Re-Submission Cover Letter

To the Editor(s),

Please find enclosed the resubmission of our manuscript, entitled "A Spatial Framework for Understanding Population Structure and Admixture, for re-consideration for publication in PLoS Genetics. We apologize for the delay in resubmission; I was working to finish and submit my PhD thesis, which I have now successfully completed.

We appreciate the opportunity to resubmit a revised version of the manuscript. The comments provided by the Guest Editor and three anonymous reviewers were very helpful, and we have made substantial revisions to the manuscript in response to their thoughtful criticism. Our detailed responses to those comments are inset below (in bold), but overall, we were able to greatly increase the clarity of the writing, and also to better and more explicitly highlight the advantages of SpaceMix over existing methods for inferring and visualizing patterns of population genetic differentiation. We hope that the paper is now acceptable for publication.

Thank you for your consideration,
Gideon Bradburd (for all of the authors).

Guest Editor's summary

Overall this paper is good and the topic is of wide enough interest to be appropriate for PLoS Genetics. However, the reviewers, particularly 1 and 2, made several points about the paper that would need to be addressed before it could be accepted. The reviewers noted a number of deficiencies in the paper that will require major revision. The most important points of the reviewers are as follows.

Reviewer 1

- I agree that the paper would be more persuasive if the real improvements over PCA, Treemix and Globetrotter could be made clearer.

  We appreciate this point, and have worked to clarify the advantages SpaceMix has over currently available methods. Specifically, we have now added comparisons between the performance of PCA, TreeMix, and SpaceMix on simulated data, and demonstrated and discussed the utility and limits of SpaceMix in these simulated scenarios.

- Why not use a third dimension, in a way similar to using PC3?

  This is a good idea, and one we've debated implementing for a while. There is no statistical reason why we couldn't implement three (or more) dimensions; however, we felt that the visualization of a more-than-two-dimension geogenetic map would be difficult and non-intuitive, especially because of the admixture arrows. For now, we are choosing to stick with two dimensions.

- Don't worry about extending the method to within-population structure unless you want to.

  We have chosen not to extend this method to explicitly compare within- and among- population structure, but we are currently working on extensions to this statistical framework

that are designed to do that.

- The comments about multiple sources of admixture is germane, especially in light of David Reich's recent papers.

  As with the comment about having more than 2 dimensions, this is a good idea, and one that we've discussed extensively. We have chosen not to implement multiple sources of admixture for each population because we have concerns about both identifiability of the model and interpretability of the resulting figures.

Reviewer 2

- The point about possible effects of LD and uneven spacing of samples are important and will have to be addressed.

  We agree that these are important points, and we appreciate the reviewer having pointed them out. To address the issue of LD between loci, we have clarified our methods (thinning for LD in the human empirical dataset), and also added a paragraph of discussion on the ways in which LD between loci in other empirical datasets could be accommodated (Lines 149–154; 715–733).

  To address the issue of uneven sampling, we randomly subsampled two of the lattice simulations (30 population simple lattice with nearest-neighbor migration and the 30 population lattice with nearest-neighbor migration and the longitudinal barrier) down to 12 populations and repeated our SpaceMix runs. The SpaceMix results of these uneven sampling scenarios (referenced in the main text Lines 229–232, Figures S3a and S3b in the SuppMat) show that the method is robust to missing or irregularly sampled population data.

- The comments made about comparison with Treemix results and the potential use of additional dimensions echo Reviewer 1's comments. Is all the useful information really contained in a two-dimensional graph? In not, what else can you say?

  As above, we agree that this is an important question. All of the primary information from a SpaceMix analysis is contained in the two-dimensional "geogenetic" graph. There is additional information in the matrix of residuals between the parametric covariance matrix and the sample covariance matrix, which gives information about samples (or, more specifically, about the relationships between samples) that are well or poorly fit by the model. In an extension to the current framework, positive residual covariance could potentially be modeled post hoc with additional admixture arrows (following TreeMix). However, we feel that the figures currently produced by the SpaceMix admixture model are sufficiently information-rich, and that adding multiple admixture arrows may be, as noted above, too much.

- Reviewer 2 made several additional technical comments that need to be addressed. Is MCMC really necessary?

