaptation process at a local scale. Despite being potentially useful for assessing the environmental variability, DEM-derived variables are largely omitted in landscape genetics. DEMs also remain underexploited regarding the large diversity of variables that can be computed from them. Our results advocate going beyond the traditional use of elevation, slope and aspect in order to cover all possible relationships between topography and habitat.

2. Is very high resolution necessary to model micro-habitat conditions encountered by plants?

One of the most important findings of this thesis is the discovery of the dependency between association models and the spatial resolution of some DEM-derived variables. We demonstrate that scale representativeness must be considered, regardless of the original resolution of the DEM and of the species studied. These examples also show that a higher resolution does not necessarily mean better explanatory power, thus contributing to an indecisive debate in the current literature. Furthermore, each DEM-derived variable responds differently to a change of resolution, which suggests that only multi-scale approaches can evaluate the role of topographic features in local adaptation. On the other hand, we found that increasing window sizes on interpolated climatic data do not substantially influence association models.

3. How does the relevance of DEM-derived variable vary in function of the extent of the study site and of the mobility of the species?

Adaptation may occur at every scale but the type of variables (climatic, topographic, soil) involved in signatures of adaptation depends on the type of organism and on the extent of the study. However, we did not observe common detections between independent approaches at a local scale, which does not allow us to confirm a higher relevance of DEM-derived variables compared

to climatic variables. Nevertheless, we report no significant models involving topography at a regional scale and few models at a large scale.

4. Are significant associations identified by means of correlative methods also detected by population genetics approaches?

Detecting genetic markers under selection with independent approaches should increase the robustness of detections by decreasing the amount of false positives. However, we notice that prerequisites of these approaches cannot always be respected. Still, we found common detections in two of the three case studies: two SNPs in the *P. major* case study and two genomic regions in the sheep & goats case study. Based on these results, we recommend to take advantage of the strong theoretical background of population genetic approaches on one hand, and to identify a potential pressure of selection in correlative approaches on the other hand. Our results also support the observation that Samβada produces a higher number of false positives. However, its efficiency in handling large datasets in a reasonable time still makes us recommend using it as an initial step, with little chance of missing potential signatures of adaptation.

5. How can spatial autocorrelation contribute to the analysis of signatures of selection?

Spatial autocorrelation is rarely measured while investigating signatures of selection. Here, we show that SA is stronger in these signatures than in neutral markers. This result adds even more controversy to an ambiguous perception regarding the nature of SA. Indeed, this topic remains ambiguous even after decades of debate and we regret that SA is often considered as a nuisance and not commonly measured in landscape genetics studies. In fact, spatial autocorrelation could be essential for better understanding evolutionary phenomena and local adaptation to environment, as advocated by Legendre (1993). We thus recommend measuring systematically SA, both at the global and local level, and consider it while investigating adaptation to the environment.

6. How does whole genome sequencing improve the detection of signatures of adaptation?

For the first time we used whole genome sequencing data in landscape genomics. Together with environmental variables, genetic data of high density allowed us to locate signatures of selection in the genome and to identify genes associated with them. By investigating the underlying functions of these genes, one can potentially relate these signatures of selection to a phenotype and better understand adaptation processes. Furthermore, analysing the spatial distribution of these genotype-phenotype associations could greatly help in conservation practices.

References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, **19**, 1655–1664.

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature reviews. Genetics*, **11**, 697–709.

Anderson CD, Epperson BK, Fortin MJ *et al.* (2010) Considering spatial and temporal scale in landscape-genetic studies of gene flow. *Molecular Ecology*, **19**, 3565–3575.

Anselin L (1995) Local indicators of spatial association — LISA. *Geographical Analysis*, **27**, 93–115.

Anselin L (1998) Exploratory spatial data analysis in a geocomputational environment. *GeoComputation*, 17–19.

Anselin L, Syabri I, Kho Y (2006) GeoDa: An introduction to spatial data analysis. *Geographical Analysis*, **38**, 5–22.

Ansell SW, Grundmann M, Russell SJ, Schneider H, Vogel JC (2008) Genetic discontinuity, breeding-system change and population history of Arabis alpina in the Italian Peninsula and adjacent Alps. *Molecular Ecology*, **17**, 2245–2257.

Antao T, Beaumont MA (2011) Mcheza: A workbench to detect selection using dominant markers. *Bioinformatics*, **27**, 1717–1718.

Aregger M, Cowling VH (2013) Human cap methyltransferase (RNMT) N-terminal non-catalytic domain mediates recruitment to transcription initiation sites. *The Biochemical journal*, **455**, 67–73.

Ashcroft MB, French KO, Chisholm LA (2011) An evaluation of environmental factors affecting species distributions. *Ecological Modelling*, **222**, 524–531.

Balkenhol N, Waits LP, Dezzani RJ (2009) Statistical approaches in landscape genetics: an evaluation of methods for linking landscape and genetic data. *Ecography*, **32**, 818–830.

Barry RG (1992) *Mountain weather and climate* (Routledge, Ed,).

Bates D, Maechler M (2009) lme4: Linear mixed-effects models using {S4} classes.{R} package version 0.999375-32.

Beaumont MA, Nichols RA (1996) Evaluating Loci for Use in the Genetic Analysis of Population Structure. *Proceedings: Biological Sciences*, **263**, 1619–1626.

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289 – 300.

Benjelloun B, Alberto FJ, Streeter I *et al.* (2015) Characterizing neutral and adaptive genomic diversity in indigenous populations of Moroccan goats (Capra hircus) using WGS data. *Frontiers in Genetics*.

Beven KJ, Kirkby MJ (1979) A physically based, variable contributing area model of basin hydrology / Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrological Sciences Bulletin*, **24**, 43–69.

Böhner J, Antonić O (2009) Chapter 8 Land-Surface Parameters Specific to Topo-Climatology. In: *Developments in Soil Science* (eds Tomislav H, Hannes IR), pp. 195–226. Elsevier.

Böhner J, Köthe R, Conrad O *et al.* (2002) Soil Regionalisation by Means of Terrain Analysis and Process Parameterisation. , **EUR 20398** .

Böhner J, McCloy KR, Strobl J (2006) SAGA – Analysis and Modelling Applications. *Göttinger Geographische Abhandlungen*, **115**, 130.

Böhner J, Selige T (2006) Spatial prediction of soil attributes using terrain analysis and climate regionalisation. *BÖHNER, J., MCCLOY, KR & J. STROBL (Eds.): SAGA–Analyses and Modelling Applications.–Göttinger Geographische Abhandlungen*, **115**, 13–28.

Bolker BM, Brooks ME, Clark CJ *et al.* (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, **24**, 127–135.

Bolliger J, Lander T, Balkenhol N (2014) Landscape genetics since 2003: status, challenges and future directions. *Landscape Ecology*, **29**, 361–366.

Borcard D, Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling*, **153**, 51–68.

Brenning A (2008) Statistical Geocomputing combining R and SAGA: The Example of Landslide susceptibility Analysis with generalized additive Models (J Böhner, T Blaschke, L Montanarella, Eds,). *SAGA —- Seconds Out*, **19**, 23–32.

Breslow NE, Clayton DG (1993) Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, **88**, 9–25.

Broxton PD, Troch PA, Lyon SW (2009) On the role of aspect to quantify water transit times in small mountainous catchments. *Water Resources Research*, **45**, W08427.

Buchanan BP, Fleming M, Schneider RL *et al.* (2013) Evaluating topographic wetness indices across central New York agricultural landscapes. *Hydrol. Earth Syst. Sci. Discuss.*, **10**, 14041–14093.

Burga CA, Krüsi B, Egli M *et al.* (2010) Plant succession and soil development on the foreland of the Morteratsch glacier (Pontresina, Switzerland): Straight forward or chaotic? *Flora - Morphology, Distribution, Functional Ecology of Plants*, **205**, 561–576.

Cavazzi S, Corstanje R, Mayr T, Hannam J, Fealy R (2013) Are fine resolution digital elevation models always the best choice in digital soil mapping? *Geoderma*, **195–196**, 111–121.

Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. *Genome Research*, **20**, 393–402.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using Environmental Correlations to Identify Loci Underlying Local Adaptation. *Genetics*, **185**, 1411–1423.

Cushman SA, Landguth EL (2010) Spurious correlations and inference in landscape genetics. *Molecular Ecology*, **19**, 3592–3602.

Cushman SA, McKelvey KS, Hayden J, Schwartz MK (2006) Gene flow in complex landscapes: testing multiple hypotheses with causal modeling. *The American naturalist*, **168**, 486–499.

Dale MRT, Fortin M-J (2002) Spatial autocorrelation and statistical tests in ecology. *Ecoscience*, **9**, 162–167.

Dale MRT, Fortin M-J (2009) Spatial autocorrelation and statistical tests: Some solutions. *Journal of Agricultural, Biological, and Environmental Statistics*, **14**, 188–206.

Dale MRT, Mah M (1998) The use of wavelets for spatial pattern analysis in ecology. *Journal of Vegetation Science*, **9**, 805–814.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Darwin C, Wallace A (1858) On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. *Journal of the Proceedings of the Linnean Society of London. Zoology*, **3**, 45–62.

Diniz-Filho JAF, Nabout JC, de Campos Telles MP, Soares TN, Rangel TFLVB (2009) A review of techniques for spatial modeling in geographical, conservation and landscape genetics. *Genetics and Molecular Biology*, **32**, 203–211.

Dobrowski SZ (2011) A climatic basis for microrefugia: The influence of terrain on climate. *Global Change Biology*, **17**, 1022–1035.

Dobson AJ, Barnett A (2008) *An Introduction to Generalized Linear Models, Third Edition*. Taylor & Francis.

Dubuis A, Giovanettina S, Pellissier L *et al.* (2013) Improving the prediction of plant species distribution and community composition by adding edaphic to topo-climatic variables. *Journal of Vegetation Science*, **24**, 593–606.

Earl D, vonHoldt B (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Elith J, Leathwick JR (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

Ellenberg H, Weber HE, Düll R *et al.* (1991) Zeigerwerte von pflanzen in Mitteleuropa.

Escoffier B, Pages J (2008) *Analyses factorielles simples et multiples* (Paris:Dunod, Ed,).

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, **14**, 2611–2620.

Fischer MC, Rellstab C, Tedder A *et al.* (2013) Population genomic footprints of selection and associations with climate in natural populations of Arabidopsis halleri from the Alps. *Molecular ecology*, **22**, 5594–5607.

Fisher RA (1930) *The Genetical Theory of Natural Selection*. At The Clarendon Press.

Foll M, Gaggiotti O (2008) A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics*, **180**, 977–993.

Fontanesi L, Beretti F, Riggio V *et al.* (2009) Missense and nonsense mutations in melanocortin 1 receptor (MC1R) gene of different goat breeds: association with red and black coat colour phenotypes but with unexpected evidences. *BMC genetics*, **10**, 47.

Fortin M-J, Dale MRT, ver Hoef J (2002) Spatial analyses in ecology. In: *Encyclopedia of environmetrics* , pp. 2051–2058.

Fournier-Level A, Korte A, Cooper MD *et al.* (2011) A Map of Local Adaptation in Arabidopsis thaliana. *Science*, **334**, 86–89.

Fraser HB (2013) Gene expression drives local adaptation in humans. *Genome Research*, **23**, 1089–1096.

Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*.

Fridley JD (2009) Downscaling climate over complex terrain: High finescale (<1000 m) spatial variation of near-ground temperatures in a montane forested landscape (Great Smoky Mountains). *Journal of Applied Meteorology and Climatology*, **48**, 1033–1049.

Fu P, Rich PM (2002) A geometric solar radiation model with applications in agriculture and forestry. *Computers and Electronics in Agriculture*, **37**, 25–35.

Gallant JC, Hutchinson MF (1996) Towards an understanding of landscape scale and structure. In: *Proceedings of the Third International Conference/Workshop on Integrating GIS and Environmental Modeling*

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*, 9.

Geiser C (2014) Genome evolution and mechanisms underlying reproductive isolation in the polyploid “Biscutella laevigata.” Univ. Neuchâtel.

Di Giulio M, Holderegger R, Tobias S (2009) Effects of habitat and landscape fragmentation on humans and biodiversity in densely populated landscapes. *Journal of Environmental Management*, **90**, 2959–2968.

Gobat JM, Duckert O, Gallandat JD (1989) *Quelques relations “microtopographie-sols-végétation” dans les pelouses pseudo-alpines du Jura suisse: exemples d’un système naturel et d'un système anthropise*. Société neuchâteloise des sciences naturelles.

Gottfried M, Pauli H, Grabherr G (1998) Prediction of vegetation patterns at the limits of plant life: A new view of the alpine-nival ecotone. *Arctic and Alpine Research*, **30**, 207.

Greenwood S, Chen J-C, Chen C-T, Jump AS (2015) Temperature and sheltering determine patterns of seedling establishment in an advancing subtropical treeline. *Journal of Vegetation Science*, n/a–n/a.

Gruber S, Peckham S (2009) *Land-surface parameters and objects in hydrology*.

Guillot G, Vitalis R, Rouzic A le, Gautier M (2014) Detecting correlation between allele frequencies and environmental variables as a signature of selection. A fast computational approach for genome-wide studies. *Spatial Statistics*, **8**, 145–155.

Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Hall L, Beissinger S (2014) A practical toolbox for design and analysis of landscape genetics studies. *Landscape Ecology*, **29**, 1487–1504.

Hanley JA, Negassa A, Edwardes MDD, Forrester JE (2003) Statistical analysis of correlated data using generalized estimating equations: An orientation. *American Journal of Epidemiology*, **157**, 364–375.

Häntzschel J, Goldberg V, Bernhofer C (2005) GIS-based regionalisation of radiation, temperature and coupling measures in complex terrain for low mountain ranges. *Meteorological Applications*, **12**, 33–42.

Hardin JW, Hilbe JM (2003) Generalized Estimating Equations. *Chapman and Hall/CRC: London*.

Hardy OJ, Vekemans X (2002) SPAGeDi: A versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2**, 618–620.

Hastings A (1983) Can spatial variation alone lead to selection for dispersal? *Theoretical Population Biology*, **24**, 244–251.

Hereford J (2009) A quantitative survey of local adaptation and fitness trade-offs. *The American naturalist*, **173**, 579–588.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hjerdt KN, McDonnell JJ, Seibert J, Rodhe A (2004) A new topographic index to quantify downslope controls on local drainage. *Water Resources Research*, **40**, W05602.

Holderegger R, Wagner HH (2008) Landscape Genetics. *BioScience*, **58**, 199–207.

Hosmer DW, Lemeshow S (2000) Introduction to the Logistic Regression Model. In: *Applied Logistic Regression* , pp. 1–30. John Wiley & Sons, Inc.

Hübner S, GÜNTHER T, FLAVELL A *et al.* (2012) Islands and streams: clusters and gene flow in wild barley populations from the Levant. *Molecular Ecology*, **21**, 1115–1129.

Jakobsson M, Rosenberg NA (2007) CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–1806.

Jones MR, Forester BR, Teufel AI *et al.* (2013) Integrating landscape genomics and spatially explicit approaches to detect loci under selection in clinal populations. *Evolution*, **67**, 3455–3468.

Joost S (2006) The geographical dimension of genetic diversity - a GIScience contribution for the conservation of animal genetic resources. EPFL.

Joost S, Bonin A, Bruford MW *et al.* (2007) A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology*, **16**, 3955–3969.

Joost S, Colli L, Baret P V *et al.* (2010) Integrating geo-referenced multiscale and multidisciplinary data for the management of biodiversity in livestock genetic resources. *Animal Genetics*, **41**, 47–63.

Joost S, Kalbermatten M, Bonin A (2008) Spatial analysis method (sam): a software tool combining molecular and environmental data to identify candidate loci for selection. *Molecular Ecology Resources*, **8**, 957–960.

Joost S, Vuilleumier S, Jensen JD *et al.* (2013) Uncovering the genetic basis of adaptive change: on the intersection of landscape genomics and theoretical population genetics. *Molecular Ecology*, n/a–n/a.

Kalbermatten M (2010) Multiscale analysis of high resolution digital elevation models using the wavelet transform. EPFL.

Kalbermatten M, Van De Ville D, Turberg P, Tuia D, Joost S (2012) Multiscale analysis of geomorphological and geological features in high resolution digital elevation models using the wavelet transform. *Geomorphology*, **138**, 352–363.

Kavanagh KD, Haugen TO, Gregersen F, Jernvall J, Vøllestad LA (2010) Contemporary temperature-driven divergence in a Nordic freshwater fish under conditions commonly thought to hinder adaptation. *BMC evolutionary biology*, **10**, 350.

Kawecki TJ, Ebert D (2004) Conceptual issues in local adaptation. *Ecology Letters*, **7**, 1225–1241.

Kimura M (1968) Evolutionary rate at the molecular level. *Nature*, **217**, 624–626.

Kleiner L, Robra JP, Gilliéron P-Y, Schaer P, Mertina C (2010) Lever de limites naturelles par scanner laser aérien (LIDAR) Evaluation et perspectives dans le cadre de la mensuration cadastrale. *Géomatique Suisse*, **4/2010**, 136–139.

Körner C (2003) *Alpine Plant Life: Functional Plant Ecology of High Mountain Ecosystems ; with 47 Tables*. Springer.

De Kort H, Vandepitte K, Bruun HH *et al.* (2014) Landscape genomics and a common garden trial reveal adaptive differentiation to temperature across Europe in the tree species Alnus glutinosa. *Molecular Ecology*, **23**, 4709–4721.

Kozak KH, Graham CH, Wiens JJ (2008) Integrating GIS-based environmental data into evolutionary biology. *Trends in ecology & evolution (Personal edition)*, **23**, 141–148.

Krieger G, Moreira A, Fiedler H *et al.* (2007) TanDEM-X: A satellite formation for high-resolution SAR interferometry. In: *IEEE Transactions on Geoscience and Remote Sensing* , pp. 3317–3340.

Lamaze FC, Sauvage C, Marie A, Garant D, Bernatchez L (2012) Dynamics of introgressive hybridization assessed by SNP population genomics of coding genes in stocked brook charr (Salvelinus fontinalis). *Molecular Ecology*, **21**, 2877–2895.

Landguth EL, Cushman SA, Schwartz MK *et al.* (2010) Quantifying the lag time to detect barriers in landscape genetics. *Molecular Ecology*, **19**, 4179–4191.

Landolt E (1977) *Ökologische Zeigerwerte zur Schweizer Flora*. Geobotan. Inst.

Landolt E, Bäumler B, Erhardt A *et al.* (2010) *Flora indicativa : ökologische Zeigerwerte und biologische Kennzeichen zur Flora der Schweiz und der Alpen = ecological indicator values and biological attributes*

of the Flora of Switzerland and the Alps. Ed. des Conservatoire et Jardin botaniques de la ville de Genève ; Haupt.

Lassueur T, Joost S, Randin CF (2006) Very high resolution digital elevation models: Do they improve models of plant species distribution? *Ecological Modelling*, **198**, 139–153.

Leempoel K, Geiser C, Daprà L *et al.* Very high resolution digital elevation models: are multi-scale derived variables ecologically relevant? *Methods in Ecology and Evolution*.

Leempoel K, Joost S (2012) Relatedness and scale dependency in very high resolution digital elevation models derivatives. In: *Open Source Geospatial Research & Education Symposium 2012* , p. 340. Lulu.com, Yverdon-les-Bains, Switzerland.

Legendre P (1993) Spatial autocorrelation: trouble or new paradigm? *Ecology*, **74**, 1659–1673.

Legendre P, Fortin M-J (2010) Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular ecology resources*, **10**, 831–844.

Levin SA (1992) The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology*, **73**, 1943–1967.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment / Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liang KY, Zeger SL (1986) Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika*, **73**, 13–22.

Lischer HEL, Excoffier L (2012) PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298–299.

Lobréaux S, Manel S, Melodelima C (2014) Development of an Arabis alpina genomic contig sequence data set and application to single nucleotide polymorphisms discovery. *Molecular Ecology Resources*, **14**, 411–418.

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, **24**, 1031–1046.

Lowry DB (2010) Landscape evolutionary genomics. *Biology Letters*, **6**, 502–504.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

Lyon SW, Troch PA, Broxton PD, Molotch NP, Brooks PD (2008) Monitoring the timing of snowmelt and the initiation of streamflow using a distributed network of temperature/light sensors. *Ecohydrology*, **1**, 215–224.

Manel S, Albert C, Yoccoz N (2012a) Sampling in Landscape Genomics. In: *Data Production and Analysis in Population Genomics SE - 1* Methods in Molecular Biology. (eds Pompanon F, Bonin A), pp. 3–12. Humana Press.

Manel S, Gugerli F, Thuiller W *et al.* (2012b) Broad-scale adaptive genetic variation in alpine plants is driven by temperature and precipitation. *Molecular Ecology*, **21**, 3729–3738.

Manel S, Holderegger R (2013) Ten years of landscape genetics. *Trends in ecology & evolution*, **28**, 614–21.

Manel S, Joost S, Epperson BK *et al.* (2010a) Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Molecular Ecology*, **19**, 3760–3772.

Manel S, Poncet BN, Legendre P, Gugerli F, Holderegger R (2010b) Common factors drive adaptive genetic variation at different spatial scales in Arabis alpina. *Molecular Ecology*, **19**, 3824–3835.

Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, **18**, 189–197.

Manel S, Segelbacher G (2009) Perspectives and challenges in landscape genetics. *Molecular Ecology*, **18**, 1821–1822.

Manton I (1937) The problem of Biscutella laevigata L. *Annals of Botany*, **51**, 439–465.

Marceau DJ, Hay GJ (1999) Remote Sensing Contributions to the Scale Issue. *Canadian Journal of Remote Sensing*, **25**, 357–366.

McKenna A, Hanna M, Banks E *et al.* (2010) The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

McVicar TR, Van Niel TG, Li L *et al.* (2007) Spatially distributing monthly reference evapotranspiration and pan evaporation considering topographic influences. *Journal of Hydrology*, **338**, 196–220.

Melodelima C, Lobréaux S (2013) Complete Arabis alpina chloroplast genome sequence and insight into its polymorphism. *Meta Gene*, **1**, 65–75.

Miller CL, Laflamme RA (1958) The digital terrain model - Theory and application. *Photogrammetric Engineering*, **24**, 433–442.

De Mita S, Thuillet A-C, Gay L *et al.* (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383–1399.

Montgomery DC, Peck EA (1982) *Introduction to linear regression analysis*. Wiley, New York.

Moore J-S, Bourret V, Dionne M *et al.* (2014) Conservation genomics of anadromous Atlantic salmon across its North American range: outlier loci identify the same patterns of population structure as neutral loci. *Molecular Ecology*, **23**, 5680–5697.

Moore ID, Grayson RB, Ladson AR (1991) Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. *Hydrological Processes*, **5**, 3–30.

Moran PAP (1950) Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23.

Morin PA, Martien KK, Taylor BL (2009) Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources*, **9**, 66–73.

Namroud MC, Beaulieu J, Juge N, Laroche J, Bousquet J (2008) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology*, **17**, 3599–3613.

Nelder JA, Wedderburn RWM (1972) Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, **135**, 370–384.

New M, Lister D, Hulme M, Makin I (2002) A high-resolution data set of surface climate over global land areas. *Climate Research*, **21**, 1–25.

Nielsen R (2005) Molecular signatures of natural selection. *Annual review of genetics*, **39**, 197–218.

Parisod C, Besnard G (2007) Glacial in situ survival in the Western Alps and polytopic autopolyploidy in Biscutella laevigata L. (Brassicaceae). *Molecular Ecology*, **16**, 2755–2767.

Parisod C, Bonvin G (2008) Fine-scale genetic structure and marginal processes in an expanding population of Biscutella laevigata L. (Brassicaceae). *Heredity*, **101**, 536–542.

Parisod C, Christin P-A (2008) Genome-wide association to fine-scale ecological heterogeneity within a continuous population of Biscutella laevigata (Brassicaceae). *New Phytologist*, **178**, 436–447.

Parisod C, Joost S (2010) Divergent selection in trailing- versus leading-edge populations of Biscutella laevigata. *Annals of Botany*, **105**, 655–660.

Patsiou TS, Conti E, Zimmermann NE, Theodoridis S, Randin CF (2014) Topo-climatic microrefugia explain the persistence of a rare endemic plant in the Alps during the last 21 millennia. *Global Change Biology*, **20**, 2286–2300.

Pepin NC, Seidel DJ (2005) A global comparison of surface and free-air temperatures at high elevations. *Journal of Geophysical Research D: Atmospheres*, **110**, 1–15.

Pérez-Figueroa A, GarcÍA-Pereira MJ, Saura M, RolÁN-Alvarez E, Caballero A (2010) Comparing three different methods to detect selective loci using dominant markers. *Journal of Evolutionary Biology*, **23**, 2267–2276.

Petren K (2013) The evolution of landscape genetics. *Evolution*, **67**, 3383–3385.

Poncet B, Herrmann D, Gugerli F (2010) Tracking genes of ecological relevance using genome scan in two independent regional population samples of Arabis alpina. *Molecular Ecology*.

Pradervand J-N, Dubuis A, Pellissier L, Guisan A, Randin C (2014) Very high resolution environmental predictors in species distribution models: Moving beyond topography? . *Progress in Physical Geography* , **38** , 79–96.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, **155**, 945–959.

Pritchard JK, Wen X, Falush D (2007) Documentation for Structure Software: Version 2.2.

Prunier JG, Kaufmann B, Fenet S *et al.* (2013) Optimizing the trade-off between spatial and genetic sampling efforts in patchy populations: Towards a better assessment of functional connectivity using an individual-based sampling scheme. *Molecular Ecology*, **22**, 5516–5530.

Purcell S, Neale B, Todd-Brown K *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, **81**, 559–575.

Randin CF, Vuissoz G, Liston GE, Vittoz P, Guisan A (2009) Introduction of Snow and Geomorphic Disturbance Variables into Predictive Models of Alpine Plant Distribution in the Western Swiss Alps. *Arctic, Antarctic, and Alpine Research*, **41**, 347–361.

Rezakhaniha R, Agianniotis A, Schrauwen JTC *et al.* (2012) Experimental investigation of collagen waviness and orientation in the arterial adventitia using confocal laser scanning microscopy. *Biomechanics and modeling in mechanobiology*, **11**, 461–473.

Richardson JL, Urban MC, Bolnick DI, Skelly DK (2014) Microgeographic adaptation and the spatial scale of evolution. *Trends in Ecology and Evolution*, **29**, 165–176.

Riley SJ, Degloria SD, Elliot R (1999) A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences*, **5**, 23–27.

Le Roux PC, Lenoir J, Pellissier L, Wisz MS, Luoto M (2012) Horizontal, but not vertical, biotic interactions affect fine-scale plant distribution patterns in a low-energy system. *Ecology*, **94**, 671–682.

Saccheri IJ, Rousset F, Watts PC, Brakefield PM, Cook LM (2008) Selection and gene flow on a diminishing cline of melanic peppered moths. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 16212–16217.

Sappington JM, Longshore KM, Thompson DB (2007) Quantifying Landscape Ruggedness for Animal Habitat Analysis: A Case Study Using Bighorn Sheep in the Mojave Desert. *The Journal of Wildlife Management*, **71**, 1419–1426.

Savolainen O (2011) The Genomic Basis of Local Climatic Adaptation. *Science*, **334**, 49–50.

Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature reviews. Genetics*, **14**, 807–20.

Scherrer D, Körner C (2011) Topographically controlled thermal-habitat differentiation buffers alpine plant diversity against climate warming. *Journal of Biogeography*, **38**, 406–416.

Schoville SD, Bonin A, François O *et al.* (2012) Adaptive genetic variation on the landscape: methods and cases. *Annual Review of Ecology, Evolution, and Systematics*, **43**, 23–43.

Schwartz MK, Luikart G, McKelvey KS, Cushman SA (2009) Landscape Genomics: A Brief Perspecitve. In: *Spatial Complexity, Informatics, and Wildlife Conservation* , pp. 165–175.

Schwartz MK, McKelvey KS (2009) Why sampling scheme matters: The effect of sampling scheme on landscape genetic results. *Conservation Genetics*, **10**, 441–452.

Segelbacher G, Cushman SA, Epperson BK *et al.* (2010) Applications of landscape genetics in conservation biology: Concepts and challenges. *Conservation Genetics*, **11**, 375–385.

Seungyong L, Wolberg G, Sung-Yong S (1997) Scattered data interpolation with multilevel B-splines. *Visualization and Computer Graphics, IEEE Transactions on*, **3**, 228–244.

Shaffer JP (1995) Multiple Hypothesis Testing. *Annual Review of Psychology*, **46**, 561–584.

Skelly DK (2004) Microgeographic countergradient variation in the wood frog, Rana sylvatica. *Evolution; international journal of organic evolution*, **58**, 160–165.

Sokal RR, Oden NL, Thomson B a. (1998) Local spatial autocorrelation in biological variables. *Biological Journal of the Linnean Society*, **65**, 41–62.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* , **100** , 9440–9445.

Storfer A, Murphy MA, Evans JS *et al.* (2006) Putting the "landscape" in landscape genetics. *Heredity*, **98**, 128–142.

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 B. EPFL.

Taberlet P, Valentini A, Rezaei HR *et al.* (2008) Are cattle, sheep, and goats endangered species? *Molecular Ecology*, **17**, 275–284.

Tachikawa T, Hato M, Kaku M, Iwasaki A (2011) Characteristics of Aster Gdem Version 2. *2011 Ieee International Geoscience and Remote Sensing Symposium (Igarss)*, 3657–3660.

Tiffin P, Ross-Ibarra J (2014) Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology & Evolution*, **29**, 673–680.

Tobler WR (1970) A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, **46**, 234–240.

Vekemans X, Hardy OJ (2004) New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology*, **13**, 921–935.

Vercauteren N, Destouni G, Dahlberg CJ, Hylander K (2012) Fine-Resolved, Near-Coastal Spatiotemporal Variation of Temperature in Response to Insolation. *Journal of Applied Meteorology and Climatology*, **52**, 1208–1220.

Van De Ville D, Sage D, Balac K, Unser M (2008) The Marr wavelet pyramid and multiscale directional image analysis. *EUSIPCO, August*, 25–29.

Wagner HH, Fortin MJ (2005) Spatial analysis of landscapes: Concepts and statistics. *Ecology*, **86**, 1975–1987.

Wagner HH, Fortin M-J (2013) A conceptual framework for the spatial analysis of landscape genetic data. *Conservation Genetics*.

Wang L, Liu H (2006) An efficient method for identifying and filling surface depressions in digital elevation models for hydrologic analysis and modelling. *International Journal of Geographical Information Science*, **20**, 193–213.

Waples RS, Gaggiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, **15**, 1419–1439.

Wilson JP, Gallant JC (2000) *Terrain Analysis: Principles and Applications* (Wiley, Ed,). Wiley.

Wood JD (1996) The geomorphological characterisation of digital elevation models.

Xiaoye Liu (2008) Airborne LiDAR for DEM generation: some critical issues. *Progress in Physical Geography* , **32** , 31–49.

Yokoyama R, Shirasawa RJ, Pike RJ (2002) Visualizing topography by openness: A new application of image processing to digital elevation models. *Photogrammetric Engineering and Remote Sensing*, **68**, 251–266.

Zevenbergen LW, Thorne CR (1987) Quantitative-Analysis of Land Surface-Topography. *Earth Surface Processes and Landforms*, **12**, 47–56.

Zimmermann NE, Kienast F (1999) Predictive mapping of alpine grasslands in Switzerland: Species versus community approach. *Journal of Vegetation Science*, **10**, 469–482.

Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM (2009) Mixed effects models and extensions in ecology with R. *Statistics*, **32**, 209–243.

