Population and Landscape Genomics Workshop – Canberra
Course on landscape genomics

Leempoel Kevin & Joost Stéphane

Friday, 28 March 2014

Content

Introduction

A brief history of Landscape genetics

Landscape genetics was originally defined by Manel in 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". Its goal is to "facilitate our understanding of how geographical and environmental features structure genetic variation at both the population and individual levels, and has implications for ecology, evolution and conservation biology" (Manel *et al.* 2003). The purpose is to identify genetic discontinuities and their correlations with environmental features such as barriers or environmental clines. Determining how much the landscape level and the environmental features are involved in the distribution of functional adaptive variation is one of the major steps in evolutionary biology (Lowry 2010).

In its original paper, (Manel *et al.* 2003) paper also insists on the necessity to use specific statistical test for LG data such as spatial autocorrelation and correlograms, interpolations, clustering approaches, PCA or mantel tests. Sampling schemes are indeed different than those of population genetics (PG) and therefore other statistical approaches can be applied. In addition, they mention the usefulness of GIS to visualize spatial genetic patterns and generate hypotheses about the cause of genetic boundaries. The geographic attributes of molecular data deserve attention and provide a view of genetic diversity and natural selection processes that complement information obtained from population genetics models. In fact, one of the main interest of GIS is the possibility to overlay genetic information with layers of physical barrier, landcover or topographical maps (Manel & Holderegger 2013). Genetic data can be also incorporated as a layer in order to understand the distribution of neutral genetic variation and gene flow (Lowry 2010).

Rapidly, LG studies started to focus mostly on gene flow and influence of habitat connectivity on genetic structure (Storfer *et al.* 2006; Joost *et al.* 2007; Holderegger & Wagner 2008; Manel *et al.* 2010a). However, landscape genetics also aims to correlate allele frequencies with the environment in order to understand its effect on the adaptive component of genetic diversity (Holderegger & Wagner 2008). To detect adaptive loci, (Joost *et al.* 2007) proposed a spatial analysis method (SAM) to use geo-referenced genetic and environmental data in a GIS environment in order to perform associations between allele frequencies and environmental variables. They worked on the large pine weevil (*Hylobius abietis*) and on sheep (*Ovis aries*) to identify commonly detected loci with PG methods, with the advantage of highlighting a potential driver of selection. It is this latter thematic that we will explore in this course.

The recent development in the detection of adaptive loci are regrouped under the term **Landscape genomics** (Luikart *et al.* 2003; Joost *et al.* 2007), defined as correlation studies between the genomic data and the environment to identify genes either potentially linked to candidate genes or the genes themselves under selection. Landscape genomics does not require phenotypic data, which can be fastidious to collect, to identify loci having adaptive significance (Joost *et al.* 2007; Manel *et al.* 2010a). Landscape genomics is still included in landscape genetics, but refers more specifically to the use of the future large amount of genetic data due to high-throughput sequencing. Landscape genomics is thus at the interface of bioinformatics, genomics, spatial statistics and landscape ecology.

Several recent papers have reviewed the advancements of landscape genetics in the last ten years. (Manel & Holderegger 2013) for example have evaluated LG short history and its future challenges, stating that the main objective of modern landscape genetics is to improve our understanding of global change influence on genetic patterns. In their review, as well as in (Petren 2013), they insist on

the fact that LG evolved both in the technical advancement of modern genomics but above all on the development of statistical and modelling approaches. In fact, after 2003, statistical approaches in LG had flourish and became more complex. Landscape genetics tools started to include population structure, landscape and historical ecology, niche modelling and conservation (Petren 2013).

Population structure is a major challenge (Lowry 2010). In fact, many methods suppose 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. However, it may be very difficult to detect outlier loci for sessile organisms above the cloud of the high Fst null distribution. These new methods should also correct the high rate of false positives and perform with large datasets in a reasonable amount of time (Manel & Segelbacher 2009). Among them we can cite TESS, sPCA, Bayenv (Coop *et al.* 2010), PCadapt or LFMM (Frichot *et al.* 2013). Also, some methods have been criticized such as mantel test or multiple regressions on distance matrices due to the non-independence in the response and predictor variables (Balkenhol *et al.* 2009; Manel & Holderegger 2013), suggesting to stop using (partial) mantel test and prefer linear correlations, regressions and canonical analysis (Legendre & Fortin 2010; Bolliger *et al.* 2014). They also suggest incorporating the covariance of allele frequencies, genetic structure, Moran's Eigenvector maps or demographic effects in mixed effect models. In addition, statistical approaches became multivariate as selected loci may be influenced by multiple environmental factors. Nevertheless, the multiplication of methods highlights the necessity to compare systematically these methods as they have conflicting results (Bolliger *et al.* 2014).

Another aspect of the diversification of LG is the increasing use of simulations. Several simulations tools, some at the individual level, are now available to model complex evolutionary processes in heterogeneous landscape (Landguth & Cushman 2010; Jones *et al.* 2013; Petren 2013).

Finally, spatial scale is a major unknown that was already mentioned in early LG studies but difficult to measure. In fact, it requires heavy fieldwork and important funding to accomplish studies at multiple scales. Few studies have analysed spatial scale. For example, (Manel *et al.* 2010b) were able to show that temperature and precipitations are the two main environmental variables to drive adaptation of *Arabis alpina* sampled in the European Alps (see case studies).

Differences between adaptive LG and population genetics approaches in the detection of neutral and adaptive variation

A major difference is that Landscape genetics ideally uses individuals as the operational unit and therefore does not require discrete populations to be defined in advance (Manel *et al.* 2003). This leads to a major issue of compatibility with studies in PG. In fact, LG requires that individuals are sampled in the entire distribution range in the study area and not just in several populations identified beforehand, sometimes with no physical barrier to justify separations. In other words, LG has the advantage that it is not necessary to work on a species that has already been studied and for which we have a priori information.

Another difference is that sampling design should be stratified across environmental variables considered in order to test specific hypotheses and should be improved to consider spatial dependence between sampled individuals (Manel *et al.* 2010a). Response to environment is here the main criteria and it is often not compatible with PG sampling strategies.

Consequences of next generation sequencing datasets on the methods used to detect natural selection

To date, most LG studies have used neutral genetic markers such as microsatellite (Bolliger *et al.* 2014) or AFLP with the inconvenient of not providing information on their genomic localisation. Today, SNPs are the most promising genetic data and allow to quickly spot regions of the chromosomes where selected loci were identified (Manel & Segelbacher 2009). But with datasets increasing in size due to both the higher number of genetic markers (millions of SNPs) and higher number of individuals, computation time of statistical tests is becoming a major issue for non-target genetic studies. It is one of the main advantages of simple correlative approaches, as they are capable of treating millions of models in a few hours, which is impossible for complex methods requiring a neutral simulation model.

