Fast spatial ancestry via flexible allele frequency surfaces

John Michael Rañola\textsuperscript{1,*}, John Novembre\textsuperscript{2} and Kenneth Lange\textsuperscript{3,*}

\textsuperscript{1}Department of Statistics, University of Washington, Seattle, WA 98195, \textsuperscript{2}Department of Human Genetics, University of Chicago, Chicago, IL 60637 and \textsuperscript{3}Department of Biomathematics, Human Genetics, and Statistics, University of California Los Angeles, Los Angeles, CA 90095, USA

ABSTRACT

Motivation: Unique modeling and computational challenges arise in locating the geographic origin of individuals based on their genetic backgrounds. Single-nucleotide polymorphisms (SNPs) vary widely in informativeness, allele frequencies change non-linearly with geography and reliable localization requires evidence to be integrated across a multitude of SNPs. These problems become even more acute for individuals of mixed ancestry. It is hardly surprising that matching genetic models to computational constraints has limited the development of methods for estimating geographic origins. We attack these related problems by borrowing ideas from image processing and optimization theory. Our proposed model divides the region of interest into pixels and operates SNP by SNP. We estimate allele frequencies across the landscape by maximizing a product of binomial likelihoods penalized by nearest neighbor interactions. Penalization smooths allele frequency estimates and promotes estimation at pixels with no data. Maximization is accomplished by a miniorize–maximize (MM) algorithm. Once allele frequency surfaces are available, one can apply Bayes' rule to compute the posterior probability that each pixel is the pixel of origin of a given person. Placement of admixed individuals on the landscape is more complicated and requires estimation of the fractional contribution of each pixel to a person's genome. This estimation problem also succumbs to a penalized MM algorithm.

Results: We applied the model to the Population Reference Sample (POPRES) data. The model gives better localization for both unmixed and admixed individuals than existing methods despite using just a small fraction of the available SNPs. Computing times are comparable to within a few hundred kilometers. Though this level of resolution is impressive, it is natural to wonder if a model-based method for spatial assignment could perform better and whether inferences could be reliably made for admixed individuals. This prompted Yang et al. (2012) to introduce spatial structure analysis (SPA), which, in fact, produces more accurate spatial assignments than PCA. In estimating allele frequency surfaces for each surveyed single-nucleotide polymorphism (SNP), SPA depends on a simple gradient function describing how allele frequencies vary with location. In practice, allele frequency surfaces can be bumpy without a dominant cline. The current article relaxes this restriction and gives more accurate reconstructions.

2 APPROACH

Our software, OriGen, adapts techniques from image reconstruction that encourages smoothness without requiring rigidly parameterized allele frequency surfaces (Chan and Shen, 2005; Lange, 1990). OriGen is model based and fast. It can infer the geographic origin of Europeans in the POPRES dataset to much less than 100 km. Its impressive speed is achieved by focusing on the most informative markers, sometimes as few as 1% of all markers, and relying on new minorization-maximization (MM) algorithms for parameter estimation. In choosing ancestry informative markers, we replace the information criterion of Rosenberg et al. (2003) by a homogeneity likelihood ratio test (LRT) that accommodates substantial differences in sample sizes.

*To whom correspondence should be addressed.
3 MATERIALS AND METHODS

3.1 A LRT criterion for SNP selection

The majority of SNPs are uninformative for ancestry and geographic localization. This fact and considerations of computational speed suggest choosing the most informative SNPs and ignoring the rest. The standard ancestry informativeness criterion of Rosenberg et al. (2003) makes the implicit assumption of equal sample sizes. The failure of this assumption in the POPRES data prompted us to turn to a homogeneity LRT. The null model of the test for a given SNP postulates that all individuals come from a single population with a unique allele frequency for the reference allele; the alternative model postulates different reference allele frequencies at the different sampling sites. Binomial sampling is in force. Suppose there are s sites with \( k_i \) sampled reference alleles and \( n_i \) sampled genes (reference alleles plus alternative alleles) at site \( i \). If \( k = \sum_{i=1}^{s} k_i \) and \( n = \sum_{i} n_i \), then the LRT statistic reduces to

\[
LRT = 2 \ln \prod_{i=1}^{s} \left( \frac{n_i}{k_i} \right) \frac{p_i^k (1 - p_i)^{n-k}}{\max_{q} q^k (1 - q)^{n-k}}
\]

where \( q = k/n \) and \( \hat{p}_i = k_i/n_i \) are the maximum likelihood estimates of the reference allele frequencies under the null and alternative models, respectively. Although small sample sizes at many sites invalidate the chi-square distribution of the LRT statistic, nothing prevents the statistic from being used as an index to rank the various SNPs. In our experience, the highest ranking SNPs are indeed the most informative.