List of figures

List of tables

List of equations

Appendix I. **Scripts**

Appendix I.a Computation of DEM-derived variables

```
#All variables script
library(RSAGA)
setwd("C:/data/RDN/MRS")
myenv <- rsaga.env(workspace=getwd(), path="C:/Program Files/SAGA-GIS_2_1")


rsaga.get.libraries()
rsaga.get.modules(c("ta_morphometry"))
rsaga.get.usage("ta_morphometry", 0)
rsaga.get.usage("ta_morphometry", 9)
rsaga.get.usage("ta_morphometry", 7)
rsaga.get.usage("ta_morphometry", 16)
rsaga.get.usage("ta_morphometry", 17)

rsaga.get.usage("ta_lighting", 3)
rsaga.get.usage("ta_lighting", 2)

rsaga.get.modules(c("ta_preprocessor"))
rsaga.get.usage("ta_preprocessor", 5)
rsaga.get.modules(c("ta_hydrology"))
rsaga.get.usage("ta_hydrology", 15)
rsaga.get.usage("ta_hydrology", 1)
rsaga.get.usage("ta_hydrology", 19)
rsaga.get.usage("ta_hydrology", 20)


filename<-c("HAGP")
res=c(16,8,4,2,1,0)
for (i in 1:6){
  if (res[i]<11){
    r0='_0'
  }else{
    r0='_'
  }

# Slope, Aspect, Curvature, Horizontal Curvature, Vertical Curvature
rsaga.geoprocessor("ta_morphometry",0,list(ELEVATION=paste(filename, '_alt', r0, res[i], ".sgrd",sep = ""),
                            SLOPE=paste(filename,"_Slo", r0, res[i], ".sgrd",sep = ""),
                            ASPECT=paste(filename,"_Asp", r0, res[i], ".sgrd",sep = ""),
                            CURV=paste(filename,"_Cu", r0, res[i], ".sgrd",sep = ""),
                            HCURV=paste(filename,"_Hcu", r0, res[i], ".sgrd",sep = ""),
                            VCURV=paste(filename,"_Vcu", r0, res[i], ".sgrd",sep = ""),
                            METHOD="5"))
#Downslope distance gradient at 5m
rsaga.geoprocessor("ta_morphometry",9,list(DEM=paste(filename, '_alt', r0, res[i], ".sgrd",sep = ""),
                             GRADIENT=paste(filename,"_DDG", r0, res[i],".sgrd",sep = ""),
                            DISTANCE="5"))


#Morphometric protection index at 1 pixel
rsaga.geoprocessor("ta_morphometry",7,list(DEM=paste(filename, '_alt', r0, res[i], ".sgrd",sep = ""),
                            PROTECTION=paste(filename,"_MPI", r0, res[i],".sgrd",sep = ""),
                            RADIUS=toString(res[i]*2)))

#Terrain ruggedness index at 1 pixel
rsaga.geoprocessor("ta_morphometry",16,list(DEM=paste(filename, '_alt', r0, res[i], ".sgrd",sep = ""),
                             TRI=paste(filename,"_TRI", r0, res[i],".sgrd",sep = ""), RADIUS=1))
```

```
#Vector ruggedness measure at 1 pixel
rsaga.geoprocessor("ta_morphometry",17,list(DEM=paste(filename, '_alt', r0, res[i], ".sgrd",sep = ""),
                            VRM=paste(filename,"_VRM", r0, res[i], ".sgrd",sep = ""),
                            RADIUS=1))

#Sky view factor and visible sky
rsaga.geoprocessor("ta_lighting",3,list(DEM=paste(filename, '_alt', r0, res[i], ".sgrd",sep = ""),
                          VISIBLE=paste(filename,"_vis", r0, res[i], ".sgrd",sep = ""),
                          SVF=paste(filename,"_SVF", r0, res[i], ".sgrd",sep = ""),
                          METHOD="1"))

#Solar radiation variables for 21 of June and 21 of December
rsaga.geoprocessor("ta_lighting",2,list(GRD_DEM=paste(filename, '_alt', r0, res[i], ".sgrd",sep = ""),
                          GRD_SVF=paste(filename,"_SVF", r0, res[i], ".sgrd",sep = ""),
                          GRD_DIRECT=paste(filename,"_Di216", r0, res[i], ".sgrd",sep = ""),
                          GRD_DIFFUS=paste(filename,"_Df216", r0, res[i], ".sgrd",sep = ""),
                          GRD_TOTAL=paste(filename,"_Ti216", r0, res[i], ".sgrd",sep = ""),
                          LATITUDE="46", PERIOD="1", DAY_A="20", MON_A="5", METHOD="0"))
rsaga.geoprocessor("ta_lighting",2,list(GRD_DEM=paste(filename, '_alt', r0, res[i], ".sgrd",sep = ""),
                          GRD_SVF=paste(filename,"_SVF", r0, res[i], ".sgrd",sep = ""),
                          GRD_DIRECT=paste(filename,"_Di2112", r0, res[i], ".sgrd",sep = ""),
                          GRD_DIFFUS=paste(filename,"_Df2112", r0, res[i], ".sgrd",sep = ""),
                          GRD_TOTAL=paste(filename,"_Ti2112", r0, res[i], ".sgrd",sep = ""),
                          LATITUDE="46", PERIOD="1", DAY_A="20", MON_A="11", METHOD="0"))

# Preprocessing of DEM for hydrology variables
rsaga.geoprocessor("ta_preprocessor",5,list(ELEV=paste(filename, '_alt', r0, res[i], ".sgrd",sep = ""),
                            FILLED=paste(filename,"_FIL", r0, res[i], ".sgrd",sep = ""),
                            MINSLOPE=0.1))

#SAGA Wetness Index
rsaga.geoprocessor("ta_hydrology",15,list(DEM=paste(filename,"_FIL", r0, res[i], ".sgrd",sep = ""),
                           C=paste(filename,"_Ca", r0, res[i], ".sgrd",sep = ""),
                           GN=paste(filename,"_CSlo", r0, res[i], ".sgrd",sep = ""),
                           CS=paste(filename,"_MCa", r0, res[i], ".sgrd",sep = ""),
                           SB=paste(filename,"_SWI", r0, res[i], ".sgrd",sep = "")))

#Total catchment area (needed for Specific catchment area)
rsaga.geoprocessor("ta_hydrology",1,list(ELEVATION=paste(filename,"_FIL", r0, res[i], ".sgrd",sep = ""),
                          CAREA=paste(filename,"_TCa", r0, res[i], ".sgrd",sep = ""),
                          Method="3"))
#Specific catchment area (needed for Topographic Wetness Index)
rsaga.geoprocessor("ta_hydrology",19,list(DEM=paste(filename,"_FIL", r0, res[i], ".sgrd",sep = ""),
                           WIDTH=paste(filename,"_FW", r0, res[i], ".sgrd",sep = ""),
                           TCA=paste(filename,"_TCa", r0, res[i], ".sgrd",sep = ""),
                           SCA=paste(filename,"_SCa", r0, res[i], ".sgrd",sep = ""),
                           METHOD="1"))
# Topographic Wetness Index
rsaga.geoprocessor("ta_hydrology",20,list(SLOPE=paste(filename,"_DDG", r0, res[i],".sgrd",sep = ""),
                           AREA=paste(filename,"_TCa", r0, res[i], ".sgrd",sep = ""),
                           TWI=paste(filename,"_TWI", r0, res[i], ".sgrd",sep = ""),
                           CONV="1",METHOD="0"))
}
```

Appendix I.b Multi-resolution computation of DEMs using a Gaussian Pyramid

```
clear all, clc
% DEM matrix of pixels
filename='C:\Data\RDN\DEM-DerivedVariables\PreProcessing\HAr0x1000_unsigned4Bytes.tif'
```

```
[OriDEM, OriR] = geotiffread(filename);
n_decomp=5;
DEM=OriDEM;
R=OriR;
for i=1:n_decomp
        gaussPyramid = vision.Pyramid('PyramidLevel', i);
        J = step(gaussPyramid, OriDEM);
        DEM=J.data;
        R.RasterSize=size(DEM);
        geotiffwrite(['HAr_' num2str(round(R.DeltaX)) '.tif'],DEM,R,'CoordRefSysCode','EPSG:26191');
        clear gaussPyramid J
        i
end
```

Appendix I.c Genetic data filtering for P.major

Initial purpose is to keep those with at least 60% coverage for now.

```
bin/./vcftools --vcf  plantago_AC6FILT.vcf  --max-missing 0.6 --recode --recode-INFO-all --out plantago_FILT.vcf
kept 5110 out of a possible 20420 Sites
Then get some statistics
#depth
bin/./vcftools --vcf plantago_FILT.vcf  --depth
#site-mean-depth
bin/./vcftools --vcf plantago_FILT.vcf  --site-mean-depth
#ld statistics
bin/./vcftools --vcf plantago_FILT.vcf --hap-r2
#heterozygosity
bin/./vcftools --vcf plantago_FILT.vcf --het
#site quality
bin/./vcftools --vcf plantago_FILT.vcf --site-quality

Then output a plink file
#output a plink file
bin/./vcftools --vcf plantago_FILT.vcf --plink plantago_FILT

In plink
#Remove individuals outside of the chosen study area
plink --file out --remove indtodel.txt --recode --out outgva479
#Keep only SNPS with MAF>0.05
plink --file outgva479 --maf 0.05 --recode --out outgva479_maf05
#Try two datasets with a maximum missingness of 0.2 or 0.1
plink --file outgva479_maf05 --geno 0.2 --recode --out gva_miss02_479
plink --file outgva479_maf05 --geno 0.1 --recode --out gva_miss01_479plink --bfile goats-filtered –missing
#Finally, keep only individuals with a maximum missingness of 50%
plink --file gva479 --mind 0.50 --recode --out gva_miss50
plink --file gva_miss50 --out gva_miss50bed --make-bed

#Compute several statistics (same statistics for sheep and goats)
plink --file gva479 --missing
plink --file gva479 --hardy
plink --file gva479 --freq
plink --file gva479 –het
```

Appendix II. **Additional results**

Appendix II.a B. laevigata

Moisture						
FCIR						
Snow						
Group1						
COORDX	COORDY = 0.93332	HAGP_alt_00 = -0.9184	HAGP_alt_01 = -0.91831	HAGP_alt_02 = -0.91822	HAGP_alt_04 = -0.91641	HAGP_alt_08 = -0.91201
minJun1528						
minJun2912						
minJul1326						
minJul2709						
miNAug2406	miNAug1023 = 0.91242					
minSep2104						
minOct0518						
maxJul1326						
maxSep2104	maxAug1023 = 0.90397	maxAug2406 = 0.9018	maxOct0518 = 0.94092			
M1pJun1528	maxJun1528 = 0.9397	meaJun1528 = 0.97113	meaJun2912 = 0.91047	RaMJun1528 = 0.97441		
meaJun2912	M1pJun1528 = 0.91047					
meaJul1326						
meaJul2709						
meaAug1023						
meaOct0518						
M1pJul1326						
M1pJul2709	M1pAug1023 = 0.90097	RaMJul2709 = 0.91531				
RaMJun2912	maxJun2912 = 0.93902	M1pJun2912 = 0.97071	RaMJul1326 = 0.92918			
RaMOct0518	M1pOct0518 = 0.92398					
HAGP_Nor_00						
HAGP_CSlo_00						
HAGP_DDG_00						
HAGP_Di216_00						
HAGP_MPI_00						
HAGP_Slo_00						
HAGP_SWI_00						
HAGP_Ti2112_00	HAGP_Eas_00 = -0.95759	HAGP_Di2112_00 = 0.99855				
HAGP_TWI_00						
HAGP_Vcu_00						
HAGP_VRM_00						
HAGP_Nor_01						
HAGP_Eas_01						
HAGP_CSlo_01	HAGP_Slo_01 = 0.93266					
HAGP_DDG_01						
HAGP_Df2112_01						
HAGP_Hcu_01						
HAGP_MPI_01						
HAGP_SWI_01						
HAGP_TWI_01						
HAGP_VRM_01						
HAGP_Nor_02						
HAGP_CSlo_02						
HAGP_DDG_02						
HAGP_Hcu_02						
HAGP_MPI_02						
HAGP_SWI_02						
HAGP_Ti2112_02	HAGP_Eas_02 = -0.9786	HAGP_Di2112_02 = 0.99689				
HAGP_TWI_02						
HAGP_VRM_02						
HAGP_Nor_04						
HAGP_Di2112_04	HAGP_Eas_04 = -0.97356	HAGP_Ti2112_04 = 0.99598				
HAGP_Hcu_04						
HAGP_MPI_04						
HAGP_TRI_04	HAGP_Cu_04 = 0.9555	HAGP_Df216_04 = -0.91729	HAGP_Df2112_04 = -0.91766	HAGP_Di216_04 = -0.93189	HAGP_Slo_04 = 0.93226	HAGP_SVF_04 = -0.91567
	HAGP_Ti216_04 = -0.93328	HAGP_Vcu_04 = 0.95104				
HAGP_TWI_04	HAGP_DDG_04 = -0.94212					
HAGP_VRM_04						
HAGP_Nor_08						
HAGP_Eas_08	HAGP_Di2112_08 = -0.95706	HAGP_Ti2112_08 = -0.94279				
HAGP_Cu_08	HAGP_Vcu_08 = 0.97237					
HAGP_Hcu_08						
HAGP_MPI_08						
HAGP_TRI_08	HAGP_Df216_08 = -0.90298	HAGP_Df2112_08 = -0.90269	HAGP_Slo_08 = 0.92886	HAGP_SVF_08 = -0.9032		
HAGP_TWI_08	HAGP_DDG_08 = -0.95033					
HAGP_VRM_08						
HAGP_Coherency_00						
HAGP_Energy_00	HAGP_Cu_00 = -0.94537	HAGP_Hcu_00 = -0.96482	HAGP_TRI_00 = 0.96307	HAGP_TRI_01 = 0.96404	HAGP_Vcu_01 = 0.90323	
HAGP_Orientation_00						
HAGP_Coherency_01						
HAGP_Energy_01	HAGP_TRI_02 = 0.91109					
HAGP_Orientation_01						
HAGP_Coherency_02						
HAGP_Energy_02						
HAGP_Orientation_02						
HAGP_Coherency_04						
HAGP_Energy_04						
HAGP_Coherency_08						
HAGP_Energy_08						
HAGP_Orientation_08						

Figure Appendix II.a.1 Correlations for selected variables in the 1st case study. Each selected variable (grey background) is shown with its highly correlated variables (>0.9).

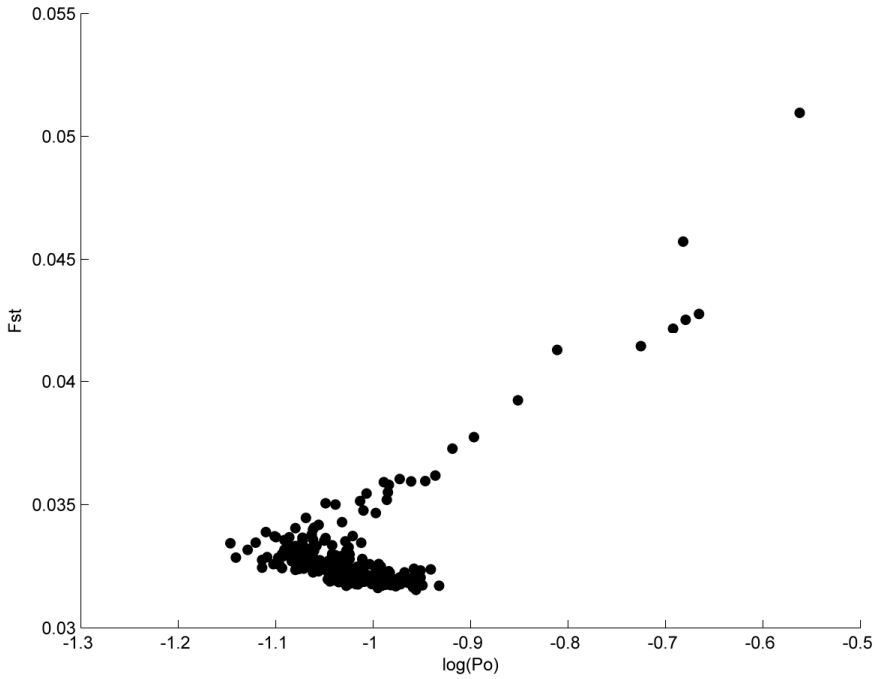


Figure Appendix II.a.2 BayeScan results for B. laevigata. The FDR threshold was set to 0.1 and corresponds to a log(PO) of 1. It is thus not shown on the graph.

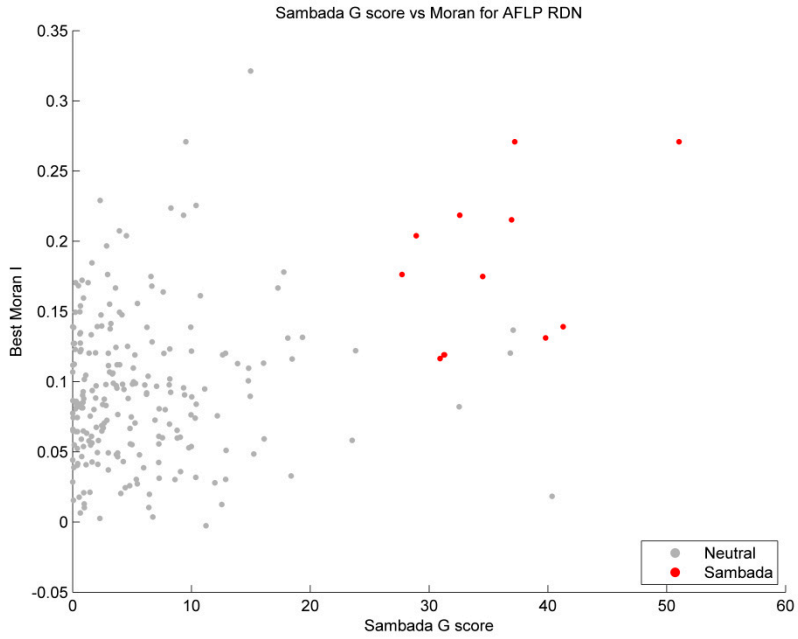


Figure Appendix II.a.2 Scatter plot showing the Moran's I (weighting scheme of 20 nearest neighbours) against Samβada's G score for ***B. laevigata****. Markers possibly under selection by environmental variables are displayed in red.*

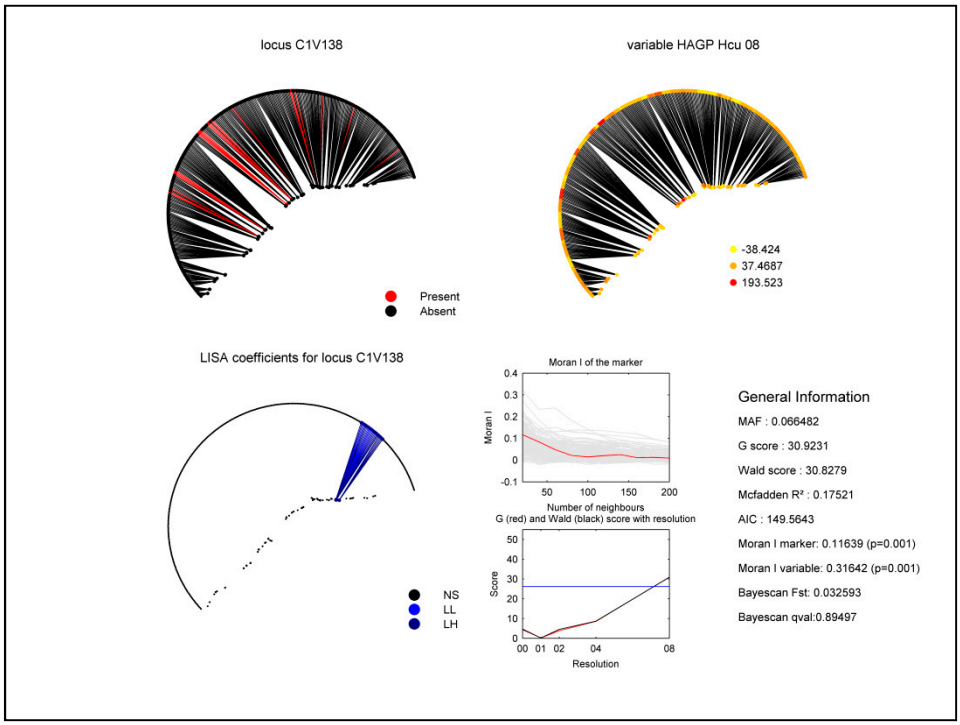


Figure Appendix II.a.3 Visualisation of the main results for the model involving marker C1V138 and variable horizontal Curvature at 8m.

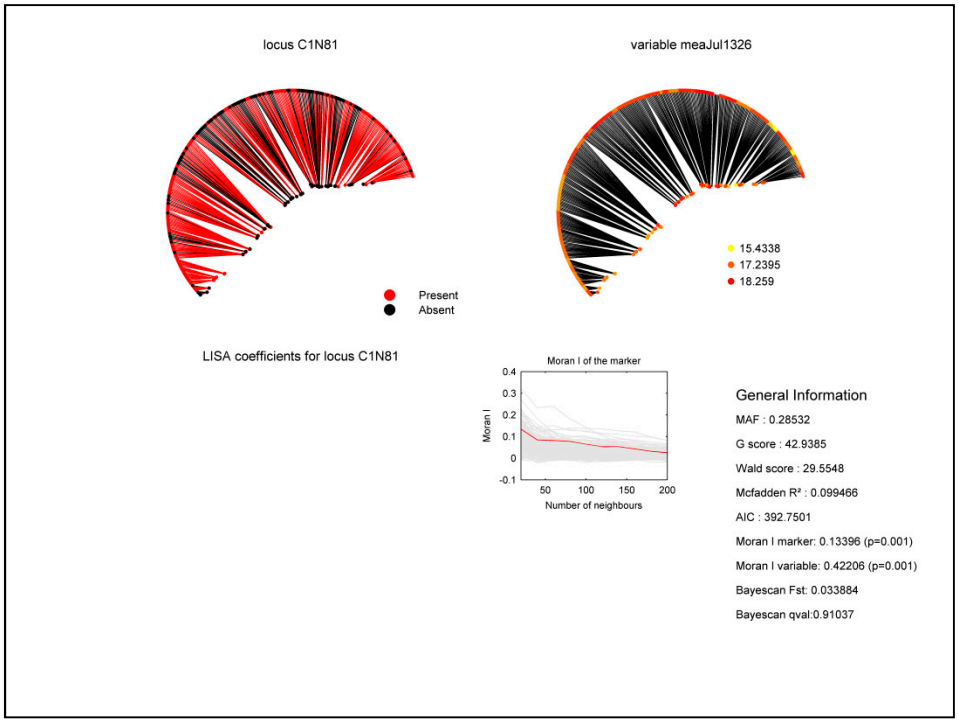


Figure Appendix II.a.4 Visualisation of the main results for the model involving marker C1N81 and mean temperature between 13 to 26 July.

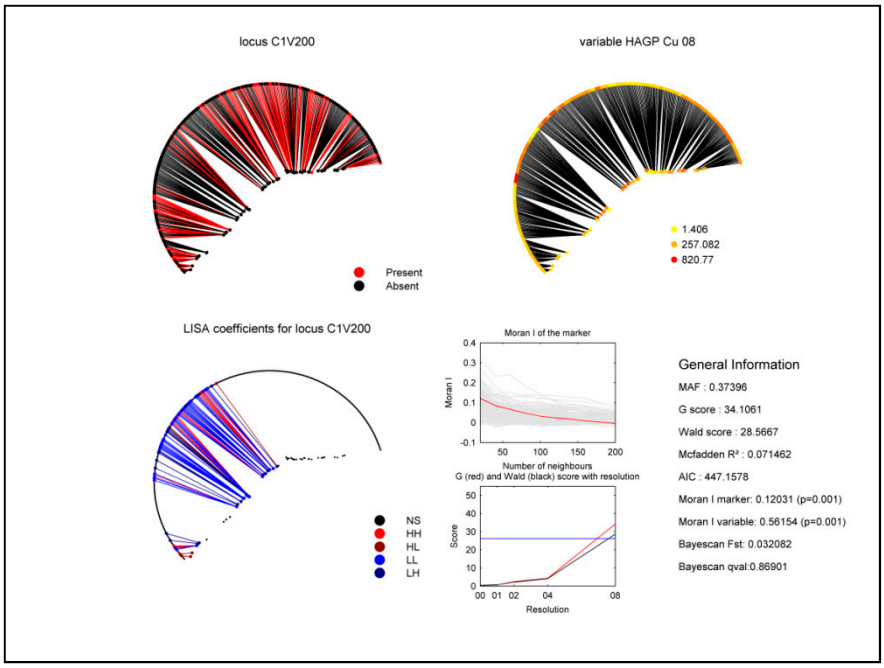


Figure Appendix II.a.5 Visualisation of the main results for the model involving marker C1V200 and variable Curvature at 8m.

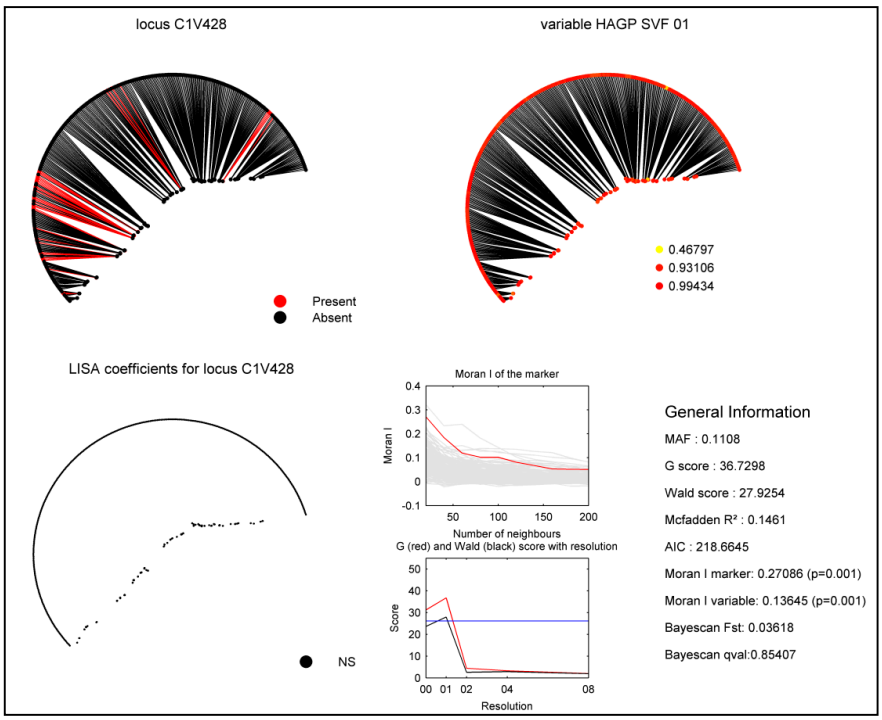


Figure Appendix II.a.6 Visualisation of the main results for the model involving marker C1V428 and Sky View Factor at 1m.

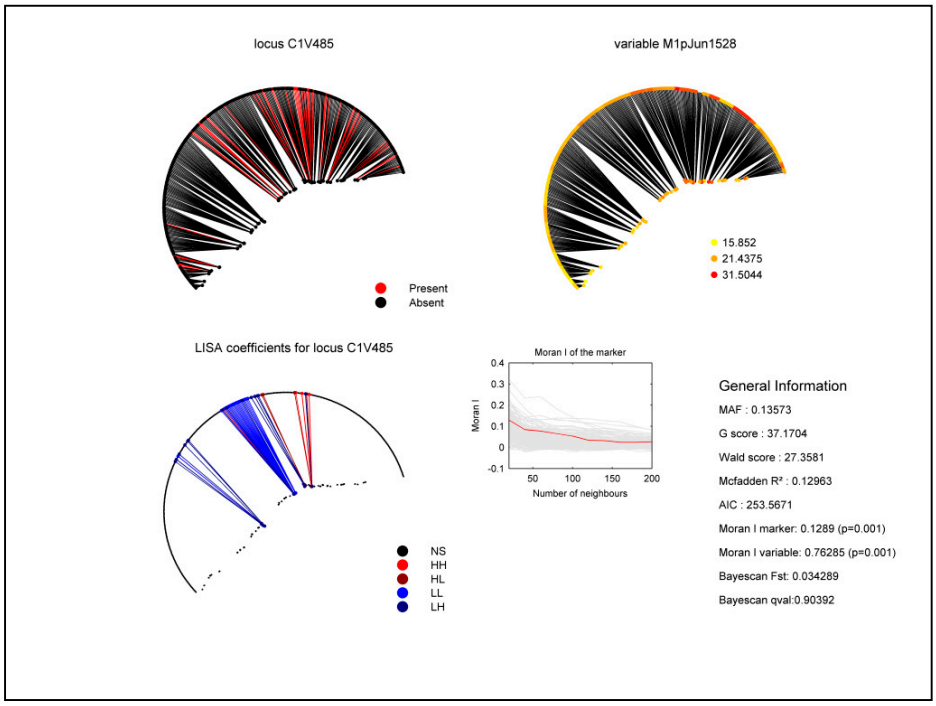


Figure Appendix II.a.7 Visualisation of the main results for the model involving marker C1V485 and variable Mean temperature at 1pm between the 15 and 28 of June.

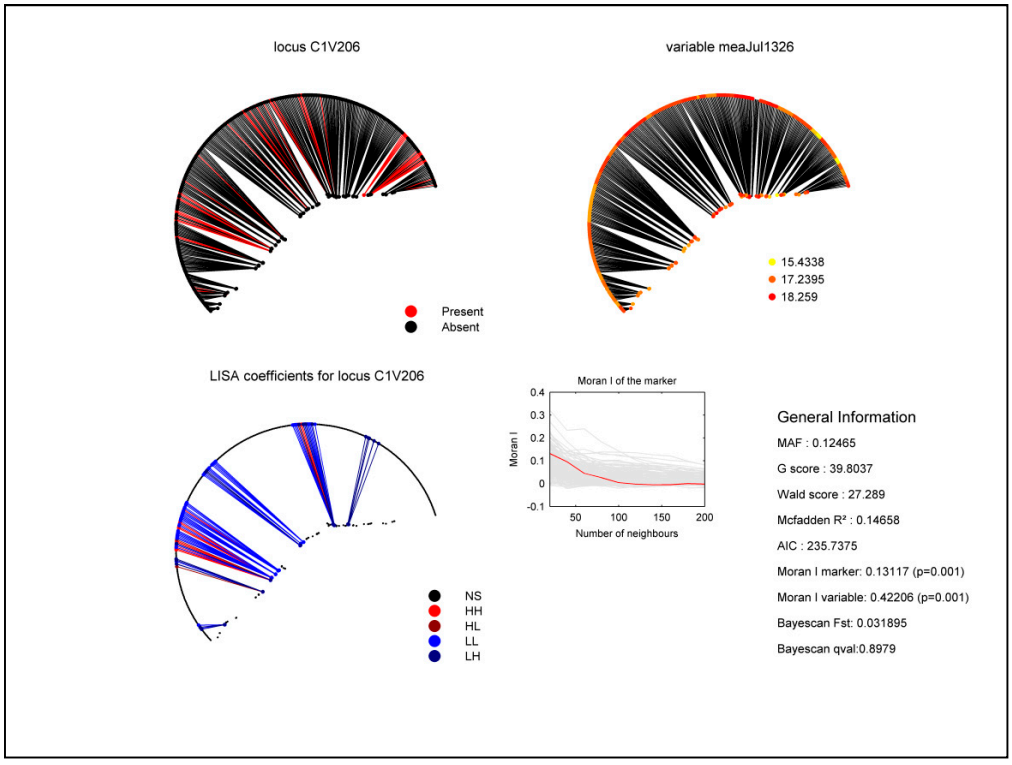


Figure Appendix II.a.8 Visualisation of the main results for the model involving marker C1V206 and variable mean temperature between the 13 and 26 of June.

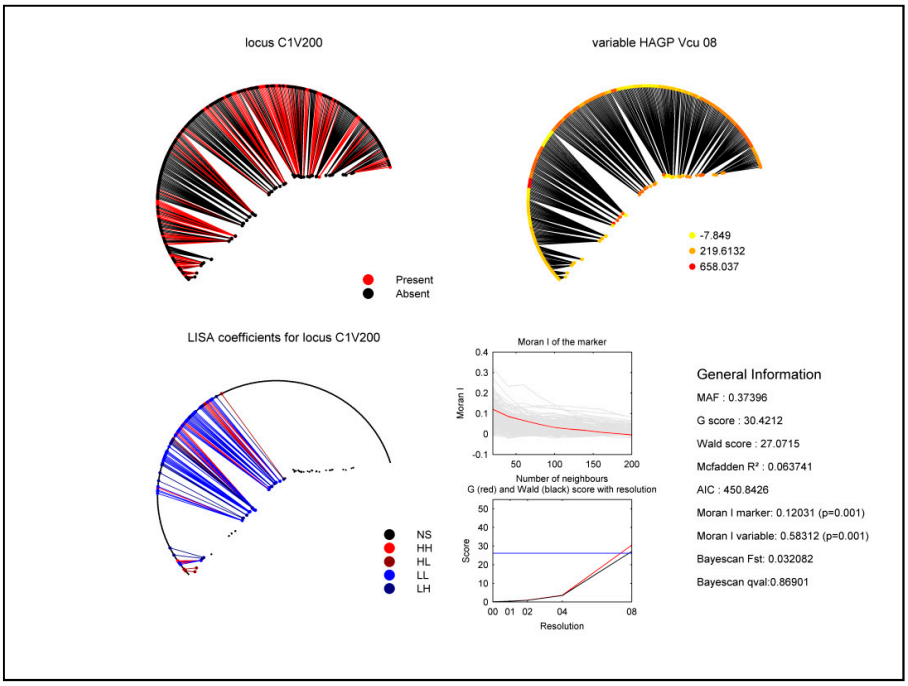


Figure Appendix II.a.9 Visualisation of the main results for the model involving marker C1V200 and variable Vertical Curvature at 8m.

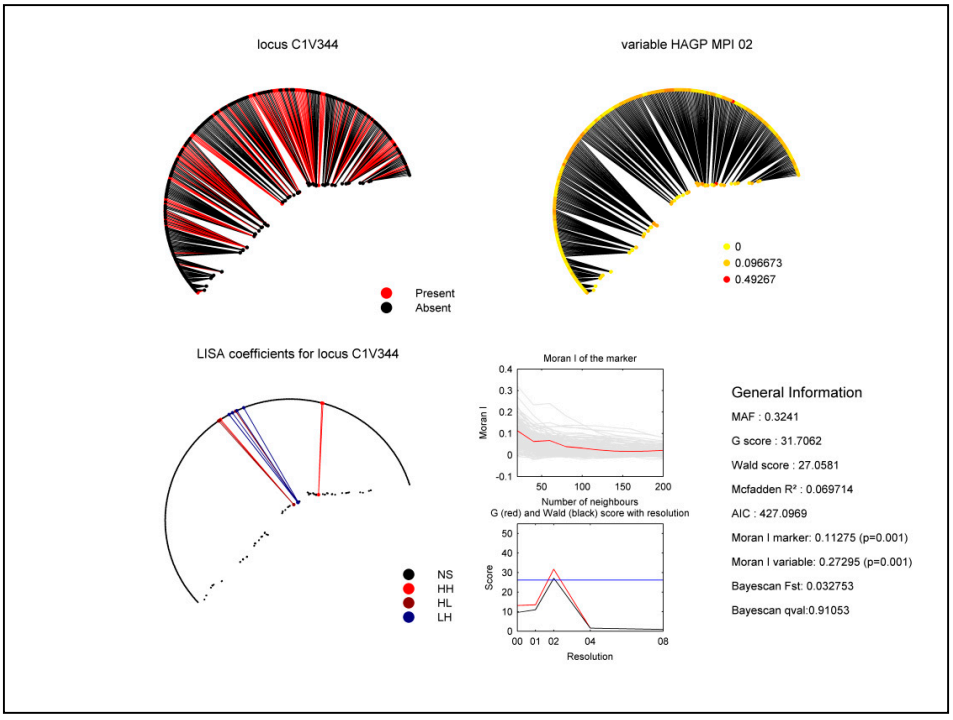


Figure Appendix II.a.10 Visualisation of the main results for the model involving marker C1V200 and variable Morphometric protection index (MPI) at 2m.

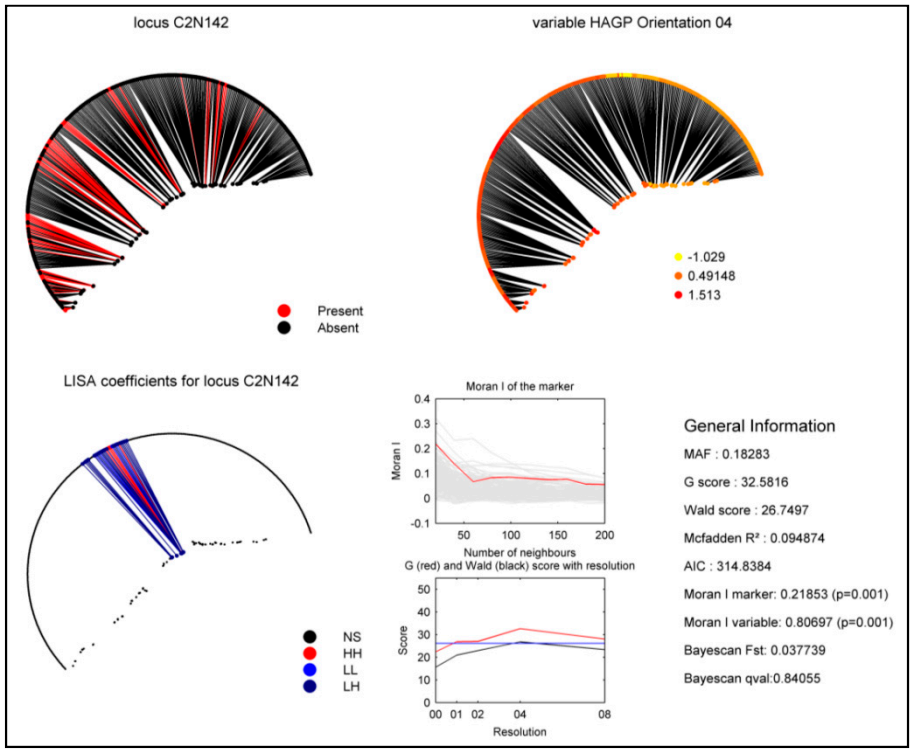


Figure Appendix II.a.11 Visualisation of the main results for the model involving marker *C2N142 and Orientation at 4m.*

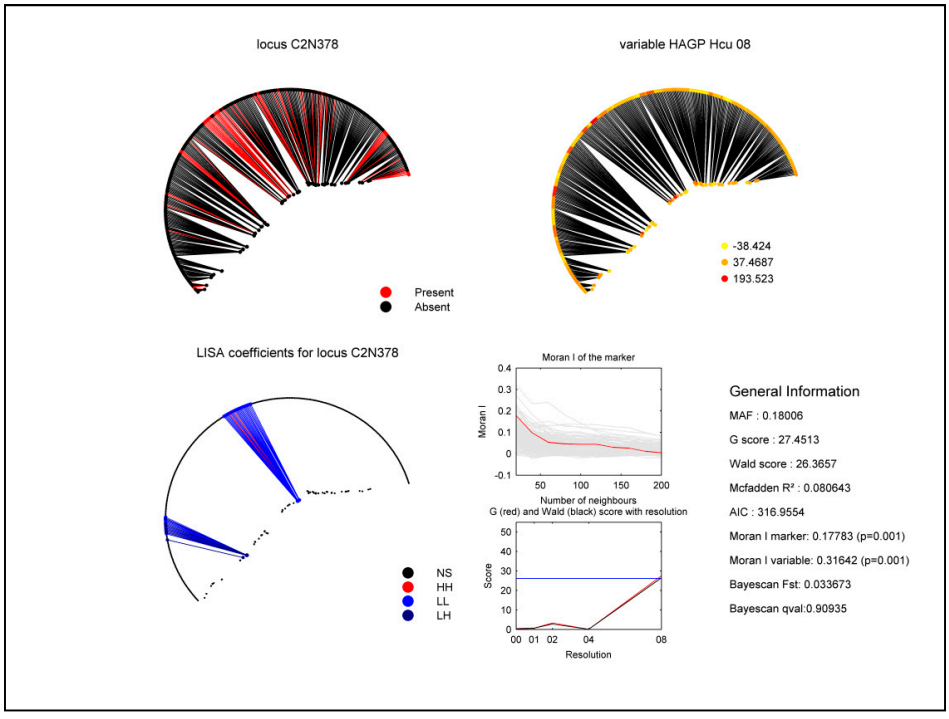


Figure Appendix II.a.12 Visualisation of the main results for the model involving marker C2N378 and variable Horizontal Curvature at 8m.