I will show in the last part of the course and during the practical work that one of the main purposes of this course is to inform participants on the drawbacks of large datasets and to advise them on different strategies that can be used to attain robust conclusions in a decent amount of time.

Materials and Methods

GIS

Landscape genetics can be combined with Geographic Information Systems (GIS) to efficiently retrieve data from large environmental datasets and allow creating, organizing and displaying georeferenced data. Among its characteristics, spatial overlay is the most important in the context of LG. For example, superimposing genetic data on a topographic or land cover layer allows users to visualize barriers to gene flow.

Data used in GIS can be categorized in two:

- Vector data: Point, lines or polygon objects that are defined by their coordinates and associated to a table of attributes (ID, landcover type, genetic data…)
- Raster data: A georeferenced matrix of evenly spaced pixels that are either multiband, such as aerial photography, or surfaces representing altitude, temperature, DEM variables for example.

GIS are often seen as complicate and expensive software but free easy-to-use alternatives are becoming major actors in the sector. GIS are also a platform to centralize spatial analysis tools such as spatial autocorrelation, clustering or shortest path distances based on ecological corridors.

The difficulty with GIS is that most of them are specialized in certain tasks. Among the most used and most appropriate for LG we can cite

- Quantum GIS: Quantum GIS is probably the easiest GIS to start with. It is mainly used for overlaying data, categorizing attributes and creating maps. In addition, since it is free, open source and compatible with Python language, a large community of developer has grown around it, producing a large library of plugins.
  We will use it during practical works to visualise and illustrate data.
- SAGA GIS is also free and open source but probably a bit less accessible and also more specific. It is mainly used to treat raster data such as remote sensing images or DEMs. It is written in C++ and can be accessed in R through RSAGA.

- Geoda: a free software Developed by the GeoDa Centre for Geospatial Analysis and Computation. Geoda is specific to the spatial analysis of vector data but has many spatial statistics tools that are not available in other GIS. Its main argument is that it is interactive; a selection on the map instantly changes the graphs in other windows (histograms, Moran scatter plot, LISA clusters maps, quantile maps etc.).

Sources of environmental data

Most studies that focus on environmental influence on evolutionary processes require large scale environmental data. It means that, in most cases, getting data from fieldwork is not feasible for wide-ranging species. In addition, fieldwork data exist only at narrow temporal scales while adaptation occurs over much longer periods of time. I recommend reading (Kozak *et al.* 2008), a recent review of available large scale environmental data.

GIS environmental data are coming usually from two sources: remote sensing and spatial interpolation.

Existing sources

Climatic data

Climatic data are based on a network of weather stations place around the world. After the data are treated to be compatible between stations, interpolations are performed for all continents. However they are not evenly spaced and some regions are poorly covered, therefore uncertainty depends on the density of stations in the region of interest. In addition, temporal availability of data varies importantly across countries as some started recording only 10-20 years ago.

CRU Climatic Research Unit (http://www.cru.uea.ac.uk/cru/data/temperature/)

The aim of the Climatic Research Unit is to improve scientific understanding in three areas:

- past climate history and its impact on humanity;
- the course and causes of climate change during the present century;
- prospects for the future.

The resolution of this dataset is 5°×5°. Data are available from 1950 to 2014.

WorldClim (http://www.worldclim.org/methods)

Interpolation of average monthly climate data from weather stations on a 30 arc-second resolution grid (often referred to as "1 km^2" resolution). Data are available from 1950 to 2000.

Interpolations use latitude, longitude and elevation. Variables included are monthly total precipitation, and monthly mean, minimum and maximum temperature, and 19 derived bioclimatic variables.

Satellite imagery, ground cover and soil maps

Multispectral satellite imagery, infrared bands or indices such as NDVI, based on red and near infrared, are the most useful satellite imagery products to measure plant activity.

Another possibility is to use supervised classification of satellite images to determine tree species, soil surface type or ground cover. Some remote sensing data are freely available. To date, the most

complete database can be found on http://earthexplorer.usgs.gov/ with free LANDSAT images, OrbView-3, Global Land Cover Characterization (GLCC), cheap declassified images from 1960-70, eMODIS (EROS Moderate Resolution Imaging Spectroradiometer).

At a large scale, geological maps can be used to determine different types of soils. This information is not based on satellite imagery but on terrain knowledge, they are most often provided by national offices of cartography.

DEM derived variables (Existing, LIDAR and stereophotogrammetry)

Digital Elevation Models have been proven useful at a regional or local scale to acquire indirect ecological factors. However, existing free datasets have to be used with cautious due to their coarse resolution and imprecision. Two models are freely available so far on http://earthexplorer.usgs.gov/

Models for the earth	SRTM3	ASTER GDEM
Acquisition date	11 days in 2000	2000-2011
Spatial resolution	3 arcsec (≈90m)	1 arcsec (≈30m)
DEM accuracy (stdev.)	±16m	±12.6m (v.2)
Coverage	60°N – 56°S	83°N – 83°S
Acquisition method	Radar interferometry	Extraction of corresponding points between images by pattern matching
Comment	Topographically steep areas causing radar shadow	Available for steep mountainous regions Missing data for regions under constant cloud cover

TANDEM-X (also named WorldDEM) will be available soon (2014) based on TANDEM SAR satellites, with a resolution of 12 meters. Price and precision are currently unknown (http://www.dlr.de/hr/en/desktopdefault.aspx/tabid-2317/).

Otherwise, national offices of cartography have often acquired a LIDAR model, characterized by a high resolution (2-5m) and a high precision. Because they are acquired from an airplane, they remain expensive and usually cover regions of human activities only. The main advantage of LIDAR point clouds is that they can be filtered according to different categories. For example, buildings, trees, vegetation, ground, power lines can be categorized using automatic treatment.

Recently, accessible methods of DEM acquisition have been made available using drones. Drones likes those of SenseFly https://www.sensefly.com/home.html can be piloted through software in which the user defines the GPS locations where he wants the drone to pass by. Afterward, stereo-photogrammetry software (http://pix4d.com/, http://www.agisoft.ru/) can be used to produce an orthophoto and a DEM. Resolution can be as high as for LIDAR data but usually less precise,

particularly in steep mountains. It's important to note that drone images are not appropriate for DEM acquisition in a forest since the point cloud would mostly cover the canopy and not the ground.

DEMs variables can be useful to characterize local habitats, particularly in mountainous areas. Some variables are simple such as slope, aspect, curvature and are often considered as proxies for hydrologic processes, snow cover, soil water content etc. But there are also more complex variables that have been developed to have a direct link to environmental features such as wetness indices, solar radiation, terrain ruggedness indices etc. See Terrain analysis modules in SAGA GIS for more information.