  This is a good point. In the current manuscript, we have shown figures using the sample from the MCMC with the highest posterior probability. We chose to only present these MAP estimates for clarity of the figures; however, in the Supplement, we show figures that take advantage of the full posterior distribution to describe both the 95% credible intervals of the population locations as well as the 95% credible intervals of admixture proportions and nuggets. We are currently exploring other inference algorithms that have lower computational costs (e.g. hill-climbing techniques to get a maximum-likelihood or MAP estimate) for use in future releases of this method. However, we believe that the full MCMC inference procedure has value for users that wish to accommodate uncertainty in inference. This point is especially germane be-

cause we anticipate great heterogeneity in the number of independent loci used to calculate the sample covariance across different empirical datasets, and therefore great disparity in the size of the credible intervals on the parameter estimates. We are currently working on functions for the publicly released package that allow for easy visualization of point estimates and the credible intervals in the geogenetic maps.

Reviewer 3

- Reviewer 3 suggests shortening and clarifying the discussion of the human data. The authors should consider the other comment about analyzing the POPRES data.

  We appreciate this feedback. We have significantly shortened, clarified, and de-anthropomorphized the discussion of the human Globetrotter dataset (see LatexDiff document for a comparison). We have considered running this method on other empirical datasets, such as the POPRES data, but we feel that the manuscript is, at this point, already quite long, and that the addition of another empirical example does not offer significant benefit.

Reviewer Comments

Reviewer 1

Bradburd et al. develop a novel method to produce "geogenetic maps" where populations are placed on the map at distances distorted to reflect rates of gene flow among the sampled populations. The authors show that in simulation their method (SpaceMix) produces results more easily interpreted and reflective of gene flow than does PCA, and they then apply SpaceMix to two datasets: one from a ring species (the greenish warbler) and one from a global sampling of human populations.

Major comments (in no particular order)

- Overall I think this method could be widely used in the field, but I think to achieve that the authors need to show more details in their applications of what (ideally, erroneous) conclusions PCA and TreeMix (or a similar tree-based method) would have on the same datasets they apply SpaceMix to. I agree in principle that PCA and tree-based methods can't fully capture admixture and/or are biased by sampling/scale of data, but I want the authors to show me how I would "go wrong" by using alternatives to SpaceMix to analyze population genetic data. Also, it wasn't totally clear to me how SpaceMix differs from Globetrotter (Hellenthal et al.), which is referenced quite late in this manuscript.

  This is a great point; we did not do many explicit comparisons between SpaceMix and other methods in the original submission. We have now included explicit comparisons between SpaceMix, PCA, and TreeMix on the same datasets, as well as a discussion of the comparisons. We did not include a comparison between the inference methods of SpaceMix and Globetrotter because they differ fundamentally in both their aims and in the aspects of the genetic data they model. Globetrotter models linkage disequilibrium due to admixture in a single focal population at a time (much like ROLLOFF and ALDER). It is, therefore, primarily designed to estimate the timing and sources of admixture of each population in turn, rather than to visualize population structure and admixture history of many populations, as SpaceMix, TreeMix, and PCA do.

- The theory underlying this paper is very nice; the fact that the covariance matrix from a population genetic dataset can be expected to be Wishart and that this allows efficient calculations of the likelihood of the data is neat. But Figure 2 was not surprising to me and I wonder why the authors limit themselves to inferring geographic locations in two dimensions - why not three and then show a series of biplots? With admixed populations, a 3rd PC is often useful in visualizing the nature of the admixture.

As discussed above, this is a good suggestion, and one that we discussed quite a bit when formulating the method. We have chosen to restrict ourselves to a 2-dimensional geogenetic map because of the ease of interpretability (all the information in the inference can be summarized in a single plot figure), and the potential difficulties in interpreting the admixture arrows if they are drawn from a greater-than-2-dimensional space. As an aside, because we are directly modeling admixture, a 3rd dimension will not usually be necessary for its visualization, as can be the case with PCA.

- I think the authors are missing an opportunity to use SpaceMix to study within-population variation. Since SpaceMix only needs 1 sample from each population, why not apply it to identify relatives or reveal structure within a population? A little bit of this happens in figure 5 with zooming in on Eurasian map but there is more that could be done; a simple application to the 1000 Genomes without filtering on relatives could produce a result with high impact.

