Open computational landscape genetics

Stéphane Joost[1], Solange Duruz[1], Estelle Rochat[1], Ivo Widmer[1]

[1] Laboratory of Geographic Information Systems (LASIG), School of Architecture, Civil and Environmental Engineering (ENAC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Corresponding author:
Stéphane Joost[1]

Email address: stephane.joost@epfl.ch

ABSTRACT

Geographical Information Systems (GIS) are considered to be applications-led technology. Consequently, geographic information scientists commonly find themselves as guest in host disciplines in order to best exploit spatial analysis tools and methods, appropriately guided by experts in the field. An example is population genetics in evolutionary biology. Genetic information being linked to living organisms can be partially characterized by geographic coordinates. A research field named landscape genetics emerged at the intersection of genetics, environmental and geographic information science. Geocomputation and programming efforts carried out with the help of open sources technologies and dedicated to the analysis of genetic data gather together a key scientific community whose goal is to extract new knowledge from the present data tsunami caused by the advent of high throughput molecular data and of new sources of high resolution environmental data. While the level of sophistication of the population genetics functions included in the analytical frameworks developed until now are cutting-edge, advanced geo-competences are also required to reinforce the spatial side of this discipline. They will be particularly useful in conservation programmes for wildlife preservation, but also in farm animal genetic resources conservation.

INTRODUCTION

GIScience is inherently interdisciplinary, being a field that provides tools useful through their application to solving problems within other disciplines. Indeed, it has long been applied for a multiplicity of uses in land survey, hydrology, archeology, anthropology, transportation, etc. In this sense, Geographical Information Systems (GIS) are considered to be applications-led technology (Longley et al., 2015). Consequently, geographic information scientists commonly find themselves as guest in host disciplines in order to best exploit spatial analysis tools and methods, appropriately guided by experts in the field.

An example is population genetics in evolutionary biology. Indeed, people, animals, plants, are dispersed in space and interact in that space. Genetic information being linked to living organisms can therefore be partially characterized by geographic coordinates. The pairing of both genetic and spatial information is very well illustrated by the «Genographic» project, launched in part by The National Geographic Society and the IBM Corporation with the goal of collecting and analyzing more than 100’000 samples of DNA in order to trace the origins and to map the

movement of humans during the last 60'000 years (https://genographic.nationalgeographic.com). The idea was also popularized by Luigi Cavalli-Sforza and colleagues in «The History and Geography of Human Genes» (Cavalli-Sforza, Menozzi & Piazza, 1994) in which they systematically relied on geographical maps to show how the frequency of human genes is evolving from one population to another across the world. But exploiting the geographic dimension of genetic data was not new. Indeed, Sewall Wright and other cofounders of the field of population genetics considered geographics in their work from the 1930s (Epperson, 2003), as they were studying the distribution of allele frequencies under the influence of evolutionary forces (natural selection, genetic drift, mutation and migration). Basically, the main use of spatial information was to calculate geographical distances for comparison to genetic distances. Since then, there has been much advance on the notion of geographic distance towards more realistic and sophisticated tools, as demonstrated by Kozak et al. (2008) in their paper on the integration of GIS-based environmental data into evolutionary biology.

Here our intention is far from proposing an exhaustive review of the research that took place at the intersection of genetics, environmental and geographic information science, and that mainly came within the scope of a broad discipline named landscape genetics (Manel et al., 2003; and see a review in Sork & Waits, 2010). Instead we shortly describe some applications to stress the importance of geocomputation in spatial genetics, and in particular of the developments carried out with the help of open sources technologies. The latter gather together a key scientific community active at the intersection of computer science, evolutionary biology (to keep it broad) and geographic information science. In 2004, Marturano & Chadwick (2004) already mentioned that there was "no new genetics without computer science"; in 2016 this is truer than ever when (georeferenced) whole genome sequence data are up to become the standard to analyze.