Fieldwork sources

Loggers (temperature, humidity, soil moisture, solar radiation)

When studying at a local or regional scale in heterogeneous landscapes, interpolated climatic data are not appropriate. In these cases it is possible to place temperature or humidity loggers that will record these variables at defined intervals. The variability they capture can be invaluable to distinguish different habitats.

In one of the case studies that we will develop later we used Ibutton (http://www.maximintegrated.com/products/ibutton/) temperature and humidity loggers as well as a soil moisture measurer.

Genetic data for landscape genomics studies

Genetic information is embedded within a geographic context. Individuals (humans, plants and animals) are directly influenced by the specific characteristics of their surrounding environment. Therefore, spatial information must be considered to understand genetic diversity, and recording of the geographic coordinates of the organisms under study is definitely valuable for further analyses.

The process of defining the geographic position of an object – georeferencing or geocoding – simply consists of attributing latitude and longitude values (and possibly altitude) to any DNA sample taken from sampled individuals and can be recorded with a GPS (Geographic Positioning System) device. The use of a GPS guarantees the required level of precision, particularly if a standard protocol is followed to avoid biases associated with different operators. For local scale studies, a higher accuracy can be obtained using Differential GPS (DGPS), usually under 10cm precision. When sampling locations have to be identified without using a GPS device, the geographical coordinates can be approximated from existing paper maps or web-accessible geodatabases like Google Maps (http://maps.google.com) or OpenStreetMap (http://www.openstreetmap.org). These tools can also prove particularly useful for attributing geographic coordinates to previously collected genetic samples, because the coordinates they provide are already in digital format.

Recoding/Filtering alleles and genotypes

After sequencing, genome comparison between individuals permits to identify polymorphic loci. Raw data contain one line per individual and two columns per locus. Filtering these data is a primordial step and can be done with PLINK for example (Purcell *et al.* 2007).

Several criteria for filtering can be applied:

- Per-individual genotyping rate is the ratio between the number of loci read and the total number of loci. Individuals with a ratio lower than the defined threshold are taken out of the analysis.

- SNP missing genotype rate. Loci with a ratio lower than the threshold are taken out as well.
- Minor allele frequency. Frequency of the minor allele determines if the loci can be considered polymorphic. Usually the frequency of the minor allele is set to a minimum of 5%.

Statistical approaches

LG combines statistical methods from landscape ecology, population genetics and spatial statistics. Detecting loci under selection requires logistic regressions to correlate presence/absence of genetic markers and quantitative environmental variables. Methods presented here differ on their consideration of population structure and spatial autocorrelation.

To illustrate these methods, we will use the results from a study on the environmental adaptation of *Bos Taurus* and *Bos indicus* in Ouganda by (Stucki 2014). We will focus on a dataset of 41000 SNPs from 804 individuals sampled over the entire country. One of the purposes of this project was to look for signs of local adaptation (climate or parasitism for example) in local breeds. In fact, local farmers have recently imported European breeds to increase production but it may be at the expense of local diversity.

Computing associations between genetic data and environmental variables

Correlative approaches (GLM) using SAMβada

SAMbada models the probability of occurrence of an allele for each individual in function of the environmental conditions in his habitat based on spatial coincidence (Goodchild 1996; Stucki 2014). SAMbada is free from any assumptions regarding the genetic models, such as hardy Weinberg and does not directly includes corrections based on population structure or spatial autocorrelation (Joost *et al.* 2007).

Logistic models are calibrated using a maximum likelihood procedure. Once the model is adjusted, SAMbada evaluates its significance with two scores: G and Wald. A model is considered significant if both tests reject H0. However, because it is mostly used for multiple models, a Bonferroni correction for multiple tests is applied. Bonferroni significance level is calculated by dividing the alpha level by the number of models computed. It is important to remember that by computing thousands or millions of models, Bonferroni correction becomes a very conservative threshold compared to FDR for example (SAMbada does not integrate a FDR procedure yet). In addition to these two scores, SAMbada also computes AIC and BIC statistics to facilitate models comparison. However, SAMbada and GLM in general are submitted to a high rate of false positives (De Mita *et al.* 2013; Jones *et al.* 2013).

SAMbada also includes multivariate models. To do so, it compares the multivariate model to its parents which are the models involving one environmental variable less that the current models. It means that SAMbada first computes a univariate model for a specific genetic marker and each environmental variable and then uses these models to compare each bivariate model to its "best" parent (the parent with the higher likelihood in order to minimize G) using a G score. Multivariate models are also assessed with Wald test: each regression parameter (except the constant) must be significant.

SAMbada can read Plink file format and convert it for its own use or to LFMM format. Environmental variables can be loaded from a distinct text file. Its output consists on a text table where each line corresponds to a model (so #models = #environmental variables X #markers) and rows correspond to

the different statistics. Model sorting is based on G scores in decreasing order. There is also an option to output a shapefile containing LISA coefficients.

Approaches considering population structure

Admixture (Alexander *et al.* 2009)

Admixture is a software using a maximum likelihood estimation of unrelated individual ancestries from multilocus autosomal SNP genotype datasets. It uses the same statistical model as STRUCTURE but with an optimized algorithm that computes estimates much more rapidly. In addition, Admixture is multithreaded; it can spread computation on all processors.

It recognizes both types of PLINK formats (Binary or Ordinary) and requires only from the user to give an estimate of the number of populations K. Therefore Admixture is often run a couple of times with different values of K and the most appropriate value of K is the one with the lowest estimated error from the cross-validation procedure. In contrary to STRUCTURE, Admixture does not attempt to estimate the model evidence but uses the cross-validation procedure to identify the value of K for which the model has best predictive accuracy, as determined by "holding out" data points.

For each K, there are two output files: Q (the ancestry fractions), and P (the allele frequencies of the inferred ancestral populations).

Bivariate models in SAMβada using population structure and environmental variables.

Since SAMbada can perform multivariate logistic models, we can add membership coefficients as variables to consider population structure. In the cattle data of Ouganda, (Stucki 2014) included one variable of membership coefficients from Admixture in the set of environmental variables. The idea being that significant multivariate models are better explained by a combination of a population structure variable and an environmental variable rather than by an univariate model based on population structure variable only, therefore reinforcing the importance of environment in the spatial distribution of loci.

*Correlative approaches including population structure (*Bayenv, LFMM*)*

Bayenv (Coop *et al.* 2010).

Bayenv is a bit different in the sense that it models allelic frequencies for a group of individuals that share the same environment. It starts from the same geographical principle that close populations usually share similar genetic background and because they are encountering similar habitats, it could lead to false detections. Therefore, it was built to compare populations that are clearly identified, like in the 1000 genomes project for example, and is not designed for a low number of populations, which is not optimal in LG studies.