3.2 Allele frequency surface estimation

To estimate the allele frequency surface for a given SNP, we divide the region of interest, say Europe, into pixels and assign a reference allele frequency \( p_i \) to each pixel \( i \). Extending our previous notation, \( k_i \) now represents the number of sampled reference alleles and \( n_i \) the number of sampled genes from pixel \( i \). For most pixels, \( k_i = n_i = 0 \). Maximizing the binomial loglikelihood

\[
L(p) = \sum \left[ \ln \left( \frac{n_i}{k_i} \right) + k_i \ln p_i + (n_i - k_i) \ln (1 - p_i) \right]
\]

would allow estimation of the reference allele frequencies if all pixels actually contained sampled people. Because this is not the case and because we desire smooth estimates across the landscape, we subtract squared difference penalties from the loglikelihood and maximize the penalized loglikelihood

\[
f(p) = L(p) - \rho \sum_{i,j} w_{ij}(p_i - p_j)^2
\]

where \( \rho = \sum_{i,j} w_{ij} \), and the inequality

\[
\rho = \sum_{i,j} w_{ij}(p_i - p_j)^2 \leq \sum_{i,j} w_{ij}(p_i) - 2p_i \rho \ln p_i + \frac{1}{\rho} \sum_{i,j} w_{ij} \ln p_i
\]

yields an equivalent cubic polynomial equation

\[
0 = a_i p_i^3 - a_i p_i^2 - (n_i + b_i p_i + k_i + b_i)
\]

where \( a_i = 2\rho \sum_{j \neq i} w_{ij} \), and \( b_i = 2p_i \rho \sum_{j \neq i} w_{ij} / p_i \). This cubic is positive when \( p_i = 0 \) and non-positive when \( p_i = 1 \). The cubic also tends to \( \pm\infty \) when \( p_i \rightarrow 0 \), hence there exists a single root on the interval \([0, 1]\). One can extract this root by one of the standard formulas for solving a cubic equation.

In practice, we add a small increment \( \eta \), say 0.1, to each sampled \( k_i \) and \( n_i \) to each sampled \( n_i \). These pseudocounts, which are similar to Laplace pseudocounts, stabilize estimation and prevent allele frequencies from converging to 0. Plotting on a log-scale, pseudocounts are mandatory. One can view pseudocounts as imposing a weak beta prior.

3.3 Localization of unknowns

Once the allele frequency surfaces for the informative SNPs are estimated by the MM algorithm, one can localize individuals of unknown origin. For person \( j \) with genotype vector \( x_j \), Bayes’ rule gives the posterior probability

\[
Pr(j \mid x_j) = Pr(x_j \mid j) \cdot Pr(j \mid \text{from pixel } i) / Pr(x_j)
\]

that \( j \) originates from pixel \( i \). Application of this rule depends on fixing a prior. Two possibilities are convenient. The simpler one is the uniform prior. A more accurate but less convenient choice is to scale the prior of a pixel by its population size. For sufficiently informative genetic data, the evidence dominates the prior, and the uniform prior is probably adequate. The likelihood term