Marturano & Chadwick also observed that open molecular data (see http://nextgen.epfl.ch/ for example) and open-source bioinformatics software were determining factors likely to enable the translation of huge amounts of data into medical or social advances. In the following sections, we selected examples to illustrate four main categories of applications in which open-source computational landscape genetic solutions can be distributed: a) the simple use of geographic coordinates for map production, distance calculation, or barrier detection; b) the simulation of spatially distributed datasets constrained by diverse biological criteria and gene flow modeling; c) the determination of population structure; and d) the detection of signatures of natural selection in the genome of investigated species.

SIMPLE USE OF GEOGRAPHIC COORDINATES

The Geographic Distance Matrix Generator (Ersts, 2016) is a simple example to illustrate what kind of geo-service biologists may need. Analyses in phylogeography for instance may require to detect patterns in the distribution of genetic variation across different spatial scales, taking into account a common process named isolation by distance (IBD) and under which genetic similarity decreases with geographic distance. The Geographic Distance Matrix Generator is a platform-independent Java application that computes all pair wise distances from a simple list of geographic coordinates. With more functionalities, GenGIS (Parks et al., 2013), is a free and open source software able to integrate biodiversity information and to display it on geographic maps. It includes calculation of alpha-diversity like the Shannon index, and geovisualization of

dissimilarity matrices. In the same category, despite it is not an open-source development strictly speaking (free software whose sources are available upon request), it is worth mentioning Barrier (Manni, Guerard & Heyer, 2004), a software to compute geographic barriers from matrices of genetic distances.

SIMULATIONS AND GENE FLOW MODELING

As landscape genetics is in part dedicated to the understanding of how geography and the environment structure genetic variation, several software include functions able to analyse processes and patterns of gene flow, and to identify genetic discontinuities and correlation between the latter and landscape features. For instance, Circuitscape (Shah & McRae, 2008) is a free and open-source software whose originality is to use algorithms from electronic circuit theory to model movement patterns and gene flow in animal and plant populations in the landscape. Although this program can be called through an ArcGIS Toolbox, Circuitscape is above all a regular Python package available from the Python packages repository. SimAdapt (Rebaudo et al., 2013) is a spatially explicit and individual-based landscape genetic simulation model. It can be combined with cellular automata to analyse evolutionary processes and population dynamics in changing landscapes. Of particular interest here is the use of the NetLogo environment, which is a free and open-source development environment for simulating natural and social phenomena (read the interesting paper by Thiele & Grimm, 2010, on the pairing of Netlogo with R). CDPOP (Landguth & Cushman, 2010) is a program to simulate gene flow, genetic drift, mutations, and also selection in complex landscapes. It is able to take into account a wide range of biological and evolutionary scenarios. CDPOP requires the Python2.7.x interpreter and uses the NumPy and SciPy packages.

And finally, one can also mention least-cost modelling approaches to provide functional landscape models, which have been a central component in the development of landscape genetics (Holderegger and Wagner, 2008). An example is «gdistance», an R package providing functionalities to calculate different distance measures and routes in heterogeneous geographic spaces represented as grids (https://cran.r-project.org/web/packages/gdistance/).

POPULATION STRUCTURE

“Population structure” means that instead of a single continuous population of a given species or breed, populations are subdivided in some way (distance, geographic barriers, etc.). When populations are subdivided, they can evolve apart and independently, and based on their proper genetic characteristics, it is possible to distinguish different population structures. Population structure is an important component of evolutionary genetics, and several software were developed to analyse it. TESS (Caye et al., 2016) is a program suited to detect genetic discontinuities in populations and to estimate individual genetic admixture proportions that vary in the geographical space. The program is based on a spatially explicit algorithm that provides an estimation of ancestry coefficients and it returns maps of geographical cluster assignments. TESS was developed using CMake (https://cmake.org/), an open-source and cross-platform family of tools designed to build package software; several R scripts come along with TESS for visualizing results. SPAGeDi (Spatial Pattern Analysis of Genetic Diversity) is a computer package also

developed with CMake whose goal is to characterise the spatial genetic structure of georeferenced individuals or populations with the help of genotype data (Hardy & Vekemans, 2002). Finally, it is important to mention adegenet (Jombart, 2008) in this section, a R package implementing a set of tools able to explore and analyse genetic data, and well illustrating the usefulness of investigating spatial patterns of genetic variability.

