



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

Guillot, Gilles; Jónsson, Hákon; Hinge, Antoine; Manchih, Nabil; Orlando, Ludovic

*Published in:*  
Bioinformatics

*Link to article, DOI:*  
[10.1093/bioinformatics/btv703](https://doi.org/10.1093/bioinformatics/btv703)

*Publication date:*  
2016

*Document Version*  
Early version, also known as pre-print

[Link back to DTU Orbit](#)

*Citation (APA):*  
Guillot, G., Jónsson, H., Hinge, A., Manchih, N., & Orlando, L. (2016). Accurate continuous geographic assignment from low- to high-density SNP data. *Bioinformatics*, 32(7), 1106-1108.  
<https://doi.org/10.1093/bioinformatics/btv703>

---

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

1                                   Supplementary material for  
2                                   *Accurate continuous geographic assignment*  
3                                   *from low- to high-density SNP data.*

4 Gilles Guillot\*, Hákon Jónsson†, Antoine Hinge\*, Nabil Manchih\*, Ludovic Orlando†

5                                   July 14, 2015

---

\*Department of Applied Mathematics and Informatics, Technical University of Denmark, 2800, Lyngby, Denmark. gigu@dtu.dk

†Centre for Geogenetics, Museum of Natural History and University of Copenhagen, Øster Voldgade 5-7, 1350 København K, Denmark

## 6 Method

### 7 Statistical model

8 We consider datasets consisting of a set of allelic counts at bi-allelic loci for a set of reference  
9 populations of known geographic locations. Additionally, genotypes for orthologous loci are  
10 available for individuals of unknown geographic origin. Our method is tailored to geolocation  
11 the latter individuals given the set of geo-referenced genetic data (hereafter referred to as  
12 training data). We denote by  $f_{sl}$  the frequency of a reference allele at locus  $l$  at geographic  
13 location  $s$ . We assume that the number of reference alleles is binomial  $B(n_{sl}, f_{sl})$  with  
14 statistical independence across loci. This amounts to assuming that individuals located around  
15 location  $s$  form a population at Hardy-Weinberg equilibrium with linkage equilibrium across  
16 markers. Our model has therefore the same likelihood function as described by Pritchard  
17 et al. (2000). We assume that spatial variation of allele frequencies can be described by  
18 a non-parametric surface in two dimensions. Following Wasser et al. (2004), we model the  
19 spatial variation of  $(f_{sl})_s$  by a set of spatially auto-correlated random variables with Gaussian  
20 distribution (a random field) denoted by  $y_{sl}$ . We assume that  $f_{sl}$  and  $y_{sl}$  relate through a  
21 logistic function  $f_{sl} = 1/[1 + \exp -(a_l + y_{sl})]$  where  $a_l$  is a locus-specific intercept. We  
22 model the spatial auto-covariance of allele frequencies by imposing a parametric form to  
23  $\text{Cov}[y_{sl}, y_{s'l}]$ .

24 We should stress that our method is designed to perform continuous assignment. There-  
25 fore, we cannot only rely on a covariance matrix, but need instead a covariance function,  
26 which models covariance variation in the continuous space. We assume that  $\text{Cov}[y_{sl}, y_{s'l}] =$   
27  $C(|s - s'|) = C(h)$  for some function  $C$ , implying that the spatial auto-covariance only de-  
28 pends on the geographical distance  $h = |s - s'|$ . As commonly assumed in spatial statistics  
29 and for reasons that will appear later, we consider that  $C$  belongs to the Matérn family i.e.  
30  $C(h) = \sigma^2(\kappa h)^\nu 2^{1-\nu} \Gamma^{-1}(\nu) K_\nu(\kappa h)$  where  $K_\nu$  is the modified Bessel function of the second  
31 kind of order  $\nu > 0$ ,  $\kappa > 0$  is a scaling parameter and  $\sigma^2$  is the marginal variance. This model  
32 can be defined either in a flat geographical domain, using straight-line distances (2D) or on  
33 the sphere using great circle distances (a sub-model referred to below as 3D model) which  
34 is more appropriate when analyzing worldwide datasets. The Matérn family of covariance

35 function is broad and flexible, it includes for example the widely used exponential covariance  
36 function  $\sigma^2 \exp(-\kappa h)$  as a particular case (Gelfand et al., 2010; Porcu et al., 2010). Under  
37 our model, the covariance between allele frequencies at geographical locations  $s$  and  $s'$  decays  
38 with the geographical distance  $|s - s'|$  and therefore models the form of population structure  
39 known as isolation-by-distance (Guillot et al., 2009; Guillot and Orlando, 2015). However,  
40 its main advantage is computational, as explained in the next section.

## 41 **Estimation within the INLA-GMRF-SPDE framework**

42 A key feature of our model is that it can be handled within the theoretical and computational  
43 framework developed by Rue et al. (2009) and Lindgren et al. (2011). The former develops  
44 a framework for Bayesian inference in a broad class of models enjoying a latent Gaussian  
45 structure. The latter bridges a gap between Markov random fields (MRF) and Gaussian  
46 random fields (GRF) theory and makes it possible to combine the flexibility of Gaussian  
47 random fields for modelling and the computational efficiency of Markov random fields for  
48 inference. The approach of Lindgren et al. (2011) is based on the observation that a Gaussian  
49 random field  $y(s)$  with a Matérn covariance function is the solution of the stochastic partial  
50 differential equation (SPDE). Solving numerically this SPDE with finite element techniques  
51 and a smart choice of basis functions makes it possible to use Markov properties. This  
52 framework can be embedded in the INLA method of Rue et al. (2009), which makes use of the  
53 Markovian structure of the model during computation. The INLA and SPDE approximate  
54 inference methods are implemented in the R-INLA package (Rue et al., 2014). See also  
55 Guillot et al. (2013) for the use of a related model in genomics.

## 56 **Practical implementation of INLA-GMRF-SPDE**

57 We now describe specific steps for casting the problem of continuous geographic assignment  
58 in the INLA-GMRF-SPDE framework. The location of samples from unknown geographical  
59 origin is estimated following three steps.

