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

Gilles Guillot\textsuperscript{1,2,*}, Hákon Jónsson\textsuperscript{2}, Antoine Hinge\textsuperscript{1}, Nabil Manchih\textsuperscript{1}, Ludovic Orlando\textsuperscript{2}

\textsuperscript{1}Department of Applied Mathematics and Computer Science, Technical University of Denmark
\textsuperscript{2}Centre for GeoGenetics, Natural History Museum of Denmark and University of Copenhagen.

Received on June 16 2015; revised on July 14 2015; accepted on XXXXX

ABSTRACT

\textbf{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.

\textbf{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 (INLA), 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 \textit{Aphelocoma coerulescens}, \textit{Arabidopsis thaliana} and humans, representing 41 to 197,146 SNPs. Our method appears to be best suited for the analysis of medium-size datasets (a few tens of thousands of loci), such as reduced-representation sequencing data that become increasingly available in ecology.

\textbf{Availability:} \url{http://www2.imm.dtu.dk/~gigu/Spasiba/}

\textbf{Contact:} gilles.b.guillot@gmail.com

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 [Waples and Gaggiotti, 2006, Schwartz et al., 2007, Kremer et al., 2012] 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]. Additionally, 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 SPASIBA (Spatial Bayesian Inference), 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 models 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.

Here, we introduce SPASIBA (Spatial Bayesian Inference), a novel method for geospatial assignment that builds on the geospatial model underlying the SCAT program [Wasser et al., 2004]. In contrast to SPA [Yang et al., 2012], this model does not expect unidirectional clinal variation and is, thus, likely to capture complex spatial patterns of allele frequencies. In SPASIBA, we exploit a recent breakthrough in statistical theory developed by Rue et al. [2009] and Lindgren et al. [2011] 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 geo-referenced 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

*to whom correspondence should be addressed
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 \((f_{st})_s\) by a set of spatially auto-correlated random variables with Gaussian distribution (a random field) denoted by \(y_{sl}\). We assume that \(f_{st}\) and \(y_{sl}\) relate through a logistic function. We model the spatial auto-covariance of allele frequencies by imposing a parametric form to \(\text{Cov}(y_{sl}, y_{s'l})\). 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 straight-line distances (2D) or on the sphere using great circle distances (a sub-model referred to below 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 \(|s-s'|\)/1 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 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 geo-referenced 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 geo-spatial 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. Additionally, our approach is free of MCMC convergence issues 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].

ACKNOWLEDGEMENT

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: This work was supported 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 (FP7 ITN-290344), the Danish Council for Independent Research, Natural Sciences, the Danish National Research Foundation (Grant DNFR 94) and Marie-Curie Actions (Career Integration Grant FP7 CIG-293845). Florida scrub jays data were generated with support from the U. S. National Science Foundation, grant #DEB-0316292

REFERENCES


R.S. Waples and O. Gaggiotti. What is a population? An empirical evaluation of some genetic methods for indentation the number of gene pools and their degree of