DETECTION OF SELECTION SIGNATURES

Based on the concept of spatial coincidence, several approaches were developed in order to detect signatures of natural selection within the genome of studied species. The principle of these correlative approaches is to use geographical coordinates to compare the variation of environmental features with the frequency of specific genomic regions. Some of them like Sambada (Stucki et al., 2014) implement simple uni- and multivariate logistic regression models enriched with spatial statistic functionalities (Moran's I or local indices of spatial association, Anselin, 1995). Sambada is written in C++ using the Scythe Statistical Library (Pemstein, Quinn & Martin, 2011) for matrix computation and probability distributions. Others are more sophisticated and take into account the structure of the populations investigated (see previous section) in addition to the variance explained by a given environmental variable. For instance, in LFMM (Frichot et al., 2013) population structure is considered in the model through unobserved variables obtained by means of a variant of Bayesian principal component analysis (latent factors mixed models). Interestingly, LFMM was developed in C and C++, but is also included in the R package named LEA (Frichot & François, 2015), which is broadly dedicated to landscape genomics and ecological association tests. Bayenv (Günther & Coop, 2013), SGLMM (Guillot et al., 2014), and more recently BayeScEnv (de Villemereuil & Gaggiotti, 2015), and Baypass (Gautier, 2015) constitute variants of the same approach.

THERE IS PLENTY OF ROOM FOR GEOSCIENTISTS!

Obviously there is plenty of room for geoscientists, not at the bottom as in the title of Richard Feynman's famous lecture, but in spatial genetics. Indeed, until now most of the programming efforts in (open) computational landscape genetics were carried out by biologists, geneticists or bioinformaticians. It is true that the main field of interest is evolutionary biology, naturally attracting biologists with programming skills. Nevertheless, migrations, gene flow, adaptation, etc. are processes that take place at the surface or the earth, with an undeniable geographic dimension. While the level of sophistication of the population genetics functions included in the analytical frameworks developed until now in landscape genetics are cutting-edge, advanced geo-competences are also required to reinforce the spatial side of this discipline. They will be particularly useful in conservation programmes for wildlife preservation, but also in farm animal genetic resources conservation where the integration of many different thematics (socio-economy, policies, demography, genetics, environment, etc.) are required, involving heterogeneous types of data at different geographical scales (Bruford et al., 2015). New knowledge will be extracted from the present data tsunami − mainly constituted by the advent of high throughput molecular data and new sources of high resolution environmental data − only if innovative, transdisciplinary and efficient computing tools are developed with the contribution of geoscientists ready to submerge themselves in evolutionary biology

REFERENCES

Anselin L. 1995. Local Indicators of Spatial Association - Lisa. Geographical Analysis 27:93–115.

Bruford MW., Ginja C., Hoffmann I., Joost S., Orozco-terWengel P., Alberto FJ., Amaral AJ., Barbato M., Biscarini F., Colli L., Costa M., Curik I., Duruz S., Ferenčaković M., Fischer D., Fitak R., Groeneveld LF., Hall SJG., Hanotte O., Hassan F., Helsen P., Iacolina L., Kantanen J., Leempoel K., Lenstra JA., Ajmone-Marsan P., Masembe C., Megens H-J., Miele M., Neuditschko M., Nicolazzi EL., Pompanon F., Roosen J., Sevane N., Smetko A., Štambuk A., Streeter I., Stucki S., Supakorn C., Telo Da Gama L., Tixier-Boichard M., Wegmann D., Zhan X. 2015. Prospects and challenges for the conservation of farm animal genomic resources, 2015-2025. Livestock Genomics:314.