Bayenv first computes a covariance matrix of neutral allelic frequencies to estimate a null hypothesis of selection signature based on demographic history. Then it uses all loci and environmental variables to detect signature of selection.

Bayenv has several inconvenient that should be considered:

- It requires a defined population structure, which is easily done only if they are geographically distinct.
- To assess associations between environmental variables and markers, separate files have to be created for each SNP.
- The computation time required to compute associations is very long (several days or weeks for millions of markers and around ten environmental variables).

- It also requires a defined set of neutral markers, which can be seen as a cyclic problem.

Also, Bayenv detected loci cannot be compared directly to other methods since the significance threshold is based on Bayes factors. (Coop *et al.* 2010) suggest an empirical approach to identify loci potentially under selection. The principle is to estimate an empiric distribution based on the covariance matrix of neutral loci and to use it as a reference to select loci with high Bayes factor values.

LFMM (Frichot *et al.* 2013)

LFMM is also a correlative approach similar to SAMbada but it introduces population structure in the model through non-observed variables.

LFMM expresses the genotype with a linear mixed model for each environmental variable and each locus and includes latent factors involving a per-individual and per-loci component. These latent factors represent the part of genetic variability that is not explained by the environment. The number of latent factors has to be defined by the user and it is recommended to use Tracy-Widom theory to estimate it (Patterson *et al.* 2006; Frichot *et al.* 2013). It is usually lower than the number of populations estimated by STRUCTURE or Admixture. (Frichot *et al.* 2013) recommend to create a p-value distribution for each K and apply a FDR correction like the one from (Benjamini & Hochberg 1995). The best value of K is the one for which FDR procedure can be best applied. Authors also plan to implement a cross validation procedure to estimate the best K in a future version.

The main advantage of LFMM is its ease of use and swiftness compared to other methods including demography (it spreads computations over all available processors). Unfortunately, the population structure that is estimated cannot be saved and compared to admixture for example.

A note on computation time

Computation time of methods considering population structure is much higher than for standard logistic regressions. For example, it took 1h to (Stucki 2014) just to recode the data for Bayenv and 42hours to perform computations in Bayenv. In comparison, univariate models in LFMM took 6h and univariate and bivariate models in SAMbada took 1.5 and 9h respectively.

Comparison of results

Results on *Bos taurus* 54K data for these methods show that several loci were detected by all methods. However, common detections are different depending on the methods compared. SAMbada detected 12782 associations. Among the 400 associations detected by Bayenv, 387 of them were also detected by SAMbada and if we look at the first 100 in SAMbada, 65 of them were also detected by Bayenv. LFMM detected 303 significant associations and 227 are common with SAMbada. However, those with the highest score in SAMbada are not the same as those with the highest score in LFMM. In total, 4 loci where detected by the 3 methods and are among the first 18 associations based on SAMbada G score.

Spatial statistics

Concerns regarding spatial autocorrelation.

Tobler's law states that "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970). Although this geographic law is central in spatial statistics, such as Spatial Autocorrelation, it violates the principle of independence between measurements which is required for most classical statistics. It should therefore be considered for any studies with spatially

dependent measurement and, if possible, should be used to disentangle associations that are due to geographical structure from those that are due to the environment.

SAMbada integrates two measures of spatial autocorrelation: Moran's I and LISA (Local index of Spatial Association). It can be used for example to localize regions where the presence of an allele in an individual is autocorrelated to the presence of the same marker in its neighbours.

Moran's I and Local Indicators of Spatial Association (LISA) (Anselin 1995)

Spatial autocorrelation of genetic data may reflect limited dispersal capabilities or local adaptation (Durand *et al.* 2009). To examine patterns of spatial dependency of candidate loci, we can measure global and local spatial autocorrelation in the program OpenGeoDa (Anselin & McCann 2009). The Spatial Autocorrelation (Global Moran's I) (MORAN 1950) measures spatial autocorrelation based on both feature locations and values simultaneously. Given a set of features and an associated attribute, it evaluates whether the pattern expressed is clustered, dispersed, or random. Moran's I values range from − 1, indicating perfect dispersion of data, to 1, indicating perfect spatial autocorrelation (clustering), and 0 indicating randomly dispersed data.

Measure of local spatial autocorrelation are carried out with Local Indicators of Spatial Association (LISA) developed by (Anselin 1995). LISA indicators are statistics that measure spatial dependence and evaluate the existence of local clusters in the spatial arrangement of a given variable and are based on the statistical index I developed by Moran. They evaluate the existence of clusters in the spatial arrangement of a given variable due to underlying autocorrelation. Their sum over the whole studied area is proportional to the global Moran's I.

LISA spatial weighting scheme should be determined on the basis of three criteria: (i) the distance for which there is no neighbourless sampling unit; (ii) a correlogram showing that Moran's I is higher for short distances before regularly decreasing for each variable investigated, and (iii) the information produced by a connectivity histogram showing the density of points within spatial lags.

The standardized scattergram of this relationship shows four distinct classes: (i) high values correlated with high weighted values; (ii) low values correlated with low weighted values; (iii) a low-high relationship, and (iv) a high–low relationship. The attribution of individuals to these four classes depends on the results of a statistical significance test. This test consists in performing random Monte Carlo permutations among the sites located in the spatial lag to compare the observed LISA to the LISA corresponding to the random permutations. If the test is significant—the observed LISA is statistically larger (or smaller in the case of a negative relationship) than the local indices resulting from all random permutations —an individual is attributed to one of the four classes based on its quadrant in the standardized Moran scatterplot. If the test is not significant, the individual remains in a neutral class (no spatial dependence).

Bivariate LISA allows identifying and map local clusters of high or low marker frequencies (the first variable) significantly correlated with environmental variables (the second variable) possibly exerting selective pressures on them. This allows to refine the identification of association signals detected by GLMs over the whole study area and to check whether these associations involve particular sub-populations only.

GWR (Fotheringham *et al.* 2002)

Geographically weighted regressions are, as the name indicates, ordinary regressions that take into account the spatial component of the data. It computes regressions for each point or polygon, using a weighting scheme defined by the user to select and weight the surrounding points to include in the local regression. However GWR were developed for quantitative data and some improvements

remain to be done regarding logistic regressions. For example models do not converge when locally selected points are either nearly all Ones or Zeroes. It also requires large datasets since it is recommended to have at least 40 measurements in the weighted regression of each points. For these reasons, GWR has not been applied often in LG.

Spatial autocorrelation analysis on Bos taurus

On the commonly detected markers for *Bos Taurus* association models, (Stucki 2014) applied LISA analysis. As expected, Correlograms based on Moran's I show a decrease in spatial autocorrelation with an increasing number of neighbours. Their results show also that the locus that was not detected by LFMM shows the highest global spatial autocorrelation. Although studies of this type are rare, it may show that loci with a strong spatial autocorrelation are probably not distributed by an environmental factor but by population structure.