60 In the first step, we estimate the parameters of the GMRF-SPDE model from the set of geo-  
61 referenced genetic data. There are three parameters  $(\sigma, \kappa, \nu)$ . However, in line with Lindgren  
62 et al. (2011) and to minimize the computational burden, we set  $\nu = 1$ . We stress that

63 the inferential difficulties reported under Markov Random field models by Sørbye and Rue  
64 (2014) bear on Intrinsic Markov Random fields (IMRF). The SPDE-GMRF model considered  
65 here differs sharply from the IMRF model and is not subject to this issue. The estimated  
66 parameters  $(\sigma, \kappa)$  of the GMRF-SPDE model summarize information on the magnitude and  
67 the spatial scale of variation of allele frequencies. This step involves processing the whole  
68 dataset jointly and can be computed for datasets consisting of typically ~500 individuals and  
69 ~1,000 loci. For larger datasets, we devised a strategy limiting computational demands and  
70 running times by picking a random subset of loci and performing inference of  $\sigma$  and  $\kappa$  on  
71 this subset. In the second step, we compute estimated geographic maps of allele frequencies  
72 for each locus using the parameters previously estimated.

73 In the third step, we assign samples of unknown origin by maximizing the likelihood that  
74 a sample comes from a specific location over the study area (in practice, the nodes of a  
75 grid which can be easily chosen to be fine enough to avoid any discretization issue). In the  
76 latter step, we maximimise the likelihood  $p(\text{genotypes}|\text{allele freq., locations})$  with respect to  
77 the geographical locations, assuming allele frequencies are perfectly estimated. The method  
78 provides therefore not only a point estimate of the unknown geographic origin but also a map  
79 informative about uncertainty in assignment and multiple putative origins, as illustrated in  
80 figure I. See (Rue et al., 2009; Lindgren et al., 2011; Simpson et al., 2012; Martins et al.,  
81 2013) for details on the INLA method and its implementation with random fields models.

82 The main competitors of SPASIBA are the SCAT program of Wasser et al. (2004) and  
83 the SPA program of Yang et al. (2012). We therefore compare our method to the latter. The  
84 accuracy of the INLA method in spatial statistics being widely validated (Lindgren et al.,  
85 2011; Simpson et al., 2012; Martins et al., 2013). Additionally, our model is very similar to  
86 that of Wasser et al. (2004). As running SCAT on a single dataset of more than 1,000 loci  
87 typically requires weeks of computations, we did not carry out full comparison of SPASIBA  
88 and SCAT. The comparison was, therefore, limited to SPASIBA and SPA. Furthermore, our  
89 focus is on medium-density SNP datasets which are becoming increasingly more common in  
90 the field of ecology. Therefore, we do not compare to recent methods that require high-density  
91 SNP data (Drineas et al., 2010; Baran et al., 2013; Rañola et al., 2014; Yang et al., 2014).  
92 We also stress that our method is tailored to perform *continuous* geographic assignment,

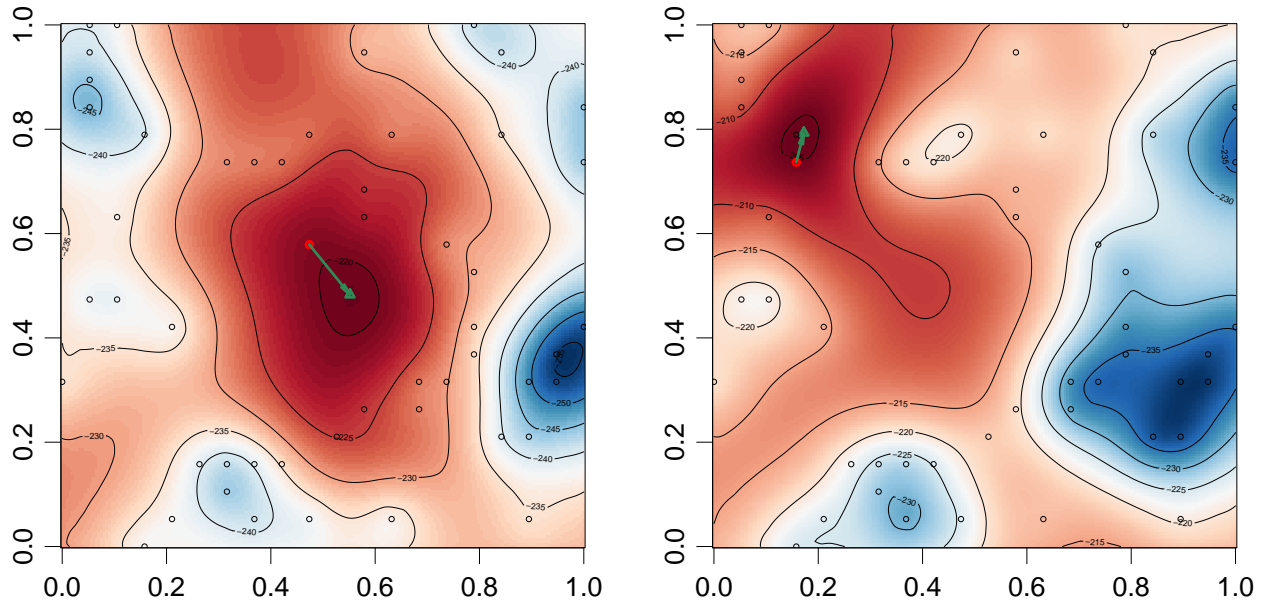


Figure I: Map of SPASIBA likelihood scores and assignment error (green arrow) recovered for one individual. Data were simulated under model underlying the SPASIBA program (50 diploid individuals with known origin, 200 SNP markers). We used SPASIBA to assign the most likely geographic origin of a given individual. The red dot indicates the true geographic position of the individuals, while the green triangle corresponds to the position inferred by SPASIBA. Typically, an individual located in an area of low spatial sampling density (left panel) is assigned with larger errors than an individual located in a area of high spatial sampling density or close to an individual of the training sample (right panel). The map relative to a specific individual can be checked for the existence of several local maxima. The various global maxima corresponding to the various individuals can be compared and help identify which individuals are assigned with low and large confidence.

93 therefore we do not compare it to methods designed to assign individuals to a set of known  
 94 populations such as GENECLASS (Piry et al., 2004).