Cavalli-Sforza LL., Menozzi P., Piazza A. 1994. The history and geography of human genes. Princeton, N. J., Etats-Unis d'Amérique: Princeton University Press.

Caye K., Deist TM., Martins H., Michel O., François O. 2016. TESS3: fast inference of spatial population structure and genome scans for selection. Molecular Ecology Resources 16:540–548.

Epperson BK. 2003. Geographical Genetics (MPB-38). Princeton University Press.

Ersts PJ. 2016.The Geographic Distance Matrix Generator - Documentation. Available at http://biodiversityinformatics.amnh.org/open_source/gdmg/documentation.php (accessed February 3, 2016).

Frichot E., Schoville SD., Bouchard G., Francois O. 2013. Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. Molecular Biology and Evolution 30:1687–1699.

Frichot E., François O. 2015. LEA: An R package for landscape and ecological association studies. Methods in Ecology and Evolution 6:925–929.

Gautier M. 2015. Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. Genetics:genetics.115.181453.

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.

Günther T., Coop G. 2013. Robust Identification of Local Adaptation from Allele Frequencies. Genetics 195:205–220.

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.

Holderegger, R. & Wagner, H.H. (2008). Landscape Genetics. BioScience, 58, 199–207.

Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405.

Kozak KH., Graham CH., Wiens JJ. 2008. Integrating GIS-based environmental data into evolutionary biology. Trends in Ecology & Evolution 23:141–148.

Landguth EL., Cushman SA. 2010. cdpop: A spatially explicit cost distance population genetics program. Molecular Ecology Resources 10:156–161.

Longley PA., Goodchild MF., Maguire DJ., Rhind DW. 2015. Geographic Information Science and Systems. Chichester: Wiley.

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

Manni F., Guerard E., Heyer E. 2004. Geographic Patterns of (Genetic, Morphologic, Linguistic) Variation: How Barriers Can Be Detected by Using Monmonier's Algorithm. Human Biology 76:173–190.

Marturano A., Chadwick R. 2004. How the Role of Computing is Driving New Genetics' Public Policy. Ethics and Information Technology 6:43–53.

Parks DH., Mankowski T., Zangooei S., Porter MS., Armanini DG., Baird DJ., Langille MGI., Beiko RG. 2013. GenGIS 2: Geospatial Analysis of Traditional and Genetic Biodiversity, with New Gradient Algorithms and an Extensible Plugin Framework. PLoS ONE 8:e69885.

Pemstein D., Quinn KM., Martin AD. 2011. The Scythe Statistical Library: An Open Source C++ Library for Statistical Computation.

Rebaudo F., Le Rouzic A., Dupas S., Silvain J-F., Harry M., Dangles O. 2013. SimAdapt: an individual-based genetic model for simulating landscape management impacts on populations. Methods in Ecology and Evolution 4:595–600.

Shah VB., McRae B. 2008. Circuitscape: A Tool for Landscape Ecology. In: Varoquaux G, Vaught T, Millman J eds. Proceedings of the 7th Python in Science Conference. Pasadena, CA USA, 62 – 65.

Sork VL., Waits L. 2010. Contributions of landscape genetics – approaches, insights, and future potential. Molecular Ecology 19:3489–3495.

Stucki S., Orozco-terWengel P., Bruford MW., Colli L., Masembe C., Negrini R., Taberlet P., Joost S., Consortium the N. 2014. High performance computation of landscape genomic models integrating local indices of spatial association. arXiv:1405.7658 [q-bio].

Thiele JC., Grimm V. 2010. NetLogo meets R: Linking agent-based models with a toolbox for their analysis. Environmental Modelling & Software 25:972–974.

de Villemereuil P., Gaggiotti OE. 2015. A new FST-based method to uncover local adaptation using environmental variables. Methods in Ecology and Evolution 6:1248–1258.