They then performed univariate and bivariate LISA analysis and showed a strong spatial structure for several loci with regions of significant positive autocorrelation and significant association with environmental variables and population structure in bivariate analysis. In fact, several variables exhibit a North-South gradient that can be mixed up with demographic structure showing why GLMs are producing a high amount of false positives.

Interestingly, the spatial distribution of locus HM-28 is similar to the prevalence of two worm parasites causing the sleeping sickness, transmitted by a fly. In addition, locus HM-28 is situated on chromosome 5 and corresponds to a transcription factor involved in the resistance to parasites. It demonstrates the usefulness of a spatial approach to form hypothesis on the causes on the variability of loci frequencies among populations.

Case studies

Multiscale analysis of *Arabis Alpina* adaptation in the Alps

Arabis Alpina has been a frequent LG case study in the European Alps. However, (Manel *et al.* 2010b) bring here 2 novelties : - they sampled the plant at 3 different spatial extent – they add broad-scaled Moran's EigenVector Maps (MEM) to the set of environmental variables to model the spatial variation of the sampled individuals not accounted by the environmental predictors included in the analysis.

At large scale, geographical and environmental variations are mostly inter-dependent, leading to respective patterns at broad spatial scales where MEMs can be included as explanatory variables in the analysis as proxies for unmeasured environmental variables. Moreover, large scale geographical effects on the structure described by genetic markers, which are also retrieved by MEM variables, relate to the influence of historical dynamics of *A. alpina*. In contrast, biotic processes such as dispersal, mating or competition mostly invoke spatial patterns at intermediate or small spatial scales.

MEM are spatial eigenfunctions, computed from the geographic coordinates of the study sites, that describe the spatial relationships among the sites at all scales that can be perceived by the sampling design. MEM analysis produces uncorrelated spatial eigenfunctions used to dissect the spatial patterns of the studied variation (allele frequencies in the present context) across a range of spatial scales. The first few MEM variables, which have large Moran's I coefficients (this is a measure of spatial autocorrelation), can be used to model broad-scale processes (e.g. environmental variation in

space), whereas subsequent MEM variables with smaller Moran's I coefficients can be used to model the spatial autocorrelation generated by biotic processes such as individual dispersal.

Their goals are to identify loci significantly correlated with environmental variables (i.e. loci linked to genes under selection and therefore of ecological relevance) at three spatial scales, i.e. large, regional, and local scales and to introduce a new approach to correlate allele frequencies derived from genome scans with a wide array of environmental variables and broad-scaled MEM variables.

The first data set was sampled over the entire European Alps (large scale) in the frame of the INTRABIODIV project (Gugerli *et al.* 2008). Three plants where sampled in each cells of a rectangular grid of 25 x 25km applied on the Alps, resulting in 385 samples in 130 cells. 140 AFLP markers were obtained.

The second one, obtained in the French Alps (regional scale) (Herrmann et al. 2010; Poncet et al. 2010), included three separate mountain massifs of the French Alps (local scale). Plants were sampled in 93 locations in the French Alps along 3 mountain massifs (Vercors, Chartreuse, southern French Alps). 3 to 9 individuals were sampled per location leading to 321 individuals in total. 712 polymorphic AFLP markers were obtained but are not the same as those from the first dataset.

Fourteen monthly and annual environmental variables related to temperature, precipitation and topography were extracted per sampling location from published GIS eco-climatic layers from 1980 to 1989 (200-m resolution; (Zimmermann & Kienast 1999). They then applied a PCA to these environmental variables to examine correlations and remove redundant variables (i.e. variables that were correlated at $|r| > 0.8$). They first identified variables correlated to each retained axis, creating groups of variables. Afterwards, they kept only one (or two) variables considered to be the most pertinent in terms of local adaptation in plants.

Multiple regressions were computed between the allele frequencies per location and the joint table comprising linear effects and MEM. Finally, they performed linear regressions between allele frequencies at each of the identified significantly correlated loci and each predictor separately in order to estimate the explanatory power provided by each environmental variable, using $adjR^2$ values.

In the results, they found that the first two axes of the PCA explained 60% of the variation in the environmental variables for the European Alps and 68% for the French Alps. At both scales, all temperature variables and elevation were highly correlated with the first PCA axis. In the French Alps, precipitation variables were also correlated with the first PCA axis but less strongly than the temperature variables. For the European Alps, the precipitation variables (except summer seasonal precipitation, prcp0608) were correlated with the second PCA axis.

In the end, they used the same eight environmental variables for the analysis at all scales: mean minimal temperature per year (tminavgty), annual mean of daily global radiation (srad), spring seasonal precipitation (prcp0305), summer seasonal precipitation (prcp0608), slope (slp), aspect (asp), integrated topographic exposure map (topo), and potential soil humidity (twi).

MEM analysis identified 11 broad-scaled MEM variables for the European Alps, 5 for the French Alps and 2 for each of the three local-scale analyses. The multiple linear regressions between allele frequencies and the joint table containing the transformed variables and these broad scaled MEMs, based on the $adjR^2$ criterion, detected 12% and 11% of all AFLP loci of ecological relevance at the large and regional scales, respectively.

At the local scale, they detected 3% of the loci as being of ecological relevance in Chartreuse, 16% in Vercors, and 17% in the southern French Alps. Nine of the 21 loci of ecological relevance with a linear response identified at regional scale in the French Alps were also detected at the scale of single massifs (eight in the southern French Alps, one in Vercors).

After accounting for spatial effects through MEM variables, tminavgty was the environmental variable with the best explanatory power. It had the highest cumulated adjR^2 value at all spatial scales, except for the large scale and for the local scale in massifs Vercors and Chartreuse where its scores came second after those of prcp0608 or prcp0305. The second major environmental driver of AFLP allele distributions was precipitation (prcp0608 or prcp0305).

Their study is an example of recurrent incompatibilities between PG and LG studies. In fact they didn't apply PG approaches since their continuous sampling design was not suitable for these approaches.

These results suggest that there may be two different types of adaptive responses acting on *A. alpina*. Many loci are probably involved in site-specific local adaptation, hence the large number of loci of ecological relevance at local scale, while other ecologically relevant loci are mainly involved in more general adaptive responses at larger geographical scales. The latter type likely reflects selective pressures consistent across scales in alpine plants, such as adaptation to altitude or frost. In contrast, adaptive fine-tuning of gene regulation and expression acts at a local scale.

One major inconvenient of their approach that they mentioned is that they didn't use logistic regressions but regressions on allele frequencies. The issue is that they often have only 3 individuals per plot and they didn't assess the representativeness of such small samples.