Appendix II.b P. major

Table Appendix II.b.1 Correlations for selected variables. Each selected variable (grey background) is shown with its highly correlated variables (>0.9).

X_PROJ					
Y_PROJ					
Group1					
G_gamst_ws08	G_gamst_ws00 = 0.99786	G_gamst_ws01 = 0.99858	G_gamst_ws02 = 0.999	G_gamst_ws04 = 0.99944	G_gamst_ws16 = 0.99842
G_tave12_ws16	G_tave12_ws00 = 0.98721	G_tave12_ws01 = 0.98799	G_tave12_ws02 = 0.98898	G_tave12_ws04 = 0.99106	G_tave12_ws08 = 0.99545
G_mind6_ws01	G_mbal6_ws00 = 0.96827	G_mbal6_ws01 = 0.9716	G_mbal6_ws02 = 0.96681	G_mbal6_ws04 = 0.93789	G_mbal6_ws08 = 0.90391
	G_mind6_ws00 = 0.98705	G_mind6_ws02 = 0.99589	G_mind6_ws04 = 0.96801	G_mind6_ws08 = 0.9264	
G_prec12_ws08	G_prec12_ws00 = 0.99893	G_prec12_ws01 = 0.99912	G_prec12_ws02 = 0.99926	G_prec12_ws04 = 0.99956	G_prec12_ws16 = 0.9987
	G_precyy_ws00 = 0.96681	G_precyy_ws01 = 0.96741	G_precyy_ws02 = 0.96815	G_precyy_ws04 = 0.96944	G_precyy_ws08 = 0.97197
	G_precyy_ws16 = 0.97562				
Coherency_04					
Orientation_04					
G_Nor_02					
G_Eas_02					
G_CSlo_02					
G_Cu_02					
G_DDG_02					
G_Nor_04					
G_Eas_04					
G_CSlo_04					
G_Cu_04					
G_DDG_04					
G_Di216_04	G_Di2112_04 = 0.90081	G_Ti216_04 = 0.9941	G_Ti2112_04 = 0.91238		
G_Nor_08					
G_Eas_08					
G_CSlo_08	G_CSlo_16 = 0.91804				
G_Cu_08					
G_DDG_08					
G_Nor_16					
G_Eas_16					
G_Cu_16					
G_DDG_16					
G_alt_32	G_ddeg300_ws00 = -0.97615	G_ddeg300_ws01 = -0.97663	G_ddeg300_ws02 = -0.97622	G_ddeg300_ws04 = -0.97488	G_ddeg300_ws08 = -0.96933
	G_ddeg300_ws16 = -0.95205	G_mbal6_ws08 = 0.90254	G_prec6_ws00 = 0.92655	G_prec6_ws01 = 0.92669	G_prec6_ws02 = 0.92664
	G_prec6_ws04 = 0.92543	G_prec6_ws08 = 0.91716	G_alt_02 = 0.99814	G_alt_04 = 0.99805	G_alt_08 = 0.99836
	G_alt_16 = 0.99817	G_Df2112_16 = 0.91138	G_Df216_32 = 0.91701	G_Df2112_32 = 0.93156	G_alt_64 = 0.99668
	G_Df216_64 = 0.93151	G_Df2112_64 = 0.94679			
G_Nor_32					
G_Eas_32					
G_CSlo_32	G_CSlo_16 = 0.90215				
G_Cu_32					
G_DDG_32					
G_Di216_32	G_Di216_16 = 0.91203	G_Di2112_32 = 0.90598	G_Di216_64 = 0.91826	G_Ti216_16 = 0.90869	G_Ti216_32 = 0.99545
	G_Ti2112_32 = 0.91559	G_Ti216_64 = 0.92147			
G_Nor_64					
G_Eas_64					
G_CSlo_64					
G_Cu_64					
G_DDG_64					
G_Di2112_64	G_Di2112_32 = 0.91437	G_Di216_64 = 0.92361	G_Ti2112_32 = 0.9107	G_Ti2112_64 = 0.99891	
G_Hcu_02					
G_MPI_02	G_Slo_02 = 0.92638	G_TRI_02 = 0.92089			
G_SVF_02	G_SVF_04 = 0.95733				
G_Hcu_04					
G_MPI_04					
G_Hcu_08					
G_MPI_08					
G_Slo_08	G_TRI_08 = 0.97952				
G_SVF_08	G_SVF_04 = 0.90962	G_SVF_16 = 0.93289			
G_Hcu_16					
G_MPI_16					
G_Hcu_32					
G_MPI_32					

G_SVF_32	G_SVF_16 = 0.95655	G_SVF_64 = 0.92082			
G_Hcu_64					
G_MCa_64	G_MCa_32 = 0.91825	G_SWI_64 = 0.92271			
G_MPI_64					
G_TCa_02	G_TWI_02 = 0.91708				
G_Ti2112_02	G_Di2112_02 = 0.99905	G_Di2112_04 = 0.90308	G_Ti2112_04 = 0.90398		
G_Vcu_02					
G_VRM_02					
G_TCa_04	G_TWI_04 = 0.91715				
G_TRI_04	G_Slo_02 = 0.90572	G_Slo_04 = 0.97527	G_TRI_02 = 0.92375		
G_Vcu_04					
G_VRM_04					
G_SWI_08	G_MCa_04 = 0.92313	G_MCa_08 = 0.98505	G_MCa_16 = 0.90209	G_SWI_04 = 0.9201	G_SWI_16 = 0.93469
G_TCa_08	G_TWI_08 = 0.92325				
G_Ti216_08	G_Di216_08 = 0.9936	G_Ti2112_08 = 0.91101			
G_Vcu_08					
G_VRM_08					
G_TCa_16	G_TWI_16 = 0.90885				
G_Ti2112_16	G_Di216_16 = 0.9148	G_Di2112_16 = 0.99887	G_Di2112_32 = 0.90967	G_Ti2112_32 = 0.91107	
G_TRI_16	G_Slo_16 = 0.98469				
G_Vcu_16					
G_VRM_16					
G_TCa_32					
G_TRI_32	G_Slo_32 = 0.98116				
G_TWI_32					
G_Vcu_32					
G_VRM_32					
G_TCa_64					
G_TRI_64	G_Slo_64 = 0.98106				
G_TWI_64					
G_Vcu_64					
G_VRM_64					

Table Appendix II.b.2 List of significant genotypes (left) and SNPs (right) detected in sambada. Columns correspond to the type of variable involved, either genotypes detected by geographic or population structure only (left), environmental variables (centre) or both (right).

Genotypes				SNPS			
	Total XYG Genotypes	Total ENV Genotypes	Common Genotypes		Total XYG SNPs	Total ENV SNPs	Common SNPs
	97 - (21 duplicate)	203 - (169 duplicate)			97 - (56 duplicate)	203 - (178	
	53	11	23		24	8	17
	MC00247904:97_AA	MC00185267:124_AG	MC03993697:21_CC		MC01041748:50	MC00185267:124	MC00247904:97
	MC00247904:97_CC	MC00827814:20_AC	MC05778243:40_AA		MC01367811:82	MC00827814:20	MC03993697:21
	MC01041748:50_CC	MC01643098:103_CT	MC06309558:73_CC		MC01467732:35	MC01643098:103	MC05778243:40
	MC01041748:50_TT	MC01643098:103_TT	MC06309558:73_CT		MC01570173:97	MC02001781:113	MC06309558:73
	MC01367811:82_AA	MC02001781:113_AG	MC06681055:73_AC		MC01591310:36	MC05667273:31	MC06681055:73
	MC01367811:82_AG	MC05667273:31_AA	MC06681055:73_CC		MC01643092:55	MC05966531:14	MC06929001:75
	MC01367811:82_GG	MC05667273:31_AG	MC06929001:75_CG		MC01775189:98	MC06085250:39	MC02192085:71
	MC01467732:35_CC	MC05966531:14_CT	MC06929001:75_GG		MC01915554:13	MC06784148:25	MC01276423:101
	MC01467732:35_TT	MC06085250:39_AG	MC02192085:71_AA		MC03957223:72		MC01301563:54
	MC01570173:97_AA	MC06784148:25_GT	MC00247904:97_AC		MC04640598:94		MC01585839:73
	MC01570173:97_GG	MC06784148:25_TT	MC01276423:101_CT		MC05872911:140		MC01651581:47
	MC01591310:36_CC		MC01301563:54_GT		MC06302512:53		MC01834138:78
	MC01591310:36_CT		MC01585839:73_CC		MC06533076:78		MC01883648:43
	MC01591310:36_TT		MC01585839:73_CT		MC06965223:91		MC02223852:142
	MC01643092:55_CC		MC01651581:47_AT		MC01068585:101		MC03690815:99
	MC01643092:55_CT		MC01834138:78_CT		MC01272757:76		MC03895169:134
	MC01643092:55_TT		MC01883648:43_GG		MC01510673:114		MC06698177:33
	MC01775189:98_CC		MC02223852:142_CG		MC01576512:22		
	MC01775189:98_CT		MC03690815:99_AA		MC01768188:74		
	MC01915554:13_AA		MC03690815:99_AT		MC02001843:31		
	MC01915554:13_CC		MC03895169:134_CT		MC02099341:85		
	MC03957223:72_CC		MC03993697:21_CG		MC02206548:70		
	MC03957223:72_CT		MC06698177:33_AG		MC03464566:154		
	MC03957223:72_TT				MC05277703:114		
	MC04640598:94_CC						
	MC04640598:94_TT						
	MC05778243:40_CC						
	MC05872911:140_AA						
	MC05872911:140_AG						
	MC05872911:140_GG						
	MC06302512:53_TT						
	MC06533076:78_AA						
	MC06533076:78_AG						
	MC06533076:78_GG						
	MC06965223:91_CC						
	MC06965223:91_CT						
	MC06965223:91_TT						
	MC01068585:101_GG						
	MC01068585:101_GT						
	MC01272757:76_GT						
	MC01272757:76_TT						
	MC01510673:114_GG						
	MC01576512:22_GT						
	MC01768188:74_CT						
	MC02001843:31_AG						
	MC02099341:85_AG						
	MC02206548:70_GG						
	MC02206548:70_GT						
	MC03464566:154_CG						
	MC04640598:94_CT						
	MC05277703:114_AA						
	MC05277703:114_AT						
	MC05778243:40_AC						

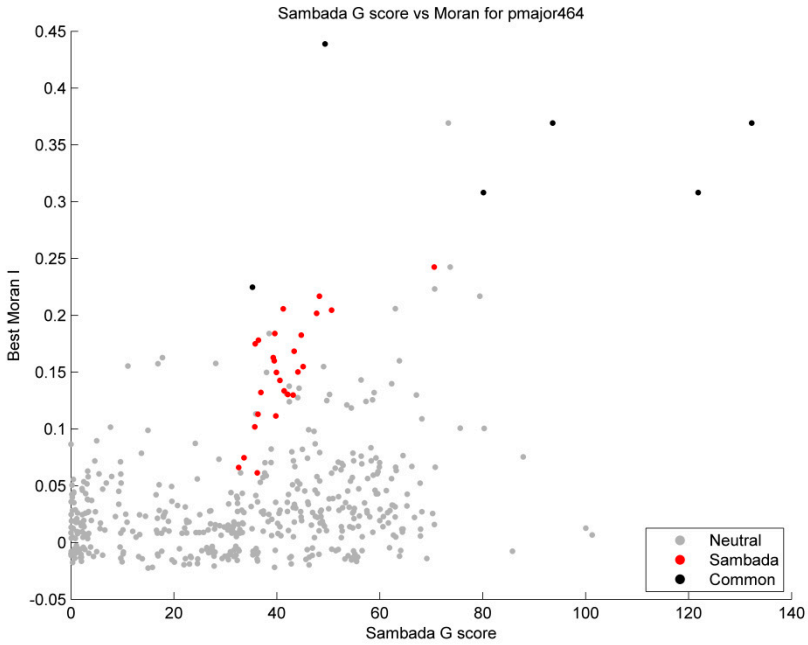


Figure Appendix II.b.1 Scatter plot showing the Moran's I (weighting scheme of 20 nearest neighbours) against Sambada's G score for ***P. major****. Markers possibly under selection by environmental variables are displayed in red.*

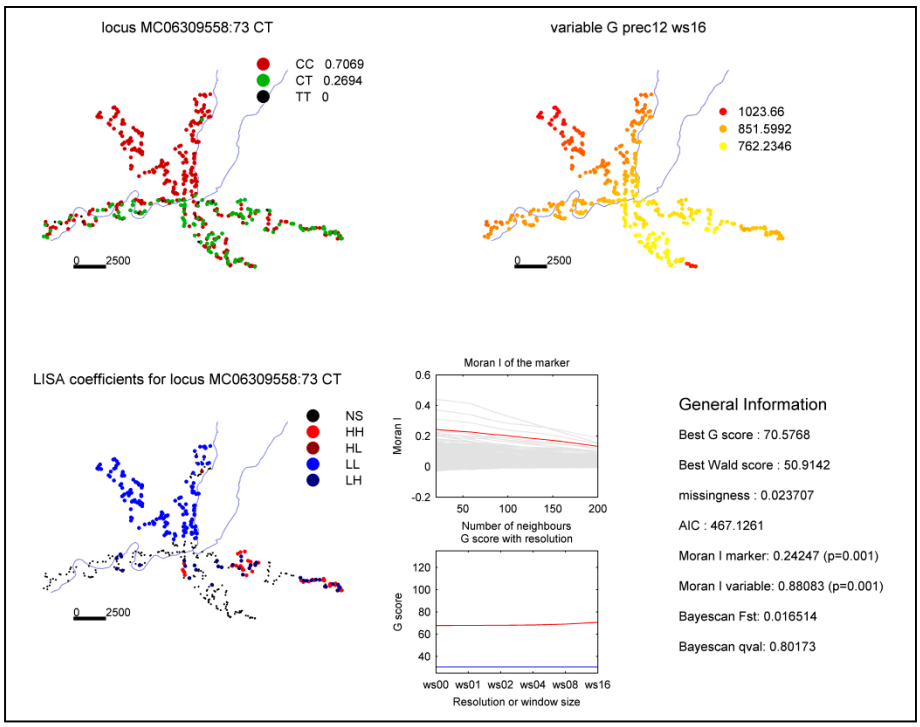


Figure Appendix II.b.1 Visualisation of the main results for the model involving genotype MC06309558:73_CT and precipitation in December.

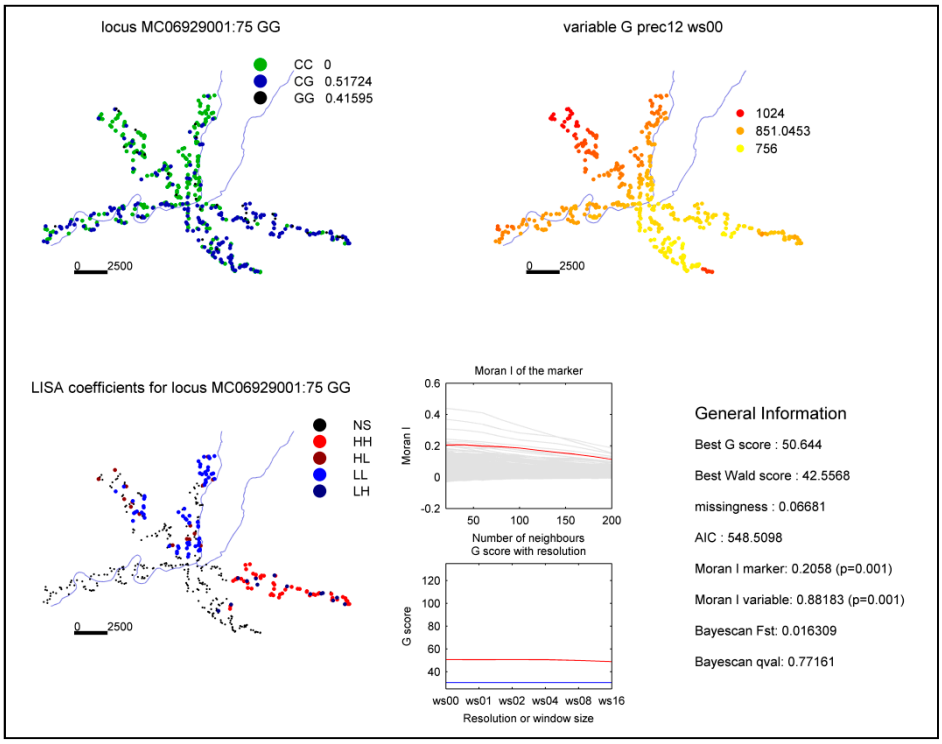


Figure Appendix II.b.2 Visualisation of the main results for the model involving genotype MC06929001:75_GG and precipitation in December.

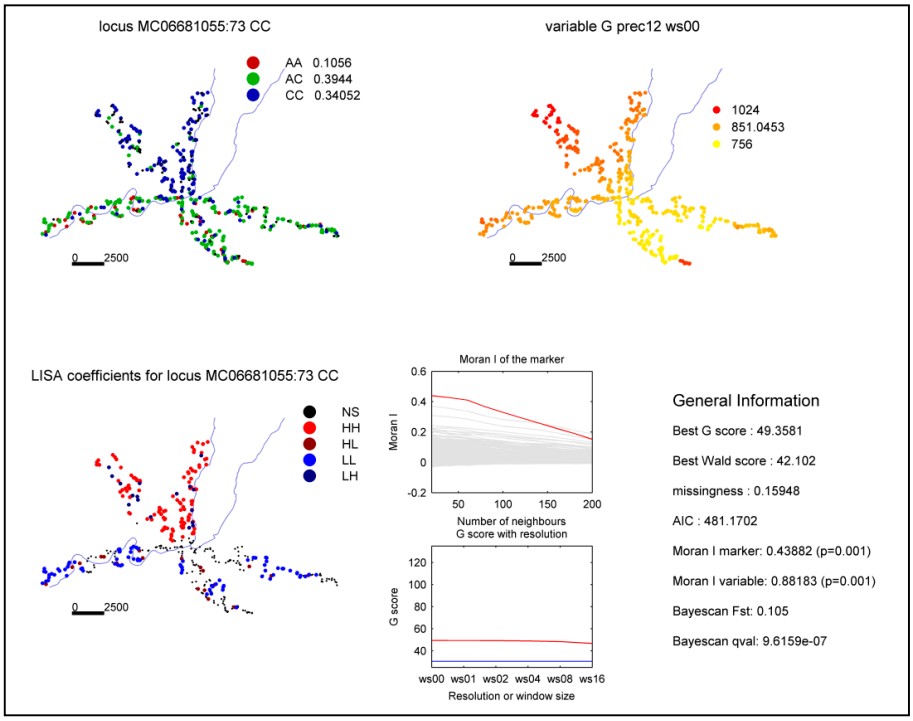


Figure Appendix II.b.3 Visualisation of the main results for the model involving genotype MC06681055:73_CC and precipitation in December.

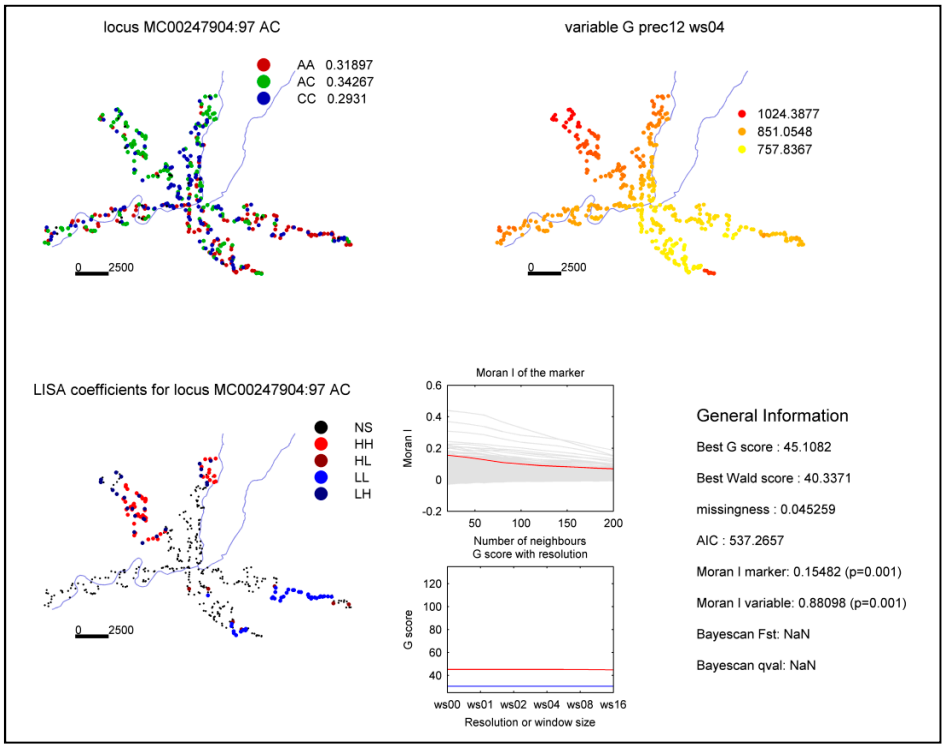


Figure Appendix II.b.4 Visualisation of the main results for the model involving genotype MC00247904:97_AC and precipitation in December.

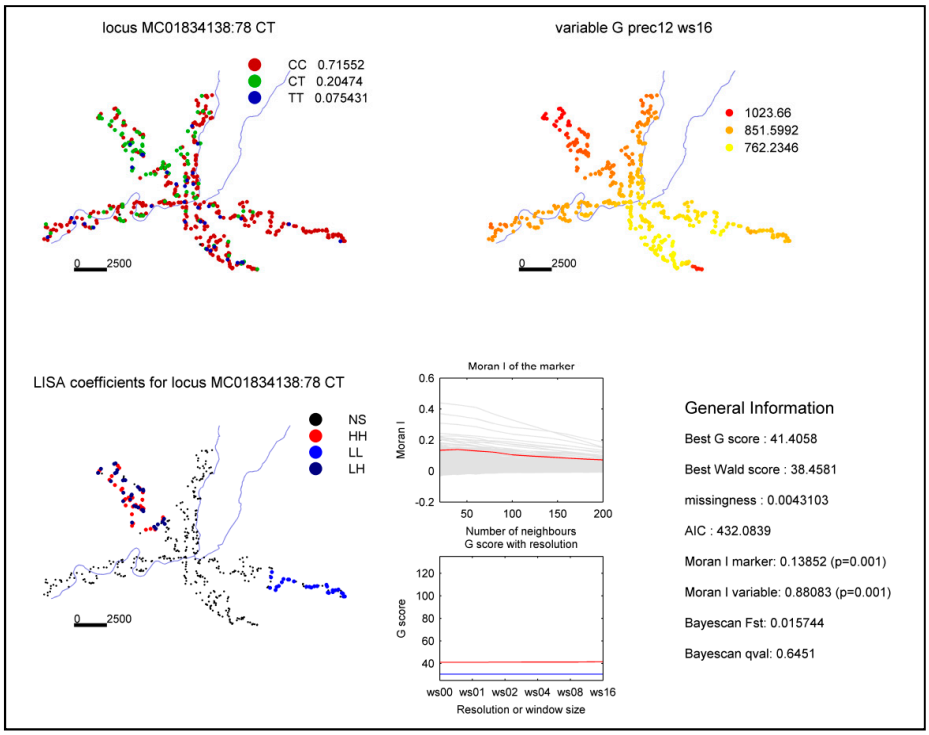


Figure Appendix II.b.5 Visualisation of the main results for the model involving genotype MC01834138:78_CT and precipitation in December.

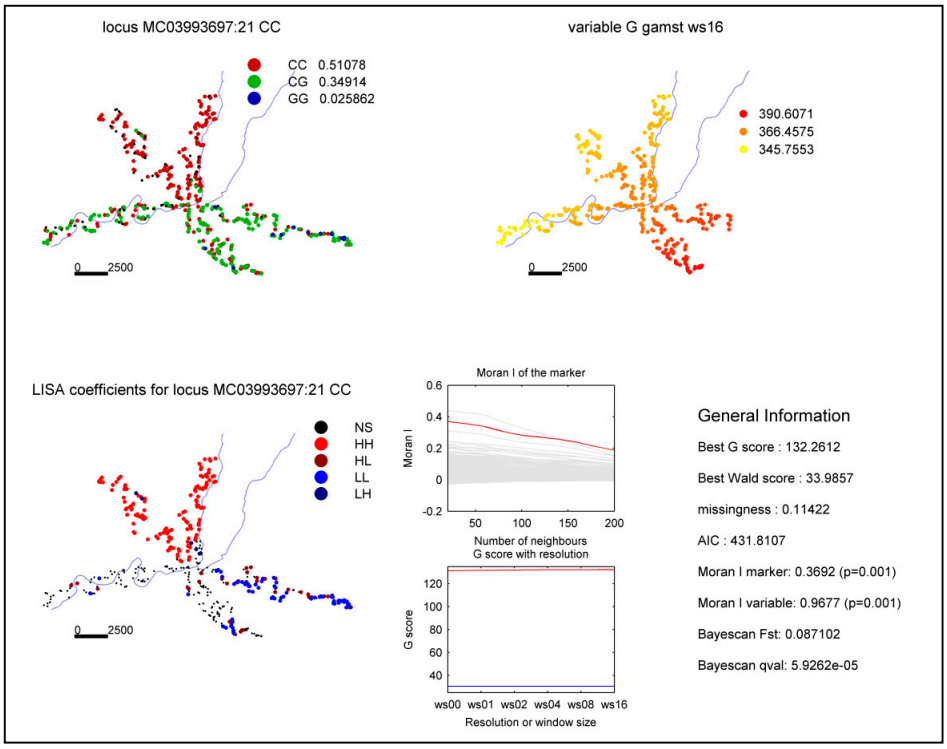


Figure Appendix II.b.6 Visualisation of the main results for the model involving genotype MC03993697:21_CC and precipitation in December.

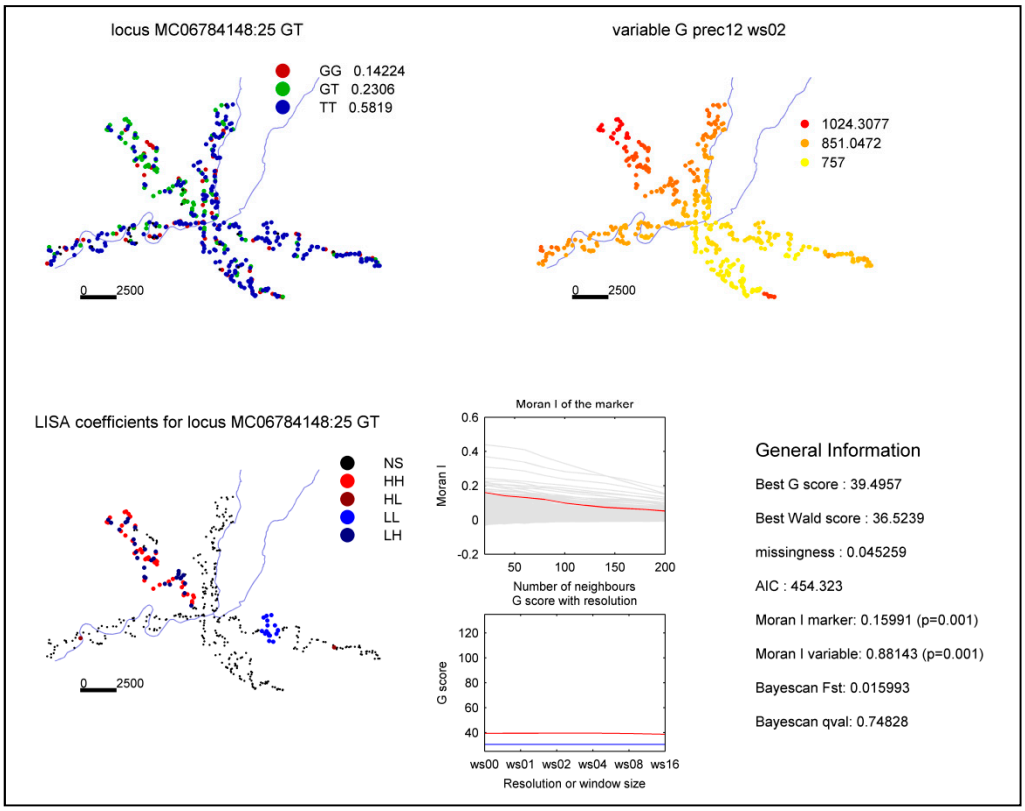


Figure Appendix II.b.7 Visualisation of the main results for the model involving genotype MC06784148:25_GT and precipitation in December.

Appendix II.c Sheep & Goats

*Table Appendix II.c.1 Selected and deleted variables for **sheep** after the selection procedure.*

Selected variable	Correlated variables for a 0.9 threshold (deleted)								
SRTM	tmin_12	tmin_11	tmin_3	tmin_2	tmin_1	tmean_12	tmean_1	bio_11	bio_6
nor									
eas									
Cu									
PlCu									
PrCu									
VRM									
CSlo	SVF	TWI	Energy200						
TI216	Di216	Df216							
TI2112	Di2112	Df2112							
Orientation200									
Coherency200									
tmin_5	tmin_9	tmin_8	tmin_7	tmin_6					
tmin_4	tmin_11	tmin_10	tmin_9	tmin_3	tmin_2	tmean_12	tmean_11	tmean_10	tmean_4
	tmean_3	tmean_2	tmean_1	tmax_12	tmax_11	bio_11	bio_1		
tmean_9	tmean_6	tmax_10	bio_1						
tmean_5	tmin_8	tmin_7	tmean_6						
tmax_9	tmean_8	tmean_7	tmax_6	tmax_5	bio_10				
tmax_8	tmean_7	tmax_7	tmax_6	bio_5					
tmax_4	tmax_10	tmax_5	tmax_3						
prec_10	prec_11	prec_3	bio_16	bio_13	bio_12				
prec_9	tmax_2	tmax_1	bio_18	bio_17					
prec_8									
prec_6	prec_5	bio_17							
prec_4	prec_5	prec_3	prec_2	bio_12					
bio_15									
bio_14	prec_7	bio_17							
bio_9									
bio_8									
bio_7	bio_4	bio_2							
bio_3									

*Table Appendix II.c.2 Selected and deleted variables for **goats** after the selection procedure.*

Selected variable	Correlated variables (deleted)								
SRTM	tmin_12	tmin_11	tmin_3	tmin_2	tmin_1	tmean_12	tmean_1	bio_11	bio_6
nor									
eas									
Slo	VRM	TWI							
Cu									
PlCu									
PrCu									
SVF	CSlo	Energy200							
TI216	Di216	Df216							
TI2112	Di2112	Df2112							
Orientation200									
Coherency200									
tmin_8	tmin_9	tmin_7	tmin_6	tmin_5	tmean_5	bio_1			
tmean_6	tmean_9	tmean_8	tmean_7	tmean_5	tmax_5	bio_10			
tmean_2	tmin_11	tmin_10	tmin_9	tmin_4	tmin_3	tmin_2	tmean_12	tmean_11	
	tmean_10	tmean_4	tmean_3	tmean_1	tmax_12	tmax_11	tmax_2	tmax_1	
	prec_9	bio_17	bio_11	bio_1					
tmax_10	tmean_10	tmean_9	tmean_4	tmax_11	tmax_4	tmax_3	tmax_2	tmax_1	bio_1
tmax_7	tmax_8	tmax_6	bio_5						
prec_8									
prec_6	prec_5	bio_17							
prec_1	prec_12	prec_11	prec_4	prec_3	prec_2	bio_19	bio_16	bio_13	bio_12
bio_15									
bio_14	prec_7	bio_17							
bio_9									
bio_8									
bio_7	bio_4	bio_2							
bio_3									