  The suggestion to analyze the 1000 genomes data is a good one. However, we have demonstrated the ability of SpaceMix to describe within-population structure in applications to individual-level data with the warbler dataset already in the paper, and, because the paper is already quite long, we feel that the inclusion of yet another empirical dataset would be excessive. We are currently exploring extensions to our approach to directly model within- and between-population structure.

- The authors assume a single source of admixture - why not have supplementary figures that at least show what happens with 2 or 3? I think this will help users compare SpaceMix to competing methods like Globetrotter.

  As discussed above, this is a great idea, and one we batted around quite a bit while making the method. We have decided

at this point to only implement a single source of admixture because we think that the geogenetic maps with admixture are already very information rich, and would be difficult to interpret with the addition of a second set of admixture arrows for each sample. We may revisit this issue in future rounds of method development.

- Line 350 – were the data generated for this paper, or previously? If the latter, more info is given here than needed.

  Thank you for pointing this out. This section has been tightened up.

- Major comment about writing: There's an anthropomorphization of populations starting around line 282 of populations that should be fixed. Examples are: "the admixed population (population 23) choose admixture from very close." –don't the authors mean the method is assigning the admixture source from very close to the true source? Another example, Line 290 – "it [population 18] explains its intermediate genetic relationships by." Line 375 – "chooses to locate itself" Line 542 – "where the Uygur have positioned themselves." Line 668 – "choose to cluster close to each other."

  Thank you for pointing this out. We have de-anthropomorphized the writing now, and believe that it is the clearer for it.

Minor comments

- Figure captions don't really explain the axis labels Northings and Eastings.

  We now explain that the axes of the geogenetic maps are presented as Northings and Eastings because the geogenetic locations no longer correspond to the latitude and longitude of the original sampling locations.

- Line 25 - Li and Durbin 2011 doesn't provide gene flow information really, but MSMC does (are the authors thinking of when they use 1 sequence from one pop and another from another?)

  Li and Durbin (2011) did have some discussion of evidence for gene flow (between West African and non-African populations), but we agree that MSMC is a better example, and now cite Schiffels & Durbin (2014)

- Line 32 - Luca et al. 1994? Fix citation.

  Fixed.

- Intro (first page) is a little repetitive in its multiple mentions of PCA – way to tighten this up?

  Fixed.

- Line 116 – "historical or ongoing migration" – vague

  This distinction is an important one. Populations that exchanged migrants in the relatively recent past, but which no longer are connected by migration, may show similar patterns of differentiation in allele frequencies as populations that continue to exchange migrants at a lower rate through the sampling time.

- Line 130 – change line end from "Within population" to "Within-population".

  Fixed.

- Subsection heading on line 176 is confusing. How about just "Simulations"?

Fixed.

- Line 266 – are heavily weighted towards small values to be conservative with respect to admixture inference. Explain.

  As given in the Table referenced in that line of the manuscript, the prior on each admixture proportion is a β distribution with shape parameters 1 and 100, so that its mean is approximately 0.01. This is a conservative prior in the sense that it is weighted toward small admixture proportions. We hope that this has adequately addressed this concern; if not, we are happy to do more to respond to this point or clarify the language in the manuscript.

- Line 455 – what does "subset of HGDP samples" means? These papers mostly deal with full HGDP.

  We meant the subset of the total dataset that was in the HGDP dataset. The wording has been fixed for clarity.

- Line 495 – typo: "the of the North"

  Fixed.

- Line 505 – the nuggets' – why quotes here and nowhere else?

  Fixed.

- Line 845 – missing a comma to indicate that migration rate is constant across all neighbors (including diagonal)

  The sentence has been edited for clarity.

Reviewer 2

In the submitted article entitled "A Spatial Framework for Understanding Population Structure and Admixture", Bradburd, Ralph and Coop present a new statistical method to describe population structure. They introduce the concept of geogenetic maps that is a projection of the genetic samples in 2 dimensions using the genetic covariance between populations or individuals. The idea is interesting and has the advantage of taking into account the decay of genetic covariance due to geographical distance, and the possibility of modeling admixture events. In addition, the method is implemented in a program SpaceMix. Because it uses the genetic covariance between populations or individuals, it is not computationally costly. However I have general and statistical comments or concerns about the method.