## 95 Results

### 96 Model validation on simulated data

97 We validated our method on datasets simulated under various spatially explicit models,  
98 in line with the validation strategy used earlier by Novembre et al. (2008) and Bradburd  
99 et al. (2013). A set of individuals is randomly selected and removed from the dataset.  
100 Remaining individuals are used to train the algorithm (training dataset) while individuals  
101 initially removed from the dataset are used as testing data for which we predict their spatial  
102 origin using genotype information only. The accuracy of each method is assessed using the  
103 average geographical distance obtained between predicted and known geographical positions.

104 We first simulated datasets under the model underlying the SPA program (Yang et al.,  
105 2012) in which variation of allele frequencies is given by a logistic function in two dimensions  
106 characterized by an origin, a slope and a direction. We considered a training set consisting  
107 of 100 diploid individuals and evaluated accuracy in assignment for 200 individuals. The  
108 locations of individuals were sampled from a uniform distribution on the unit square, the  
109 direction of the cline was sampled uniformly on  $[-\pi, \pi]$  and the slope was sampled uniformly  
110 on  $[1, 10]$ . This type of simulation can be seen as the best-case scenario for the SPA method.  
111 We then simulated data under the geostatistical random field model underlying the SPA-  
112 SIBA program. The data simulated here display far more variability than those generated  
113 under the SPA model. We considered a training set consisting of 100 diploid individuals and  
114 evaluated accuracy in assignment for 200 individuals. The marginal variance of the random  
115 field was set to one and the scale parameter to  $10/3$  on a unit square domain.

116 Lastly, we used the MS program (Hudson, 2002)) to simulate data under a two-dimensional  
117 stepping stone model. This approach was selected because it explicitly accounts for demo-  
118 graphic and mutational processes and therefore provides spatial genetic structure. Import-  
119 tantly, it does not rely on any of the assumptions underlying the SPA and the SPASIBA  
120 program. Data were simulated for haploid individuals on a  $20 \times 20$  grid with training and  
121 testing sets of size 380 and 20 individuals respectively. In all cases the mutation and migra-  
122 tion were controlled by setting mutation rate  $4N\mu = 1$  and the migration rate  $4Nm = 0.4$ .  
123 Simulations were performed for a number of loci varying from 20 to 5,000. Results reported

124 for each condition are obtained as averages over five independent datasets. Results for the  
125 three types of simulations are summarized on figure II.

126 For data simulated under the logistic curve underlying the SPA program, our method  
127 performed similarly or better than the SPA method, as long as a large number of loci was  
128 considered (superior to 1,000). For smaller datasets, SPASIBA achieved better accuracy  
129 than SPA, with for example an average error twice smaller for 20 loci (Fig. II top panel).

130 For data simulated under the geostatistical model underlying the SPASIBA program, the  
131 assignment errors are typically larger than those observed for data simulated under the SPA  
132 model, which reflects the greater spatial complexity in the genetic variation simulated. In  
133 such conditions, the SPASIBA method outperforms the SPA method regardless of the num-  
134 ber of loci analyzed (Fig. II middle panel).

135 In our attempts to implement the SPA program on the stepping-stone data, we faced numer-  
136 ous cases where the assignment error appears of several orders of magnitude larger than the  
137 size of the geographic domain considered. This phenomenon becomes increasingly important  
138 with increasing numbers of loci (Tab. I). Even when discarding such problematic datasets  
139 from the analysis, the assignment error of the SPA method is larger (up to about 10-fold  
140 over the range of loci considered) than that of SPASIBA (Fig. II bottom panel).

141 As SPASIBA provided great performance in simulated settings, we next applied our  
142 method to three real datasets, selected to represent a range of possible biological situations.



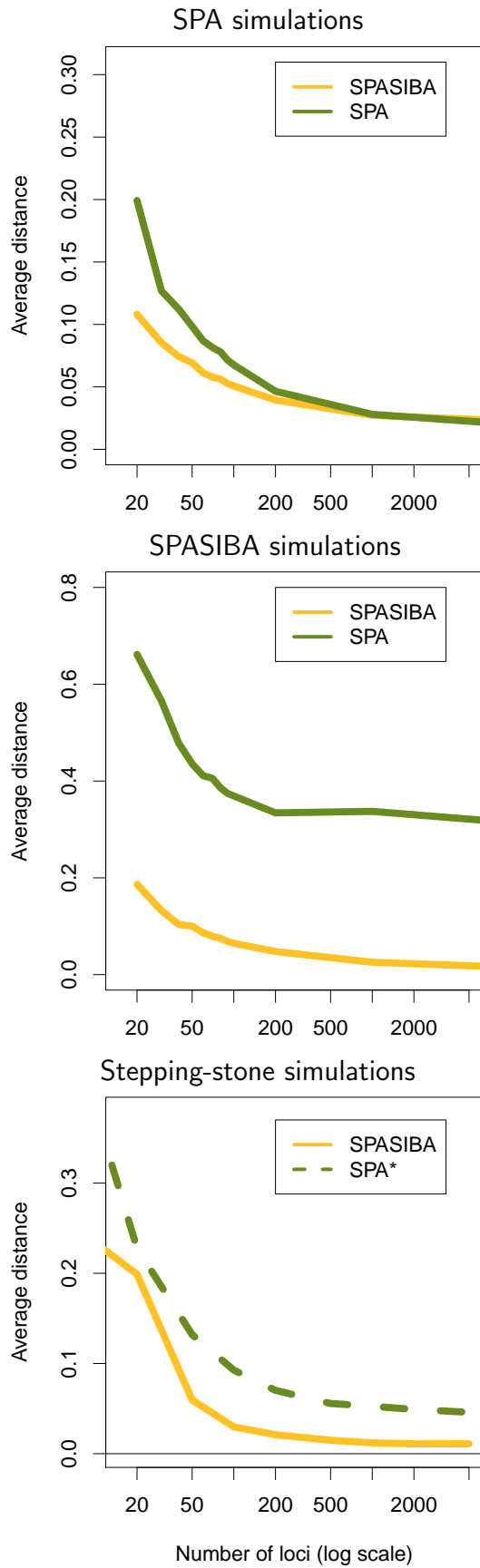


Figure II: Assignment error on simulated data. We simulated spatially explicit genetic datasets using three methods (Top: SPA, Middle: SPASIBA, Bottom: MS). In the bottom plot, the curve for the SPA method corresponds to the subset of data where SPA did not fail, see text for detail.