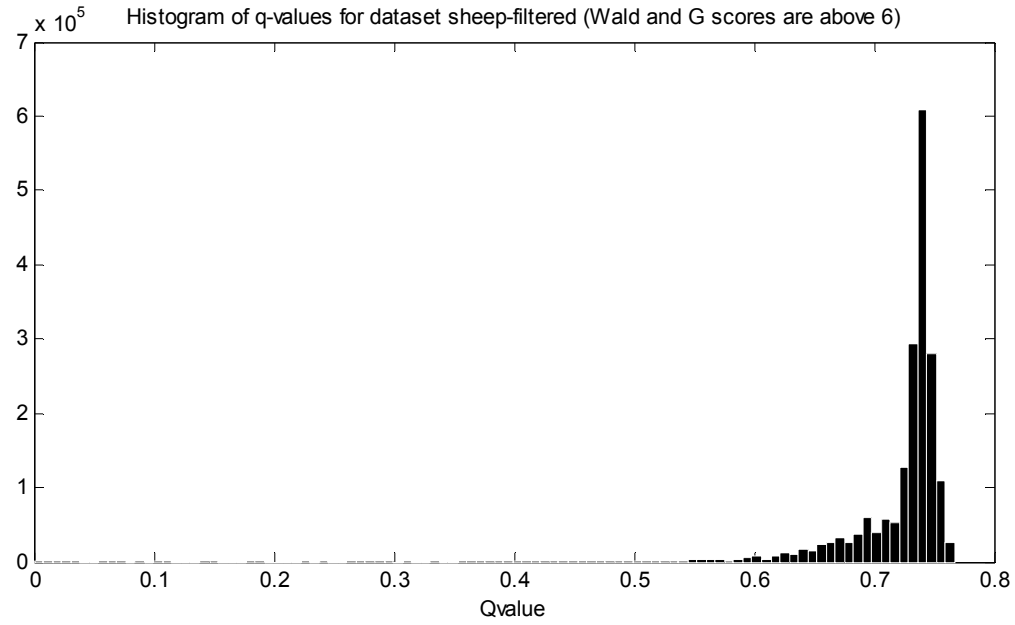


Figure Appendix II.c.1 Distribution of Q-values from Samβada for sheep

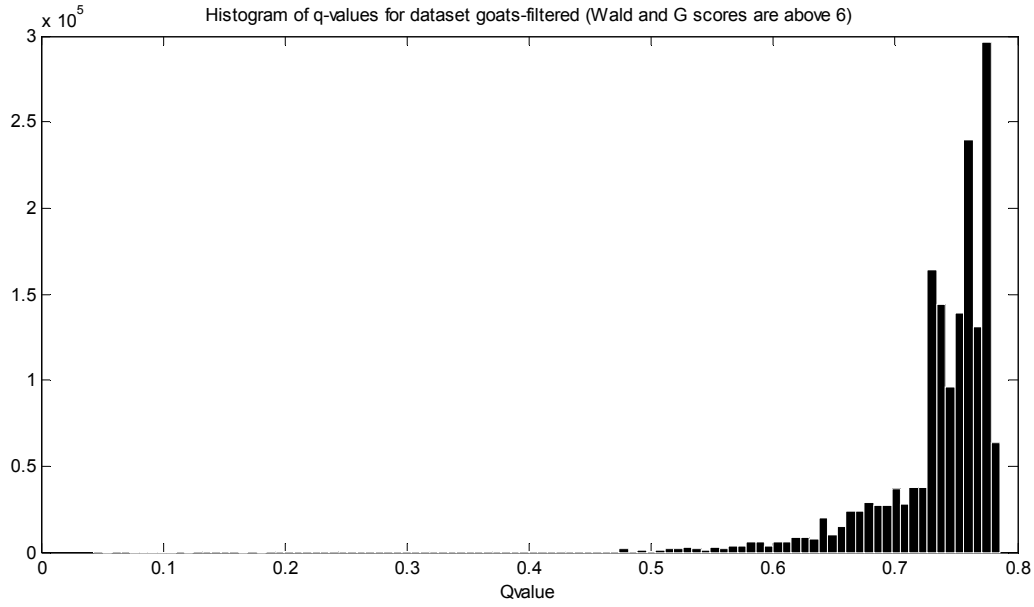


Figure Appendix II.c.1 Distribution of Q-values from Samβada for goats

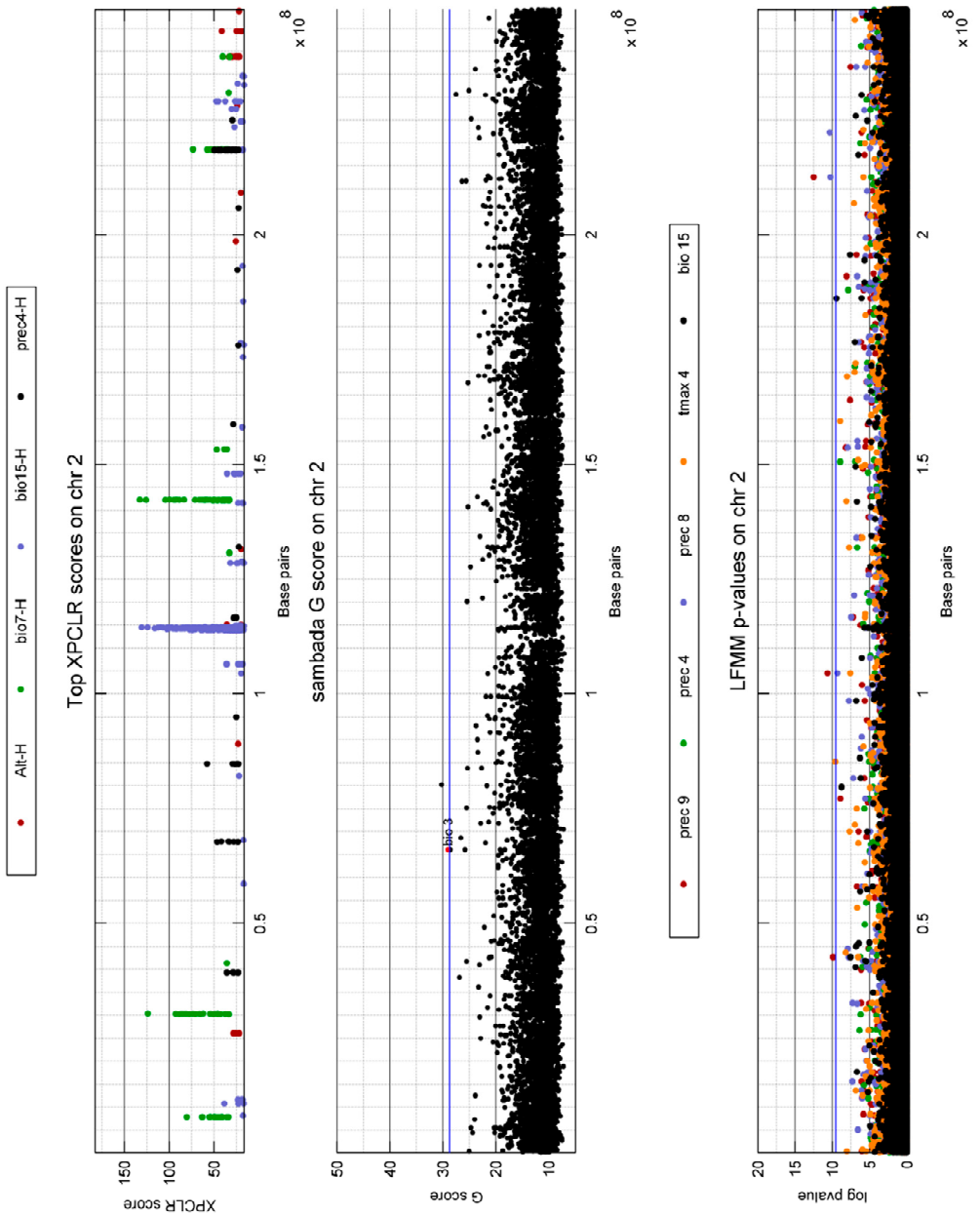


*Figure Appendix II.c.2 Comparison between significant XP-CLR, Sambada and LFMM results according to their position on chromosome 2 for **sheep**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Sambada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

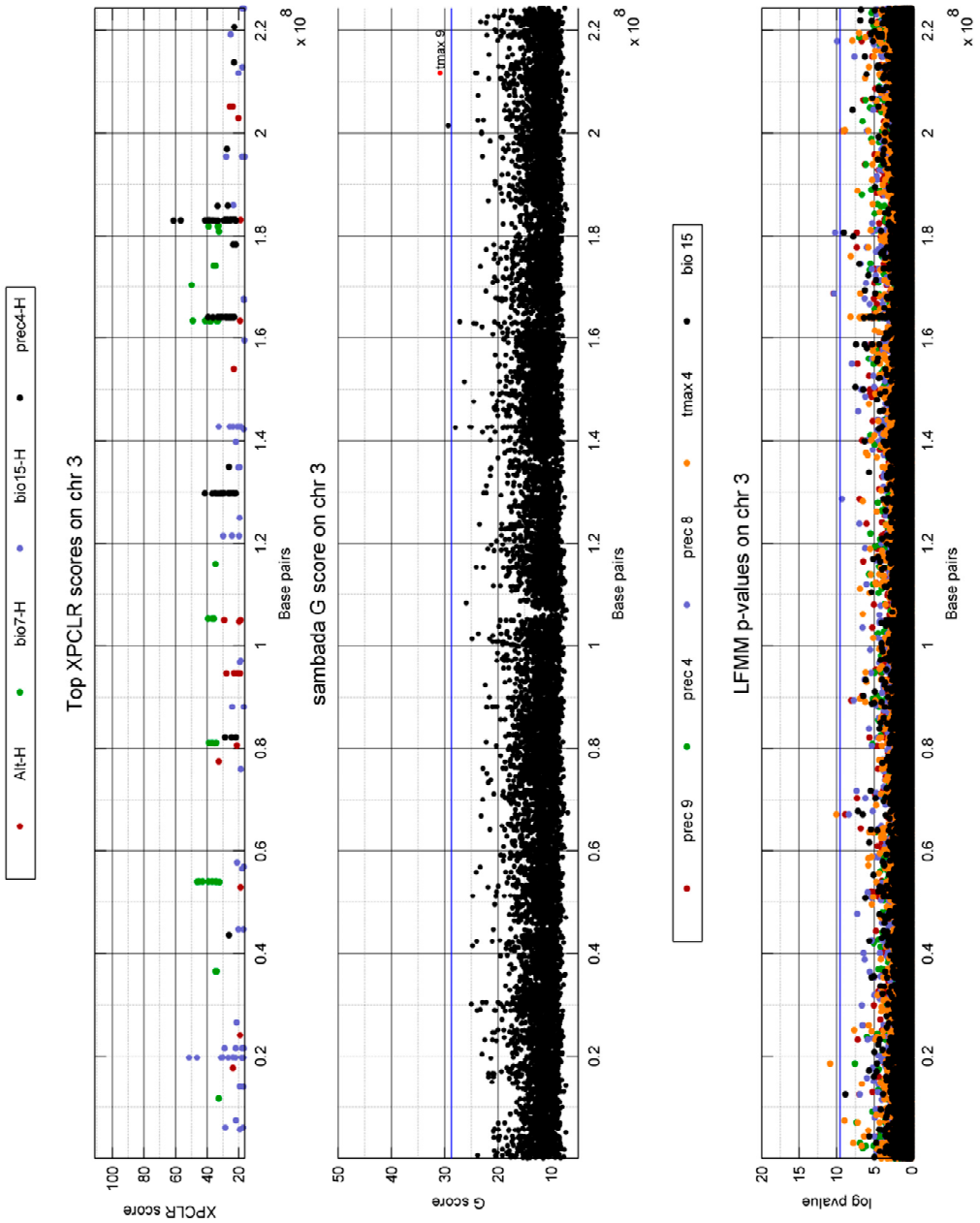


*Figure Appendix II.c.3 Comparison between significant XP-CLR, Samβada and LFMM results according to their position on chromosome 3 for **sheep**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Samβada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

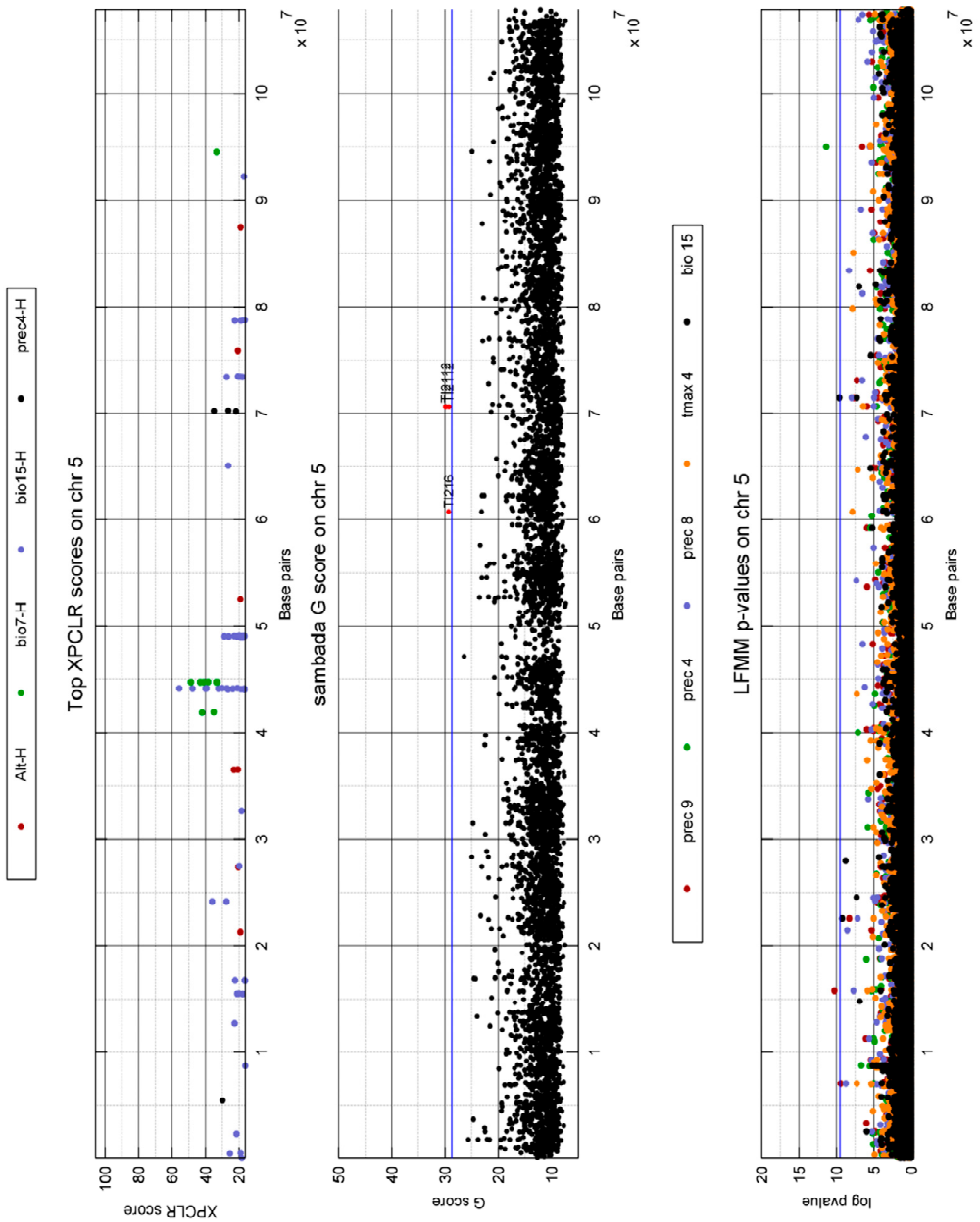


*Figure Appendix II.c.4 Comparison between significant XP-CLR, Samβada and LFMM results according to their position on chromosome 5 for **sheep**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Samβada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

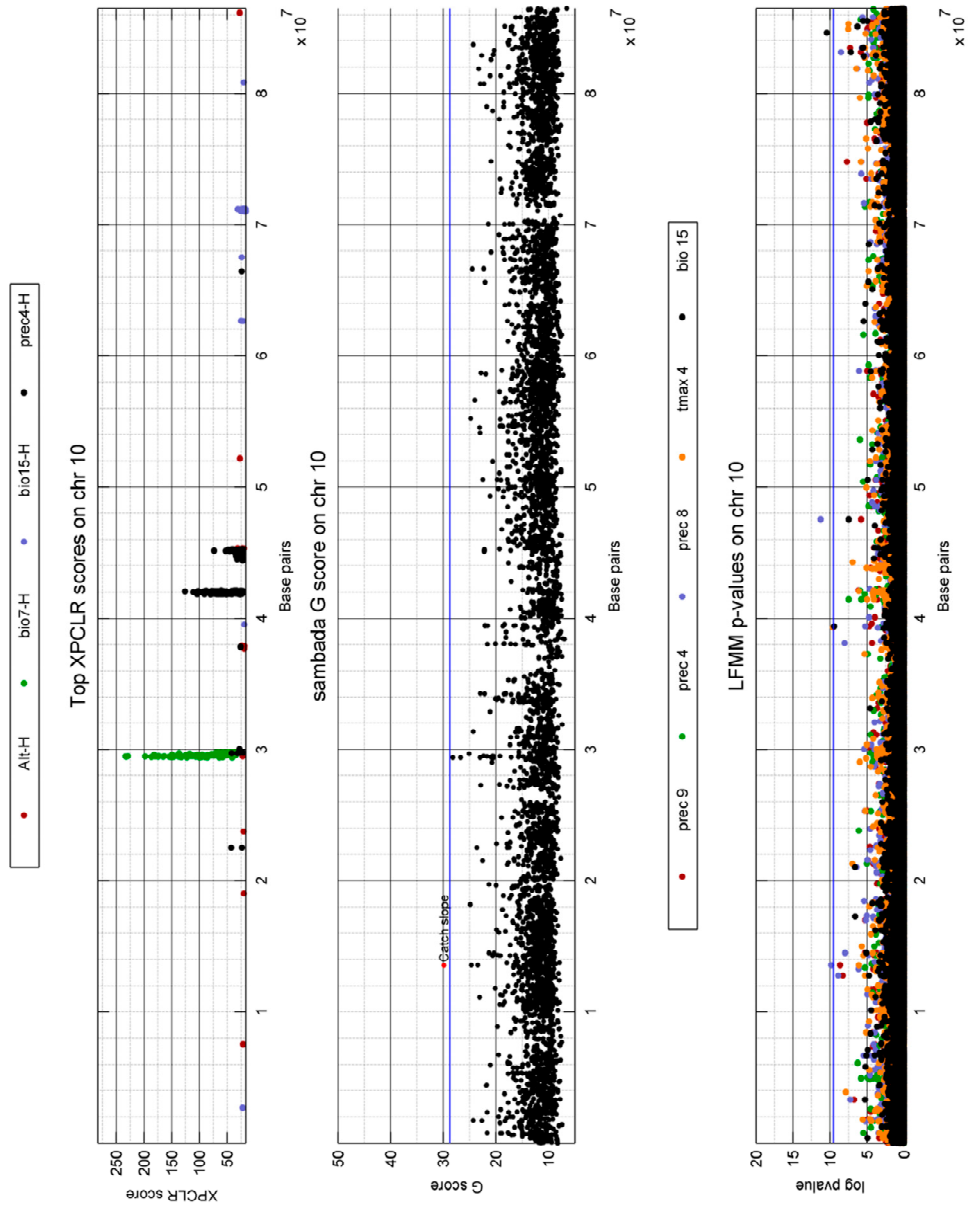


*Figure Appendix II.c.5 Comparison between significant XP-CLR, Sambada and LFMM results according to their position on chromosome 10 for **sheep**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Sambada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

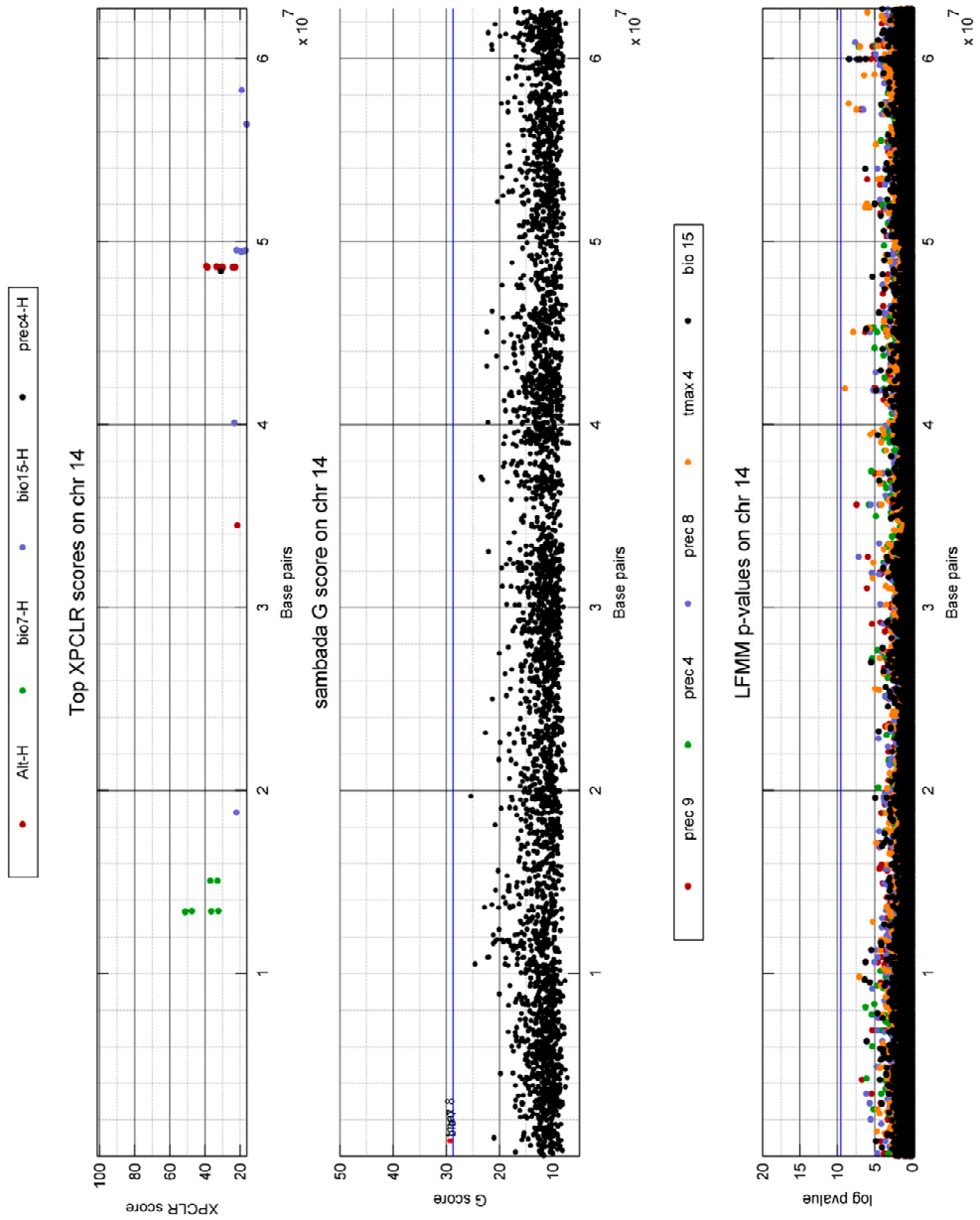


*Figure Appendix II.c.6 Comparison between significant XP-CLR, Sambada and LFMM results according to their position on chromosome 14 for **sheep**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Sambada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

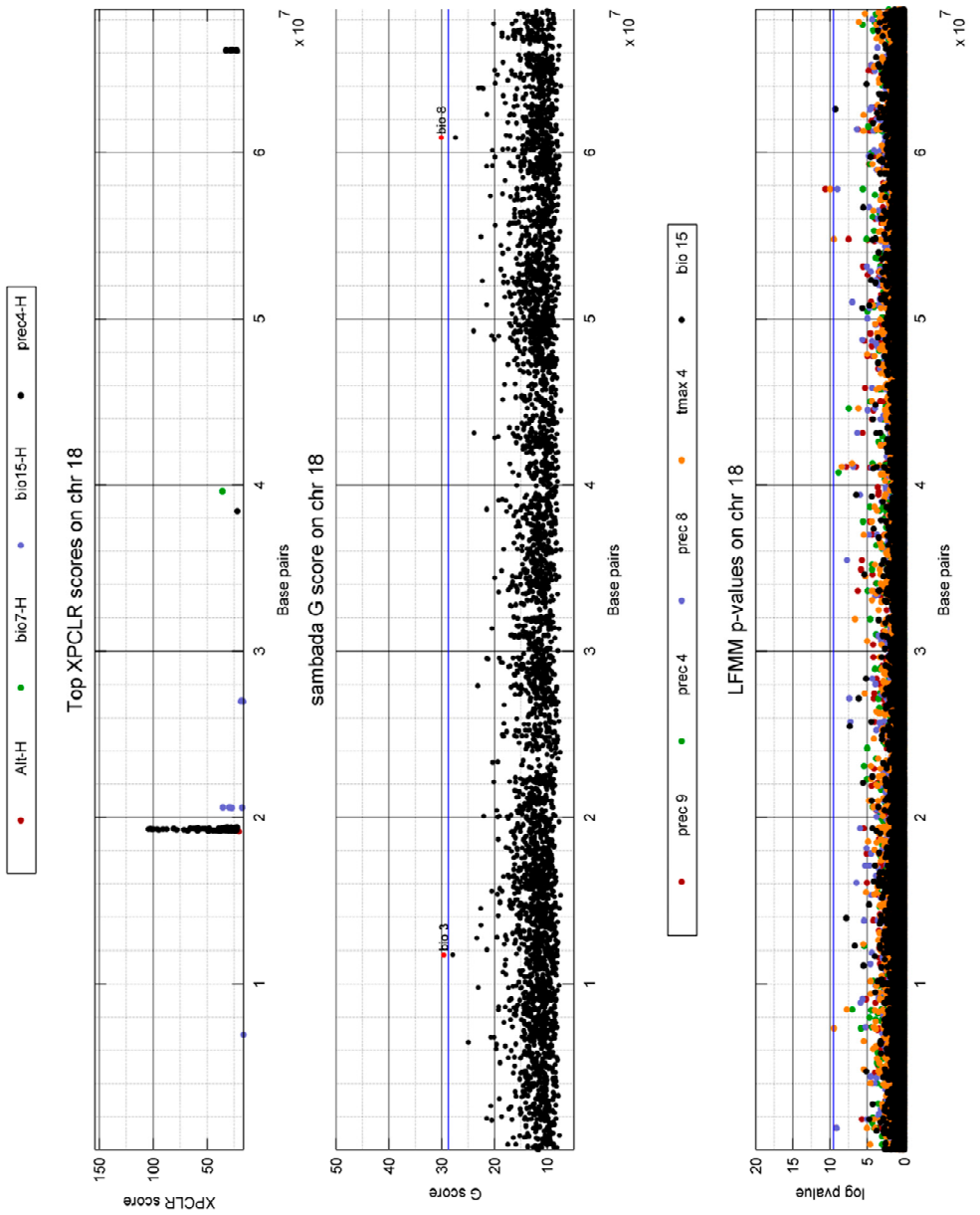


*Figure Appendix II.c.7 Comparison between significant XP-CLR, Sambada and LFMM results according to their position on chromosome 18 for **sheep**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Sambada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

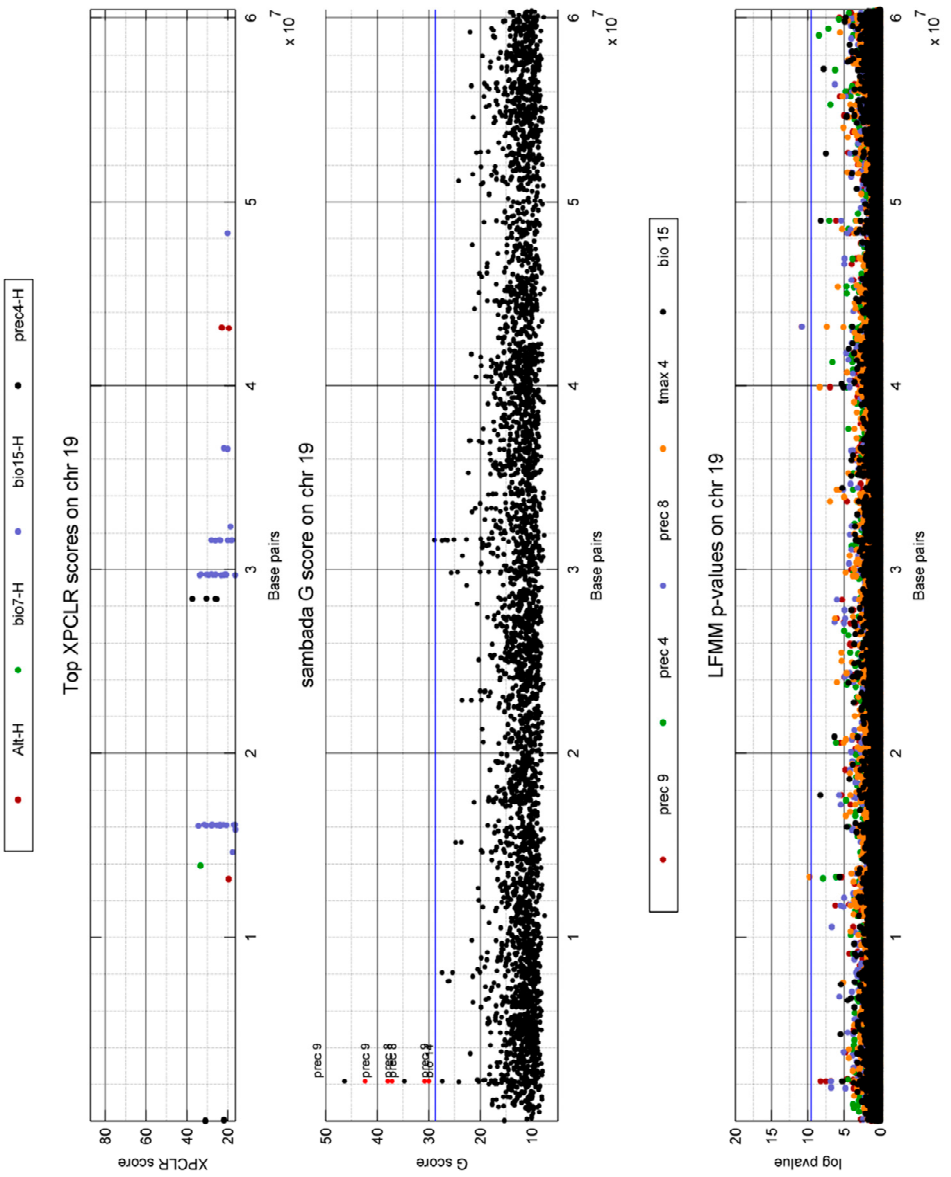


*Figure Appendix II.c.8 Comparison between significant XP-CLR, Sambada and LFMM results according to their position on chromosome 19 for **sheep**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Sambada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

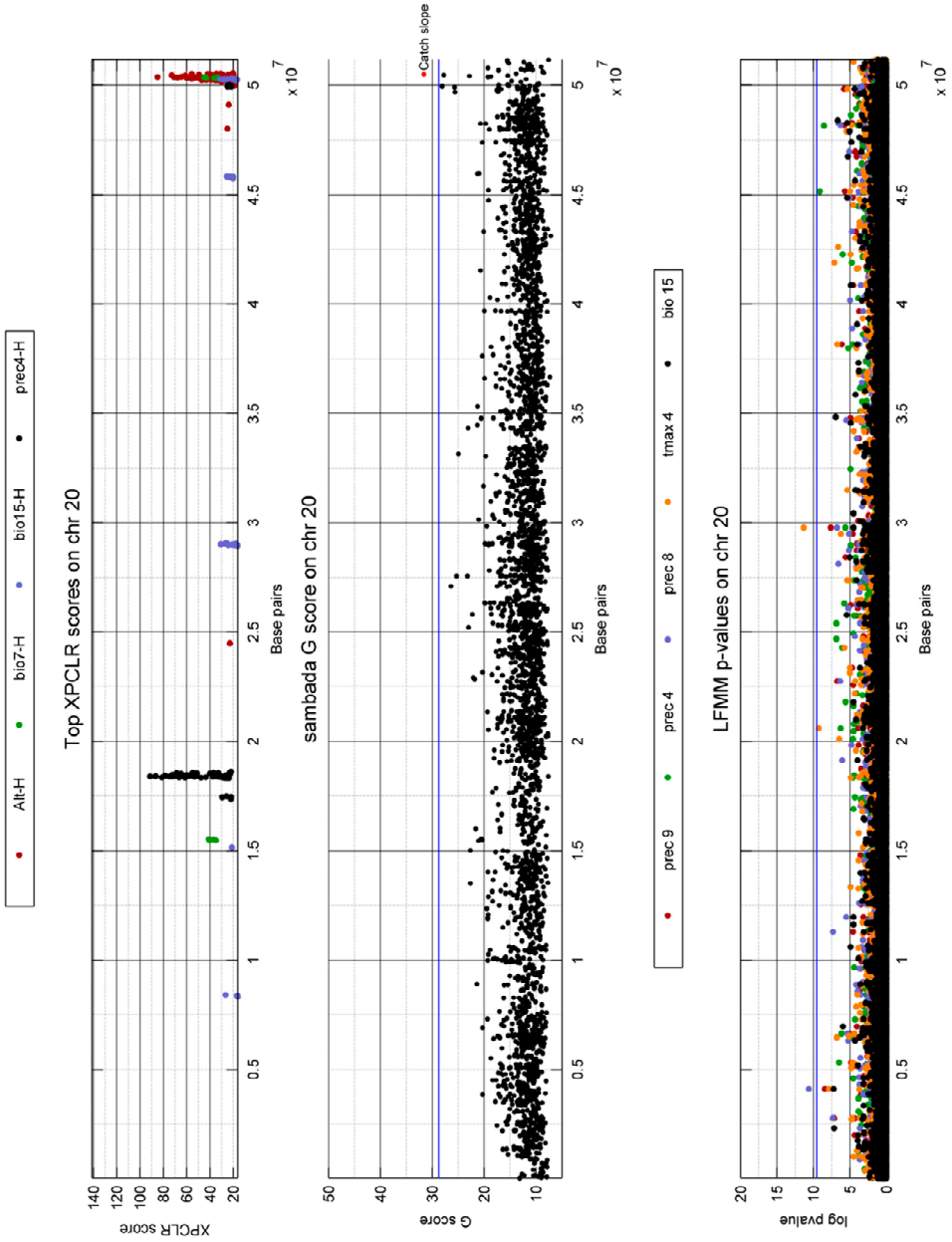


*Figure Appendix II.c.9 Comparison between significant XP-CLR, Samβada and LFMM results according to their position on chromosome 20 for **sheep**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Samβada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

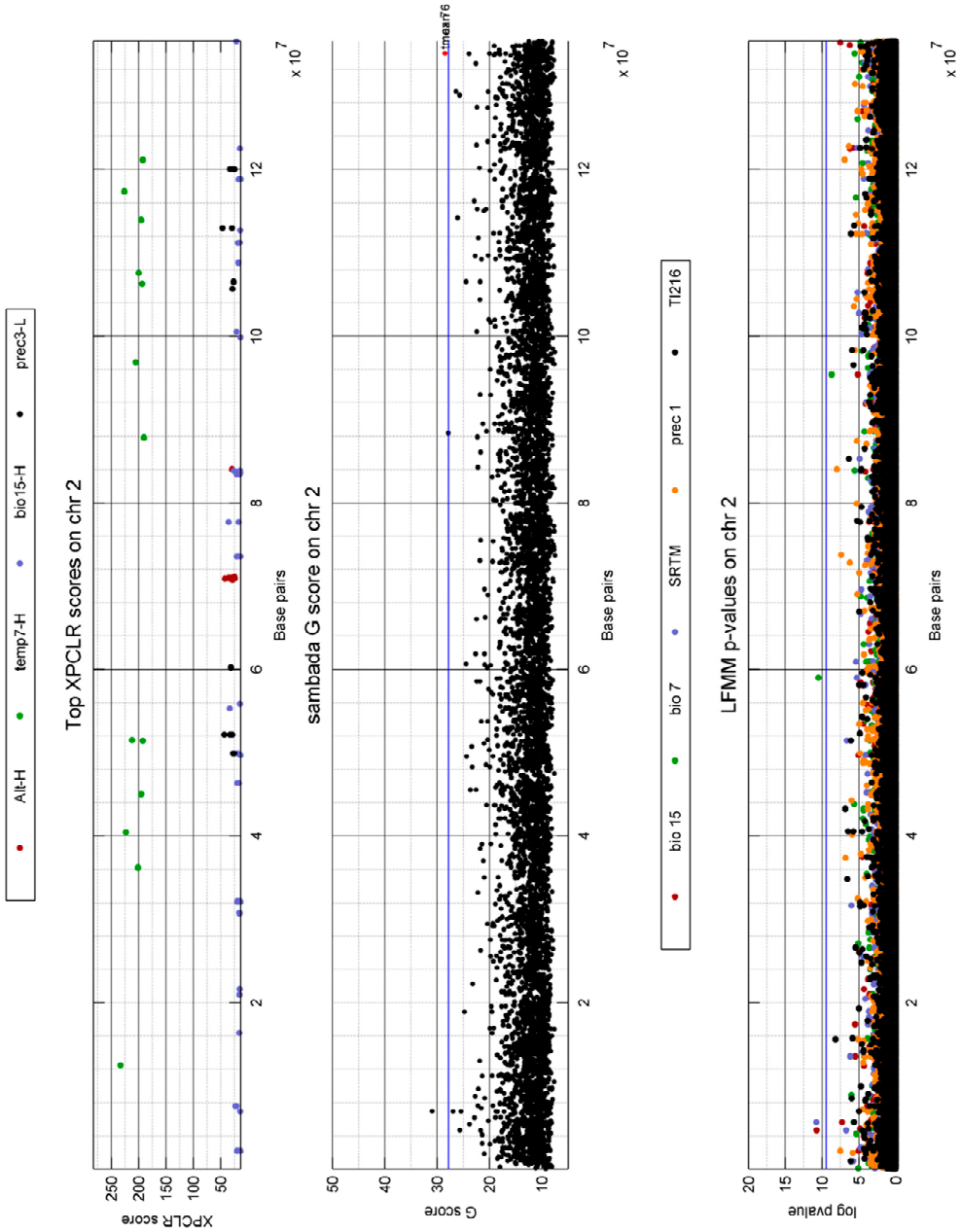


*Figure Appendix II.c.10 Comparison between significant XP-CLR, Samβada and LFMM results according to their position on chromosome 2 for **goats**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Samβada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