Another is that the limited differences between the local scale and the regional and large scale may be due to the sampling strategy. Indeed, species may be found in various types of habitats at a local scale, such as steep slopes, rocky outcrops, moist or eutrophic sites. Their analysis did not consider such small-scale micro site variation, which may be the reason why topography-related environmental variables, e.g. topo or twi, revealed low values of adjR^2.

Local adaptation of *Biscutella Laevigata* at Les Rochers-de-Naye (CH) using DEMs

Recent studies, however, have challenged this belief, showing that molecular adaptation is often local (Conover et al. 2006), a view shared by most ecologists. In reality, adaptive processes probably form a complex multi-scale continuum, with natural selection being the result of this complex continuum of scales.

The buckler mustard (*Biscutella laevigata* L.) is a small Alpine plant that grows in small patches in many places in the European Alps. But at Les Rochers-de-Naye, its distribution is peculiar as most individuals grow very close to a cliff (1-10m from it). Also, this ridge is quite short (1.5km long) and *B. laevigata* has an isolated population in that area.

This peculiar distribution allows us to study a plant at a very local scale and to evaluate the relevance of high resolution DEM variables. The question of finding an appropriate scale to study adaptation emerged only recently in landscape genomics, while in landscape ecology its components (extent and resolution or grain) have been addressed in many papers (Wiens 1989; Levin 1992; Wilson & Gallant 2000). In fact, in landscape genetics only a couple of studies have considered different extents to

study adaptation - (Manel *et al.* 2010b) that was mentioned in the introduction for example - but the spatial resolution of environmental variables was not considered. However, when these variables are derived from digital elevation models (DEMs), the question of resolution becomes crucial.

For this study we used a signal processing generalization technique to produce DEMs at multiple scales starting from a very high resolution, allowing for a continuous representation of the landscape. We then investigated the adaptive response of the buckler mustard (*Biscutella laevigata* L.) with the help of environmental variables derived from these DEMs.

Firstly, we produced a set of 266 AFLP markers in 361 individuals sampled and georeferenced on the ridge. Secondly, we acquired a very high resolution DEM (0.5m) for the same area. On this VHR DEM, we applied a multilevel spline generalization algorithm (Seungyong *et al.* 1997) implemented in SAGA GIS to compute DEMs at resolutions of 1m to 10m with intervals of 1m. Thirdly we produced DEM variables, also in SAGA GIS, at all resolutions and extracted their values at the coordinates of the sampling locations (Leempoel & Joost 2012). Finally we used a correlative method based on generalized linear models to calculate univariate association models between the genetic markers and the variables.

We found several significant models involving different variables such as Terrain Wetness Index, Vector Ruggedness Measure, and slope at several resolutions. In fact, most associations showed their strongest p-value at resolutions between 4 and 8m. In total, 98 models were significant with a 99% confidence level involving 14 different markers (5.3% of the 266 markers).

These results show that resolution matters at a local scale and that a higher resolution does not necessarily implies a stronger association. However, the interpretation of an optimal resolution is tricky, particularly because of the high variability of significance from one resolution to the other.

Practical work: Loblolly Pine association to aridity in the US

In this study, (Eckert *et al.* 2010) used SNPs genotyped across 3059 functional genes to study patterns of population structure and identify loci associated with aridity across the natural range of Loblolly pine (*Pinus taeda L.*).

Loblolly pine is distributed throughout the south-eastern United States. Its wide range is divided primarily by the Mississippi River Valley, with 60% of the distribution range located east of the river. The structure of this discontinuity is consistent with a dual Pleistocene refugial model, which has also been used to explain differential growth abilities, disease resistance, and concentrations of secondary metabolites among families located across this discontinuity.

The canonical approach to searching for environmental associations would be to correlate measures of genetic diversity to environmental variation related to drought stress. However, they state that one limitation to this approach is that environment is likely to be confounded with geography and, by proxy, overall genetic structure. Therefore, a correction for population structure should be applied.

Their hypotheses are: (1) what are the patterns of population structure across the range of loblolly pine? (2) What is the degree of confounding between environmental variation and population structure due to geography? (3) Which loci are associated with water availability across the range of loblolly pine? (4) Are loci with strong genotypic correlations with aridity also those with extreme allele-frequency differences among populations?

Needle tissue was collected from 907 largely unrelated trees sampled. These samples are georeferenced by county and the average number of sampled trees per county was 4 (range: 1–67). The first set of genetic markers comprises 23 unlinked nuclear SSR markers selected from the PtTX marker set for medium to high polymorphism rate and full coverage of the linkage map. The second set comprises 23,000 SNPs, of which they chose 7216 for genotyping, that were identified through the resequencing of 7535 uniquely expressed sequence tag (EST) contigs in 18 loblolly pine.

Climate data were gathered from the WORLDCLIM 2.5-min. The temperature and precipitation data were used to estimate potential evapotranspiration (PET) with the method of Thornthwaite (1948). An aridity index (AI) was defined as the ratio of precipitation to PET, with this ratio being defined quarterly.

Population genetic structure was analysed by means of principal component analysis (PCA) on genotypes from individual trees. The significance of PCs was determined by comparing the value of each standardized eigenvalue to a Tracy–Widom distribution. For comparison, they also used the program STRUCTURE (Pritchard *et al.* 2000) using the 23 nuclear SSR markers and the optimal value of K was determined using the ΔK method (Evanno et al. 2005).

Loci associated with environment were identified using a standard association mapping approach, substituting aridity for phenotype. PCA analysis was used to correct for spurious associations due to confounding of ancestry and aridity. Two vectors of ancestry-corrected residuals are obtained by multiple linear regression on environmental and genotypic values, using the k significant genetic PCs as independent variables. SNP loci showing the strongest association with different aridity indices were identified by Q–Q analysis of P-values. The magnitude of environmental differences among SNPs was evaluated using a general linear model with environment as a dependent variable and corrected genotypic values as explanatory variables.

They compared these detected loci with those detected by fdist (Beaumont & Nichols 1996) using the populations identified using PCA or STRUCTURE.

PCA analysis of population structure on the SNP data revealed the presence of seven significant PCs defining eight genetic clusters. Visual inspection of the eigenvalues, however, shows the presence of two major PCs explaining 2.4% of the total variation (56% of the significant variation), which indicates the presence of three clearly differentiated clusters. These three clusters are largely divided along the Mississippi River Valley, with a further division of the eastern cluster into Gulf and Atlantic Coast clusters.

Association analysis resulted in the identification of five loci with significant correlations to aridity. The strongest associations were between four loci and aridity during the second quarter (AI2), with subsets of these loci also associating with aridity during the first (AI1) and fourth quarters (AI4). All five significant SNPs were located in loci with high sequence similarity to coding sequences in Arabidopsis that primarily affect abiotic and pathogenic stress responses. Two of these five SNPs are mapped to linkage group 3. Three of the five SNPs are located in synonymous positions, while the remaining two are located in an intron and a 39 UTR. There was no overlap between the loci associated with aridity gradients and those identified as outlier, loci associated with aridity gradients had values of Fst within the range of the mean value across loci.