Nb loci \ Index sim	1	2	3	4	5
10	0	0	2	0	0
20	0	0	0	0	0
50	0	0	0	0	0
100	0	0	0	0	0
200	0	0	0	0	0
500	0	0	13	11	0
1000	0	1	0	6	13
2000	8	0	3	0	9
5000	8	10	20	12	9
10000	7	11	9	13	11

Table I: Summary about problematic runs with the SPA program on data simulated under a stepping stone model: number of individuals with outlier estimated coordinates. These are defined conventionally as those larger than  $10^{64}$ .

## 143 **Florida scrub jays**

144 We consider here a dataset consisting of 1,311 Florida scrub jay birds (*Aphelocoma cærulescens*),  
145 which are known for their short dispersal distances (Woolfenden and Fitzpatrick, 1984, 1996;  
146 Fitzpatrick et al., 1999). For example, Coulon et al. (2010) reported dispersal distances of  
147 the order of 1.3-4.2 km (depending on sex and habitat). This species is therefore expected  
148 to show strong geographical population structure, which should facilitate geospatial assign-  
149 ment. The species was sampled extensively over Florida and genotyped for a limited number  
150 of SNP markers (for a total of 41). This allowed us to explore how the method performs  
151 with types of datasets that are classical for ecological surveys and population monitoring.  
152 The population density and the spatial sampling strategy are both characterized by the  
153 absence of clusters, which are known to be problematic for traditional population-based as-  
154 signment methods (Manel et al., 2005). We investigated the assignment accuracy of our  
155 method by splitting the dataset into a random training set of 1,000 individuals, the 311  
156 remaining individuals being used as a testing set. Running the SPA program on the same  
157 training and testing dataset returned non-sensical results with a large proportion of individu-  
158 als assigned at locations farther than several thousands of kilometers away from Florida. For  
159 SPASIBA outputs, we computed the distance between the predicted origin and the sampling  
160 location and used this as a genuine measure of the assignment error. This distance has a  
161 median of 26.4 km, a 75% quantile of 76.6 km and a maximum of 274.5 km. The distribu-  
162 tion of the distance between predicted origin and sampling location is displayed on figure III.  
163 This, together with the short dispersal distances of Florida scrub jays, suggests that even if a  
164 dispersal event occurred for individuals of our testing set, at the scale of Florida, our method  
165 is able to detect their birthplace with relatively high accuracy. This is particularly striking  
166 as only 41 SNPs were considered and those had not been pre-selected for the purpose of  
167 making assignment, not even for their ability to a priori reflect population structure.

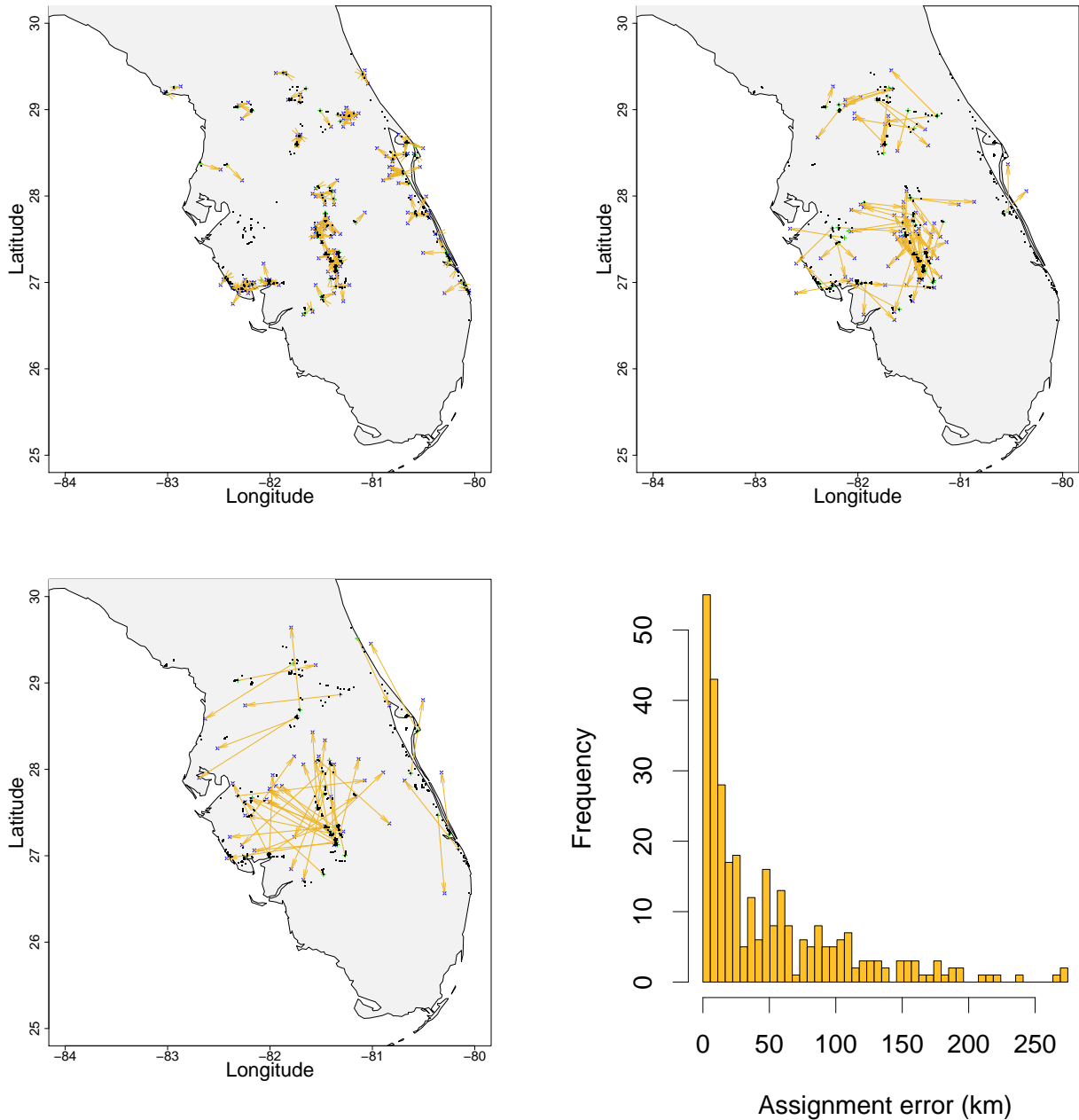
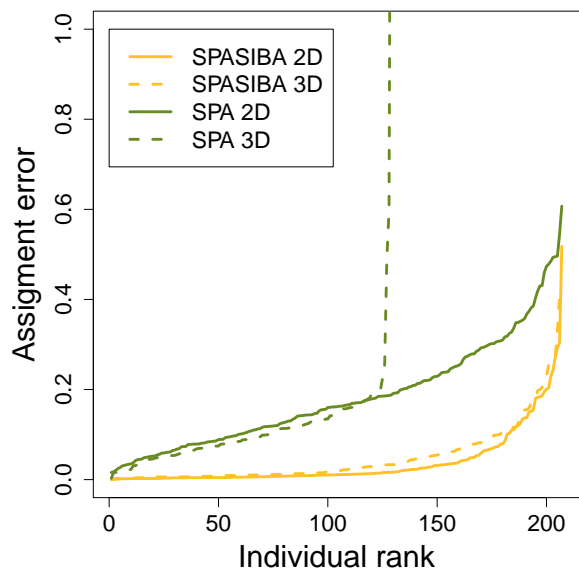