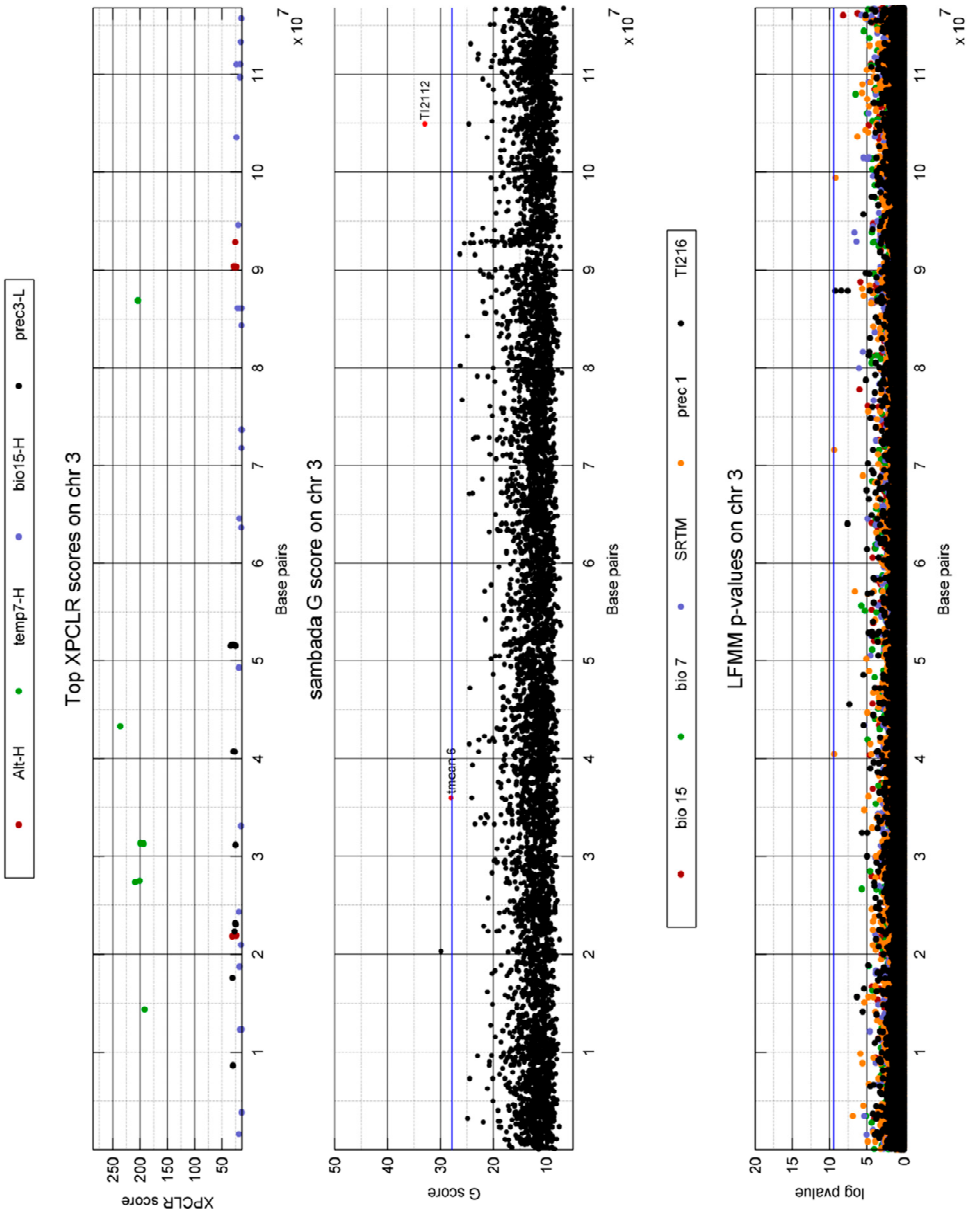


*Figure Appendix II.c.11 Comparison between significant XP-CLR, Sambada and LFMM results according to their position on chromosome 3 for **goats**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Sambada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

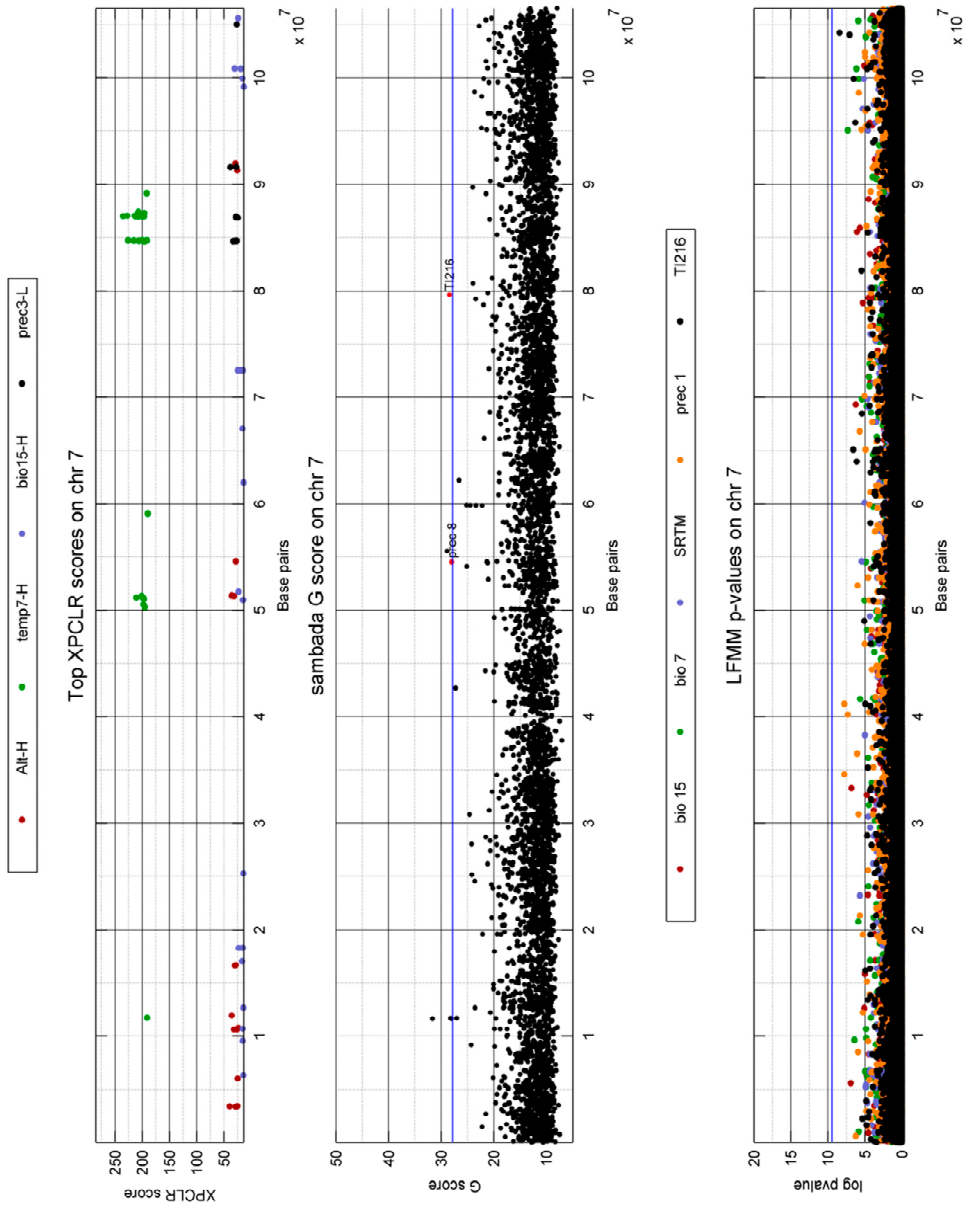


*Figure Appendix II.c.12 Comparison between significant XP-CLR, Samβada and LFMM results according to their position on chromosome 7 for **goats**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Samβada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

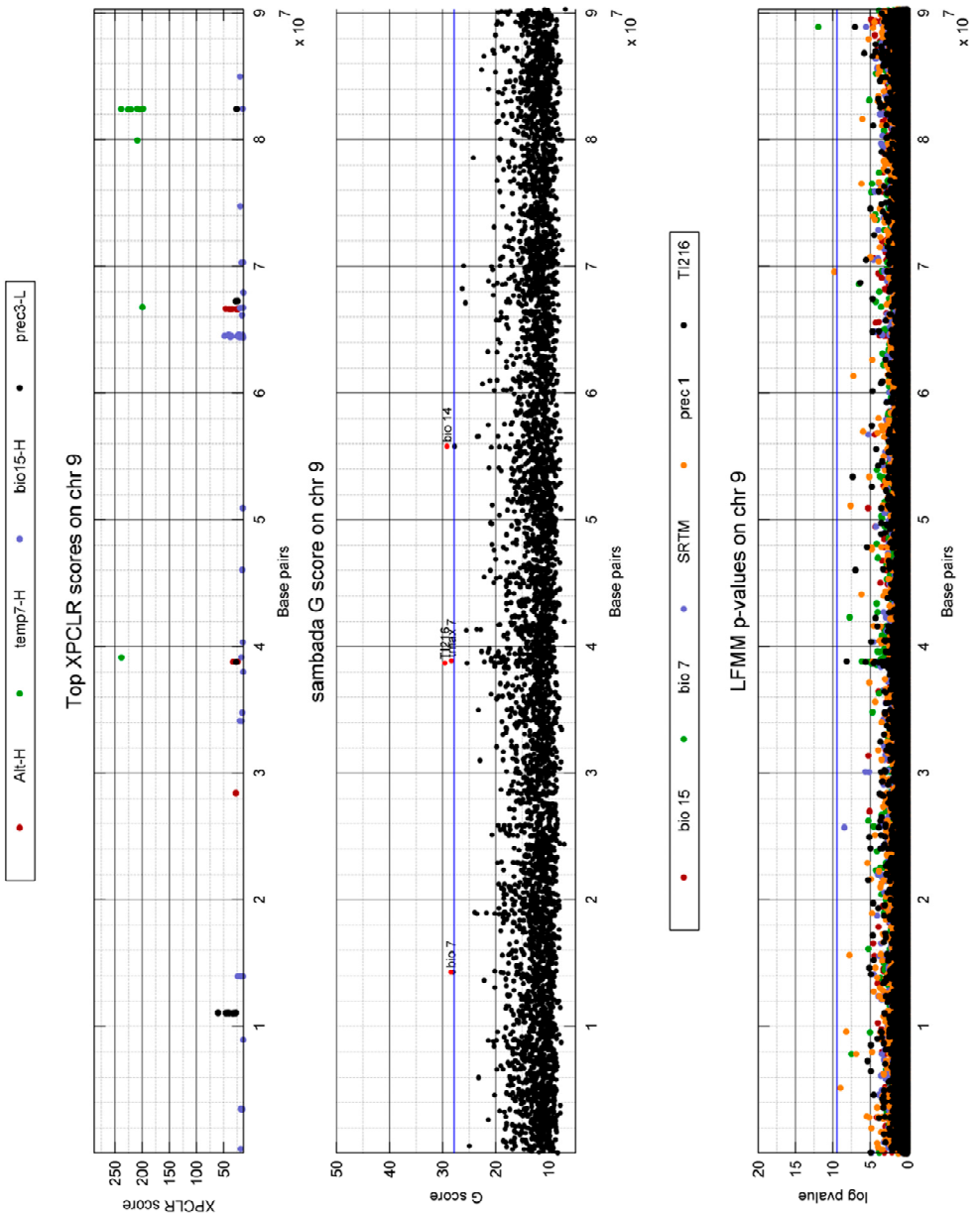


*Figure Appendix II.c.13 Comparison between significant XP-CLR, Sambada and LFMM results according to their position on chromosome 9 for **goats**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Sambada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

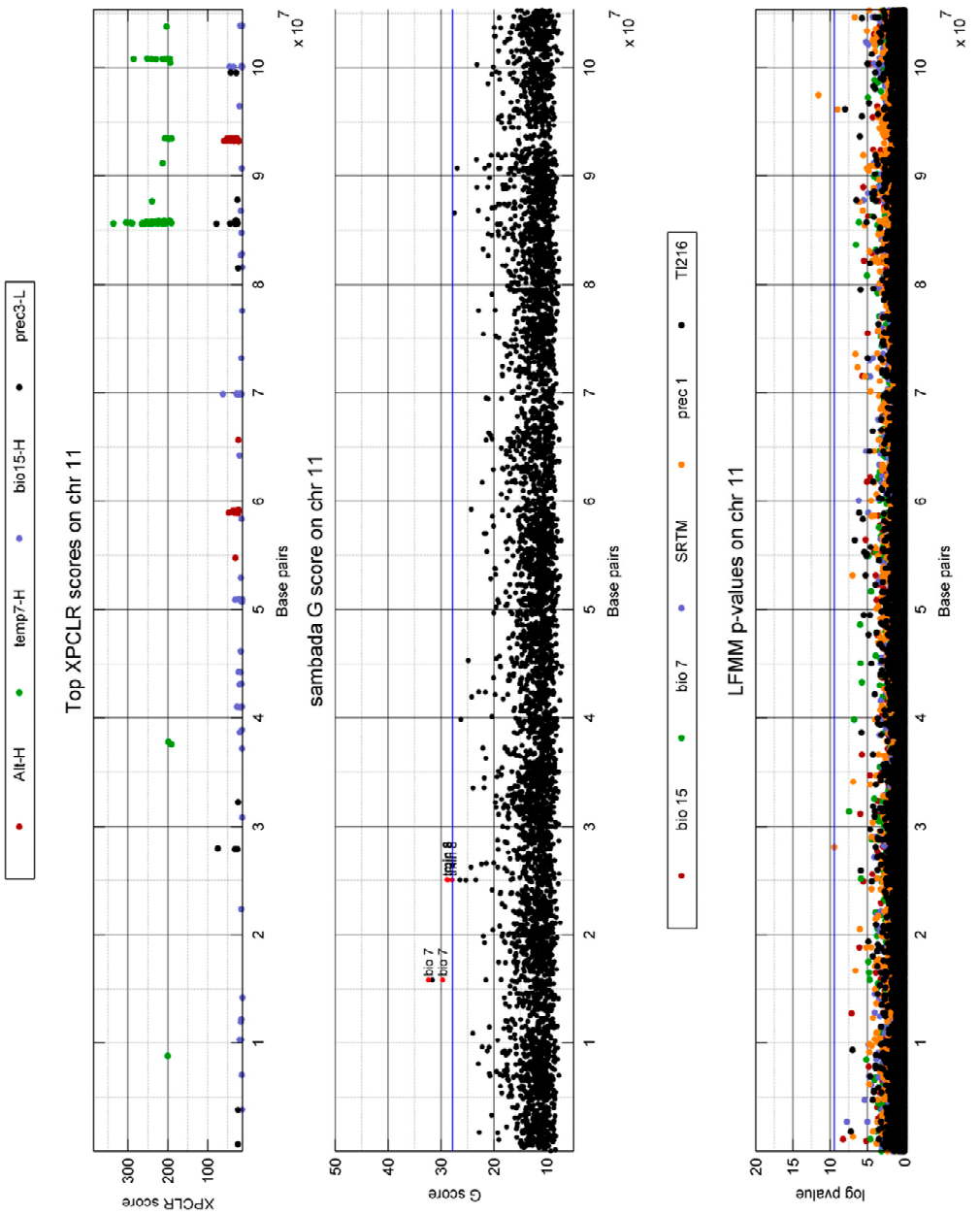


*Figure Appendix II.c.14 Comparison between significant XP-CLR, Samβada and LFMM results according to their position on chromosome 11 for **goats**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Samβada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

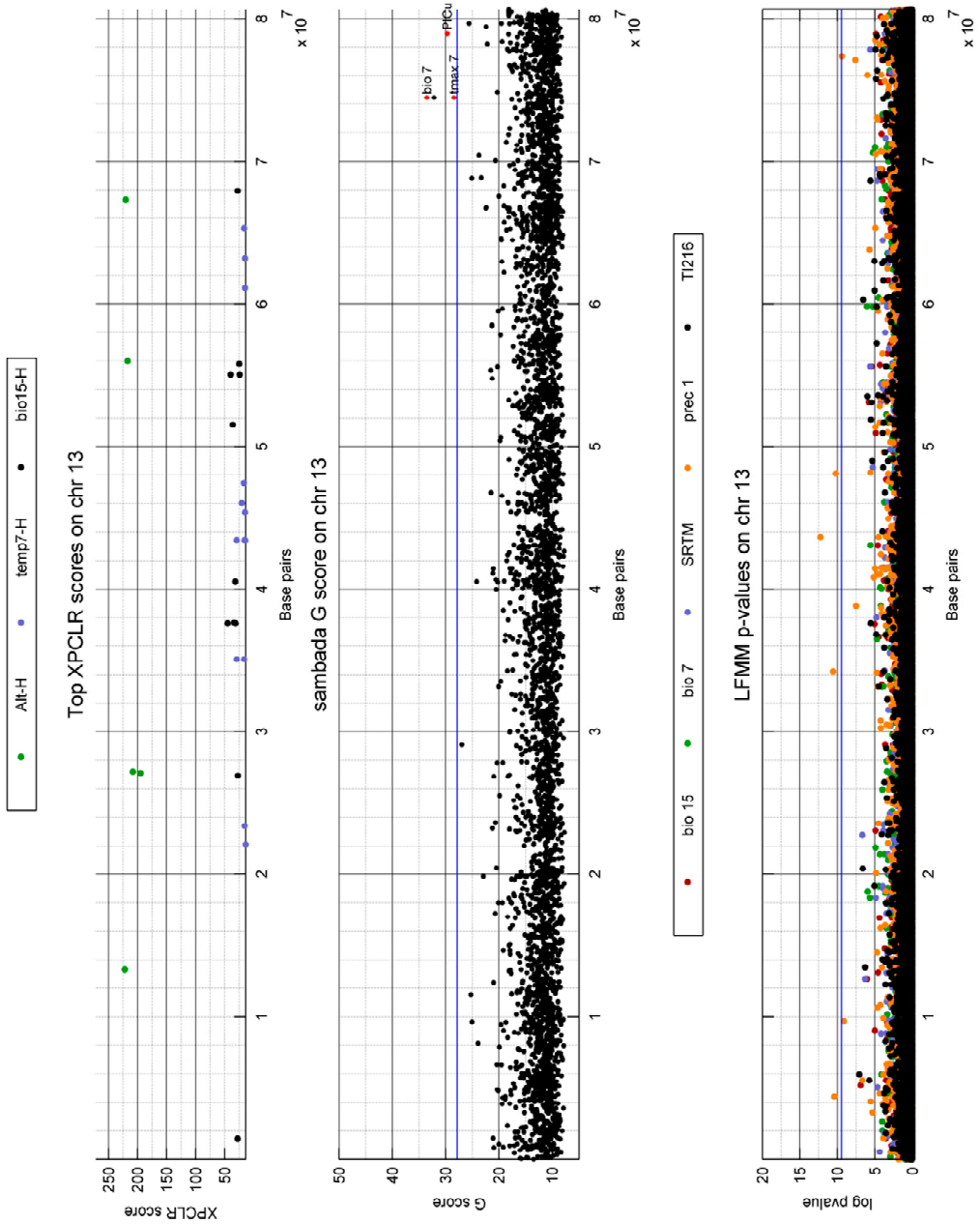


*Figure Appendix II.c.15 Comparison between significant XP-CLR, Samβada and LFMM results according to their position on chromosome 13 for **goats**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Samβada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

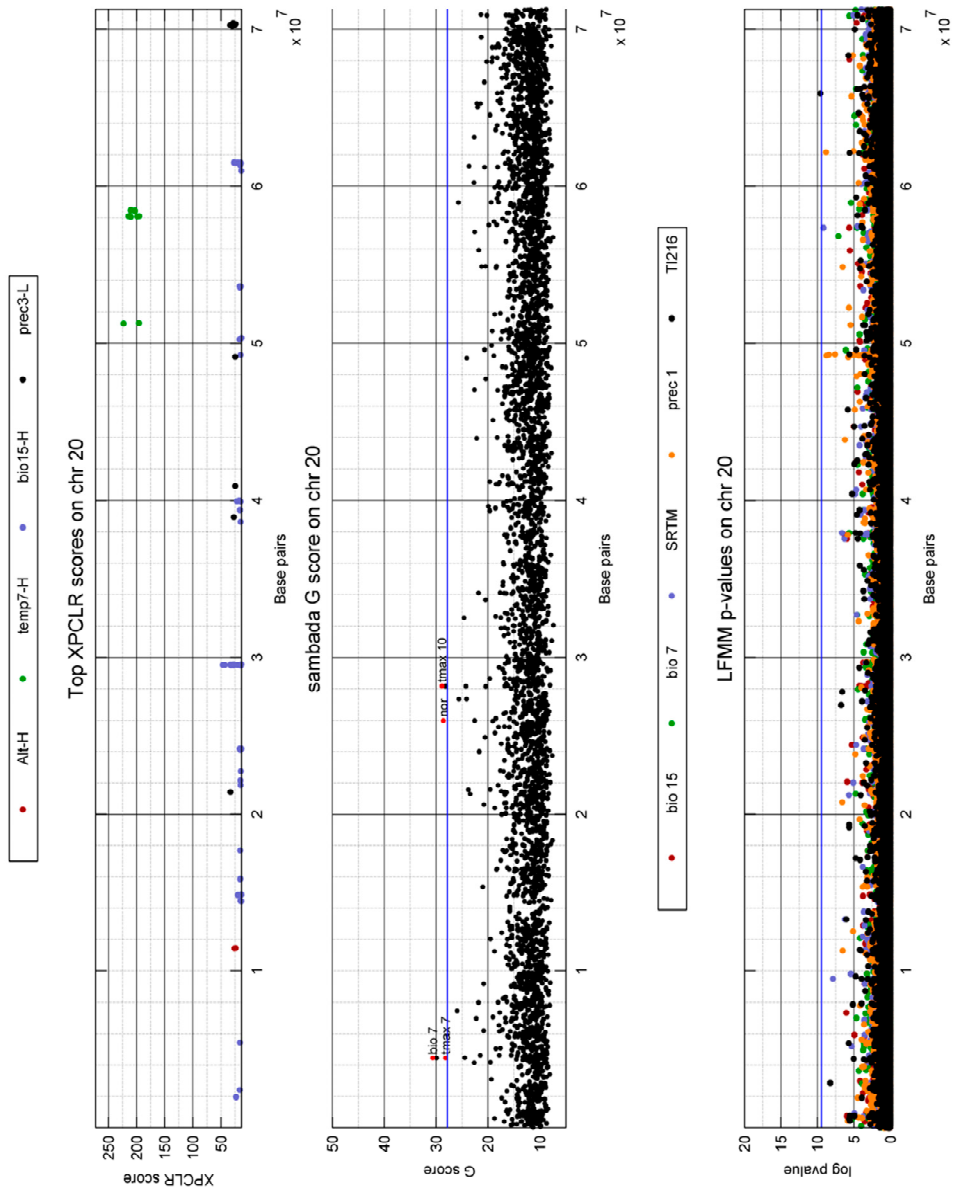


*Figure Appendix II.c.16 Comparison between significant XP-CLR, Samβada and LFMM results according to their position on chromosome 20 for **goats**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Samβada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

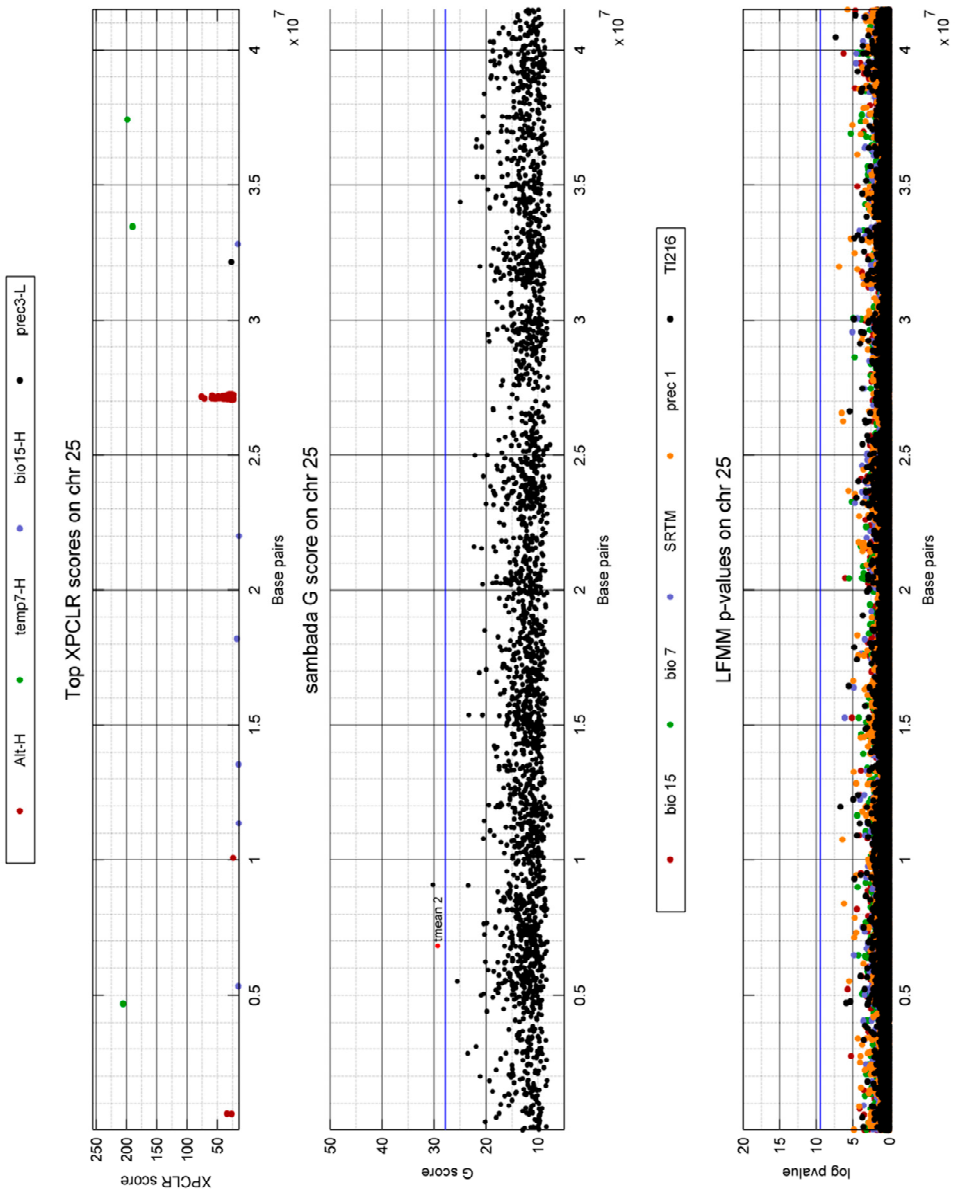


*Figure Appendix II.c.17 Comparison between significant XP-CLR, Sambada and LFMM results according to their position on chromosome 25 for **goats**. For XP-CLR, only significant windows are shown and color-codes depend on the variable. Significant genotypes in Sambada are shown in red and their associated variable are given in as labels. The blue bar represents the threshold of significance. LFMM's SNPs are color-coded according to their associated variable. The blue bar represents the threshold of significance.*

Appendix III. **Papers**

Very high resolution digital elevation models: are multi-scale derived variables ecologically relevant?

Short running title: Ecological relevance of VHR DEM-derived environmental variables

Kevin Leempoel[1], Christian Parisod[2], Céline Geiser[2], Lucas Daprà[2], Pascal Vittoz[3], Stéphane Joost[1]

[1]Laboratory of Geographic Information Systems (LASIG), School of Civil and Environmental Engineering (ENAC), École Polytechnique Fédérale de Lausanne (EPFL), Bâtiment GC, Station 18, 1015 Lausanne, Switzerland

[2]Laboratory of evolutionary botany, University of Neuchâtel, CH-2000 Neuchâtel, Switzerland

[3]Department of Ecology and Evolution, University of Lausanne, Biophore, CH-1015 Lausanne, Switzerland

Abstract

- Digital Elevation Models (DEMs) are often used in landscape ecology to retrieve elevation or first derivative terrain attributes such as slope or aspect in the context of species distribution modelling. However, DEM-derived variables are scale-dependent and, given the increasing availability of very high resolution (VHR) DEMs, their ecological relevance must be assessed for different spatial resolutions.
- In a study area located in the Swiss Western Alps, we derived VHR DEMs-derived variables related to morphometry, hydrology and solar radiation. Based on an original spatial resolution of 0.5 meters, we generated DEM-derived variables at 1m, 2m and 4m spatial resolutions, applying a Gaussian Pyramid. Their associations with local climatic factors, measured by sensors (direct and ambient air temperature, air humidity and soil moisture) as well as ecological indicators derived from species distribution, were assessed with multivariate Generalized Linear Models (GLM) and Mixed Models (GLMM).
- Specific VHR DEM-derived variables showed significant associations with climatic factors. In addition to slope, aspect and curvature, the underused wetness and ruggedness indices predicted measured ambient humidity and soil moisture, respectively. Remarkably, spatial resolution of VHR DEM-derived variables had a significant influence on models' strength, with coefficients of determination decreasing with coarser resolutions or showing a local optimum with a 2m resolution, depending on the variable considered.

- These results support the relevance of using multi-scale DEM variables to provide surrogates for important climatic variables such as humidity, moisture and temperature, offering suitable alternatives to direct measurements for evolutionary ecology studies at a local scale.

Keywords: Digital Elevation Models, Multi-scale analysis, Very High Spatial Resolution, Temperature and Humidity Loggers, Landolt's Ecological Indicators, Generalized Linear Models, Local Scale

Introduction

Digital elevation models (DEMs) are widely used in landscape and evolutionary ecology to understand the distribution of species and their genetic variation (Kozak *et al.* 2008). Their most common use in ecology consists in retrieving elevation, or in computing primary terrain attributes (i.e. slope, aspect and curvature), which underlie biophysical processes at local or regional scales, especially in mountainous areas (Elith & Leathwick 2009; Manel *et al.* 2010a). In many studies, primary attributes have been used as proxies for factors such as solar radiation (Fu & Rich 2002), evapotranspiration (Guisan & Zimmermann 2000), overland and subsurface flow (Broxton *et al.* 2009), soil water content (Moore *et al.* 1991), wind, erosion/deposition rate, soil characteristics (Wilson & Gallant 2000), climatic variables as well as snow accumulation and melt (Lyon *et al.* 2008; Dobrowski 2011). Their accuracy and increasing availability turned them into accessible indicators of topographic variability, though not necessarily those with the highest predictive potential (Guisan & Zimmermann 2000; Pradervand *et al.* 2014).

A large variety of DEM-derived variables can be computed. Conventionally primary terrain attributes are calculated on the basis of 3x3 moving window (Wilson & Gallant 2000; Böhner *et al.* 2002), but more complex variables have been developed over the last two decades to model hydrological processes, solar radiation or local morphometry (Wilson & Gallant 2000; Kalbermatten *et al.* 2012). Named secondary topographic attributes, they are often a combination of primary attributes calculated using a moving window of varying size. Solar radiation for example combines slope, aspect, sunshine duration and adjacent relief (Wilson & Gallant 2000). The higher predictive power of secondary topographic attributes such as wetness indices (Beven & Kirkby 1979), stream power (Moore *et al.* 1991), terrain ruggedness (Riley *et al.* 1999) or temperature (Wilson & Gallant 2000) may be of particular interest for assessing ecological patterns related to specific processes at a landscape scale. For example, Böhner & Selige (2006) used two secondary topographic attributes - a wetness index and a solifluction index - to predict soil pH and snow cover. Secondary topographic attributes were also developed for specific purposes, such as differentiating habitats across different mountain ranges using the Vector Ruggedness Measure (VRM) developed by Sappington *et al.* (2007). Despite convincing examples of their usefulness, DEM-derived variables diversity is rarely potentiated in species distribution models or landscape genetics.

Commonly used DEMs show a moderate to coarse resolution (≈30m for ASTER GDEM, ≈90m for SRTM) and a poor accuracy (Tachikawa *et al.* 2011). In addition, most studies would only consider DEMs at their original resolution or use GPS measurement to compute slope and aspect (Patsiou *et al.* 2014; Greenwood *et al.* 2015). However, the gradual emergence of very high resolution (VHR, ≈1m) elevation data offered unprecedented level of details for exploring the morphological characteristics of landscape and promoted new applications (see Lassueur *et al.* 2006; Kalbermatten *et al.* 2012 and references therein). Indeed, high resolution provides several general advantages. First, it permits to get safer projections on the persistence of species, like in refugia, and more accurate estimations of species distribution in response to global changes (Dobrowski 2011). Climate experienced by an organism is indeed a combination of regional climate pattern and local terrain influence, which shape the habitat constraints an organism is facing. For example, cold air drainage, elevation, topographic position, slope and aspect are the main terrain factors influencing the coupling between local and regional climate (Barry 1992).

In particular, the use of VHR invites reconsidering a number of scale issues raised 20 years ago by Levin (1992). Among them, it is crucial to remember that a high spatial resolution (a small grain) does not necessarily imply better models. Accordingly, it is key to understand the scale-dependency of topographic features and thus to evaluate the usefulness of VHR DEM-derived environmental variables for studies at local scales (≈1 km²) in the light of multi-scale analysis. It is indeed necessary to use spatial resolutions matching the geographic distribution of phenomena under study and the accuracy of sampling's georeferencing. Accordingly, evaluating the influence of scale on the computation of environmental variables is essential. In particular, to what extent VHR likely evidence micro-relief and related micro-climate physical phenomena that may not be grasped at coarser resolutions remains poorly known (Levin 1992; Marceau & Hay 1999; Cavazzi *et al.* 2013). Furthermore, no consensus has emerged yet on the benefits and drawbacks of VHR and this is well illustrated by the multi-resolution approaches of Pradervand *et al.* (2013) that did hardly improve species distribution models of alpine plants at a regional scale, although the distribution of some plants known to live in microhabitats was significantly better predicted. Even though the relationship between species' occurrences and a given environmental variable does not necessarily hold across scales, most studies in ecology use variables at a single resolution with no consideration of scale representativeness. However, scale-dependency is likely central to reservations raised about the contribution of DEMs to ecological modelling and thus deserves additional investigations.

The present work integrates the methodological constraints mentioned above to illustrate how VHR DEM-derived variables can be used to characterize mosaic habitats along a 2 km long alpine ridge encompassing the subalpine-alpine ecotone (Parisod & Christin 2008). Given the steep alpine configuration of this landscape, topography was assumed to be a major driver of air temperature and humidity, as well as soil moisture, thus ruling the distribution of plants (Körner 2003). Accordingly, our aims were to (i) assess the ecological relevance of VHR DEM-derived variables by modelling the relationship between primary as well as secondary VHR DEM-derived environmental variables (e.g. direct solar radiation, wetness index, vector ruggedness measure) and climatic variables measured in the field, and (ii) to identify relevant scales by computing VHR DEM-derived variables at spatial resolutions of 0.5, 1, 2 and 4 meters and assessing the goodness-of-fit and significance of corresponding models. Climatic variables were obtained from different sources;

105 loggers were distributed along the ridge to measure temperature and humidity at high temporal resolution during several months. In addition, we obtained one time measurements of soil moisture at high spatial density. Finally, we modelled the relationship between the same VHR DEM-derived variables and a series of ecological indicators derived from plant species composition (Landolt *et al.* 2010).