The association approach used here assesses the effect of natural selection along specific environmental gradients, while Fst outlier methods aim to identify loci influenced by natural selection driving allele-frequency differences among populations (i.e., ancestral groups) and are thus agnostic about the environmental gradients driving the extreme values of Fst. Significant associations

with an environmental gradient likely represent those polymorphisms underlying functional responses to that gradient. In contrast, post hoc interpretations of environmental differences are attributed to the cause of Fst outliers.

In their case, the process of defining populations affects the identification of Fst outliers. The identification of genetic clusters may be difficult or inappropriate for species such as loblolly pine that are distributed continuously across large geographical expanses and for which paleobotanical and population genetic evidence suggests historical fluctuations in population size. Invocation of the island model in these cases to derive the null distribution of test statistics will result in undesirable statistical behaviours.

In contrast, the association approach has low power when there is complete confounding between environment gradients and axes of ancestry. They note however that their model is likely more biologically plausible than the standard model used in Fst outlier analyses and in fact explains the observed data as well as, if not better than, drift or the island model alone. In addition, all five loci have putative orthologs in Arabidopsis that are responsive to abscisic (ABA) or jasmonic (JA) acid, two plant hormones with well-documented correlations to abiotic stress responses.

Finally, they mention that one of the major drawback of their dataset is their sampling scheme. Indeed, Within-county sampling should be more effective for association analyses than the county level approach because fine-scale environmental variation may be of considerable biological relevance.

Conclusion and perspectives

Concluding remark on common findings and differences between datasets

These examples show that a comparative analysis of different methods with different pre-requisites can increase the robustness of a study. The purpose is not simply to find commonly detected marker and see with which variables they are related to but also to understand their spatial repartition and formulate hypothesis on why they were detected by some methods and not by others. Indeed, the approaches we presented start with different information regarding the number of population, sets of neutral markers, pre-defined population structure, number of neighbours to consider. Analysis of many different results take time but allow to focus on several loci of interest for which population structure, multivariate models and local spatial autocorrelation can help to formulate hypothesis on their repartition among populations and space. Finally, the ultimate purpose is locating them on the genome and identifying their roles, hoping to find coherent signs of adaptation to the environment.

Challenges in landscape genomics in the coming years

The main topics of these last years will continue to be central in the coming years. Development of new methods to detect adaptive loci is on-going at an increasing rhythm. New approaches are likely to try merging information from both adaptive and neutral genetic variation which would assess how genes under selection disperse across landscape or how gene flow counterbalances local adaptation (Manel & Holderegger 2013; Bolliger *et al.* 2014). But it will require a more systematic comparison of these approaches. However, an important point to improve is to try to make LG samplings compatible with PG sampling by using similar sampling schemes (Joost *et al.* 2013). Genetic data extraction technologies will also continue to develop and LG will be helped by the reducing costs of high-throughput next generation sequencing. This will allow us to sample more individuals and gain greater power to detect causal associations as the number of semi-replicated populations surveyed

will increase (Petren 2013). Still one of the main issue is that few studies have investigated the relevance of identified loci (Manel & Holderegger 2013).

As stated by (Manel & Holderegger 2013), LG case studies are likely to continue focusing on one hand on long term consequences of global change on genetic diversity and adaptation and on the other to assess the effectiveness of conservation measures (Bolliger *et al.* 2014). An advantage of landscape genomics is that it can determine which population should be given priority in conservation of adaptive genetic variation.

Improvement of methodologies will also allow more studies to have either multiple species or multiple landscapes (Manel & Holderegger 2013). To date, only a few examples exist such as (Amos *et al.* 2012) who worked on ten woodland birds or (Manel *et al.* 2012) who identified common environmental variables that drive adaptation of several alpine plants.

Similarly, reduced cost will permit to study the same species at several scales to answer questions such as "what is the relationship between global and local patterns of adaptation?" or "How important are spatial scale and habitat heterogeneity in maintaining adaptive genetic variation?" (Manel & Holderegger 2013).

But current research should not forget to get off beaten tracks and consider habitats in the world that haven't been considered a lot in LG. For example, studies should focus more on tropical species, only 10% of the studies have been made there so far (Storfer *et al.* 2010; Manel & Holderegger 2013). Another environment that has been poorly studied is urban areas. It is important to assess connectivity in urban environment for plants, insects or birds. It is also a challenge to characterize urban environment as many variables could be relevant (noise, pollution, building height, night light…). LG could also be interesting to measure gene escape from genetically modified organisms. Finally, historical habitat properties are starting to be considered. Wetland size for example explains better allelic richness in a wetland plant that recent habitat (Bolliger *et al.* 2014). Similar conclusions could exist for temperature by exploiting ancient temperature estimations through paleoclimatic data.

Bibliography

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

Amos, J.N., Bennett, A.F., Mac Nally, R., Newell, G., Pavlova, A., Radford, J.Q., Thomson, J.R., White, M. & Sunnucks, P. (2012). Predicting landscape-genetic consequences of habitat loss, fragmentation and mobility for multiple species of woodland birds. *PLoS ONE*, **7**.

Anselin, L. (1995). Local indicators of spatial association — LISA. *Geographical Analysis*, **27**, 93–115. Retrieved from www.drs.wisc.edu/people/faculty/curtis/documents/RS977/Anselin1995.pdf\n www.spatialanalysisonline.com/output/html/LocalindicatorsofspatialassociationLISA.html

Anselin, L. & McCann, M. (2009). OpenGeoDa, Open Source Software for the Exploration and Visualization of Geospatial Data. *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* pp. 550–551. GIS '09. ACM, New York, NY, USA. Retrieved from http://doi.acm.org/10.1145/1653771.1653871

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

Beaumont, M.A. & Nichols, R.A. (1996). Evaluating Loci for Use in the Genetic Analysis of Population Structure. *Proceedings: Biological Sciences*, **263**, 1619–1626. Retrieved from http://www.jstor.org/stable/50648

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. Retrieved from http://www.jstor.org/stable/2346101

Bolliger, J., Lander, T. & Balkenhol, N. (2014). Landscape genetics since 2003: status, challenges and future directions. *Landscape Ecology*, **29**, 361–366. Retrieved from http://dx.doi.org/10.1007/s10980-013-9982-x

Coop, G., Witonsky, D., Di Rienzo, A. & Pritchard, J.K. (2010). Using Environmental Correlations to Identify Loci Underlying Local Adaptation. *Genetics*, **185**, 1411–1423. Retrieved from http://www.genetics.org/content/185/4/1411.abstract

Durand, E., Jay, F., Gaggiotti, O.E. & François, O. (2009). Spatial inference of admixture proportions and secondary contact zones. *Molecular biology and evolution*, **26**, 1963–1973.