\[
Pr(x_j \mid j \text{ from pixel } i) = \prod_k Pr(x_{jk} \mid j \text{ from pixel } i)
\]
factors into a product of likelihoods at the canvassed SNPs under the assumption of linkage and Hardy–Weinberg equilibrium. The likelihood \( \Pr(\mathbf{x}_h | \mathbf{x}_i) \) from pixel \( i \) at SNP \( k \) equals one of the three genotype probabilities \( p_{id}^2, 2p_{id}(1-p_{id}) \) or \((1-p_{id})^2\) depending on \( j \)-type genotype at SNP \( k \). In practice, it is advisable to work with the logarithms of these quantities to avoid computer underflows. Although the pixel with the highest posterior probability provides the most likely localization, it is a good idea in practice to assign an average latitude and longitude and highlight the set of pixels on the domain \([0, \delta]\) and by the constant \( \delta \) on the domain \([-\delta, \infty)\). We previously used a variant of the current admixture model and penalty to estimate haplotype frequencies. Rather than repeat the mathematical derivation of the same penalized MM algorithm here, we refer the reader to the reference (Ayers and Lange, 2008) for details. Suffice it to say that with parameters separated, the MM updates require solving a simple quadratic equation for each component \( f_c \). Generic extrapolation techniques for MM and similar algorithms permit convergence acceleration beyond that afforded by our specific penalization (Zhou et al., 2011).

Because highly admixed individuals are nearly impossible to characterize fully, we choose \( \delta = \frac{\pi}{4} \). Beyond this value no further penalty is exerted to eliminate pixels with little evidence of admixture. The total strength \( \lambda \) of the penalty is chosen to minimize the expected geodesic distance

\[
e(\delta) = \sum_k \sum f_k d_k^2
\]

between the true and estimated centers of a person’s admixture distribution over many simulated admixed people. Here, \( d_k \) is the geodesic distance between the centers of pixels \( i \) and \( j \). Although somewhat ad hoc, these choices of \( \delta \) and \( \lambda \) perform well in practice.

### 4 RESULTS

#### 4.1 A LRT criterion for SNP selection

The utility of SNPs in identifying ancestral origins varies widely. The Rosenberg et al. (2003) criterion for ranking SNPs implicitly assumes equal sample sizes at the different sampling sites. In practice this assumption is usually violated. As an alternative, we turned to a LRT statistic for testing homogeneity of allele frequencies across sites. The LRT statistic compares the best loglikelihood of the data under the null hypothesis of homogeneity to the best loglikelihood of the data under the alternative hypothesis of complete heterogeneity. Figure 1 allows us to compare the value of the two different methods of ranking SNPs. The vertical axis of the figure represents the average distance under cross-validation between the true location of the POPRES individuals and their estimated locations under OriGen. The horizontal axis represents the number of SNPs used, with SNPs taken in their order of informativeness. Although the three curves document the value of ancestry informative SNPs in geographical projection, it is obvious that the LRT criterion performs better than the information criterion.

#### 4.2 Allele frequency surfaces

Accurate allele frequency surfaces are the primary reason for OriGen’s superior performance. OriGen surfaces are more adaptable and less rigidly parameterized. Figure 2 depicts the estimated allele frequency surfaces of the six most informative SNPs of the POPRES data. The figure also plots the maximum likelihood estimates for each sampled site as a filled-in circle at the appropriate location. For comparison, a figure in the Supplementary Material depicts the surfaces for the same SNPs generated by SPA. The figures demonstrate that OriGen surfaces match the sampled allele frequencies (represented by the shading of the circles at each sample location) better than the SPA surfaces. SPA appears to be too heavily influenced by outlier sites and less adaptable overall.
4.3 Ancestral origin inference

Spatial assignment is the main application of OriGen. To showcase OriGen’s accuracy, we computed average localization error by leave-one-out cross-validation. Figure 3 displays the results for OriGen versus SPA. The lower curve for SPA emphasizes the benefits of exploiting LRT ordered SNPs. Examination of the figure shows that OriGen using 1% of the SNPs achieves better accuracy than SPA using all of the SNPs. Using 5% of the SNPs, OriGen is nearly perfect at the pixel level in its localizations. The same point can be made by comparing OriGen’s results to the results in Table 1 of the SPA paper (Yang et al., 2012). Given the nature of the table, a fair comparison requires using OriGen to estimate the optimal ancestral origin of each person and then assigning the person to the closest sampling site as measured by geodesic distance. Overall, OriGen was more than twice as accurate as PCA and SPA based on just 1% of the data. With 5% of the data, OriGen maps individuals to sampled pixels with 99% accuracy. In Figure 4, we show the localization results of OriGen obtained from cross-validation using only 2% of the SNPs. With this small amount of data, it can already be seen that OriGen does well in placing individuals at their true origin.