Material and Methods

a) Study area and sampling design

The focal study area is a narrow ridge located in the Swiss Western Alps, at "Les Rochers de-Naye" (N46°26'00'' E6°58'50''), covering an elevation range included between 1864 and 2043 m. Locally adapted ecotypes of the plant *Biscutella laevigata* were shown to grow within a distance of less than 10 meters from the cliff in contrasted microsites (Parisod & Bonvin 2008; Parisod & Joost 2010) and this area was thus selected as a suitable model landscape to highlight mosaic habitats across the local subalpine-alpine ecotone.

In order to assess the ecological relevance of VHR DEM-derived environmental variables, the design and the georeferencing of sampling locations are key elements since the precision of their location has to exactly match the highest resolution of the DEM described in the next section. Therefore, sampling locations were selected following a random cluster sampling guided by the population density of the focal species and guaranteeing that all data points are located within pixels representing 0.5x0.5m in the field, resulting in 60 4x4m areas with at least five individuals of B. laevigata (see resulting distribution in Figure 1A). Briefly, direct air temperature was measured with 60 uncovered temperature loggers placed at the centre of each area as well as 20 additional ones installed at random locations along the ridge. Ambient temperature was measured with 25 temperature and humidity covered loggers, placed next to one uncovered logger over three. Soil moisture was measured at 201 sampling locations representative of the focal species (Figure 2B). Furthermore, species composition was assessed in 452 plots of 0.2 m × 0.2 m at the corners of 1m × 1m squares located within the 60 areas as well as 53 additional ones randomly located along the ridge (Appendix S1).

Details on these measurements can be found in the next sub-section.

All sampling points and loggers were geo-referenced with a differential GPS unit (TOPCON-HIPer Pro, http://www.topcon.com.sg/survey/hiperpro.html) offering a horizontal accuracy of ~2-3cm and a vertical accuracy of ~3-4cm.

b) Temperature, humidity and soil moisture data

Air temperature and humidity

Direct air temperature (DT) was measured with uncovered IButton loggers (1922L) from Maxim Integrated (http://www.maximintegrated.com/) placed 15cm above the ground. Furthermore,

covered temperature and humidity loggers (IButton 1923) measured ambient temperature (AT) and humidity (HU) at 15cm above the ground (. These loggers were covered with a white shield pierced with several holes to avoid stagnant air. Loggers were set to record information with a frequency of 30 minutes during 126 days, from June 15, 2013 to October 18, 2013, with an accuracy level of 0.5 degrees C° and 5% for humidity. These 126 days were grouped in 9 periods of 14 days (P1: June 15 to 28; P2: June 29 to July 12; P3: July 13 to 26; P4: July 27 to August 9; P5: August 10 to 23; P6: August 24 to September 6; P7: September 7 to 20; P8: September 21 to October 4; P9: October 5 to 18).

The following descriptive statistics were computed for DT, AT and HU during each period: minimum (MIN), maximum (MAX), mean (MEA), standard deviation (SD), median (MED), mean value at 1am (M1A), mean value at 1pm (M1P), mean daily range (MDR).

Soil moisture

The soil volumetric water content was evaluated once with a FieldScout TDR 300 Soil Moisture Meter (Spectrum Technologies, Aurora, USA, http://www.specmeters.com/). Following le Roux *et al.* (2012), soil moisture values are highly correlated among distinct sampling events and a singly measurement taken more than 24 hours after rainfall was assumed to yield reliable soil moisture values (MSM).

c) Ecological indicators

Species composition was assessed in 452 plots (Appendix S1), with species cover estimated as the proportions of the plot covered by the species. Landolt's ecological indicator values (Landolt et al. 2010) were used to provide an expert-based ecological characterization of sampling plots from their composition in plant species. Landolt's indicators specify tolerance of species of the Swiss flora to climatic or soil conditions, including competitive interactions between species. They are better adapted to the alpine flora than the more commonly used Ellenberg's ecological indicators (Ellenberg *et al.* 1991). The mean value of indicators, weighted by the square-rooted abundance of species, was estimated at the plot level, providing a set of five soil indicators, *LDT-colloidal_dispersion* (soil aeration), *LDT-moisture*, *LDT–humus* (humus proportion), *LDT-nutritive_substances* (soil fertility, mainly nitrogen), *LDT-pH_reaction* (soil pH), and three climate indicators, *LDT-continentality*, *LDT-light*, and *LDT-temperature*.

d) DEM acquisition and processing

We acquired a VHR DEM based on Airborne LIDAR (LIght Detection And Range) technology. A Riegl VQ-480 laser scanner (http://www.riegl.com/) was installed on a helicopter in October 2011 by the HELIMAP Company (http://www.helimap.ch/) to get an average density of 25 soil points/m². The raw point cloud was then processed with the TERRASCAN software (TERRASOLID Ltd, Helsinki; http://www.terrasolid.fi/) to filter buildings, vegetation and all other surface elements in order to obtain a terrain model (Xiaoye Liu 2008). The final density of the ground class

was 10 points/m² on average and the spatial resolution of the DEM was set to 50cm. A few void locations (no data) were filled with the help of a 1m resolution model obtained from the State of Vaud (ASIT-VD; http://www.asitvd.ch/), and using a Multilevel B-Spline Interpolation in SAGA GIS (Seungyong *et al.* 1997).

A multi-scale analysis framework was used to understand how important micro-habitat conditions are and what level of detail is necessary to optimally correlate climatic variables with topographic related variables. Our approach is based on the work of Kalbermatten (2010) and Kalbermatten *et al.* (2012), who showed that a wavelet transform pipeline is a clever way to generalize topography and demonstrated the usefulness of B-splines, a generalization of Bezier curve, to model arbitrary functions, such as DEMs. Therefore, we took advantage of the Gaussian Pyramid algorithm implemented in MATLAB (MATLAB Version 12b. Natick, Massachusetts: The MathWorks Inc., 2010) to approximate topography at multiple resolutions. The original VHR DEM (50cm) was thus generalized to 1, 2 and 4 meters to constitute the multi-scale DEM datasets.

We used SAGA GIS (Böhner *et al.* 2006) and the R package RSAGA (Brenning 2008) to compute and automate the production of DEM-derived variables. We initially computed 16 DEM variables related to morphometry, hydrology and solar radiation, for which details are provided in Appendix S2. Solar radiation variables were computed during one month of the growing season (June).

e) Selection of independent DEM-derived variables

Correlation between each pair of variable was assessed (Appendix S3) and specific variables were omitted from subsequent analyses according to the following rules: (i) the maximum correlation threshold was set to 0.6, (ii) secondary attributes that where highly correlated (>0.6) with primary attributes (i.e. slope and eastness/northness) were deleted, and (iii) the remaining choice between eastness and northness was decided at random due to the high correlation between these two variables. In the end, eight independent variables were retained (Table 1): altitude (alt), terrain wetness index (twi), sine of aspect or eastness (eas), downslope distance gradient (ddg), slope (slo), horizontal curvature (hcu), vertical curvature (vcu), and vector ruggedness measure (vrm).

Given the limited number of observations available for covered ambient temperature (AT) and air humidity (HU) variables (n=25), correlations between retained DEM variables where higher than for uncovered loggers locations and we had to limit the study to 5 independent DEM-derived variables (Appendix S4): altitude (alt), eastness (eas), slope (slo), horizontal curvature (hcu) and terrain wetness index (twi).

f) Regression analysis

Multivariate regression models were performed to explain the variability of climatic variables and ecological factors measured in the field, for each spatial resolution. We used a Step Generalized Linear Models (SGLM; Nelder & Wedderburn 1972) with a Gaussian family and controlled the addition or removal of a term based on the Akaike Information Criterion (AIC). After model completion, co-linearity between variables was controlled using Variance Inflation Factors (VIF;

Montgomery & Peck 1982), based on the threshold >3 (Zuur *et al.* 2009). Models with variables having VIF>3 were processed again, excluding the inflating variables. Landolt factors were log-transformed to fit at normal distribution and all variables were standardized. Adjusted R^2 ((N-1)/(N-k-1) where N = number of observations and k = number of predictors) were calculated for each model.

Instead of GLMs, Generalized linear mixed models (GLMMs) (Breslow & Clayton 1993; Bolker *et al.* 2009) were performed on the dataset of soil moisture and Landolt's indicators to take into account the possible effect of spurious spatial autocorrelation. These variables were indeed collected in plots and the merging by plot was thus considered as a random effect. GLMMs were performed with the lme4 R package (Bates & Maechler 2009). As the package does not support step procedure, we used the resulting DEM-derived variables from SGLMs procedures as fixed effects in the GLMMs.

g) Conventions for variables abbreviations

To facilitate understanding of the following chapters, the conventions used for abbreviations are here-below summarized.

Environmental variables from loggers are written in Upper case and with two letters (DT, AT, HU). Landolt indicators are written in upper case with three letters in italic (ex: *LDT-moisture*) and measured soil moisture with three letters (MSM).

For DT, AT and HU models, measured variables are written in upper case with three letters (MEA, MED, MIN, MAX, MDR, M1A, M1P).

Finally, all DEM-derived variables are written in lower case (alt, slo, twi, vrm, eas, hcu, vcu, ddg).

Results

The distribution of average direct air temperature (DT) over the whole sampling period provides a global view on climatic conditions during summer 2013 (mean 12.1°C; Figure 1B). We focused here on four among the nine periods of 14 days representative of contrasted weather conditions at such altitude: P1 and P9 are representative of the beginning and the end of the growing season and present a cold and a snowy episode, respectively, whereas P3 and P6 are representative of early and late summer conditions, respectively, and are characterized by warm averages with high standard deviations.

Together with altitude (alt), terrain wetness index (twi), vector ruggedness measure (vrm), eastness (eas) and slope (slo) are the DEM-derived variables that best explain the variance of measured environmental variables. Hereunder, we present the VHR DEM-derived variables showing the best model' fit to explain the variability of measured environmental variables and ecological factors, depending on different spatial resolutions and periods of time.

a) Direct air temperature (DT)

Among all DT models, twi is the most frequently significant DEM-derived variable (47% of the models). It is positively correlated with measured variables related to high temperatures (M1P, MAX, MDR) and negatively correlated with those related to cold temperatures (M1A, MIN, MEA) (see Table 2 and Appendix S5). Similarly, alt is also frequently significant (55% of the models), but mainly with measured variables related to cold temperatures (M1A, MED, MEA, MIN). Other DEM-derived variables such as slope, eastness and ddg are less frequently significant.

Significance of DEM-derived variables varies considerably with spatial resolution, whereas it remains relatively constant at all resolutions for elevation. Although the significance for twi is lower when computed at 0.5 or 1m than at coarser resolutions (Appendix S5), adjusted R^2 (aR^2) are usually highest in models at 0.5 or 2m resolution and almost systematically lower at 4m. Noticeably, aR^2 are higher for all measured variables (except for mean range) during periods P1 and P9, which correspond to the two coldest periods among the four analysed.

b) Ambient temperature (AT)

Significant contributions of DEM-derived variables in AT models are much less frequent (49% of the models that converged) than for previously presented DT models (91%; Appendix S6). However, relevant predictors are the same as for DT models, except that horizontal curvature (hcu) is significant at a 2 m resolution (Table 3). Like DT models, twi is positively correlated with measured variables related to high temperatures, and negatively correlated with cold temperatures. Altitude also remains a good predictor and is involved in the models with the highest R^2, particularly during the snow episode (P9).

c) Ambient humidity (HU)

Among the 112 HU models computed, only 35 (40%) showed at least one significant predictor (Appendix S7), contrasting with prior models for DT (90%) and AT (70%). This is likely related to the rare significance of altitude and of DEM-derived variables such as eastness, slo and hcu in HU models (5% of them). On the other hand, twi is the DEM-derived variable with most frequently and highly significant models (37%). It is significant for all categories of measured variables and all periods analysed, except during the snowy episode (P9). Like DT models, resolution influences twi significance and models have an aR^2 optimum at 1 or 2m (Table 4).

To assess the importance of the time-period for the three categories of environmental variables (DT, AT, HU), we computed models between DEM-derived variables and measured variables over the entire fieldwork season (i.e. 15 June to 18 October) (Appendix S8). Although the same DEM variables are significant for almost the same measured climatic variables, our results show that periods of cold, cloud cover (P1) or snow cover (P9) contrasted with those of sunshine (P3, P6). Indeed, a stronger significance of eas, slo, twi and a weaker significance of altitude are observed during those sunshine periods. In addition, the use of several measured variables is justified in

order to distinguish different ecological conditions, as recommended by (Ashcroft et al. 2011; Vercauteren et al. 2012).

d) Soil moisture

In soil moisture models, vector ruggedness measure (vrm) was the only DEM-derived variable that had a significant contribution across resolutions (Table 5). However, its contribution was dependent on resolution, as models were less and less significant with coarser resolutions. Given that alt showed a stable contribution though scales, the highest aR^2 was obtained at 0.5m resolution.

e) Ecological indicators

Determination coefficients of models including Landolt's ecological indicators were low at all resolutions. Only LDT-moisture and LDT-nutritive_substances showed aR^2 above 0.15. Two DEM-derived variables, twi and slope, showed a significant contribution to LDT-moisture across scales (Table 6). Unlike other models, GLMM's aR^2 values for LDT-moisture were stable through resolutions.

Discussion

Variables derived from DEMs are crucial for species distribution models or landscape genetics, but their ecological relevance remains subject to caution (Lassueur et al. 2006; Dubuis et al. 2013). In particular, the relationship between DEM-derived variables and ecological features does not necessarily hold across spatial scales and appears highly dependent on the spatial resolution. In order to foster application of DEMs in ecology and evolution, their relevance to approximate environmental features must be evaluated and suitable approaches should be further developed. Our results validate two essential concerns regarding DEMs: i) multi-scale approaches are valuable when facing topographic heterogeneity, and ii) it is crucial to investigate a large diversity of DEM-derived variables in order to evaluate all topographic aspects that might influence climatic variability. Using a specific area with challenging features at the interface between subalpine and alpine conditions, we were able to show that DEM-derived variables can be used as relevant surrogates for environmental variables and to better understand relationships with local topography. Indeed, physiological activity and adaptation of plants are affected by temperature, humidity and soil characteristics (Körner 2003; Böhner & Selige 2006; Manel et al. 2012b).

Our models consistently report decreased aR^2 at 4m spatial resolution, supporting the hypothesis that VHR provides better predictions in heterogeneous areas such as mountains. However, our models did not generally converge towards a clear optimal resolution and reveal that the most suitable resolution depends on the type of DEM-derived variable considered. This is particularly well illustrated by vrm, showing highest significance at 0.5m and highlighting that soil characteristics are best grasped when initially computed with as much details as possible, whereas hydrology variables, such as twi, reach optima at different resolutions (Böhner & Selige 2006; Buchanan et al. 2013). Variation in the model fit across scales highlights the necessity of implementing multi-scale approaches in ecological studies involving DEM-derived variables. The computation of such variables at multiple scales should be systematically considered to model micro-climatic variables

such as temperature, humidity and soil moisture in a mountainous area. Furthermore, we argue that using DEMs at their original grid resolution, without consideration of scale representativeness, likely leads to an underestimated role of topographic features in ecological models. In fact, a too fine resolution may hold an excess of details and generate too much noise, while too coarse resolution would only show generalized properties of the landscape and lose explanatory power (Cavazzi *et al.* 2013). Although most studies using DEMs at their original resolution often ended up with a minor contribution of topography in their models (Zimmermann & Kienast 1999; Manel *et al.* 2010b; Vercauteren *et al.* 2012; Patsiou *et al.* 2014), we show here that coupling VHR DEMs with a multi-scale approach generates variables with a high predictive power. Accordingly, acquiring high or very high resolution DEMs and performing multi-scale analysis further on represent a suitable approach for local scale studies in ecology and evolution. At the moment, LIDAR represents the best DEM acquisition technology, providing great precision and high resolution across hardly accessible terrains, but still expensive (Xiaoye Liu 2008). Although they do not show the same level of precision like LIDAR, stereo-photogrammetry from Unmanned Aerial Vehicles (UAV) constitutes a less powerful but suitable and cheaper alternative subject to intense research (Leempoel & Joost 2012).

Our results further bring advantages of using a large panel of DEM-derived variables. On the one hand, terrain wetness index (twi) showed the highest explanatory power among the DEM-derived variable here tested, highlighting a relevant proxy for dryness across the studied landscape (Figure 2A). In addition, models including more variables such as eastness and slope best predicted temperature, probably because these primary attributes have a high influence on radiation and wind exposure (Wilson & Gallant 2000; McVicar *et al.* 2007; Appendix S5). For instance, in our specific study area, twi partially accounted for the distance to the ridge as well as for the protection from wind, which could further contribute to temperature and humidity variability. In fact, distance to ridge and twi were moderately correlated at high resolution (i.e. 0.6 at 0.5m and 0.7 at 1m) and dropped to 0.3 at coarser resolutions. Although such correlations are inevitable and likely blur interpretations, our models showed that most of the significant contribution of twi were obtained at 0.5 and 2m, when the correlation between twi and distance to ridge were not the strongest. This, again, highlights the relevance of a multi-scale analysis.

Among other overlooked DEM-derived variables in the literature, vector ruggedness measure (vrm) appeared as the most important predictor of soil moisture (MSM), suggesting that vrm at such high resolution is a suitable proxy for the distribution of stony soils along the ridge and thus for soils with different porosities. Accordingly, the negative coefficients observed here support this hypothesis that high roughness highlight stony soils implying low soil moisture, whereas low roughness reflects developed soils retaining higher moisture. This vrm variable, measuring vector dispersion across the central pixel rather than being a derivative of slope, represents a much better proxy than related proxies such as Terrain Ruggedness Index (Appendix S3&4), as previously stressed by Sappington *et al.* (2007). Nevertheless, the present models demonstrate a variety of DEM-derived variables as suitable or complementary surrogates to *in situ* measurements for characterization of plant habitats and we recommend to go beyond their traditional use of elevation, slope and aspect (Dobrowski 2011).

In addition, DEM-derived variables are easy-to-compute proxies of environmental features, involving limited fieldwork but good knowledge of Geographic Information Systems, DEM-derived variables should thus be widely used as proxies of environmental features in ecology and evolution (Kozak *et al.* 2008). Furthermore, open source GIS alternatives (e.g. SAGA GIS, Quantum GIS and GRASS) provide algorithms to process a variety of secondary terrain attributes.

The distribution of the focal species along an apparently homogeneous ridge, showing a constant slope and slight changes in orientation, in fact turned out to be highly heterogeneous at a high resolution. Prior work on ecotypes of *Biscutella laevigata* (Parisod & Christin 2008) suggested a mosaic distribution of subalpine and alpine habitats, and the use of VHR DEM-derived variables here brought clear evidence of topographic control on micro-climatic patterns. Our results indeed show a significant contribution of micro-topography to model micro-habitat, even though unmeasured factors may play a major role. For instance, it is generally admitted that high elevation and exposed sites are more likely to be coupled with free air environment as compared with low elevation sites that are protected (Pepin & Seidel 2005). However, we observed 5°C difference in ranges for AT and up to 8°C for DT. Such important temperature variability over short distances cannot only be due to large scale effects and support our evidences for a micro-topographic control (Fridley 2009). In addition, VHR DEM-derived variables in our models highlighted the lower relevance of elevation as compared with studies at regional or continental scale. Despite a correlation of -0.99 reported between temperature and elevation across Switzerland (Zimmermann & Kienast 1999), we here showed that the 0.5°C decrease per 100m elevation increase did not hold at a local scale. Therefore, the important variability of temperature observed here is likely valid in various mountainous areas, even when microhabitats variability is only partially distinguished from large scale factors. Our results thus confirm that proxies other than elevation can - and in fact probably better - account for temperature variability in as mountainous areas.

On top of micro-climatic factors, meso-climatic ones might affect climatic variables in the study area. For instance, varying wind patterns and cloud cover across the studied ridge could impact on the variability of local climates. The results obtained here for micro-topography are however not disqualified by meso-climatic patterns. In contrast to common cloudiness on the highest part of the study area early and late during the growth season, the contribution of DEM-derived variables appeared consistently significant at different time periods, demonstrating a substantial effect of micro-topography. In addition, several DEM variables such as protection index, sky view factor or ruggedness might expected to be surrogates of protection from wind at a micro-climatic level. Noticeably, temperatures measured during the snow episode provide an indirect measure of snow cover, as loggers situated under the snow during that period did not show a daily cycle of temperature at sampling locations. Therefore, modelling of snow cover heterogeneity could be improved by combining topographic variables (Gottfried *et al.* 1998; Randin *et al.* 2009) with the daily cycles of loggers. Our results thus highlight the role of micro-topographic effects and the need to consider different measured variables and temporal variability at a scale pertinent for plants, as previously reported by Körner (2003) and Scherrer & Körner (2011).

Noticeably, variables derived from VHR DEM predicted Landolt indicators derived from species distribution with less accuracy than climatic variables. Insufficient variability in this biological dataset compared to extension of Landolt's indicator values (attributed to species across the whole Alps (Landolt *et al.* 2010) certainly explains such limited relevance of micro-topography to a large extent. Our data are indeed restricted to a single site and may thus not show sufficient variation for indicators such as temperature (here, only alpine belt), continentality (only oceanic conditions), light (only open, alpine grasslands), soil pH (only calcareous soils), humus and aeration (mainly humic and silty soils). Furthermore, Landolt's indicators include biotic interactions such as competition that were not taken into consideration by DEM-derived variables. Although the exact reasons underlying the relatively low adjusted R^2 in models derived from biotic data remain elusive, this work shows that models using VHR DEM-derived variable were generally significant for ecological indicators showing a high variability at local scale in mountainous environment (Körner 2003). Variables retained in models (i.e. wetness index, ruggedness, slope and curvature) were indeed highly coherent with factors related to micro-topography and to slope, such as lower soil humidity on steep slopes leading to higher drainage and in superficial soils likely developing on mounds rather than in hollows (Gobat *et al.* 1989; Burga *et al.* 2010) .

Although they are not directly linked to ecological features, DEM-derived variables are relevant proxies and easily accessible sources of environmental variability. We demonstrated that DEMs cannot be used without consideration for their scale representativeness and that only a multi-scale approach can detect these features. In fact, a VHR DEMs is mandatory to model properly local topographic features and allowed us to perform multi-scale analysis to show a strong effect of resolution to model climatic variables, models in which higher resolution does not necessarily mean better explanatory power. Finally, we noted that DEMs are underexploited regarding the large diversity of variables that can be computed from them and we strongly recommended going beyond the traditional use of elevation, slope and aspect.

Acknowledgments

This work was funded by a grant from the Velux Stiftung (Project 705 to CP). We thank Philippa Griffin for her help producing and improving R scripts. We thank our fieldwork helpers: Amélie Bardil, Daniela Biosa, Benjamin Dauphin, Timothée Produit and Ivo Widmer.

References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, **19**, 1655–1664.

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature reviews. Genetics*, **11**, 697–709.

Anderson CD, Epperson BK, Fortin MJ *et al.* (2010) Considering spatial and temporal scale in landscape-genetic studies of gene flow. *Molecular Ecology*, **19**, 3565–3575.

Anselin L (1995) Local indicators of spatial association — LISA. *Geographical Analysis*, **27**, 93–115.

Anselin L (1998) Exploratory spatial data analysis in a geocomputational environment. *GeoComputation*, 17–19.

Anselin L, Syabri I, Kho Y (2006) GeoDa: An introduction to spatial data analysis. *Geographical Analysis*, **38**, 5–22.

Ansell SW, Grundmann M, Russell SJ, Schneider H, Vogel JC (2008) Genetic discontinuity, breeding-system change and population history of Arabis alpina in the Italian Peninsula and adjacent Alps. *Molecular Ecology*, **17**, 2245–2257.

Antao T, Beaumont MA (2011) Mcheza: A workbench to detect selection using dominant markers. *Bioinformatics*, **27**, 1717–1718.

Aregger M, Cowling VH (2013) Human cap methyltransferase (RNMT) N-terminal non-catalytic domain mediates recruitment to transcription initiation sites. *The Biochemical journal*, **455**, 67–73.

Ashcroft MB, French KO, Chisholm LA (2011) An evaluation of environmental factors affecting species distributions. *Ecological Modelling*, **222**, 524–531.

Balkenhol N, Waits LP, Dezzani RJ (2009) Statistical approaches in landscape genetics: an evaluation of methods for linking landscape and genetic data. *Ecography*, **32**, 818–830.

Barry RG (1992) *Mountain weather and climate* (Routledge, Ed,).

Bates D, Maechler M (2009) lme4: Linear mixed-effects models using {S4} classes.{R} package version 0.999375-32.

Beaumont MA, Nichols RA (1996) Evaluating Loci for Use in the Genetic Analysis of Population Structure. *Proceedings: Biological Sciences*, **263**, 1619–1626.

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289 – 300.

Benjelloun B, Alberto FJ, Streeter I *et al.* (2015) Characterizing neutral and adaptive genomic diversity in indigenous populations of Moroccan goats (Capra hircus) using WGS data. *Frontiers in Genetics*.

Beven KJ, Kirkby MJ (1979) A physically based, variable contributing area model of basin hydrology / Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrological Sciences Bulletin*, **24**, 43–69.

Böhner J, Antonić O (2009) Chapter 8 Land-Surface Parameters Specific to Topo-Climatology. In: *Developments in Soil Science* (eds Tomislav H, Hannes IR), pp. 195–226. Elsevier.

Böhner J, Köthe R, Conrad O *et al.* (2002) Soil Regionalisation by Means of Terrain Analysis and Process Parameterisation. , **EUR 20398** .

Böhner J, McCloy KR, Strobl J (2006) SAGA – Analysis and Modelling Applications. *Göttinger Geographische Abhandlungen*, **115**, 130.

Böhner J, Selige T (2006) Spatial prediction of soil attributes using terrain analysis and climate regionalisation. *BÖHNER, J., MCCLOY, KR & J. STROBL (Eds.): SAGA–Analyses and Modelling Applications.–Göttinger Geographische Abhandlungen*, **115**, 13–28.

Bolker BM, Brooks ME, Clark CJ *et al.* (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, **24**, 127–135.

Bolliger J, Lander T, Balkenhol N (2014) Landscape genetics since 2003: status, challenges and future directions. *Landscape Ecology*, **29**, 361–366.

Borcard D, Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling*, **153**, 51–68.

Brenning A (2008) Statistical Geocomputing combining R and SAGA: The Example of Landslide susceptibility Analysis with generalized additive Models (J Böhner, T Blaschke, L Montanarella, Eds,). *SAGA –- Seconds Out*, **19**, 23–32.

Breslow NE, Clayton DG (1993) Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, **88**, 9–25.

Broxton PD, Troch PA, Lyon SW (2009) On the role of aspect to quantify water transit times in small mountainous catchments. *Water Resources Research*, **45**, W08427.

Buchanan BP, Fleming M, Schneider RL *et al.* (2013) Evaluating topographic wetness indices across central New York agricultural landscapes. *Hydrol. Earth Syst. Sci. Discuss.*, **10**, 14041–14093.

Burga CA, Krüsi B, Egli M *et al.* (2010) Plant succession and soil development on the foreland of the Morteratsch glacier (Pontresina, Switzerland): Straight forward or chaotic? *Flora - Morphology, Distribution, Functional Ecology of Plants*, **205**, 561–576.

Cavazzi S, Corstanje R, Mayr T, Hannam J, Fealy R (2013) Are fine resolution digital elevation models always the best choice in digital soil mapping? *Geoderma*, **195–196**, 111–121.

Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. *Genome Research*, **20**, 393–402.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using Environmental Correlations to Identify Loci Underlying Local Adaptation. *Genetics*, **185**, 1411–1423.

Cushman SA, Landguth EL (2010) Spurious correlations and inference in landscape genetics. *Molecular Ecology*, **19**, 3592–3602.

Cushman SA, McKelvey KS, Hayden J, Schwartz MK (2006) Gene flow in complex landscapes: testing multiple hypotheses with causal modeling. *The American naturalist*, **168**, 486–499.

Dale MRT, Fortin M-J (2002) Spatial autocorrelation and statistical tests in ecology. *Ecoscience*, **9**, 162–167.

Dale MRT, Fortin M-J (2009) Spatial autocorrelation and statistical tests: Some solutions. *Journal of Agricultural, Biological, and Environmental Statistics*, **14**, 188–206.

Dale MRT, Mah M (1998) The use of wavelets for spatial pattern analysis in ecology. *Journal of Vegetation Science*, **9**, 805–814.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Darwin C, Wallace A (1858) On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. *Journal of the Proceedings of the Linnean Society of London. Zoology*, **3**, 45–62.

Diniz-Filho JAF, Nabout JC, de Campos Telles MP, Soares TN, Rangel TFLVB (2009) A review of techniques for spatial modeling in geographical, conservation and landscape genetics. *Genetics and Molecular Biology*, **32**, 203–211.

Dobrowski SZ (2011) A climatic basis for microrefugia: The influence of terrain on climate. *Global Change Biology*, **17**, 1022–1035.

Dobson AJ, Barnett A (2008) *An Introduction to Generalized Linear Models, Third Edition*. Taylor & Francis.

Dubuis A, Giovanettina S, Pellissier L *et al.* (2013) Improving the prediction of plant species distribution and community composition by adding edaphic to topo-climatic variables. *Journal of Vegetation Science*, **24**, 593–606.

Earl D, vonHoldt B (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Elith J, Leathwick JR (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

Ellenberg H, Weber HE, Düll R *et al.* (1991) Zeigerwerte von pflanzen in Mitteleuropa.

Escoffier B, Pages J (2008) *Analyses factorielles simples et multiples* (Paris:Dunod, Ed,).

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, **14**, 2611–2620.

Fischer MC, Rellstab C, Tedder A *et al.* (2013) Population genomic footprints of selection and associations with climate in natural populations of Arabidopsis halleri from the Alps. *Molecular ecology*, **22**, 5594–5607.

Fisher RA (1930) *The Genetical Theory of Natural Selection*. At The Clarendon Press.

Foll M, Gaggiotti O (2008) A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics*, **180**, 977–993.

Fontanesi L, Beretti F, Riggio V *et al.* (2009) Missense and nonsense mutations in melanocortin 1 receptor (MC1R) gene of different goat breeds: association with red and black coat colour phenotypes but with unexpected evidences. *BMC genetics*, **10**, 47.

Fortin M-J, Dale MRT, ver Hoef J (2002) Spatial analyses in ecology. In: *Encyclopedia of environmetrics* , pp. 2051–2058.

Fournier-Level A, Korte A, Cooper MD *et al.* (2011) A Map of Local Adaptation in Arabidopsis thaliana. *Science*, **334**, 86–89.

Fraser HB (2013) Gene expression drives local adaptation in humans. *Genome Research*, **23**, 1089–1096.

Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*.

Fridley JD (2009) Downscaling climate over complex terrain: High finescale (<1000 m) spatial variation of near-ground temperatures in a montane forested landscape (Great Smoky Mountains). *Journal of Applied Meteorology and Climatology*, **48**, 1033–1049.

Fu P, Rich PM (2002) A geometric solar radiation model with applications in agriculture and forestry. *Computers and Electronics in Agriculture*, **37**, 25–35.

Gallant JC, Hutchinson MF (1996) Towards an understanding of landscape scale and structure. In: *Proceedings of the Third International Conference/Workshop on Integrating GIS and Environmental Modeling*

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*, 9.

Geiser C (2014) Genome evolution and mechanisms underlying reproductive isolation in the polyploid "Biscutella laevigata." Univ. Neuchâtel.

Di Giulio M, Holderegger R, Tobias S (2009) Effects of habitat and landscape fragmentation on humans and biodiversity in densely populated landscapes. *Journal of Environmental Management*, **90**, 2959–2968.

Gobat JM, Duckert O, Gallandat JD (1989) *Quelques relations "microtopographie-sols-végétation" dans les pelouses pseudo-alpines du Jura suisse: exemples d'un système naturel et d'un système anthropise*. Société neuchâteloise des sciences naturelles.

Gottfried M, Pauli H, Grabherr G (1998) Prediction of vegetation patterns at the limits of plant life: A new view of the alpine-nival ecotone. *Arctic and Alpine Research*, **30**, 207.

Greenwood S, Chen J-C, Chen C-T, Jump AS (2015) Temperature and sheltering determine patterns of seedling establishment in an advancing subtropical treeline. *Journal of Vegetation Science*, n/a–n/a.

Gruber S, Peckham S (2009) *Land-surface parameters and objects in hydrology*.

Guillot G, Vitalis R, Rouzic A le, Gautier M (2014) Detecting correlation between allele frequencies and environmental variables as a signature of selection. A fast computational approach for genome-wide studies. *Spatial Statistics*, **8**, 145–155.

Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Hall L, Beissinger S (2014) A practical toolbox for design and analysis of landscape genetics studies. *Landscape Ecology*, **29**, 1487–1504.

Hanley JA, Negassa A, Edwardes MDD, Forrester JE (2003) Statistical analysis of correlated data using generalized estimating equations: An orientation. *American Journal of Epidemiology*, **157**, 364–375.

Häntzschel J, Goldberg V, Bernhofer C (2005) GIS-based regionalisation of radiation, temperature and coupling measures in complex terrain for low mountain ranges. *Meteorological Applications*, **12**, 33–42.

Hardin JW, Hilbe JM (2003) Generalized Estimating Equations. *Chapman and Hall/CRC: London*.

Hardy OJ, Vekemans X (2002) SPAGeDi: A versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2**, 618–620.

Hastings A (1983) Can spatial variation alone lead to selection for dispersal? *Theoretical Population Biology*, **24**, 244–251.

Hereford J (2009) A quantitative survey of local adaptation and fitness trade-offs. *The American naturalist*, **173**, 579–588.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hjerdt KN, McDonnell JJ, Seibert J, Rodhe A (2004) A new topographic index to quantify downslope controls on local drainage. *Water Resources Research*, **40**, W05602.

Holderegger R, Wagner HH (2008) Landscape Genetics. *BioScience*, **58**, 199–207.

Hosmer DW, Lemeshow S (2000) Introduction to the Logistic Regression Model. In: *Applied Logistic Regression* , pp. 1–30. John Wiley & Sons, Inc.

Hübner S, GÜNTHER T, FLAVELL A *et al.* (2012) Islands and streams: clusters and gene flow in wild barley populations from the Levant. *Molecular Ecology*, **21**, 1115–1129.

Jakobsson M, Rosenberg NA (2007) CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–1806.

Jones MR, Forester BR, Teufel AI *et al.* (2013) Integrating landscape genomics and spatially explicit approaches to detect loci under selection in clinal populations. *Evolution*, **67**, 3455–3468.

Joost S (2006) The geographical dimension of genetic diversity - a GIScience contribution for the conservation of animal genetic resources. EPFL.

Joost S, Bonin A, Bruford MW *et al.* (2007) A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology*, **16**, 3955–3969.

Joost S, Colli L, Baret P V *et al.* (2010) Integrating geo-referenced multiscale and multidisciplinary data for the management of biodiversity in livestock genetic resources. *Animal Genetics*, **41**, 47–63.

Joost S, Kalbermatten M, Bonin A (2008) Spatial analysis method (sam): a software tool combining molecular and environmental data to identify candidate loci for selection. *Molecular Ecology Resources*, **8**, 957–960.

Joost S, Vuilleumier S, Jensen JD *et al.* (2013) Uncovering the genetic basis of adaptive change: on the intersection of landscape genomics and theoretical population genetics. *Molecular Ecology*, n/a–n/a.