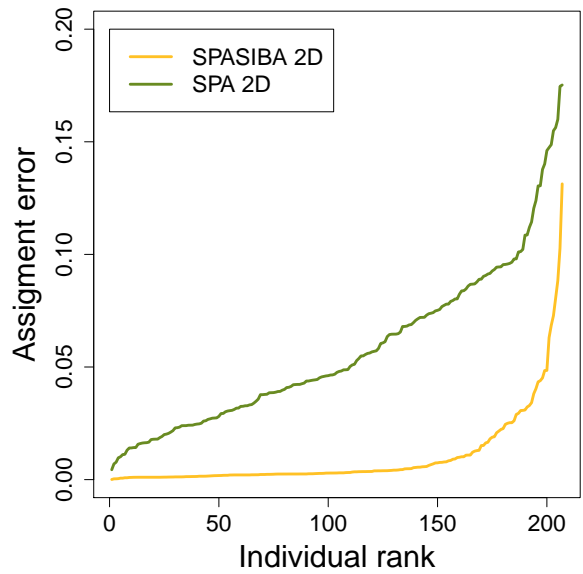
Figure III: SPASIBA geo-spatial assignments of Florida scrub jays with the SPASIBA method. Arrows originate from the true sampling site and point towards the estimated origin and provide a measure of assignment errors. They are displayed for different quantiles: Top left, 0-median; top right, median- $q_{0.75}$ ; bottom left,  $q_{0.75}-q_{0.9}$ . The full distribution of assignment errors is indicated for the 311 individuals of the testing set in the bottom right panel.

## 168 *Arabidopsis thaliana* in Europe

169 We further explore the performance of our method using a large genetic dataset of *Arabidopsis*  
170 *thaliana*, which represents an extensively studied model organism. We consider here a subset  
171 of the data from Horton et al. (2012), consisting of the 1,007 samples located in Eurasia  
172 with longitude between 20°W and 100°E. We perform assignment on random training sets  
173 of eight hundreds specimens at random subsets of  $L = 100$  then  $L = 1,000$  loci. Geospatial  
174 assignment was performed in each case using the remaining 207 samples. Because the data  
175 are sampled at large scale, we investigate both the 2D and 3D versions of these programs. In  
176 many runs of SPA in the 3D option, the output was non-sensical, showing samples assigned  
177 to geographic regions located at the antipodes of the sampling area. We therefore limited  
178 our exploration of the 3D option to  $L = 100$ . We found that SPASIBA was more accurate  
179 than SPA for all cases considered, and predicted the geographic position of a large number of  
180 specimens to be extremely close to their known positions (Fig. IV). More specifically, three  
181 quarters of the samples were assigned within 375 kilometers of their exact geographic origin,  
182 when using 100 loci and within 93 kilometers when using 1,000 loci.



$L=100$  loci



$L=1,000$  loci

Figure IV: Assignment errors estimated using datasets of 100 SNPs and 1,000 SNPs on *A. thaliana* data. Eight hundreds specimens were used as a training dataset and geospatial assignment was performed using the remaining 207 samples. Assignment errors are indicated in increasing order. On the vertical axis, the assignment error is expressed as a fraction of the distance between two most remote points of the geographical sampling window (7,500 km).

## 183 Geographic assignment of Europeans

184 Lastly, we explore the performance of our method in a case where extensive genetic infor-  
185 mation is available for a large number of individuals. More specifically, we consider here  
186 the subset of the Population Reference Sample (POPRES Nelson et al., 2008), used by  
187 Novembre et al. (2008) which consists of 1,385 individuals with grandparents of similar an-  
188 cestry. We use genotypes at 197,146 loci (after pruning tightly linked loci). In this dataset,  
189 the exact geographic origin of individuals is unknown and each individual is conventionally  
190 geo-referenced to the centre of its reported country of origin (except for a few countries for  
191 which another location was considered more reflective of the origins of these individuals).  
192 This implies that the uncertainty in the known geographic origin of samples varies with the  
193 size of the country of origin, ranging from around 80 km in Macedonia up to thousands of  
194 kilometres in Russia.

195 To assess the accuracy of methods on this dataset, we proceeded in two different ways to  
196 compute predicted maps of allele frequencies. In a first assessment, we used the whole dataset  
197 to compute these maps and estimated origins of each individual using these maps. This is  
198 likely to produce unrealistically low estimates of assignment errors. Therefore, to assess the  
199 accuracy of the two methods in a more realistic setting, and following a strategy taken by  
200 Wasser et al. (2004), we removed all individuals of a country at a time from the dataset, then  
201 computed predicted maps of allele frequency with a training set of geo-referenced genotypes  
202 consisting of individuals from all other countries only (which we refer below to as 'leave-one-  
203 population-out') and estimated origins of remaining individuals from these maps. The detail  
204 of estimated origins is displayed in figure V.