Eckert, A.J., van Heerwaarden, J., Wegrzyn, J.L., Nelson, C.D., Ross-Ibarra, J., González-Martínez, S.C. & Neale, D.B. (2010). Patterns of population structure and environmental associations to aridity across the range of loblolly pine (Pinus taeda L., Pinaceae). *Genetics*, **185**, 969–982.

Fotheringham, A.S., Brunsdon, C. & Charlton, M. (2002). *Geographically weighted regression: the analysis of spatially varying relationships*. Wiley. Retrieved from http://books.google.ch/books?id=mPWlG-0G3rQC

Frichot, E., Schoville, S.D., Bouchard, G. & François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*. Retrieved from http://mbe.oxfordjournals.org/content/early/2013/03/29/molbev.mst063.abstract

Goodchild, M.F. (1996). Geographic Information Systems and spatial analysis in the social sciences. *Anthropology, space, and Geographic Information Systems*.

Gugerli, F., Englisch, T., Niklfeld, H., Tribsch, A., Mirek, Z., Ronikier, M., Zimmermann, N.E., Holderegger, R. & Taberlet, P. (2008). Relationships among levels of biodiversity and the relevance of intraspecific diversity in conservation – a project synopsis. *Perspectives in Plant Ecology, Evolution and Systematics*, **10**, 259–281. Retrieved from http://www.sciencedirect.com/science/article/pii/S1433831908000528

Holderegger, R. & Wagner, H.H. (2008). Landscape Genetics. *BioScience*, **58**, 199–207. Retrieved from http://dx.doi.org/10.1641/B580306

Jones, M.R., Forester, B.R., Teufel, A.I., Adams, R. V, Anstett, D.N., Goodrich, B.A., Landguth, E.L., Joost, S. & Manel, S. (2013). INTEGRATING LANDSCAPE GENOMICS AND SPATIALLY EXPLICIT APPROACHES TO DETECT LOCI UNDER SELECTION IN CLINAL POPULATIONS. *Evolution*, **67**, 3455–3468. Retrieved from http://dx.doi.org/10.1111/evo.12237

Joost, S., Bonin, A., Bruford, M.W., Després, L., Conord, C., Erhardt, G. & Taberlet, P. (2007). A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology*, **16**, 3955–3969.

Joost, S., Vuilleumier, S., Jensen, J.D., Schoville, S., Leempoel, K., Stucki, S., Widmer, I., Melodelima, C., Rolland, J. & Manel, S. (2013). Meeting review. Uncovering the genetic basis of adaptive change: on the intersection of landscape genomics and theoretical population genetics. *Molecular ecology*, **22**, 3659–65. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24003454

Kozak, K.H., Graham, C.H. & Wiens, J.J. (2008). Integrating GIS-based environmental data into evolutionary biology. *Trends in ecology & evolution (Personal edition)*, **23**, 141–148. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0169534708000426

Landguth, E.L. & Cushman, S.A. (2010). cdpop: A spatially explicit cost distance population genetics program. *Molecular Ecology Resources*, **10**, 156–161. Retrieved from http://dx.doi.org/10.1111/j.1755-0998.2009.02719.x

Leempoel, K. & Joost, S. (2012). Relatedness and scale dependency in very high resolution digital elevation models derivatives. *Proc of The OGRE Symposium, October 24-26*, 340. Retrieved from http://ogrs2012.org/\nhttp://www.ogrs2012.org/public/ogrs2012/OGRS_LIVRE_2012_FINAL.pdf

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

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

Luikart, G., England, P.R., 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.

Manel, S., Gugerli, F., Thuiller, W., Alvarez, N., Legendre, P., Holderegger, R., Gielly, L., Taberlet, P. & IntraBioDiv, C. (2012). Broad-scale adaptive genetic variation in alpine plants is driven by temperature and precipitation. *Molecular Ecology*, **21**, 3729–3738. Retrieved from http://dx.doi.org/10.1111/j.1365-294X.2012.05656.x

Manel, S. & Holderegger, R. (2013). Ten years of landscape genetics. *Trends in ecology & evolution*, **28**, 614–21. Retrieved February 23, 2014, from http://www.sciencedirect.com/science/article/pii/S0169534713001341

Manel, S., Joost, S., Epperson, B.K., Holderegger, R., Storfer, A., Rosenberg, M.S., Scribner, K.T., Bonin, A. & Fortin, M.-J. (2010a). Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Molecular Ecology*, **19**, 3760–3772.

Manel, S., Poncet, B.N., 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. Retrieved from <Go to ISI>://000281285200023

Manel, S., Schwartz, M.K., 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. Retrieved from http://dx.doi.org/10.1111/j.1365-294X.2009.04151.x

De Mita, S., Thuillet, A.-C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J. & Vigouroux, Y. (2013). Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383–1399. Retrieved from http://dx.doi.org/10.1111/mec.12182

MORAN, P.A.P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23.

Patterson, N., Price, A.L. & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, **2**, 2074–2093.

Petren, K. (2013). THE EVOLUTION OF LANDSCAPE GENETICS. *Evolution*, **67**, 3383–3385. Retrieved from http://dx.doi.org/10.1111/evo.12278

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, **155**, 945–959. Retrieved from http://www.genetics.org/content/155/2/945.abstract

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. & Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, **81**, 559–575.

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.

Storfer, A., Murphy, M.A., Evans, J.S., Goldberg, C.S., Robinson, S., Spear, S.F., Dezzani, R., Delmelle, E., Vierling, L. & Waits, L.P. (2006). Putting the “landscape” in landscape genetics. *Heredity*, **98**, 128–142.

Storfer, A., Murphy, M.A., Spear, S.F., Holderegger, R. & Waits, L.P. (2010). Landscape genetics: where are we now? *Molecular Ecology*, **19**, 3496–3514.

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. Retrieved from http://infoscience.epfl.ch/record/196921/files/EPFL_TH6014.pdf

Tobler, W.R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, **46**, 234–240 CR – Copyright © 1970 Clark Universi. Retrieved from http://www.jstor.org/stable/143141

Wiens, J.A. (1989). Spatial scaling in ecology. *Functional ecology*, **3**, 385–397.

Wilson, J.P. & Gallant, J.C. (2000). *Terrain Analysis : Principles and applications* (Wiley, Ed.). New York.

Zimmermann, N.E. & Kienast, F. (1999). Predictive mapping of alpine grasslands in Switzerland: Species versus community approach. *Journal of Vegetation Science*, **10**, 469–482. Retrieved from <Go to ISI>://000082992300005