Kalbermatten M (2010) Multiscale analysis of high resolution digital elevation models using the wavelet transform. EPFL.

Kalbermatten M, Van De Ville D, Turberg P, Tuia D, Joost S (2012) Multiscale analysis of geomorphological and geological features in high resolution digital elevation models using the wavelet transform. *Geomorphology*, **138**, 352–363.

Kavanagh KD, Haugen TO, Gregersen F, Jernvall J, Vøllestad LA (2010) Contemporary temperature-driven divergence in a Nordic freshwater fish under conditions commonly thought to hinder adaptation. *BMC evolutionary biology*, **10**, 350.

Kawecki TJ, Ebert D (2004) Conceptual issues in local adaptation. *Ecology Letters*, **7**, 1225–1241.

Kimura M (1968) Evolutionary rate at the molecular level. *Nature*, **217**, 624–626.

Kleiner L, Robra JP, Gilliéron P-Y, Schaer P, Mertina C (2010) Lever de limites naturelles par scanner laser aérien (LIDAR) Evaluation et perspectives dans le cadre de la mensuration cadastrale. *Géomatique Suisse*, **4/2010**, 136–139.

Körner C (2003) *Alpine Plant Life: Functional Plant Ecology of High Mountain Ecosystems ; with 47 Tables*. Springer.

De Kort H, Vandepitte K, Bruun HH *et al.* (2014) Landscape genomics and a common garden trial reveal adaptive differentiation to temperature across Europe in the tree species Alnus glutinosa. *Molecular Ecology*, **23**, 4709–4721.

Kozak KH, Graham CH, Wiens JJ (2008) Integrating GIS-based environmental data into evolutionary biology. *Trends in ecology & evolution (Personal edition)*, **23**, 141–148.

Krieger G, Moreira A, Fiedler H *et al.* (2007) TanDEM-X: A satellite formation for high-resolution SAR interferometry. In: *IEEE Transactions on Geoscience and Remote Sensing* , pp. 3317–3340.

Lamaze FC, Sauvage C, Marie A, Garant D, Bernatchez L (2012) Dynamics of introgressive hybridization assessed by SNP population genomics of coding genes in stocked brook charr (Salvelinus fontinalis). *Molecular Ecology*, **21**, 2877–2895.

Landguth EL, Cushman SA, Schwartz MK *et al.* (2010) Quantifying the lag time to detect barriers in landscape genetics. *Molecular Ecology*, **19**, 4179–4191.

Landolt E (1977) *Ökologische Zeigerwerte zur Schweizer Flora*. Geobotan. Inst.

Landolt E, Bäumler B, Erhardt A *et al.* (2010) *Flora indicativa : ökologische Zeigerwerte und biologische Kennzeichen zur Flora der Schweiz und der Alpen = ecological indicator values and biological attributes of the Flora of Switzerland and the Alps*. Ed. des Conservatoire et Jardin botaniques de la ville de Genève ; Haupt.

Lassueur T, Joost S, Randin CF (2006) Very high resolution digital elevation models: Do they improve models of plant species distribution? *Ecological Modelling*, **198**, 139–153.

Leempoel K, Geiser C, Daprà L *et al.* Very high resolution digital elevation models: are multi-scale derived variables ecologically relevant? *Methods in Ecology and Evolution*.

Leempoel K, Joost S (2012) Relatedness and scale dependency in very high resolution digital elevation models derivatives. In: *Open Source Geospatial Research & Education Symposium 2012* , p. 340. Lulu.com, Yverdon-les-Bains, Switzerland.

Legendre P (1993) Spatial autocorrelation: trouble or new paradigm? *Ecology*, **74**, 1659–1673.

Legendre P, Fortin M-J (2010) Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular ecology resources*, **10**, 831–844.

Levin SA (1992) The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology*, **73**, 1943–1967.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment / Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liang KY, Zeger SL (1986) Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika*, **73**, 13–22.

Lischer HEL, Excoffier L (2012) PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298–299.

Lobréaux S, Manel S, Melodelima C (2014) Development of an Arabis alpina genomic contig sequence data set and application to single nucleotide polymorphisms discovery. *Molecular Ecology Resources*, **14**, 411–418.

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, **24**, 1031–1046.

Lowry DB (2010) Landscape evolutionary genomics. *Biology Letters*, **6**, 502–504.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

Lyon SW, Troch PA, Broxton PD, Molotch NP, Brooks PD (2008) Monitoring the timing of snowmelt and the initiation of streamflow using a distributed network of temperature/light sensors. *Ecohydrology*, **1**, 215–224.

Manel S, Albert C, Yoccoz N (2012a) Sampling in Landscape Genomics. In: *Data Production and Analysis in Population Genomics SE - 1* Methods in Molecular Biology. (eds Pompanon F, Bonin A), pp. 3–12. Humana Press.

Manel S, Gugerli F, Thuiller W *et al.* (2012b) Broad-scale adaptive genetic variation in alpine plants is driven by temperature and precipitation. *Molecular Ecology*, **21**, 3729–3738.

Manel S, Holderegger R (2013) Ten years of landscape genetics. *Trends in ecology & evolution*, **28**, 614–21.

Manel S, Joost S, Epperson BK *et al.* (2010a) Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Molecular Ecology*, **19**, 3760–3772.

Manel S, Poncet BN, Legendre P, Gugerli F, Holderegger R (2010b) Common factors drive adaptive genetic variation at different spatial scales in Arabis alpina. *Molecular Ecology*, **19**, 3824–3835.

Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, **18**, 189–197.

Manel S, Segelbacher G (2009) Perspectives and challenges in landscape genetics. *Molecular Ecology*, **18**, 1821–1822.

Manton I (1937) The problem of Biscutella laevigata L. *Annals of Botany*, **51**, 439–465.

Marceau DJ, Hay GJ (1999) Remote Sensing Contributions to the Scale Issue. *Canadian Journal of Remote Sensing*, **25**, 357–366.

McKenna A, Hanna M, Banks E *et al.* (2010) The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

McVicar TR, Van Niel TG, Li L *et al.* (2007) Spatially distributing monthly reference evapotranspiration and pan evaporation considering topographic influences. *Journal of Hydrology*, **338**, 196–220.

Melodelima C, Lobréaux S (2013) Complete Arabis alpina chloroplast genome sequence and insight into its polymorphism. *Meta Gene*, **1**, 65–75.

Miller CL, Laflamme RA (1958) The digital terrain model - Theory and application. *Photogrammetric Engineering*, **24**, 433–442.

De Mita S, Thuillet A-C, Gay L *et al.* (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383–1399.

Montgomery DC, Peck EA (1982) *Introduction to linear regression analysis*. Wiley, New York.

Moore J-S, Bourret V, Dionne M *et al.* (2014) Conservation genomics of anadromous Atlantic salmon across its North American range: outlier loci identify the same patterns of population structure as neutral loci. *Molecular Ecology*, **23**, 5680–5697.

Moore ID, Grayson RB, Ladson AR (1991) Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. *Hydrological Processes*, **5**, 3–30.

Moran PAP (1950) Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23.

Morin PA, Martien KK, Taylor BL (2009) Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources*, **9**, 66–73.

Namroud MC, Beaulieu J, Juge N, Laroche J, Bousquet J (2008) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology*, **17**, 3599–3613.

Nelder JA, Wedderburn RWM (1972) Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, **135**, 370–384.

New M, Lister D, Hulme M, Makin I (2002) A high-resolution data set of surface climate over global land areas. *Climate Research*, **21**, 1–25.

Nielsen R (2005) Molecular signatures of natural selection. *Annual review of genetics*, **39**, 197–218.

Parisod C, Besnard G (2007) Glacial in situ survival in the Western Alps and polytopic autopolyploidy in Biscutella laevigata L. (Brassicaceae). *Molecular Ecology*, **16**, 2755–2767.

Parisod C, Bonvin G (2008) Fine-scale genetic structure and marginal processes in an expanding population of Biscutella laevigata L. (Brassicaceae). *Heredity*, **101**, 536–542.

Parisod C, Christin P-A (2008) Genome-wide association to fine-scale ecological heterogeneity within a continuous population of Biscutella laevigata (Brassicaceae). *New Phytologist*, **178**, 436–447.

Parisod C, Joost S (2010) Divergent selection in trailing- versus leading-edge populations of Biscutella laevigata. *Annals of Botany*, **105**, 655–660.

Patsiou TS, Conti E, Zimmermann NE, Theodoridis S, Randin CF (2014) Topo-climatic microrefugia explain the persistence of a rare endemic plant in the Alps during the last 21 millennia. *Global Change Biology*, **20**, 2286–2300.

Pepin NC, Seidel DJ (2005) A global comparison of surface and free-air temperatures at high elevations. *Journal of Geophysical Research D: Atmospheres*, **110**, 1–15.

Pérez-Figueroa A, GarcÍA-Pereira MJ, Saura M, RolÁN-Alvarez E, Caballero A (2010) Comparing three different methods to detect selective loci using dominant markers. *Journal of Evolutionary Biology*, **23**, 2267–2276.

Petren K (2013) The evolution of landscape genetics. *Evolution*, **67**, 3383–3385.

Poncet B, Herrmann D, Gugerli F (2010) Tracking genes of ecological relevance using genome scan in two independent regional population samples of Arabis alpina. *Molecular Ecology*.

Pradervand J-N, Dubuis A, Pellissier L, Guisan A, Randin C (2014) Very high resolution environmental predictors in species distribution models: Moving beyond topography? . *Progress in Physical Geography* , **38** , 79–96.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, **155**, 945–959.

Pritchard JK, Wen X, Falush D (2007) Documentation for Structure Software: Version 2.2.

Prunier JG, Kaufmann B, Fenet S *et al.* (2013) Optimizing the trade-off between spatial and genetic sampling efforts in patchy populations: Towards a better assessment of functional connectivity using an individual-based sampling scheme. *Molecular Ecology*, **22**, 5516–5530.

Purcell S, Neale B, Todd-Brown K *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, **81**, 559–575.

Randin CF, Vuissoz G, Liston GE, Vittoz P, Guisan A (2009) Introduction of Snow and Geomorphic Disturbance Variables into Predictive Models of Alpine Plant Distribution in the Western Swiss Alps. *Arctic, Antarctic, and Alpine Research*, **41**, 347–361.

Rezakhaniha R, Agianniotis A, Schrauwen JTC *et al.* (2012) Experimental investigation of collagen waviness and orientation in the arterial adventitia using confocal laser scanning microscopy. *Biomechanics and modeling in mechanobiology*, **11**, 461–473.

Richardson JL, Urban MC, Bolnick DI, Skelly DK (2014) Microgeographic adaptation and the spatial scale of evolution. *Trends in Ecology and Evolution*, **29**, 165–176.

Riley SJ, Degloria SD, Elliot R (1999) A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences*, **5**, 23–27.

Le Roux PC, Lenoir J, Pellissier L, Wisz MS, Luoto M (2012) Horizontal, but not vertical, biotic interactions affect fine-scale plant distribution patterns in a low-energy system. *Ecology*, **94**, 671–682.

Saccheri IJ, Rousset F, Watts PC, Brakefield PM, Cook LM (2008) Selection and gene flow on a diminishing cline of melanic peppered moths. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 16212–16217.

Sappington JM, Longshore KM, Thompson DB (2007) Quantifying Landscape Ruggedness for Animal Habitat Analysis: A Case Study Using Bighorn Sheep in the Mojave Desert. *The Journal of Wildlife Management*, **71**, 1419–1426.

Savolainen O (2011) The Genomic Basis of Local Climatic Adaptation. *Science*, **334**, 49–50.

Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature reviews. Genetics*, **14**, 807–20.

Scherrer D, Körner C (2011) Topographically controlled thermal-habitat differentiation buffers alpine plant diversity against climate warming. *Journal of Biogeography*, **38**, 406–416.

Schoville SD, Bonin A, François O *et al.* (2012) Adaptive genetic variation on the landscape: methods and cases. *Annual Review of Ecology, Evolution, and Systematics*, **43**, 23–43.

Schwartz MK, Luikart G, McKelvey KS, Cushman SA (2009) Landscape Genomics: A Brief Perspecitve. In: *Spatial Complexity, Informatics, and Wildlife Conservation* , pp. 165–175.

Schwartz MK, McKelvey KS (2009) Why sampling scheme matters: The effect of sampling scheme on landscape genetic results. *Conservation Genetics*, **10**, 441–452.

Segelbacher G, Cushman SA, Epperson BK *et al.* (2010) Applications of landscape genetics in conservation biology: Concepts and challenges. *Conservation Genetics*, **11**, 375–385.

Seungyong L, Wolberg G, Sung-Yong S (1997) Scattered data interpolation with multilevel B-splines. *Visualization and Computer Graphics, IEEE Transactions on*, **3**, 228–244.

Shaffer JP (1995) Multiple Hypothesis Testing. *Annual Review of Psychology*, **46**, 561–584.

Skelly DK (2004) Microgeographic countergradient variation in the wood frog, Rana sylvatica. *Evolution; international journal of organic evolution*, **58**, 160–165.

Sokal RR, Oden NL, Thomson B a. (1998) Local spatial autocorrelation in biological variables. *Biological Journal of the Linnean Society*, **65**, 41–62.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* , **100** , 9440–9445.

Storfer A, Murphy MA, Evans JS *et al.* (2006) Putting the "landscape" in landscape genetics. *Heredity*, **98**, 128–142.

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 B. EPFL.

Taberlet P, Valentini A, Rezaei HR *et al.* (2008) Are cattle, sheep, and goats endangered species? *Molecular Ecology*, **17**, 275–284.

Tachikawa T, Hato M, Kaku M, Iwasaki A (2011) Characteristics of Aster Gdem Version 2. *2011 Ieee International Geoscience and Remote Sensing Symposium (Igarss)*, 3657–3660.

Tiffin P, Ross-Ibarra J (2014) Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology & Evolution*, **29**, 673–680.

Tobler WR (1970) A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, **46**, 234–240.

Vekemans X, Hardy OJ (2004) New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology*, **13**, 921–935.

Vercauteren N, Destouni G, Dahlberg CJ, Hylander K (2012) Fine-Resolved, Near-Coastal Spatiotemporal Variation of Temperature in Response to Insolation. *Journal of Applied Meteorology and Climatology*, **52**, 1208–1220.

Van De Ville D, Sage D, Balac K, Unser M (2008) The Marr wavelet pyramid and multiscale directional image analysis. *EUSIPCO, August*, 25–29.

Wagner HH, Fortin MJ (2005) Spatial analysis of landscapes: Concepts and statistics. *Ecology*, **86**, 1975–1987.

Wagner HH, Fortin M-J (2013) A conceptual framework for the spatial analysis of landscape genetic data. *Conservation Genetics*.

Wang L, Liu H (2006) An efficient method for identifying and filling surface depressions in digital elevation models for hydrologic analysis and modelling. *International Journal of Geographical Information Science*, **20**, 193–213.

Waples RS, Gaggiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, **15**, 1419–1439.

Wilson JP, Gallant JC (2000) *Terrain Analysis: Principles and Applications* (Wiley, Ed,). Wiley.

Wood JD (1996) The geomorphological characterisation of digital elevation models.

Xiaoye Liu (2008) Airborne LiDAR for DEM generation: some critical issues. *Progress in Physical Geography* , **32** , 31–49.

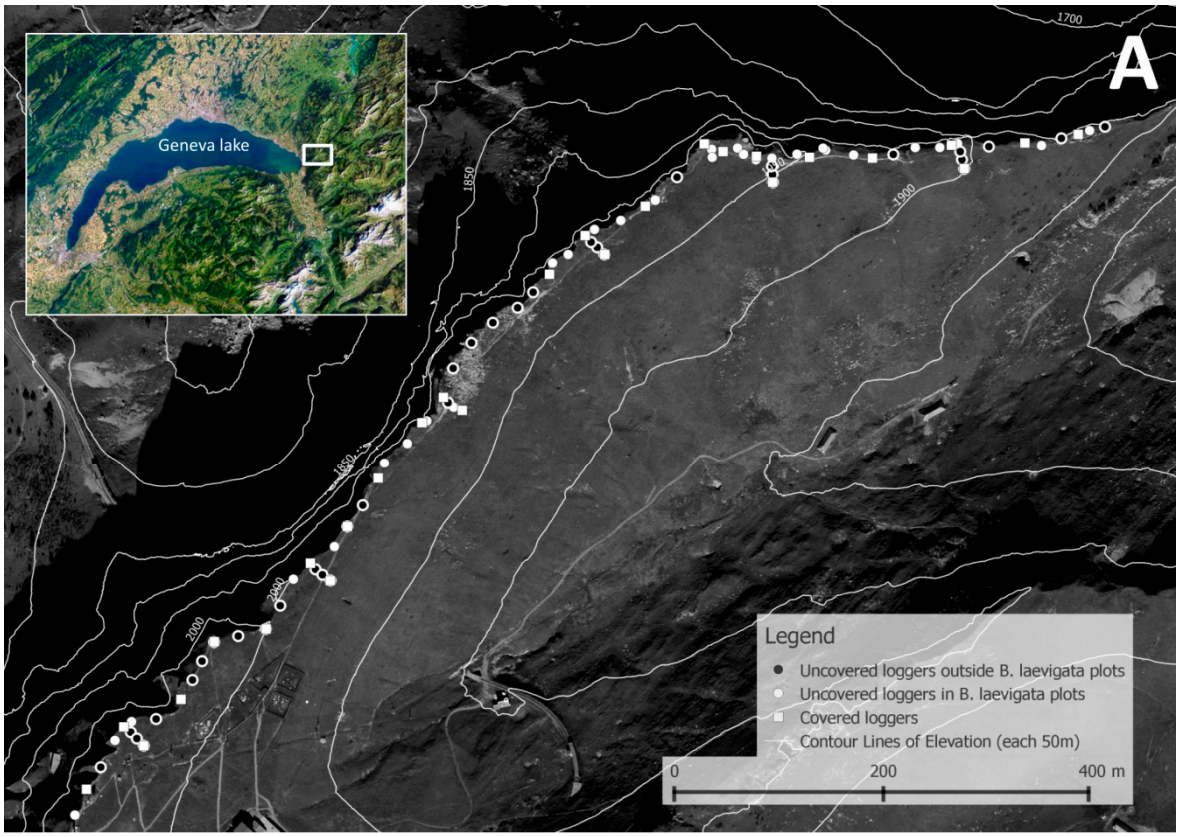
Yokoyama R, Shirasawa RJ, Pike RJ (2002) Visualizing topography by openness: A new application of image processing to digital elevation models. *Photogrammetric Engineering and Remote Sensing*, **68**, 251–266.

Zevenbergen LW, Thorne CR (1987) Quantitative-Analysis of Land Surface-Topography. *Earth Surface Processes and Landforms*, **12**, 47–56.

Zimmermann NE, Kienast F (1999) Predictive mapping of alpine grasslands in Switzerland: Species versus community approach. *Journal of Vegetation Science*, **10**, 469–482.

Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM (2009) Mixed effects models and extensions in ecology with R. *Statistics*, **32**, 209–243.

Figure 1 (A) Study zone and sampling locations for loggers on the ridge of Les Rochers-de-Naye in the Swiss Western Alps. Loggers were disposed at and between Biscutella laevigata locations (not shown). Uncovered and covered loggers were used to measure direct air temperature and ambient temperature respectively. (Background image with 50 m isoelevation lines: Swissimage © 2013 swisstopo (JD100064)). (B) Mean daily direct air temperature and standard deviation (in grey) from the 15 June to the 18 October 2013, measured with uncovered loggers set 15 cm above soil level. Vertical lines delimit the defined periods. Retained periods for following analyses are in bold.



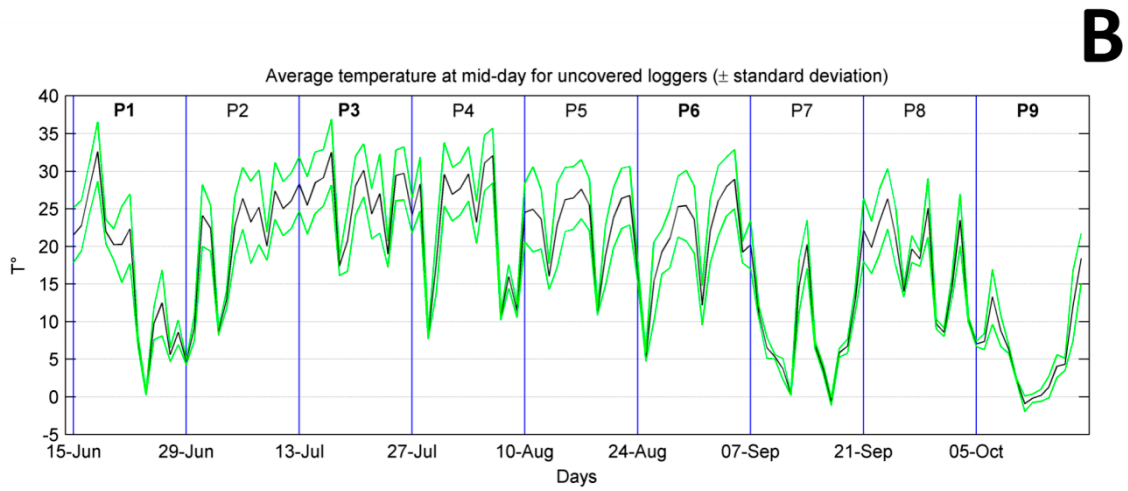


Figure 2 (A) Map of the mean Direct air Temperature (DT) at 1am (M1A) during period P6 (August 24 to September 6). Terrain Wetness Index at 1m resolution computed from the DEM is in the background with 50 m iso-elevation lines. Additional zoom on the ridge to distinguish the loggers and visualize the correlation between the measured variable and the twi. (B) Map of one-time measurements of soil moisture (in percent) with Vector Ruggedness Measure at a 0.5m resolution computed from the DEM is in the background with 50 m iso-elevation lines. Additional zoom on the ridge to distinguish the loggers and visualize the correlation between soil moisture and the vrm. For more details on these results, refer to table 2 and 5.

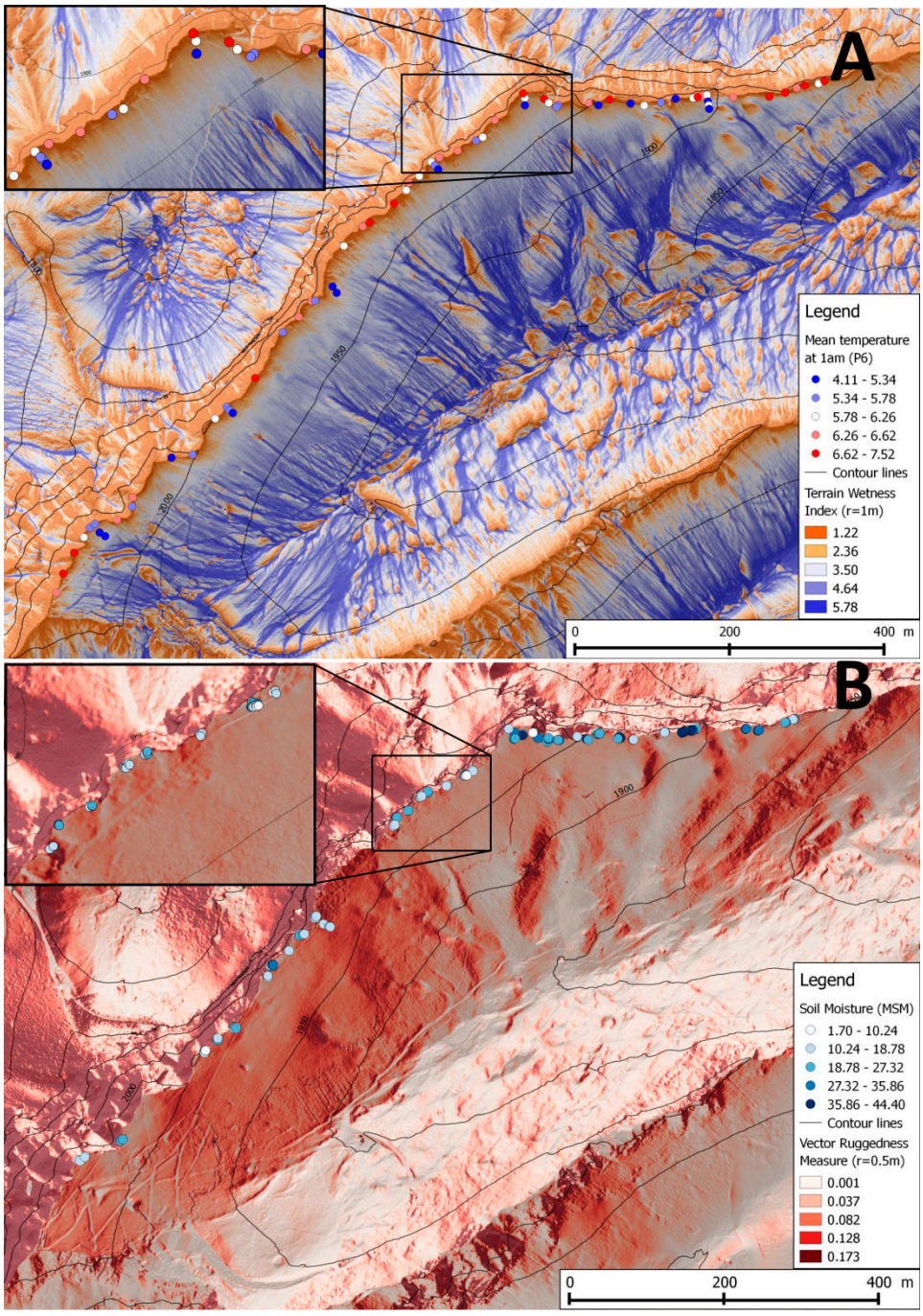


Table 1: Description and parameters of selected DEM variables computed at each resolution (i.e. 0.5, 1, 2, 4m). The full table can be found in Appendix S2.

	Variable	Abreviation	Description	Units	Parameters/ Reference
Primary attributes	Altitude	alt	DEM Altitude	m	
	Slope	slo	Proxies for water flow, snow movements, erosion, solar radiation	radians	Method= Zevenbergen and Thorne, 1987
	Sinus of Aspect (eastness)	eas		radians	
	Profile curvature	vcu		radians/m	
	Plan curvature	hcu		radians/m	
	Downslope distance gradient	ddg	Quantify downslope controls on local drainage	radians	Vertical distance = 2m (Hjerdt et al., 2004)
Secondary attributes	Vector Ruggedness Measure	vrm	Quantifies ruggosity with less correlation to slope	no unit	Radius = 1 pixel (Sappington et al., 2007)
	Terrain Wetness Index	twi	Quantifies topographic control on hydrological processes	Where a is the specific catchment area and S the ddg	$W = \frac{a}{ln(S)}$

Table 2 Summary of multivariate generalized linear models sorted by adjusted R^2(aR^2) in decreasing order for DIRECT AIR TEMPERATURE (DT), measured with uncovered loggers at 15 cm above soil level. First column is the abbreviation of the model showed, with different measured variables and time periods. The second column tells at which resolution (Res) the highest aR^2 was found. Coefficients of each variable are showed when significant and significance is expressed with "" where p-values <0.001 correspond to ***, <0.01: **, <0.05: *. All models at all resolutions can be found in supporting information of this paper. Abbreviations are the following. Measured variables: minimum (MIN), maximum (MAX), mean (MEA), median (MED), mean temperature at 1am (M1A), mean temperature at 1pm (M1P), mean daily range (MDR). Time periods: P1=15 to 28 June, P3=13 to 26 July, P6=24 August to 06 September, P9=05 to 18 October. DEM-derived variables : Altitude (alt), Terrain Wetness Index (twi), Vector Ruggedness Measure (vrm), Eastness (eas), Slope (slo), Horizontal Curvature (hcu), Vertical Curvature (vcu), Downslope Distance Gradient (ddg)*

Model	Res	aR^2	alt	twi	vrm	eas	slo	hcu	vcu	ddg
DT-M1A-P9	0.5	0.69	−0.71***	0.17*	−0.21*					
DT-MIN-P9	2	0.50					0.28**			
DT-M1A-P6	1	0.46	−0.49***	−0.81***		0.25**		−0.20*		
DT-MED-P3	2	0.37	−0.40***	−0.57***						
DT-MEA-P6	2	0.32	−0.35**	−0.80***			0.41**			−0.45*
DT-MDR-P3	0.5	0.22	0.25*	0.47***		−0.41***				
DT-MDR-P1	2	0.19	−0.25*		−0.38***					
DT-MIN-P1	0.5	0.13	−0.37**							

Table 3 Summary of multivariate generalized linear models sorted by adjusted R^2(aR^2) in decreasing order for AMBIENT TEMPERATURE (AT), measured with uncovered loggers at 15 cm above soil level. First column is the abbreviation of the model showed, with different measured variables and time periods. The second column tells at which resolution (Res) the highest aR^2 was found. Coefficients of each variable are showed when significant and significance is expressed with "" where p-values <0.001 correspond to ***, <0.01: **,*

*<0.05: *. All models at all resolutions can be found in supporting information of this paper. Abbreviations as in*

Table 4.1.

Model	Res	aR^2	alt	twi	eas	slo	hcu
AT-MED-P9	0.5	0.89	−0.94***	−0.35**			
AT-MED-P6	4	0.80	−0.74***	−0.44**			
AT-MDR-P3	2	0.49	0.43*		0.52**		−0.69***
AT-MAX-P6	2	0.43				0.48*	−0.44*
AT-M1A-P3	2	0.40	−0.74***				0.48*
AT-MIN-P1	2	0.38	−0.81***	0.87**	−0.75**		0.55*
AT-MDR-P6	1	0.31				0.58*	
AT-MDR-P9	0.5	0.31				0.58*	

Table 4 Summary of multivariate generalized linear models sorted by adjusted R^2 (aR^2) in decreasing order for AMBIENT HUMIDITY (HU), measured with uncovered loggers at 15 cm above soil level. First column is the abbreviation of the model showed, with different measured variables and time periods. The second column tells at which resolution (Res) the highest aR^2 was found. Coefficients of each variable are showed when significant and significance is expressed with "" where p-values <0.001 correspond to ***, <0.01: **, <0.05: *. All models at all resolutions can be found in supporting information of this paper. Abbreviations as in*

Table 4.1

Table 4.1.

Model	Res	aR^2	alt	twi	eas	slo	hcu
HU-M1A-P6	1	0.76		0.82***		0.48**	0.54**
HU-MDR-P1	2	0.48		−0.75***	0.42*		
HU-MED-P3	2	0.47		0.70**			
HU-M1P-P6	0.5	0.38		0.55*		−0.53*	
HU-M1P-P1	2	0.28		0.59**			
HU-MDR-P6	0.5	0.27		−0.63*		0.51*	
HU-M1P-P9	1	0.23		−0.47*			
HU-MDR-P3	1	0.19		−0.76*			

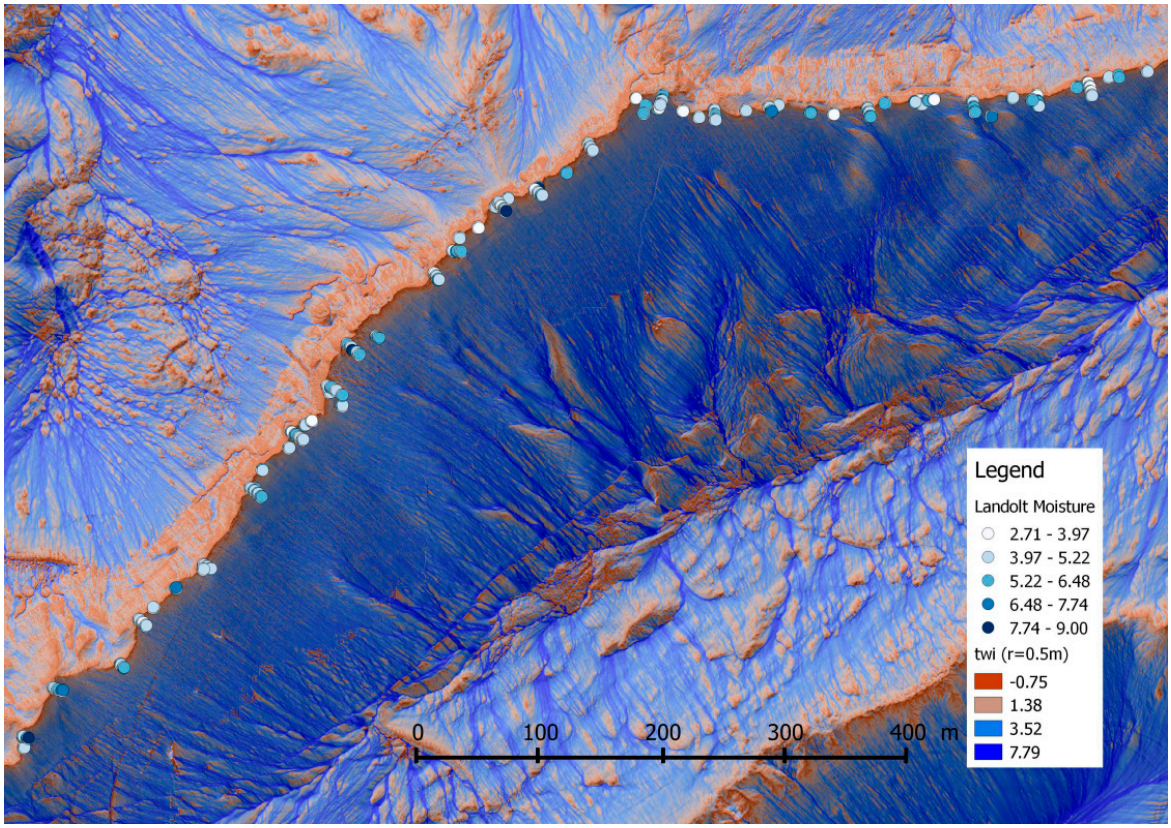
Table 5 Summary of multivariate GLMMs on one-time measurements of SOIL MOISTURE sorted by adjusted R^2 (aR^2). Coefficients of each variable are showed when significant and significance is expressed with "" where p-values <0.001 correspond to ***, <0.01: **, <0.05: *. Abbreviations as in*

Table 4.1.