- The covariance estimated is assumed to follow a Wishart distribution. This is the case if the allele frequencies are independent Gaussian vectors. However at the genome-wide level, Linkage Disequilibrium would completely distort the Wishart distribution, just like summing correlated squared Gaussian variables distorts a chi-square distribution. Especially when working on data with spatial structure, or potential admixed populations, where the level of LD is expected to be stronger. The authors mention l.92 "unlinked" and LD in a brief paragraph in the discussion. I suggest to emphasize this issue earlier in the article, as it would be important for SpaceMix users to know that. In addition, subsample of the SNPs is done in the empirical analysis, does it mean that r^2 measures between SNPs are very low? Can the authors detail?

  This is an excellent point. We have now written more detailed descriptions of the procedure we use for dealing with LD in our analyses of empirical data (Lines 369–370; 469–473), as well as possible method extensions to deal with LD (Lines 149–154; 715–733).

- The simulated examples are useful to understand geogenetic maps. However all the sampling schemes are regular grids. It would be of interest to see how robust these maps are to uneven sampling. This is the most common case in empirical analysis. Did the authors make sure that missing samples on the grid do not distort the resulting geo-

genetic maps?

This is a great suggestion. We have responded to this comment in detail above, but, briefly, we randomly subsampled two of our lattice simulations (with and without a barrier) and SpaceMix successfully recovers the configuration of the subset. The relevant figures are SI Figs S3a and S3b

- There are already tools for l.66 "visualizing patterns of population differentiation" that could be mentioned. To assess the power and the novelty of the method, the authors should compare the geogenetic maps to what is already done. There is no need to compare to all the previous methods, but at least with the very common Multi dimensional scaling (which would give the same projections as PCA, but philosophically closer to the method (1)). The only comparison made is for a 1D stepping stone model in Fig. S13, where the population structure needs only 1 dimension to be described. With comparisons the authors could assess the advantage of the statistical modeling of the Isolation by distance pattern, and the admixture. In the same spirit, the method mentions TreeMix that estimates admixture, but does not compare the estimated admixture proportion w to TreeMix results. This is unfortunate, because it could show how taking geography into account helps estimate admixture.

  This is a very fair critique. We have responded in detail to this point above, but, briefly, we now perform a more extensive comparison between SpaceMix and PCA, and we also perform a comparison between SpaceMix and TreeMix.

- My main concern is rather general than statistical. Although the authors claim it is a simple and intuitive way to visualize population structure, the interpretation does not always seem obvious. For example in the expansion scenario of Figure 1, the populations where expansion took place have closer geogenetic coordinates than populations with only migrations. But based on the coordinates only, it seems unlikely to know if the geogenetic pattern is due to expansion, higher

migration rates, strong edge effect, or other. Would it be possible to disentangle the sources of population structure without a priori knowledge? In the Human data analysis, It would be difficult to interpret the 2 independent expansions for Oceania and America with no prior knowledge. The fact that Native American populations are northmost population is not intuitive, when with PCA, they would be separated on PC3 or PC4. It feels like too much information is being summarized in 2 dimensions. When adding admixture, the many arrows make the visualization tricky (Figure 9). Wouldn't it be better to separate the G and G* on two different maps?

There are two important points raised here. The first is that it is frequently difficult to infer process from the pattern SpaceMix infers (e.g., geogenetic proximity between Oceania and eastern Eurasia could be do to a recent expansion from the latter to the former, or an older split with some low rate of ongoing migration. In this case, archaeological, (bio)geographic, and other genetic evidence helps disentangle these scenarios, but obviously not all systems will be so information-rich.). This is a fair criticism, but also one that could be applied to many other methods for inferring and visualizing patterns of relatedness, including TreeMix and PCA. We believe this method still has utility, and we are very explicit in our caveats about the limits of SpaceMix's inference and its interpretation (see Lines 623–638; 687–745).

The second important point raised is that it the geogenetic maps with admixture are too information rich for easy interpretation. We agree that this is the case with some of applications, especially in the dataset of human samples (this is one of the reasons we are hesitant to introduce multiple inferred sources of admixture). However, we also feel that such complexity will be not be present in all systems, and, in systems that are sufficiently complex to merit multiple plots (e.g., one for G and one for G^*), the output of SpaceMix contains all the information for them to do so. We are also currently working on an interactive graphics framework for more elegant exploration of the results, which will be implemented in

future releases.

