Bioinformatics, 32(7), 2016, 1106–1108
doi: 10.1093/bioinformatics/btv703
Advance Access Publication Date: 28 November 2015
Applications Note

OXFORD

Genetics and population analysis

Accurate continuous geographic assignment from low- to high-density SNP data

Gilles Guillot[1,2,*], Hákon Jónsson[2], Antoine Hinge[1], Nabil Manchih[1] and Ludovic Orlando[2]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark and [2]Centre for GeoGenetics, Natural History Museum of Denmark and University of Copenhagen, Copenhagen, Denmark

*To whom correspondence should be addressed.
Associate Editor: Janet Kelso

Received on 16 June 2015; revised on 17 September 2015; accepted on 25 November 2015

Abstract

Motivation: Large-scale genotype datasets can help track the dispersal patterns of epidemiological outbreaks and predict the geographic origins of individuals. Such genetically-based geographic assignments also show a range of possible applications in forensics for profiling both victims and criminals, and in wildlife management, where poaching hotspot areas can be located. They, however, require fast and accurate statistical methods to handle the growing amount of genetic information made available from genotype arrays and next-generation sequencing technologies.
Results: We introduce a novel statistical method for geopositioning individuals of unknown origin from genotypes. Our method is based on a geostatistical model trained with a dataset of georeferenced genotypes. Statistical inference under this model can be implemented within the theoretical framework of Integrated Nested Laplace Approximation, which represents one of the major recent breakthroughs in statistics, as it does not require Monte Carlo simulations. We compare the performance of our method and an alternative method for geospatial inference, SPA in a simulation framework. We highlight the accuracy and limits of continuous spatial assignment methods at various scales by analyzing genotype datasets from a diversity of species, including Florida Scrub-jay birds *Aphelocoma coerulescens, Arabidopsis thaliana* and humans, representing 41–197,146 SNPs. Our method appears to be best suited for the analysis of medium-sized datasets (a few tens of thousands of loci), such as reduced-representation sequencing data that become increasingly available in ecology.
Availability and implementation: http://www2.imm.dtu.dk/~gigu/Spasiba/
Contact: gilles.b.guillot@gmail.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Inferring the geographic origin of living organisms from their genetic information is of great interest for many applications in biology. It can provide information about gene flow, migration patterns and connectivity in natural populations (Kremer *et al.*, 2012; Schwartz *et al.*, 2007; Waples and Gaggiotti, 2006) but can also help inform wildlife managers about illegal animal translocations and poaching hotspots (Manel *et al.*, 2005; Ogden *et al.*, 2009). As such, this information can complement the arsenal of DNA-based fraud detection methods, aiming at detecting derivatives of endangered and trade-restricted species (Coghlan *et al.*, 2012). In addition, DNA-informed geospatial localization can reveal the geographic source of pathogens during epidemiological outbreaks (Sloan *et al.*, 2009) or the geographic origin of plants and animals used in the industrial manufacture of food products (Lees,

2003). In forensics, DNA-informed geospatial localization can help profiling criminals and identifying the origin of unidentified bodies (Primorac and Schanfield, 2014). Here, we introduce Spatial Bayesian Inference (SPASIBA), a novel method for geospatial assignment. The premise of the SPASIBA method is that in most natural contexts, spatial patterns of allele frequencies are complex and are likely to be well captured by a geostatistical model such as the one implemented in the SCAT program (Wasser *et al.*, 2004), but here, we leverage the power of a recent breakthrough statistical theory developed by (Rue *et al.* 2009) and (Lindgren *et al.* 2011). This allows us to make MCMC-free inference in only a fraction of the time required by SCAT.

2 Methods

We consider datasets consisting of a set of allelic counts at bi-allelic loci for a set of reference populations of known geographic locations. Individuals of unknown geographic origin are genotyped for the complete set of orthologous loci. Our method is tailored to geographically assign the latter individuals given the set of georeferenced genetic data (hereafter referred to as training data). We denote by f_{sl} the frequency of a reference allele at locus l at geographic location s. We assume that the number of reference alleles is binomial $B(n_{sl}, f_{sl})$ with statistical independence across loci. This amounts to assuming that individuals located around location s form a population at Hardy–Weinberg equilibrium with linkage equilibrium across markers. Our model has therefore the same likelihood function as the one described by (Pritchard *et al.* 2000). We assume that spatial variation of allele frequencies can be described by a non-parametric surface in two dimensions. Following (Wasser *et al.* 2004), we model the spatial variation of $(f_{sl})_s$ by a set of spatially auto-correlated random variables with Gaussian distribution (a random field) denoted by y_{sl}. We assume that f_{sl} and y_{sl} relate through a logistic function. We model the spatial auto-covariance of allele frequencies by imposing a parametric form to $\mathrm{Cov}[y_{sl}, y_{s'l}]$. We should stress that our method is designed to perform continuous assignment. Therefore, we cannot only rely on a covariance matrix, but need instead a covariance function, which models covariance variation in the continuous space. This model can be defined either in a flat geographic domain, using straightline distances (2D) or on the sphere using great circle distances (a sub-model referred to be low as 3D model, better appropriate to analyze worldwide datasets). Under our model, the covariance between allele frequencies at geographic locations s and s' decays with the geographic distance $|s - s'|$ and therefore captures the form of population structure known as isolation-by-distance (Guillot *et al.*, 2009). A key feature of our model is that it can be handled within the Integrated Nested Laplace Approximation (INLA) framework. The location of samples from unknown geographic origin is estimated following three steps. In the first step, we estimate the parameters of the covariance model from the set of georeferenced genetic data, which summarize information on the magnitude and the spatial scale of variation of allele frequencies. In the second step, we compute estimated geographic maps of allele frequencies for each locus using the parameters previously estimated. In the third step, we assign samples of unknown origin by maximizing the likelihood that a sample comes from a specific location over the study area (discretized over a fine grid). Our method is described in full detail in the Supporting material.