Res	aR^2	alt	twi	vrm	eas	slo	hcu	vcu	ddg
0.5	0.46	−0.26**		−0.43***					
1	0.43	−0.45***		−0.19**					
2	0.41	−0.46***		−0.20*					
4	0.35	−0.44***				−0.23**			

Supporting information

Appendix S1 Map of *LDT-moisture* (background: Terrain Wetness Index at 0.5m resolution)



Appendix S2 Description and parameters of DEM-derived variables computed at each resolution (i.e. 0.5, 1, 2, 4m)

	Variable	Abreviation	Description	Units	Parameters/ Reference
	Altitude	alt	DEM Altitude	m	/
Primary attributes	Slope	slo	Proxies for water flow, snow movements, erosion, solar radiation	radians	Method= Zevenbergen and Thorne, 1987
	Sinus of Aspect (eastness)	eas		radians	
	Profile curvature	vcu		radians/m	
	Plan curvature	hcu		radians/m	
	Downslope distance gradient	ddg	Quantify downslope controls on local drainage	radians	Vertical distance : 2m Hjerdt et al. (2004)
Secondary attributes	Morphometric protection index	mpi	Expresses the protection of a point from the surrounding relief. It is based on the maximum angle found at the zenith or at nadir from the point over a defined radius	no unit Value is negative when the point is not protected and positive when it is.	Radius = 1 pixel (Yokohama, 2002)
	Terrain ruggedness index	tri	Quantitative measure of topographic heterogeneity	no unit	Radius = 1 pixel (Bönher & Antonic, 2009)
	Vector Ruggedness Measure	vrm	Quantifies ruggosity with less correlation to slope	no unit	Radius = 1 pixel (Sappington et al. 2007)
	Visible Sky	vis	Ratio of the sky area over the obstructed area	no unit	Default Parameters (Riley et al., 1999)
	Sky-view factor	svf	Ratio of the radiation received by a planar surface to the radiation emitted by the entire hemispheric environment	no unit	
	Diffuse Solar radiation in June	df6	Diffuse insolation	kwh/m²	Latitude=46°; Time Period=30 day; Time resolution=0.5h; Day of year=01/06 -> 30/06; Atomspheric effects=Lumped Atmospheric Transmittance (Boehner & Antonic, 2009; Wilson & Gallant, 2000).
	Direct Solar radiation in June	di6	Direct insolation	kwh/m²	
	Total Solar radiation in June	ti6	Sum of direct and diffuse insolation.	kwh/m²	
	Terrain Wetness Index	twi	Quantifies topographic control on hydrological processes	Where a is the specific catchment area and S the ddg	$W = \frac{a}{ln\ S}$

Appendix S3 Table of Spearman correlations coefficients calculated between DEM-derived variables showing a spatial resolution of 0.5m at Landolt sampling locations. Selected variables for all models are displayed on a grey background in the headers. Grey level in the table is lighter when approaching 0 and darker when approaching 1 or -1. Coefficients with an absolute value above 0.6 are written in bold. DEM-derived variables abbreviations : Altitude (alt), sinus of aspect (eas), cosine of aspect (nor), downslope distance gradient (ddg), horizontal curvature (hcu), morphometric protection index (mpi), slope (slo), terrain ruggedness index (tri), vertical curvature (vcu), vector ruggedness measure (vrm), diffuse insolation in June (df6), direct insolation in June (di6), total insolation in June (ti6), sky view factor (svf), visible sky (vis), terrain wetness index (twi).

	alt	eas	ddg	hcu	mpi	slo	tri	vcu	vrm	df6	di6	ti6	svf	vis	twi
alt	**1**	-0.21	0.149	0.178	-0.29	-0.21	-0.2	0.067	0.145	0.409	-0.02	0.013	0.201	0.339	-0.5
eas	-0.21	**1**	0.23	0.141	-0.21	-0.14	-0.11	0.058	-0.25	0.051	0.409	0.363	0.086	0.138	-0.16
ddg	0.149	0.23	**1**	-0.12	0.4	**0.632**	0.56	-0.04	0.587	-0.48	-0.39	-0.4	**-0.62**	-0.36	-0.17
hcu	0.178	0.141	-0.12	**1**	**-0.71**	-0.42	-0.39	**0.791**	-0.41	0.406	0.329	0.359	0.423	**0.647**	-0.59
mpi	-0.29	-0.21	0.4	**-0.71**	**1**	**0.841**	**0.8**	-0.58	**0.664**	**-0.82**	**-0.64**	**-0.67**	**-0.83**	**-0.88**	0.505
slo	-0.21	-0.14	**0.632**	-0.42	**0.841**	**1**	**0.987**	-0.29	**0.804**	**-0.9**	**-0.82**	**-0.84**	**-0.96**	**-0.67**	0.298
tri	-0.2	-0.11	0.56	-0.39	**0.8**	**0.987**	**1**	-0.24	**0.835**	**-0.93**	**-0.86**	**-0.88**	**-0.97**	**-0.63**	0.261
vcu	0.067	0.058	-0.04	**0.791**	-0.58	-0.29	-0.24	**1**	-0.23	0.255	0.166	0.207	0.288	**0.668**	**-0.7**
vrm	0.145	-0.25	0.587	-0.41	**0.664**	**0.804**	**0.835**	-0.23	**1**	**-0.71**	**-0.82**	**-0.83**	**-0.85**	-0.54	0.169
df6	0.409	0.051	-0.48	0.406	**-0.82**	**-0.9**	**-0.93**	0.255	**-0.71**	**1**	**0.752**	**0.791**	**0.947**	**0.723**	-0.35
di6	-0.02	0.409	-0.39	0.329	**-0.64**	**-0.82**	**-0.86**	0.166	**-0.82**	**0.752**	**1**	**0.993**	**0.824**	0.444	-0.18
ti6	0.013	0.363	-0.4	0.359	**-0.67**	**-0.84**	**-0.88**	0.207	**-0.83**	**0.791**	**0.993**	**1**	**0.85**	0.495	-0.23
svf	0.201	0.086	**-0.62**	0.423	**-0.83**	**-0.96**	**-0.97**	0.288	**-0.85**	**0.947**	**0.824**	**0.85**	**1**	**0.718**	-0.3
vis	0.339	0.138	-0.36	**0.647**	**-0.88**	**-0.67**	**-0.63**	**0.668**	-0.54	**0.723**	0.444	0.495	**0.718**	**1**	**-0.71**
twi	-0.5	-0.16	-0.17	-0.59	0.505	0.298	0.261	**-0.7**	0.169	-0.35	-0.18	-0.23	-0.3	**-0.71**	**1**

Appendix S5 Summary of multivariate generalized linear models for DIRECT AIR TEMPERATURE (DT), measured with uncovered loggers at 15 cm above soil level. First column is the abbreviation of the model shown, with different measured variables and time periods. Coefficients of each variable are shown when significant and significance is expressed with "*" where p-values <0.001 correspond to ***, <0.01: **, <0.05: *. Abbreviations are the following. Measured variables: minimum (MIN), maximum (MAX), mean (MEA), median (MED), mean temperature at 1am (M1A), mean temperature at 1pm (M1P), mean daily range (MDR). Time periods: P1=15 to 28 June, P3=13 to 26 July, P6=24 August to 06 September, P9=05 to 18 October. DEM-derived variables abbreviations as in Appendix S3.

Model	Res	aR²	alt	twi	vrm	eas	slo	hcu	vcu	ddg
DT-M1A-P1	0.5	0.49	-0.77***	-0.5***		0.22*				
DT-M1A-P1	1	0.52	-0.79***	-0.52***						
DT-M1A-P1	2	0.52	-0.64***	-0.49***						
DT-M1A-P1	4	0.49	-0.6***	-0.54***						
DT-M1A-P3	0.5	0.39	-0.53***	-0.62***		0.32**				
DT-M1A-P3	1	0.44	-0.54***	-0.75***		0.18*				
DT-M1A-P3	2	0.45	-0.38***	-0.67***						
DT-M1A-P3	4	0.32	-0.27**	-0.78***					-0.48*	
DT-M1A-P6	0.5	0.28	-0.46***	-0.58***		0.27*				
DT-M1A-P6	1	0.46	-0.49***	-0.81***		0.25**		-0.2*		
DT-M1A-P6	2	0.42	-0.32**	-0.69***		0.28*				
DT-M1A-P6	4	0.31	-0.3**						-0.48*	
DT-M1A-P9	0.5	0.69	-0.71***	0.17*	-0.21*					
DT-M1A-P9	1	0.65	-0.72***				0.23**			
DT-M1A-P9	2	0.66	-0.73***					-0.2*		0.29*
DT-M1A-P9	4	0.65	-0.69***			0.17*				
DT-M1P-P1	0.5	0.21	-0.27*	0.35**						
DT-M1P-P1	1	0.19	-0.26*	0.28*						
DT-M1P-P1	2	0.25	-0.39***		-0.33**					
DT-M1P-P1	4	0.21	-0.36**							-0.5**
DT-M1P-P3	0.5	0.18		0.38**		-0.37**				
DT-M1P-P3	1	0.10		0.26*						
DT-M1P-P3	2	0.12								-0.33**
DT-M1P-P3	4	0.01								
DT-M1P-P6	0.5	0.11				-0.26*	0.26*			
DT-M1P-P6	1	0.08					0.25*			
DT-M1P-P6	2	0.08					0.27*			
DT-M1P-P6	4	0.00								
DT-M1P-P9	0.5	0.23		0.38**	-0.33*		0.44*			
DT-M1P-P9	1	0.16					0.45***			
DT-M1P-P9	2	0.14					0.31*			-0.29**
DT-M1P-P9	4	0.04								

Model	Res	aR²	alt	twi	vrm	eas	slo	hcu	vcu	ddg
DT-MAX-P1	0.5	0.10		0.27*						
DT-MAX-P1	1	0.05			-0.25*					
DT-MAX-P1	2	0.11			-0.34**					
DT-MAX-P1	4	0.07								-0.96*
DT-MAX-P3	0.5	0.11		0.32**		-0.27*				
DT-MAX-P3	1	0.10		0.27*						
DT-MAX-P3	2	0.13								-0.35**
DT-MAX-P3	4	0.00								
DT-MAX-P6	0.5	0.00								
DT-MAX-P6	1	0.00								
DT-MAX-P6	2	0.00								
DT-MAX-P6	4	0.00								
DT-MAX-P9	0.5	0.16		0.38**			0.27*			
DT-MAX-P9	1	0.11		0.24*			0.24*			
DT-MAX-P9	2	0.16					0.29*			-0.35**
DT-MAX-P9	4	0.06								-0.27*
DT-MEA-P1	0.5	0.32	-0.44***	0.3*						
DT-MEA-P1	1	0.31	-0.42***	0.25*						
DT-MEA-P1	2	0.37	-0.53***		-0.3**					
DT-MEA-P1	4	0.34	-0.5***							-0.48**
DT-MEA-P3	0.5	0.15					0.38*		0.24*	
DT-MEA-P3	1	0.13	-0.34**	-0.31*			0.26*			
DT-MEA-P3	2	0.09	-0.24*	-0.34*						
DT-MEA-P3	4	0.11	-0.23*	-0.29*						
DT-MEA-P6	0.5	0.30	-0.4***				0.46**	-0.3*	0.29*	
DT-MEA-P6	1	0.27	-0.48***	-0.34**			0.33**			
DT-MEA-P6	2	0.32	-0.35**	-0.8***			0.41**			-0.45*
DT-MEA-P6	4	0.24	-0.34**	-1.26***						-1.17**
DT-MEA-P9	0.5	0.48	-0.41***	0.31**	-0.36**		0.37*			0.36*
DT-MEA-P9	1	0.40	-0.4***				0.37***			
DT-MEA-P9	2	0.36	-0.45***				0.26*		-0.22*	
DT-MEA-P9	4	0.28	-0.39***			0.24*				
DT-MED-P1	0.5	0.58	-0.75***							
DT-MED-P1	1	0.59	-0.74***				0.17*			
DT-MED-P1	2	0.59	-0.75***							
DT-MED-P1	4	0.56	-0.74***							
DT-MED-P3	0.5	0.32	-0.54***	-0.51***		0.27*				
DT-MED-P3	1	0.35	-0.57***	-0.6***						
DT-MED-P3	2	0.37	-0.4***	-0.57***						
DT-MED-P3	4	0.23	-0.32**	-0.37**						
DT-MED-P6	0.5	0.32	-0.56***	-0.65***		0.25*				
DT-MED-P6	1	0.47	-0.56***	-0.77***		0.2*		-0.21*		
DT-MED-P6	2	0.46	-0.42***	-0.67***		0.21*				
DT-MED-P6	4	0.39	-0.4***	-0.55*						
DT-MED-P9	0.5	0.79	-0.84***				0.2**		0.11*	
DT-MED-P9	1	0.77	-0.86***	-0.14*			0.2**			
DT-MED-P9	2	0.75	-0.79***		0.19*				-0.18*	
DT-MED-P9	4	0.74	-0.82***							

Model	Res	aR²	alt	twi	vrm	eas	slo	hcu	vcu	ddg
DT-MIN-P1	0.5	0.13	-0.37**							
DT-MIN-P1	1	0.12	-0.42***							
DT-MIN-P1	2	0.13	-0.39**							
DT-MIN-P1	4	0.12	-0.37**							
DT-MIN-P3	0.5	0.18	-0.27*	-0.55***		0.23*				
DT-MIN-P3	1	0.00								
DT-MIN-P3	2	0.28	-0.24*							
DT-MIN-P3	4	0.29	-0.28**						-0.57*	0.96***
DT-MIN-P6	0.5	0.11		-0.38**		0.31*				
DT-MIN-P6	1	0.16	-0.28*	-0.5***				-0.24*		
DT-MIN-P6	2	0.12		-0.42***		0.27*				
DT-MIN-P6	4	0.03								
DT-MIN-P9	0.5	0.44		0.82***				0.27*		0.4***
DT-MIN-P9	1	0.33		0.49***			0.26*			
DT-MIN-P9	2	0.50					0.28**			
DT-MIN-P9	4	0.30		0.4*		0.47**				
DT-MDR-P1	0.5	0.14		0.42***		-0.24*				
DT-MDR-P1	1	0.12		0.38**						
DT-MDR-P1	2	0.19	-0.25*		-0.38***					
DT-MDR-P1	4	0.14								-0.53**
DT-MDR-P3	0.5	0.22	0.25*	0.47***		-0.41***				
DT-MDR-P3	1	0.15	0.25*	0.39**			0.25*			
DT-MDR-P3	2	0.14					0.24*			-0.37**
DT-MDR-P3	4	0.03								
DT-MDR-P6	0.5	0.14		0.39**		-0.29*				
DT-MDR-P6	1	0.11		0.38**						
DT-MDR-P6	2	0.16					0.25*			-0.36**
DT-MDR-P6	4	0.06								-0.63*
DT-MDR-P9	0.5	0.13					0.47**		0.33*	
DT-MDR-P9	1	0.06					0.32*			
DT-MDR-P9	2	0.04					0.3*			
DT-MDR-P9	4	0.00								
DT-RAS-P1	0.5	0.20		0.48***		-0.29*				
DT-RAS-P1	1	0.14		0.4***						
DT-RAS-P1	2	0.19								
DT-RAS-P1	4	0.17	-0.23*	0.39**						
DT-RAS-P3	0.5	0.19		0.41***		-0.32**				
DT-RAS-P3	1	0.15		0.43***						
DT-RAS-P3	2	0.19								-0.54***
DT-RAS-P3	4	0.07								-0.29*
DT-RAS-P6	0.5	0.10	0.31*	0.35*		-0.25*				
DT-RAS-P6	1	0.12	0.33**	0.38**						
DT-RAS-P6	2	0.12			-0.32**					
DT-RAS-P6	4	0.00								
DT-RAS-P9	0.5	0.13		0.34**			0.28*			
DT-RAS-P9	1	0.12		0.33*						
DT-RAS-P9	2	0.17					0.28*			-0.36**
DT-RAS-P9	4	0.00								

Model	Res	aR²	alt	twi	vrm	eas	slo	hcu	vcu	ddg
DT-STD-P1	0.5	0.13		0.48***						
DT-STD-P1	1	0.11		0.35**						
DT-STD-P1	2	0.15								
DT-STD-P1	4	0.13								-0.58**
DT-STD-P3	0.5	0.17		0.41**		-0.37**				
DT-STD-P3	1	0.13		0.34*			0.27*			
DT-STD-P3	2	0.11					0.25*			-0.31**
DT-STD-P3	4	0.03								
DT-STD-P6	0.5	0.13		0.3*		-0.3*				
DT-STD-P6	1	0.11		0.28*		-0.25*				
DT-STD-P6	2	0.13					0.28*			-0.42**
DT-STD-P6	4	0.04								-0.57*
DT-STD-P9	0.5	0.13		0.31*			0.49**			
DT-STD-P9	1	0.08					0.32*			
DT-STD-P9	2	0.09					0.24*			-0.27*
DT-STD-P9	4	0.00								

Appendix S6 Summary of multivariate generalized linear models for AMBIENT TEMPERATURE (AT), measured with uncovered loggers at 15 cm above soil level. First column is the abbreviation of the model shown, with different measured variables and time periods. Coefficients of each variable are shown when significant and significance is expressed with "*" where p-values <0.001 correspond to ***, <0.01: **, <0.05: *. Measured variables, DEM-derived variables and periods abbreviations as in Appendix S3&4.

Model	Res	aR²	alt	twi	vrm	eas	slo	hcu	vcu	ddg
AT-M1A-P1	0.5	0.34	-0.82**							
AT-M1A-P1	1	0.37	-0.88**	-0.45*						
AT-M1A-P1	2	0.38	-0.73**	-0.41*						
AT-M1A-P1	4	0.37	-0.6**	-0.39*						
AT-M1A-P3	0.5	0.21	-0.51*							
AT-M1A-P3	1	0.33	-0.69**	-0.64*						
AT-M1A-P3	2	0.40	-0.74***					0.48*		
AT-M1A-P3	4	0.18	-0.46*							
AT-M1A-P6	0.5	0.45	-0.71**							
AT-M1A-P6	1	0.61	-0.67**	-0.56*						
AT-M1A-P6	2	0.58	-0.56**	-0.49*						
AT-M1A-P6	4	0.60	-0.6**	-0.47**						
AT-M1A-P9	0.5	0.42	-0.71**							
AT-M1A-P9	1	0.57	-0.66**	-0.38*			0.53*			
AT-M1A-P9	2	0.48	-0.53*	-0.74*						
AT-M1A-P9	4	0.37	-0.63**							
AT-M1P-P1	0.5	0.29								
AT-M1P-P1	1	0.31	-0.56*							
AT-M1P-P1	2	0.31								
AT-M1P-P1	4	0.32								
AT-M1P-P3	0.5	0.23					0.5*			
AT-M1P-P3	1	0.24								
AT-M1P-P3	2	0.42				0.45*		-0.66**		
AT-M1P-P3	4	0.00								
AT-M1P-P6	0.5	0.23					0.52*			
AT-M1P-P6	1	0.34					0.6**			
AT-M1P-P6	2	0.07								
AT-M1P-P6	4	0.07								
AT-M1P-P9	0.5	0.33					0.6**			
AT-M1P-P9	1	0.50					0.53*			
AT-M1P-P9	2	0.37				0.43*				
AT-M1P-P9	4	0.27				0.55*				
AT-MAX-P1	0.5	0.13		0.42*						
AT-MAX-P1	1	0.03								
AT-MAX-P1	2	0.17					0.45*			
AT-MAX-P1	4	0.12								
AT-MAX-P3	0.5	0.06								
AT-MAX-P3	1	0.14								
AT-MAX-P3	2	0.19		-0.81*		0.64*		-0.81**		
AT-MAX-P3	4	0.00								
AT-MAX-P6	0.5	0.35					0.55**			
AT-MAX-P6	1	0.15		0.51*		-0.48*				
AT-MAX-P6	2	0.43					0.48*	-0.44*		
AT-MAX-P6	4	0.12								
AT-MAX-P9	0.5	0.41					0.55*			
AT-MAX-P9	1	0.37								
AT-MAX-P9	2	0.29								
AT-MAX-P9	4	0.25						-0.5*		

Model	Res	aR²	alt	twi	vrm	eas	slo	hcu	vcu	ddg
AT-MEA-P1	0.5	0.41	-0.7**							
AT-MEA-P1	1	0.41	-0.68**							
AT-MEA-P1	2	0.37	-0.72**							
AT-MEA-P1	4	0.42	-0.58*							
AT-MEA-P3	0.5	0.19								
AT-MEA-P3	1	0.14								
AT-MEA-P3	2	0.32				0.63**		-0.61*		
AT-MEA-P3	4	0.07								
AT-MEA-P6	0.5	0.27					0.55*			
AT-MEA-P6	1	0.35					0.61**			
AT-MEA-P6	2	0.00								
AT-MEA-P6	4	0.00								
AT-MEA-P9	0.5	0.41					0.53*			
AT-MEA-P9	1	0.55					0.64**			
AT-MEA-P9	2	0.32	-0.58*							
AT-MEA-P9	4	0.20	-0.48*							
AT-MED-P1	0.5	0.35	-0.7**							
AT-MED-P1	1	0.36	-0.7**							
AT-MED-P1	2	0.35	-0.7**							
AT-MED-P1	4	0.36	-0.71**							
AT-MED-P3	0.5	0.16	-0.43*							
AT-MED-P3	1	0.22	-0.58*							
AT-MED-P3	2	0.21	-0.46*							
AT-MED-P3	4	0.16	-0.43*							
AT-MED-P6	0.5	0.66	-0.84***							
AT-MED-P6	1	0.72	-0.82***	-0.5*						
AT-MED-P6	2	0.73	-0.72***	-0.4*						
AT-MED-P6	4	0.80	-0.74***	-0.44**						
AT-MED-P9	0.5	0.89	-0.94***	-0.35**						
AT-MED-P9	1	0.86	-0.89***	-0.35*						
AT-MED-P9	2	0.87	-0.92***	-0.25*						
AT-MED-P9	4	0.84	-0.88***							
AT-MIN-P1	0.5	0.15								
AT-MIN-P1	1	0.27	-0.55*					0.6*		
AT-MIN-P1	2	0.38	-0.81***	0.87**		-0.75**		0.55*		
AT-MIN-P1	4	0.12	-0.4*							
AT-MIN-P3	0.5	0.10								
AT-MIN-P3	1	0.10								
AT-MIN-P3	2	0.41	-0.8***	0.66*		-0.51*		0.92**		
AT-MIN-P3	4	0.10								
AT-MIN-P6	0.5	0.00								
AT-MIN-P6	1	0.00								
AT-MIN-P6	2	0.00								
AT-MIN-P6	4	0.00								
AT-MIN-P9	0.5	0.13								
AT-MIN-P9	1	0.19						-0.42*		
AT-MIN-P9	2	0.43				0.67**		-0.89*		
AT-MIN-P9	4	0.26				0.83*				

Model	Res	aR²	alt	twi	vrm	eas	slo	hcu	vcu	ddg
AT-MDR-P1	0.5	0.14		0.41*						
AT-MDR-P1	1	0.19		0.49*						
AT-MDR-P1	2	0.15								
AT-MDR-P1	4	0.18				0.54*				
AT-MDR-P3	0.5	0.19					0.46*			
AT-MDR-P3	1	0.21					0.48*			
AT-MDR-P3	2	0.49	0.43*			0.52**		-0.69***		
AT-MDR-P3	4	0.00								
AT-MDR-P6	0.5	0.26					0.54*			
AT-MDR-P6	1	0.31					0.58*			
AT-MDR-P6	2	0.08								
AT-MDR-P6	4	0.10								
AT-MDR-P9	0.5	0.31					0.58*			
AT-MDR-P9	1	0.26					0.54*			
AT-MDR-P9	2	0.22								
AT-MDR-P9	4	0.26						-0.53*		
AT-RAS-P1	0.5	0.38		0.73***						
AT-RAS-P1	1	0.28		0.58**						
AT-RAS-P1	2	0.30						-0.45*		
AT-RAS-P1	4	0.24				0.52*				
AT-RAS-P3	0.5	0.15								
AT-RAS-P3	1	0.22	0.54*					-0.49*		
AT-RAS-P3	2	0.59	0.75***			0.48*		-0.96***		
AT-RAS-P3	4	0.05								
AT-RAS-P6	0.5	0.40	0.59*				0.54*			
AT-RAS-P6	1	0.35	0.56*	0.62*						
AT-RAS-P6	2	0.36		0.53*						
AT-RAS-P6	4	0.22								
AT-RAS-P9	0.5	0.37								
AT-RAS-P9	1	0.39								
AT-RAS-P9	2	0.38		0.68**						
AT-RAS-P9	4	0.33								
AT-STD-P1	0.5	0.10								
AT-STD-P1	1	0.09								
AT-STD-P1	2	0.08								
AT-STD-P1	4	0.11				0.47*				
AT-STD-P3	0.5	0.17					0.44*			
AT-STD-P3	1	0.18					0.45*			
AT-STD-P3	2	0.41	0.4*			0.51**		-0.62**		
AT-STD-P3	4	0.00								
AT-STD-P6	0.5	0.27					0.55*			
AT-STD-P6	1	0.31					0.58*			
AT-STD-P6	2	0.15								
AT-STD-P6	4	0.10								
AT-STD-P9	0.5	0.29								
AT-STD-P9	1	0.28		0.51*						
AT-STD-P9	2	0.33		0.64**						
AT-STD-P9	4	0.25								

Appendix S7 Summary of multivariate generalized linear models for AMBIENT HUMIDITY (HU), measured with uncovered loggers at 15 cm above soil level. First column is the abbreviation of the model shown, with different measured variables and time periods. Coefficients of each variable are shown when significant and significance is expressed with "*" where p-values <0.001 correspond to ***, <0.01: **, <0.05: *. Measured variables, DEM-derived variables and periods abbreviations as in Appendix S3&4.

Model	Res	aR²	alt	twi	vrm	eas	slo	hcu	vcu	ddg
HU-M1A-P1	0.5	0.18		0.47*						
HU-M1A-P1	1	0.12								
HU-M1A-P1	2	0.12								
HU-M1A-P1	4	0.15		0.44*						
HU-M1A-P6	0.5	0.47		0.56**						
HU-M1A-P6	1	0.76		0.82***			0.48**	0.54**		
HU-M1A-P6	2	0.44		0.71**						
HU-M1A-P6	4	0.62								
HU-M1A-P9	0.5	0.16								
HU-M1A-P9	1	0.18								
HU-M1A-P9	2	0.06								
HU-M1A-P9	4	0.20				0.55*				
HU-M1P-P1	0.5	0.27		0.96**				0.59*		
HU-M1P-P1	1	0.21		0.76**						
HU-M1P-P1	2	0.28		0.59**						
HU-M1P-P1	4	0.10								
HU-M1P-P3	0.5	0.11		0.63*						
HU-M1P-P3	1	0.27		0.88**				0.54*		
HU-M1P-P3	2	0.10		0.84*						
HU-M1P-P3	4	0.00								
HU-M1P-P6	0.5	0.38		0.55*			-0.53*			
HU-M1P-P6	1	0.16								
HU-M1P-P6	2	0.28		0.7*			-0.8*			
HU-M1P-P6	4	0.00								
HU-M1P-P9	0.5	0.22								
HU-M1P-P9	1	0.23		-0.47*						
HU-M1P-P9	2	0.30								
HU-M1P-P9	4	0.15								

Model	Res	aR²	alt	twi	vrm	eas	slo	hcu	vcu	ddg
HU-MEA-P1	0.5	0.17		0.56*						
HU-MEA-P1	1	0.14		0.43*						
HU-MEA-P1	2	0.30		0.61**						
HU-MEA-P1	4	0.15		0.45*						
HU-MEA-P3	0.5	0.12		0.58*						
HU-MEA-P3	1	0.31		0.74**		-0.45*				
HU-MEA-P3	2	0.35		0.75**						
HU-MEA-P3	4	0.09								
HU-MEA-P6	0.5	0.40	0.45*	0.77**						
HU-MEA-P6	1	0.42	0.46*	0.76**						
HU-MEA-P6	2	0.34		0.64**						
HU-MEA-P6	4	0.38		0.62**						
HU-MEA-P9	0.5	0.00								
HU-MEA-P9	1	0.00								
HU-MEA-P9	2	0.00								
HU-MEA-P9	4	0.00								
HU-MED-P3	0.5	0.16		0.63*						
HU-MED-P3	1	0.35		0.78**						
HU-MED-P3	2	0.47		0.7**						
HU-MED-P3	4	0.11								
HU-MED-P6	0.5	0.49	0.58**	0.67*						
HU-MED-P6	1	0.63	0.53**	0.72**						
HU-MED-P6	2	0.59		0.72***						
HU-MED-P6	4	0.65	0.34*	0.7***						
HU-MIN-P1	0.5	0.10		0.58*						
HU-MIN-P1	1	0.07								
HU-MIN-P1	2	0.24						-0.56**		
HU-MIN-P1	4	0.00								
HU-MIN-P3	0.5	0.06								
HU-MIN-P3	1	0.12		0.56*						
HU-MIN-P3	2	0.16		0.74*						
HU-MIN-P3	4	0.11								
HU-MIN-P6	0.5	0.09								
HU-MIN-P6	1	0.00								
HU-MIN-P6	2	0.22						0.53*		
HU-MIN-P6	4	0.09								

Model	Res	aR²	alt	twi	vrm	eas	slo	hcu	vcu	ddg
HU-MDR-P1	0.5	0.27	-0.62*	-0.7*		0.42*				
HU-MDR-P1	1	0.16								
HU-MDR-P1	2	0.48		-0.75***		0.42*				
HU-MDR-P1	4	0.27					0.42*			
HU-MDR-P3	0.5	0.09								
HU-MDR-P3	1	0.19		-0.76*						
HU-MDR-P3	2	0.07								
HU-MDR-P3	4	0.00								
HU-MDR-P6	0.5	0.27		-0.63*			0.51*			
HU-MDR-P6	1	0.20		-0.66*						
HU-MDR-P6	2	0.24		-0.94*			0.7*			
HU-MDR-P6	4	0.00								
HU-MDR-P9	0.5	0.00								
HU-MDR-P9	1	0.00								
HU-MDR-P9	2	0.00								
HU-MDR-P9	4	0.00								
HU-RAS-P1	0.5	0.24	-0.52*							
HU-RAS-P1	1	0.21	-0.57*							
HU-RAS-P1	2	0.37	-0.55*				0.51*			
HU-RAS-P1	4	0.22	-0.58*							
HU-RAS-P3	0.5	0.08								
HU-RAS-P3	1	0.08								
HU-RAS-P3	2	0.25						-0.63**		
HU-RAS-P3	4	0.00								
HU-RAS-P6	0.5	0.56	0.81***							
HU-RAS-P6	1	0.57	0.77***	0.47*						
HU-RAS-P6	2	0.53	0.63**							
HU-RAS-P6	4	0.56	0.56**							
HU-RAS-P9	0.5	0.40		0.63**						
HU-RAS-P9	1	0.29		0.52*						
HU-RAS-P9	2	0.25		0.55*						
HU-RAS-P9	4	0.26		0.54*						
HU-STD-P1	0.5	0.20		-0.59*						
HU-STD-P1	1	0.17		-0.68*						
HU-STD-P1	2	0.27		-0.58**						
HU-STD-P1	4	0.14		-0.43*						
HU-STD-P3	0.5	0.11		-0.64*						
HU-STD-P3	1	0.16		-0.76*						
HU-STD-P3	2	0.00								
HU-STD-P3	4	0.00								
HU-STD-P6	0.5	0.25		-0.51*						
HU-STD-P6	1	0.07								
HU-STD-P6	2	0.18								
HU-STD-P6	4	0.07								
HU-STD-P9	0.5	0.00								
HU-STD-P9	1	0.00								
HU-STD-P9	2	0.00								
HU-STD-P9	4	0.09								

Appendix IV. **Curriculum Vitae**

CONTACT INFORMATION

École Polytechnique Fédérale de Lausanne	kevin.leempoel@epfl.ch
Environmental Engineering Institute - LASIG	k.leempoel@gmail.com
GC D2 413	Phone: +41 (21) 693 57 84
Station 18, CH-1015 Lausanne, Switzerland	Mobile: +41 79 953 50 49

PROFESSIONAL INTERESTS

- Conservation biology
- Geography of diseases
- Human geography
- Environmental policies
- Sustainable development
- Open-source technologies

PROFESSIONAL EXPERIENCE

Feb 2011 – present **PhD Student**

École Polytechnique Fédérale de Lausanne (EPFL), Environmental Engineering Institute, Lausanne, Switzerland

http://people.epfl.ch/kevin.leempoel?lang=fr

Thesis topic: Relevance of Geographic information systems and digital elevation models to study local adaptation of plants and animals.

Supervisor: François Golay

Co-supervisor: Stéphane Joost

Additional responsibilities:

- Head of the EPFL Geo-database
- Teaching assistant in spatial analysis and Geovisualisation
- Assistant to student projects

Sep 2008 – Sep 2010 **M.Sc. in Organisms' Biology.**
Université Libre de Bruxelles (Free University of Brussels), Belgium

- Ecology-Ethology section
- Laboratory of Complexity and Dynamics of Tropical Systems
- Professor Farid Dahdouh-Guebas

Master's thesis : Heterogeneity of the spatial structure in mangroves using GeoEye-1 imagery and GIS-analyses : a case study in the Zhanjiang Mangrove National Nature Reserve (China)

Grade: "Great distinction" (16/20)

Sep 2008 – Feb 2009 Erasmus study exchange program – *Università di Bologna (Italy)*

Sep 2005 - Sep 2008 **B.Sc. in Biology**
Université Libre de Bruxelles (Free University of Brussels), Belgium

Sep 2004 - Sep 2005 1st year of B.Sc. in Physics
Université Libre de Bruxelles (Free University of Brussels), Belgium

Sep 1999 - Sep 2004 Secondary school (High School)
Athénée Royal d'Uccle 2, Brussels, Belgium
Section: Mathematics and sciences.

AWARDS

Sep 2010 Award: Troisième prix AScBr (obtained for the master thesis)

SKILLS & ACTIVITIES

Skills Landscape Genetics, Molecular Ecology, Conservation Genetics, Mangrove Ecology, Geographic Information Systems, Remote Sensing, Matlab, R

Langages French (native), English (fluent), Italian (fair knowledge)

Hobbies Travelling, Ski, Hiking, Cinema, Video games, Animal behaviour

JOURNAL PUBLICATIONS

Leempoel K, Geiser C, Daprà L et al. Very high resolution digital elevation models: are multi-scale derived variables ecologically relevant? Methods in Ecology and Evolution (Accepted).

K. Leempoel, B. Satyaranayana, C. Bourgeois, J. Zhang, M. Chen, J. Wang, J. Bogaert, F. Dahdouh-Guebas: *Dynamics in mangroves assessed by high-resolution and multi-temporal satellite data: a case study in Zhanjiang Mangrove National Nature Reserve (ZMNNR), P. R. China*. Biogeosciences 08/2013; 10(8):5681-5689.

Stéphane Joost, Séverine Vuilleumier, Jeffrey D. Jensen, Sean Schoville, Kevin Leempoel, Sylvie Stucki, Ivo Widmer, Christelle Melodelima, Jonathan Rolland, Stéphanie Manel: *Uncovering the genetic basis of adaptive change: on the intersection of landscape genomics and theoretical population genetics*. Molecular Ecology 06/2013; 22(14):3659-3665.

CONFERENCE PROCEEDINGS

Ivo Widmer, Estelle Rochat, Kevin Leempoel, Alain Clémence, Olivier Ertz, Daniel Rappo, Jens Ingensand, Jean-Marc Theler, Idris Guessous, Stéphane Joost: *Biodiversity dynamics and the effect of urban environment on the distribution of genetic variation in the Geneva cross-border area*. First Annual Meeting in Conservation Genetics – Science and Practice, Birmensdorf, Switzerland; 01/2015

K Leempoel, S Stucki, S Joost: *Subsampling as an economic consequence of using whole genome sequence data in landscape genomics: how to maximize environmental information from a reduced number of locations?* Livestock Genomic Resources in a Changing World, Cardiff, UK; 06/2014

S Stucki, K Leempoel, S Joost: *Riding the whole-genome data tsunami: a landscape genomic study of local adaptation in Moroccan sheep and goats*. Livestock Genomic Resources in a Changing World, Cardiff, UK; 06/2014

Stephane Joost, Sylvie Stucki, Kevin Leempoel: *Geocomputational approaches for the analysis of Next-Generation Sequencing (NGS) and multi-scale data in landscape genomics*. 11th Swiss Geoscience Meeting, Lausanne; 11/2013

Kevin Leempoel, Stéphane Joost: *Relatedness and scale dependency in very high resolution digital elevation models derivatives*. Proceedings of the second Open Source Geospatial Research & Education Symposium; 10/2012

Kevin Leempoel, Sylvie Stucki, Christian Parisod, Stéphane Joost: *Very high resolution digital elevation models (VHR DEMs) and multi-scale landscape genomics analysis applied to an alpine plant species*. SIGSPATIAL Special. 11/2011; 3:10-14.