- From a statistical point of view, perform a mapping $R^2 \rightarrow R^2$ based on a covariance matrix and variogram analysis is something that has been studied (2, 3). A connection between the present work and the statistical literature is interesting to place the work in a broader context of warping study.

  We agree that it would be exciting to explore other methods of visualizing covariance matrices (and also to place this work in a broader statistical context) in future work on SpaceMix. We now discuss this issue in the manuscript, in lines 670–675.

- In the article, several models are introduced. The inference of geognetic maps can be done with or without stationary population location, and with or without admixture. The geogenetic coordinates estimated by the different models are presented for simulated data and empirical data, and are different (Figure 2, 6, 8, 9). The model selection approach is here unclear. The model with the less parameters is advised l.275, why? How can a user know which model is the most relevant? An intuitive approach such as running the with admixture model and looking for large admixture proportion w does not seem a good idea as the true admixture proportion value may be outside the 95% credibility interval (l.286) in the simulations.

  W e have now included a detailed discussion of when users may wish to employ the different models described in lines 797–808.

- I don't understand why the authors run an MCMC algorithm if they are only interested in a MAP. The authors could use an efficient gradient algorithm such as the Conjugate gradient algorithm, or a variational approach to get a MAP (4). These options would be much faster than running a long Markov Chain. The credibility interval would not be returned, but the G coordinates would be the same if the MAP is actually reached. It would avoid computing chains with millions of steps.

This is a good point. We have responded above.

Minor

- l.146 "The likelihood of the data". One should say "likelihood" or "probability of the data".

  Fixed.

- l.191 "and and"

  Fixed.

- l.245 "between $X_{k,l}$ and $X_{k,l}$"

  Fixed.

- l.446 "the interpretation our results"

  Fixed.

References

1. MARDIA, Kantilal Varichand, KENT, John T., et BIBBY, John M. Multivariate analysis. Academic press, 1979. Chap. 14
2. SAMPSON, Paul D. et GUTTORP, Peter. Nonparametric estimation of nonstationary spatial covariance structure. Journal of the American Statistical Association, 1992, vol. 87, no 417, p. 108-119.
3. BOOKSTEIN, Fred L. Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Transactions on pattern analysis and machine intelligence, 1989, vol. 11, no 6, p. 567-585.

4. RUSTAGI, Jagdish S. Optimization techniques in statistics. Elsevier, 2014.

Reviewer 3

This paper describes a new method for summarizing and visualizing genetic data. It uses an isolation by distance model as a null and tries to infer the positions of populations in two dimensional space, adding admixture edges as necessary. I like the approach very much. Overall the paper is well written, the method is sound, and I don't have any major criticisms. Below are a few suggestions that might improve the paper.

- I thought that the analysis of the human data was the weakest part of the paper. It's clear that the IBD model is a poor fit to the worldwide data and, and I don't know how to interpret many of the admixture edges. As the authors write in another context - it looks a bit tortured. What does it mean that the Brahui apparently have Bantu-like admixture, for example? I agree with the authors' discussion that this might be an artifact due to multiple sources of admixture or unsampled populations, but I felt like this whole discussion was a bit long and over-interpreted some of these analyses. When the analysis is restricted to the Eurasian samples it looks cleaner, but then seems to find very few admixture edges compared to what I'd expect. I don't have any specific suggestions except that this discussion could be shortened a bit.

  These are great points; we have responded above, but, briefly, we have significantly re-written the section on the interpretation of the human dataset analyses (see the LatexDiff document).

- It would have been very interesting to see the result of running SpaceMix on a dataset which fits the IBD model better – for example the POPRES data. This would also have the advantage that we could compare directly with the PCA and SPA analyses of this dataset.

  We have responded to this point above, but, briefly, we think

this is a good idea, but that the manuscript is already so long that the addition of another empirical analysis would be excessive.

- Maybe I'm misunderstanding this, but on line 807 the methods say that the move is rejected if it lies outside the range of the prior. Doesn't that mean that the proposal distribution is not actually symmetric?

  The proposal is not repeated until an "in-bounds" move is proposed (which would not be symmetric). Instead, the proposed "out-of-bounds" move is rejected and the current value of the parameter is logged and reused in the next iteration of the MCMC, which does generate a symmetric proposal distribution. The wording of the sentence has been edited for greater clarity.

- Line 108 – "roughly unit variance in some sense" – perhaps this could be more specific.

  Fixed.