205 In the approach using the whole dataset to obtain allele frequencies maps, the median dis-  
206 tance of the estimated origins to the centre of the country is 72.8 km for SPASIBA (187 km  
207 for SPA) and the bias (defined as the mean distance of the per-country average estimated lo-  
208 cation to the country center) is 7.9 km for SPASIBA (21.8 km for SPA). Therefore, under this  
209 validation scheme, both methods show great accuracy, albeit SPASIBA consistently shows  
210 slightly better performance than SPA. Under the leave-one-population-out strategy, these  
211 statistics are respectively 696 km and 45.8 km for SPASIBA (543 km and 75km for SPA).

212 This suggests that the accuracy of both methods is extremely reduced when the training  
213 dataset does not include a population from the same genetic background as the test indi-  
214 viduals. Importantly, while SPA appears to perform better than SPASIBA in this setting,  
215 the assignment errors of SPASIBA appear to be homogeneously distributed geographically  
216 in contrast to those of SPA, which all appear to converge to the center of the study domain.

## 217 **Miscellaneous remarks**

218 The statistical model underlying our method is largely reminiscent of the SCAT program  
219 (Wasser et al., 2004, 2007). However, taking advantage of INLA instead of MCMC allowed  
220 us to significantly reduce computing times typically by several orders of magnitudes. Addi-  
221 tionally, our approach is free from MCMC convergence issues that can increase considerably  
222 the computation burden. In the Florida Scrub-jay dataset (1,311 individuals, 41 SNPs),  
223 SPASIBA achieved a full analysis in about ten minutes using a single 3GHz-CPU. SCAT  
224 required about a week of computation, while SPA provided results within a few seconds.  
225 These computing times scale linearly with the number of loci. With such running times  
226 and the accuracy levels demonstrated above, SPASIBA appears well tailored for the routine  
227 analysis of SNP datasets for non-model species consisting of a few tens of thousands of loci.  
228 In particular, it appears to be an ideal method for the analysis of reduced-representation  
229 sequencing data that become increasingly available in ecology. However, for a larger number  
230 of loci, SPASIBA is best carried on a computer cluster where the predictive maps of allele  
231 frequencies can be computed in parallel. Implementing this strategy on the POPRES data  
232 on a 80-CPU cluster, allowed us to carry out the analysis in 24-48 hours.

233 The algorithm underlying SPA and SPASIBA are essentially deterministic, while SCAT is  
234 stochastic. Defining a computing time for an MCMC-based like SCAT is impossible as com-  
235 putations are usually carried out over a number of iterations, larger than what is assumed  
236 to be necessary, and it is checked a posteriori and over several independent runs that the  
237 MCMC algorithm did not experience any convergence issue.

238 In SPA, all computations are locus-specific, therefore the computing time scales linearly  
239 with the number of loci. In SPASIBA, the computing time for the inference of the parameters