We also performed some small-scale comparisons with SCAT (Wasser et al., 2004). SCAT is computationally demanding and so more ambitious comparisons were impossible to implement.

Fig. 1. Average distance between the geographic origin of the POPRES individuals and their OriGen estimated origins as a function of the number of SNPs used. The figure reflects leave-one-out cross-validation. The solid curve relies on the Rosenberg et al. (2003) information content, the dashed curve relies on no ordering and the dotted curve relies on the LRT criterion.

Fig. 2. Allele frequency surfaces generated by OriGen with tuning parameter $\rho = 0.1$ for the six most informative SNPs. These surfaces are overlaid with filled-in circles to convey the MLE estimates for each sampled site. For the sake of comparison, the same SNP surfaces are depicted in the Supplementary Material for SPA.
Table 2 records the average distance to the true origin and the run times for the 100 most informative SNPs. OriGen and SCAT place individuals almost 2000 km closer to their true origin than SPA. SPA is the fastest (1 min) of the three programs, followed closely by OriGen (2 min) and distantly by SCAT (362 minutes). Thus, the current version of OriGen delivers good placement with competitive execution times. The current formulation of SCAT is unable to handle large numbers of SNPs.

4.4 Estimating proportions of admixed origins

Many individuals have mixed ancestry. PCA tends to localize individuals with parents of different ethnicities in between their parents’ regions of origin. SPA has the capacity to localize each parent separately, but the user must inform the program beforehand how many different ancestries contribute to a given individual. Because this information is often unavailable, it would be preferable for admixture detection and origin selection to be more agnostic. OriGen can estimate admixture fractions on a pixel-by-pixel basis. For example, when applied to a person with a German parent and an Italian parent, ideally OriGen should deliver 50% German ancestry, 50% Italian ancestry and 0% other ancestry. This would make OriGen comparable with the program ADMIXTURE (Alexander et al., 2009), with the benefit of using more accurate allele frequencies and covering small countries with no sampled people at all.

In admixture mode, OriGen exploits the same allele frequency surfaces that it does in normal mode. However, instead of applying Bayes’ rule to find the posterior probability of origin of each pixel, it estimates an admixture fraction for each pixel by penalized maximum likelihood estimation. OriGen is not only able to select the two contributing populations, it is also able to estimate their proportions well. As Figure 5 illustrates, OriGen takes admixture estimation a step further by estimating the fractions at each pixel instead of each population. OriGen allows one to place individuals at locations with no sampled data. In the figure, the true locations of the individual’s grandparents are highlighted, while OriGen’s results are written as text.

Table 1. Comparison of localization by population

<table>
<thead>
<tr>
<th>Geographic origin</th>
<th>Number of individuals</th>
<th>Accuracy</th>
<th>OriGen</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PCA</td>
<td>SPA 1% of data</td>
</tr>
<tr>
<td>Italy</td>
<td>219</td>
<td>0.70 ± 0.03</td>
<td>0.74 ± 0.03</td>
</tr>
<tr>
<td>UK</td>
<td>200</td>
<td>0.44 ± 0.04</td>
<td>0.53 ± 0.04</td>
</tr>
<tr>
<td>Spain</td>
<td>136</td>
<td>0.71 ± 0.04</td>
<td>0.69 ± 0.04</td>
</tr>
<tr>
<td>Portugal</td>
<td>128</td>
<td>0.20 ± 0.04</td>
<td>0.38 ± 0.04</td>
</tr>
<tr>
<td>Switzerland-French</td>
<td>125</td>
<td>0.26 ± 0.04</td>
<td>0.33 ± 0.04</td>
</tr>
<tr>
<td>France</td>
<td>89</td>
<td>0.70 ± 0.05</td>
<td>0.66 ± 0.05</td>
</tr>
<tr>
<td>Switzerland-German</td>
<td>84</td>
<td>0.23 ± 0.05</td>
<td>0.27 ± 0.05</td>
</tr>
<tr>
<td>Germany</td>
<td>71</td>
<td>0.25 ± 0.05</td>
<td>0.28 ± 0.05</td>
</tr>
<tr>
<td>Ireland</td>
<td>61</td>
<td>0.28 ± 0.06</td>
<td>0.28 ± 0.06</td>
</tr>
<tr>
<td>Yugoslavia</td>
<td>44</td>
<td>0.25 ± 0.07</td>
<td>0.30 ± 0.07</td>
</tr>
<tr>
<td>Mean</td>
<td>115.7</td>
<td>0.40 ± 0.05</td>
<td>0.45 ± 0.05</td>
</tr>
</tbody>
</table>