3 Results

In Supporting material, we assess the performance of our method and SPA (Yang *et al.*, 2012), the most-commonly used method in geospatial assignment. We evaluated the accuracy of both methods using real and simulated datasets, spanning a range of possible applications in biology. The three application cases considered included organisms characterized by very diverse vagility and dispersal behaviors, spatial scales ranging from the regional to the continental scale and genotyping information ranging from 41 to 197 146 SNPs. Simulations were performed under a series of statistical models, selected to uncover different underlying biological processes. In most situations, our method was found to outperform SPA, showing assignment errors corresponding to only a fraction of those measured in SPA. The difference between both methods was most pronounced when a limited number of loci were considered.

4 Conclusion

The statistical model underlying our method is largely reminiscent of the SCAT program (Wasser *et al.*, 2004, 2007). However, building on INLA instead of MCMC allowed us to significantly reduce computing times by typically several orders of magnitudes. In addition, our approach is free of MCMC convergence issues that can considerably increase the computation burden. In the Florida Scrub-jay dataset (1311 individuals, 41 SNPs), SPASIBA achieved a full analysis in ~10 min using a single 3–GHz CPU. SCAT required about a week of computation, while SPA provided results within a few seconds. These computing times scale linearly with the number of loci. With such running times and the accuracy levels demonstrated above, SPASIBA appears appropriate for the routine analysis of SNP datasets consisting of a few tens of thousands of loci. In particular, it appears to be an ideal method for the analysis reduced-representation sequencing data that become increasingly available in ecology, including for non-model organisms (Davey *et al.*, 2011).

Acknowledgements

Access to the POPRES dataset was granted by the Data Access Committee of the NCBI dbGaP Data Access request system at the National Institute of Health. We thank John W. Fitzpatrick, Reed Bowman, Aurelie Coulon, the Archbold Biological Station and the Cornell Lab of Ornithology for permission to use their extensive sample of georeferenced genetic data on Florida Scrub-jays.

Funding

The Danish e-Infrastructure for Computing, the working group on Computational Landscape Genomics at the National Institute for Mathematics and Biology Synthesis, a Marie-Curie Initial Training Network EUROTAST [Grant number FP7ITN-290344], the Danish Council for Independent Research, Natural Sciences, the Danish National Research Foundation [Grant number DNFR94] and Marie-Curie Actions [Career Integration Grant number FP7CIG-293845]. Florida Scrub-jays data were generated with support from the U.S. National Science Foundation [Grant number DEB-0316292].

Conflict of interest: none declared.

References

Coghlan,M. *et al.* (2012) Deep sequencing of plant and animal DNA contained within traditional Chinese medicines reveals legality issues and health safety concerns. *PLoS Genet.*, **8**, e1002657.

Davey,J.W. *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.*, **12**, 499–510.

Guillot,G. *et al.* (2009) Statistical methods in spatial genetics. *Mol. Ecol.*, **18**, 4734–4756.

Kremer,A.O. *et al.* (2012) Long-distance gene flow and adaptation of forest trees to rapid climate change. *Ecol. Lett.*, **15**, 378–392.

Lees,M. (2003) *Food Authenticity and Traceability*. Elsevier, Amsterdam.

Lindgren,F. *et al*. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. R. Stat. Soc., Ser. B*, **73**, 423–498.

Manel,S. *et al*. (2005) Assignment methods: matching biological questions with appropriate techniques. *Trends Ecol. Evol*., **20**, 136–142.

Ogden,R. *et al.* (2009) Wildlife DNA forensics-bridging the gap between conservation genetics and law enforcement. *Endanger. Species Res*., **9**, 179–195.

Primorac,D. and Schanfield,M. (2014) *Forensic DNA Applications: An Interdisciplinary Perspective*. CRCPress, BocaRaton, FL.

Pritchard,J. *et al*. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Rue,H. *et al*. (2009) Approximate Bayesian inference for latent Gaussian models by using Integrated Nested Laplace Approximations. *J. R. Stat. Soc., Ser. B*, **71**, 1–35.

Schwartz,M. *et al*. (2007) Genetic monitoring as a promising tool for conservation and management. *Trends Ecol. Evol*., **22**, 25–33.

Sloan,C.D. *et al*.(2009) Ecogeographic genetic epidemiology. *Genet. Epidemiol*., **33**, 281–289.

Waples,R. and Gaggiotti,S.O. (2006) What is a population? An empirical evaluation of some genetic methods for indentifying the number of gene pools and their degree of connectivity. *Mol. Ecol*., **15**, 1419–1439.

Wasser,S.K. *et al*. (2004) Assigning African elephants DNA to geographic region of origin: applications to the ivory trade. *Proc. Natl Acad. Sci. USA*, **101**, 14847–14852.

Wasser,S. *et al*. (2007) Using DNA to track the origin of the largest ivory seizure since the 1989 trade ban. *Proc. Natl Acad. Sci. USA*, **104**, 4228–4233.

Yang,W. *et al*. (2012) A model-based approach for analysis of spatial structure in genetic data. *Nat. Genet*., **44**, 725–731.