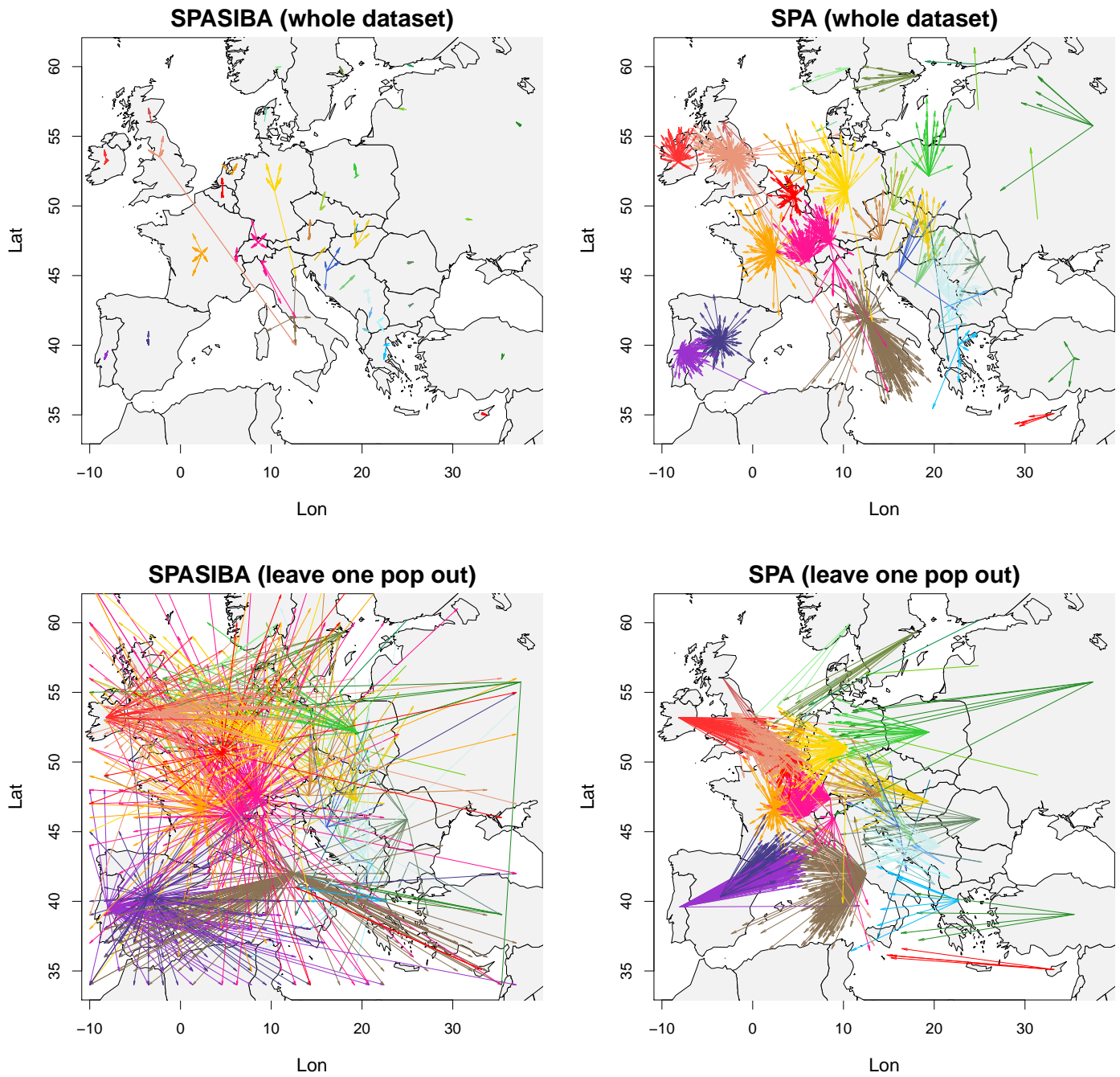


Figure V: Predicted geographic origins of Europeans. We used the POPRES data and evaluated the assignment error of SPA and SPASIBA using the whole dataset approach (top panels, using the whole dataset), or a leave-on-population-out approach (bottom panels, leave-one-pop-out).

240 of the random field scales non-linearly with the size of the data matrix (whose dimension is  
241 given by the product of the number of geographic sampling sites and the number of loci).  
242 The task of computing predicted allele frequency maps scales linearly with the number of  
243 loci.

244 In the tasks above, deterministic algorithms seek to optimize one criterion until a condi-  
245 tion is fulfilled. For the reasons described above, we are reluctant to provide exact computing  
246 times for the various methods discussed here. However, in our computations we observed  
247 that computations with SPA are in the order of hundred times faster than those with SPA-  
248 SIBA, which are themselves in the order of hundred times faster than those with SCAT. We  
249 note however that SCAT is the only program that handles micro-satellite data.

## 250 **Limitations of the SPASIBA method**

251 A potential advantage of SCAT over our SPASIBA method is the computer implementation  
252 that allows SCAT to restrict geographic assignments to a set of polygonal areas. Imple-  
253 menting this feature in SPASIBA would be straightforward and could increase accuracy  
254 in assignment when the spatial sampling window includes areas known to be non-suitable  
255 habitats. We note however that in the Florida scrub jay case, SPASIBA assigned only a  
256 handful of individuals a few kilometers away from the landmass (Fig. III), even though the  
257 assignment was not restricted to any specific area of the rectangular domain encompassing  
258 Florida.

## 259 **Lesser accuracy of the SPA method**

260 The SPA method is based on the assumption that allele frequencies vary logistically on the  
261 plan or the sphere, displaying essentially a nearly linear behavior in a central region and no  
262 variation elsewhere with frequencies fixed to 0 or 1. This may be a reasonable approximation  
263 for the data used earlier to assess the SPA method, namely human data in Europe and at  
264 the synoptic scale. At smaller scales, spatial patterns of genetic variation also likely reflect  
265 the processes of local genetic drift, migration and relatedness, which presumably features  
266 more spatial complexity. Additionally, the logistic model underlying SPA has the property  
267 of being invariant under shifts orthogonal to the main axis of variation. We believe that

268 a combination of these factors explain the lesser accuracy observed for SPA and also its  
269 propensity to numerical instabilities, as observed here with the *Arabidopsis thaliana* dataset  
270 (especially under the 3D option), the Florida scrub jay dataset and MS simulations.

## 271 **Limitations of current continuous assignment methods**

272 The interpolation of alleles frequencies between reference populations assumes a model of  
273 isolation-by-distance, however in reality, many biological populations display restricted gene  
274 flow due to a range of barriers that disrupt this relationship. These includes habitat variation  
275 and physical dispersal barriers (Wang and Bradburd, 2014). This is not handled by any of  
276 the continuous assignment methods and may affect the accuracy obtained.

277 Related to the point above, current continuous assignment methods assume marker neu-  
278 trality. While this is likely to be true for smaller microsatellite and SNP panels selected  
279 at random, genome-wide SNP panels, such as those produced by whole-genome or reduced-  
280 representation sequencing are likely to include loci under selection where the change in allele  
281 frequency may be completely disconnected from geographic distance. A recent study by  
282 Nielsen et al. (2012) suggests that such loci are highly informative for geographic assign-  
283 ment. However, the latter study is not based on an isolation-by-distance model and how  
284 the information gained from the use of highly informative loci will be offset by the use of a  
285 model that does not fit these loci, has still to be assessed.

## 286 **Re-appraisal of assignment results on the POPRES dataset**

287 The POPRES population reference sample has become an invaluable resource in many areas  
288 of human genetics, including pharmacogenetics and population genetics (Nelson et al., 2008).  
289 Here, we were able to bring the assignment error down to 72.8km but we caution that  
290 this figure only represents a lower bound for assignment errors. We note, however, that  
291 removing all individuals from a country from the training data (the leave-one-population-  
292 out approach) resulted in substantially larger assignment errors (696 km and 543 km for  
293 SPASIBA and SPA, respectively). Additionally, SPASIBA was characterized by relatively  
294 isotropic errors while SPA systematically biased predicted geo-spatial assignments towards  
295 the centre of the study area. Our leave-one-population-out approach revealed that none

296 of the two methods is robust to uneven population sampling in the training dataset and  
297 are particularly inefficient at estimating the country of origin of an individual whose true  
298 country of origin is not represented in the training dataset. It opens avenues for novel  
299 statistical approaches reducing the impact of uneven training sets on spatial assignments.