Note: Population of origin was predicted for each individual using leave-one-out cross-validation. Accuracy ± SD is the proportion of individuals from each population correctly assigned to their true population. The values listed for OriGen represent either 1% of the data (2 K SNPs) or 5% of the data (10 K SNPs). To make the values from OriGen comparable with PCA and SPA, the most likely location of each individual was estimated, and the population closest in distance to that point was chosen as the population of origin. The results for PCA and SPA are taken from Table 1 of the paper (Yang et al., 2012).
their respective locations. The results presented in Figure 5 for admixed individuals are typical of many reconstructions.

5 DISCUSSION
Motivated by advances in image reconstruction, we have presented a probability model for the estimation of complex allele frequency surfaces. Our model captures not only linear clines, but also multiple local peaks on a landscape. Allele frequency estimates represent a compromise between locally sampled genotypes and smoothness. The degree of smoothness is determined empirically by cross validation. Spatial assignment exploits the allele frequency surfaces of the most informative SNPs. To no one’s surprise, the ancestry informative SNPs drive projection. In ranking SNPs our homogeneity LRT statistic outperforms the information criterion of Rosenberg et al. (2003), which assumes equal sample sizes at the sampled sites. Our computing times are vastly superior to SCAT and competitive with PCA and SPA.

We have also proposed a model for spatial assignment of admixed individuals. Our model assigns an admixture coefficient to each pixel. To avoid over-parameterization, we impose a penalty that enforces parsimony and focuses attention on those pixels with the greatest explanatory power. Estimation of both allele frequency surfaces and admixture coefficients benefits from the MM principle. The MM algorithms generated are simple to code and automatically enjoy the ascent property. Convergence can be slow, but standard extrapolation techniques accelerate convergence dramatically. On the negative side of the balance sheet, our software OriGen requires more storage per SNP than SPA, which characterizes an allele frequency surface by just three parameters. The accuracy of OriGen is also limited by the number of pixels. For the POPRES dataset, we enclosed Europe in a square with 70 pixels on a side. Smaller pixels make little discernible difference in resolution at the expense of considerably more computation. The heatmaps of posterior probabilities and admixture coefficients afforded by the pixels are a decided plus. The ability to exclude infeasible pixels over oceans is another advantage.

Modeling is an art. The best models combine realism with computational efficiency. The injection of ideas and techniques from image reconstruction is a major contribution of OriGen. Dividing regions into pixels and nearest neighbor interactions offer a logical framework for estimation. MM algorithms are also ubiquitous in imaging. Our admixture model is directly motivated by genetic considerations. It cleanly circumvents the need for specifying which ancestors of an admixed person should
be taken as geographically localized. Finally, our SNP selection criterion is probably better suited to identifying ancestry informative SNPs than abstract information criterion. Readers will doubtless think of many other ways of improving the current model. For example, a reviewer suggested that it might be useful to incorporate standard errors of allele frequency estimates into localization heatmaps. Our preliminary testing of this plausible idea finds no improvement, probably because localization averages across so many SNPs. Science, like product design, is usually an iterative process of successive refinement.

**Funding:** NIH grants from the National Human Genome Research Institute (HG006139, HG007089, T32 HG00035) and the National Institute of General Medical Sciences (GM053275).

**Conflict of interest:** none declared.

**REFERENCES**


---

**Fig. 5.** Admixture coefficients for four simulated Europeans with grandparents from locations highlighted in lighter colors. The numbers listed are the estimated admixture coefficients at their respective pixels based on 40 K SNPs; values <1% are omitted. In the top left is a simulated individual with four grandparents coming from the UK. On his right is an individual with two grandparents from Germany and two from Poland. On the bottom left is an admixed individual with two grandparents from Portugal and one each from France and Poland. Finally, on the bottom right is an individual whose four grandparents come from Spain, France, UK and Poland.