# References

- 300
- 301 Y. Baran, I. Quintela, Á. Carracedo, B. Pasaniuc, and E. Halperin. Enhanced localization of genetic samples through linkage-  
302 disequilibrium correction. *The American Journal of Human Genetics*, 92(6):882–894, 2013.
- 303 G.S. Bradburd, P.L. Ralph, and G.M. Coop. Disentangling the effects of geographic and ecological isolation on genetic differentiation.  
304 *Evolution*, 67(11):3258–3273, 2013.
- 305 A. Coulon, J.W. Fitzpatrick, R. Bowman, and I. J. Lovette. Effects of habitat fragmentation on effective dispersal of Florida Scrub-Jays.  
306 *Conservation Biology*, 24(4):1080–1088, 2010.
- 307 Petros Drineas, Jamey Lewis, and Peristera Paschou. Inferring geographic coordinates of origin for Europeans using small panels of  
308 ancestry informative markers. *PLoS One*, 5(8):e11892, 2010.
- 309 J. W. Fitzpatrick, G. E. Woolfenden, and Bowman R. Dispersal distance and its demographic consequences in the florida scrub-jay. In  
310 N. J. Adams and R. H. Slotow, editors, *22nd international ornithological congress*, pages 2465–2479, Johannesburg., 1999. BirdLife  
311 South Africa.
- 312 A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes, editors. *Handbook of Spatial Statistics*. Handbooks of Modern Statistical  
313 Methods. Chapman & Hall/CRC, Boca Raton, 2010.
- 314 G. Guillot and L. Orlando. *Oxford Bibliographies in Evolutionary Biology*, chapter Population Structure. Oxford University Press,  
315 New York, 2015.
- 316 G. Guillot, R. Leblois, A. Coulon, and A. Frantz. Statistical methods in spatial genetics. *Molecular Ecology*, 18:4734–4756, 2009.
- 317 G. Guillot, R. Vitalis, A. le Rouzic, and M. Gautier. Detection of correlation between genotypes and environmental variables. A fast  
318 computational approach for genomewide studies. *Spatial Statistics*, 8:145–155, 2013.
- 319 M. W. Horton, A. M. Hancock, Y. S. Huang, C. Toomajian, S. Atwell, A. Auton, N. W. Mulyati, A. Platt, F. G. Sperone, B. J.  
320 Vilhjálmsson, et al. Genome-wide patterns of genetic variation in worldwide arabidopsis thaliana accessions from the regmap panel.  
321 *Nature Genetics*, 44(2):212–216, 2012.
- 322 R.R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.
- 323 F. Lindgren, H. Rue, and E. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic  
324 partial differential equation approach. *Journal of the Royal Statistical Society, series B*, 73(4):423–498, 2011.
- 325 T. G. Martins, D. Simpson, F. Lindgren, and H. Rue. Bayesian computing with INLA : New features. *Computational Statistics and  
326 Data Analysis*, 67:68–83, 2013.
- 327 M.R. Nelson, K. Bryc, K.S. King, A. Indap, A. R. Boyko, J. Novembre, L.P. Briley, Y. Maruyama, D.M. Waterworth, G. Waeber,  
328 et al. The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research. *The  
329 American Journal of Human Genetics*, 83(3):347–358, 2008.
- 330 E.E. Nielsen, A. Cariani, E. Mac Aoidh, G. E. Maes, I. Milano, R. Ogden, M. Taylor, J. Hemmer-Hansen, M. Babbucci, L. Bargelloni,  
331 et al. Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. *Nature Communications*, 3:851,  
332 2012.
- 333 J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A.R. Boyko, A. Indap A. Auton, K.S. King, S. Bergman, M.R. Nelson, M. Stephens,  
334 and C.D. Bustamante. Genes mirror geography within Europe. *Nature*, 456:98–101, 2008.
- 335 A. Piry, S. Alapetite, J.M. Cornuet, D. Paetkau, L. Baudoin, and A. Estoup. GeneClass2: A software for genetic assignment and  
336 first-generation migrant detection. *Journal of Heredity*, 95(6):536–539, 2004.
- 337 E. Porcu, J.M. Montero, and M. Schlather, editors. *Advances and Challenges in Space-time Modelling of Natural Events*. Springer,  
338 Heidelberg Dordrecht London New York, 2010.
- 339 J.K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959,  
340 2000.
- 341 J.M. Rañola, D.H. Alexander, and K. Lange. Fast spatial ancestry via flexible allele frequency surfaces. *Bioinformatics*, 2014. URL  
342 doi:10.1093/bioinformatics/btu418.
- 343 H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace  
344 approximations. *Journal of the Royal Statistical Society, series B*, 71(2):1–35, 2009.
- 345 H. Rue, S. Martino, F. Lindgren, D. Simpson, A. Riebler, and E. Krainski. *INLA: Functions which allow to perform full Bayesian  
346 analysis of latent Gaussian models using Integrated Nested Laplace Approximation*, 2014. <http://www.r-inla.org/>.

- 347 D. Simpson, F. Lindgren, and H. Rue. Think continuous : Markovian gaussian models in spatial statistic. *Spatial Statistics*, 1:16–29,  
348 2012.
- 349 S. H. Sørbye and H. Rue. Scaling intrinsic gaussian markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51, 2014.
- 350 I J Wang and G S Bradburd. Isolation by environment. *Molecular ecology*, 23(23):5649–5662, 2014.
- 351 S.K. Wasser, A.M. Shedlock, K. Comstock, E.A. Ostrander, B. Mutayoba, and M. Stephens. Assigning African elephants DNA to  
352 geographic region of origin: applications to the ivory trade. *Proceedings of the National Academy of Sciences*, 101(41):14847–  
353 14852, 2004.
- 354 S.K. Wasser, C. Mailand, R. Booth, B. Mutayoba, E. Kisamo, and M. Stephens. Using DNA to track the origin of the largest ivory  
355 seizure since the 1989 trade ban. *Proceedings of the National Academy of Sciences*, 104(10):4228–4233, 2007.
- 356 G. E. Woolfenden and J. W. Fitzpatrick. *The Florida Scrub Jay-demography of a cooperative-breeding bird*. Princeton University  
357 Press, 1984.
- 358 G. E. Woolfenden and J. W. Fitzpatrick. *Birds of North America*, chapter Florida Scrub-Jay (*Aphelocoma coerulescens*). The Academy  
359 of Natural Sciences, Washington, D.C., and The American Ornithologists' Union, Philadelphia, Pennsylvania, 1996.
- 360 W.Y Yang, J. Novembre, E. Eskin, and E. Halperin. A model-based approach for analysis of spatial structure in genetic data. *Nature*  
361 *Genetics*, 44(6):725–731, 2012.
- 362 W.Y. Yang, A. Platt, C. W.K Chiang, E. Eskin, J. Novembre, and B. Pasaniuc. Spatial localization of recent ancestors for admixed  
363 individuals. *Genes, Genomes, Genetics*, 2014. doi:10.1534/g3.114.014274.