Journal of Biogeography (J. Biogeogr.) (2016) **43**, 1464–1476

SPECIAL PAPER

Quantifying the similarity between genes and geography across Alaska's alpine small mammals

L. Lacey Knowles[1]*, Rob Massatti[1], Qixin He[1], Link E. Olson[2] and Hayley C. Lanier[3]

[1]*Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 41809-1079, USA,* [2]*University of Alaska Museum, University of Alaska, Fairbanks, AK 99775, USA,* [3]*Department of Zoology and Physiology, University of Wyoming at Casper, Casper, WY 82601, USA*

ABSTRACT

Aim Quantitatively evaluate the similarity of genomic variation and geography in five different alpine small mammals in Alaska, and use this quantitative assessment of concordance as a framework for refining hypotheses about the processes structuring population genetic variation in either a species-specific or shared manner.

Location Alaska and adjacent north-western Canada.

Methods For each taxon we generated 3500–7500 single-nucleotide polymorphisms and applied a Procrustes analysis to find an optimal transformation that maximizes the similarity between principal components analysis maps of genetic variation and geographical maps of sample locations. We generate stability maps using projected distributions from ecological niche models of the Last Glacial Maximum and the present.

Results Significant similarity between genes and geography exists across taxa. However, the extent to which geography is predictive of patterns of genetic variation not only differs among taxa, but the correspondence between genes and geography varies over space. Geographical areas where genetic structure aligns poorly with the geographical coordinates are of particular interest because they indicate regions where processes other than isolation by distance (IBD) have influenced genetic variation. The clustering of individuals according to their sample location does not support suppositions of admixture, despite the presumed high vagility of some species (e.g. arctic ground squirrels).

Main conclusions Genomic data indicate a more nuanced biogeographical history for the taxa than suggested by previous studies based on mtDNA alone. These include departures from IBD that are shared among taxa, which suggest some shared processes structuring genetic variation, including new potential ancestral source populations. In addition, some regions fit expectations of IBD where incremental migration and gene flow play a strong role in population structure, despite any ecological difference among taxa. Differences in dispersal capabilities do not result in different species-specific local patterns of population structure, at least at the sampling scale examined here. We highlight how the general fit to, as well as departures from, expectations for patterns of genetic variation based on the Procrustes analyses can be used to generate hypotheses about the underlying processes.

*Correspondence: L. Lacey Knowles, University of Michigan, 1109 Geddes Ave., Ann Arbor, MI 48109-1079, USA.
E-mail: knowlesl@umich.edu

Keywords

Alaska, climate change, isolation by distance, mammal, next generation sequencing, phylogeography, Procrustes analyses, RADseq

http://wileyonlinelibrary.com/journal/jbi
doi:10.1111/jbi.12728

INTRODUCTION

The geographical structure of population genetic variation is the foundation of phylogeography. Such spatial structure provides a basis for investigating the history of migration and/or colonization routes. The most common method applied for assessing the correspondence between genes and geography is a test of isolation by distance (IBD), where a pairwise measure of population similarity/dissimilarity is compared to a pairwise measure of geographical distance separating populations. Although the exact metric used in such tests might vary (e.g. the pairwise genetic distance or a measure of F_{ST} might be used to characterize the axis of genetic divergence, and the Euclidean distance or some rescaled distance based on the habitat suitability separating populations might be used to characterize the axis of geographical distance; see Wang & Bradburd, 2014), the utility of such tests is a statistical evaluation of the extent to which geography predicts patterns of genetic variation. However, a drawback to these broadly applied tests is that, because they do not retain information about the relative positions of populations across space, they do not contain information for interpreting the relative deviations from an IBD of specific sampled populations (Papadopoulou & Knowles, 2015a).

Principal components analysis (PCA) and non-metric multidimensional scaling provide a means for visualizing summaries of genetic data, decomposing the high dimensionality of genomic data into a reduced number of axes to qualitatively investigate the clustering patterns of genetic variation. When single-nucleotide polymorphisms (SNPs) are adequately sampled across populations, distances between population clusters on a two-dimensional (2D) principal components (PC) space are proportional to their pairwise F_{ST} measures (McVean, 2009). Therefore, the positioning of population clusters in PC space matches exactly with their geographical distributions under a strict IBD model. A recently developed Procrustes analysis approach to studying genetic variation makes use of this important feature of PCA analysis by statistically quantifying the association between genetic PCA and geographical maps (Wang *et al.*, 2012). Specifically, this approach finds an optimal transformation that maximizes the similarity between genes and geography by minimizing the sum of squared Euclidean distances between a PCA map of genetic variation and geographical coordinates while preserving relative pairwise distances among points within the genetic and geographical maps. When coupled with permutations, the statistical significance of the similarity between genes and geography can be evaluated. In addition, the relative deviations of sampled individuals from expectations based on their geographical location can be visualized, identifying both the magnitude of deviations and also the general direction of such deviations (e.g. individuals that are more closely related to those at different latitudes or longitudes than expected based on where they were sampled; see Papadopoulou & Knowles, 2015a).

We apply a Procrustes analysis approach to not only systematically test for an association between genes and geography among sampled populations in different species of alpine small mammals across Alaska and adjacent regions of north-western Canada, but also to systematically assess how this association differs across taxa. Moreover, by using this approach, we can visualize the role of geography in explaining the genetic similarity of populations from different locations, and identify regions that correspond more (or less) to expectations based on the geographical locality of sampled individuals. The five mammal taxa – collared pika [*Ochotona collaris* (Nelson, 1893)], hoary marmot [*Marmota caligata* (Eschscholtz, 1829)], singing vole [*Microtus miurus* (Osgood, 1901)], brown lemming [*Lemmus trimucronatus* (Richardson, 1825)] and arctic ground squirrel [*Spermophilus parryii* (Richardson, 1825)] – are broadly distributed across much of Alaska, but differ in the latitudinal extent of their ranges (e.g. the ranges of the arctic ground squirrel, brown lemming and singing vole extend to higher latitudes compared to the hoary marmot and collared pika). A comparative phylogeographical study based on mtDNA has identified seemingly concordant east-west splits in all five species, which may be indicative of a shared refugial history (Lanier *et al.*, 2015a). Given their overlapping ranges in the region and reliance on alpine or tundra habitat, they provide an ideal context for testing the extent to which taxa from similar geographical areas, some of which have a dynamic history that includes extensive glacial impacts, show common versus species-specific spatial patterns of genetic variation, especially given differences in the dispersal capabilities and habitat affinities of the taxa (Lanier *et al.*, 2015a). We discuss how deviations of individuals in the PC maps from predictions based on geography can be used to generate hypotheses about underlying processes and consider the challenges with interpreting spatial patterns in PCA maps (see Novembre *et al.*, 2008; Francois *et al.*, 2010). In addition, we consider a set of refined hypotheses based on comparing the patterns from the Procrustes analyses with projections of distributional stability inferred from ecological niche models (ENMs) for the present and the past (specifically, the Last Glacial Maximum, LGM) (Alvarado-Serrano & Knowles, 2014). Lastly, through a series of sequential exclusion of sampled populations, the robustness of the similarity between geography and genes, as well as between the genetic PCAs with and without particular populations, is assessed in each species.

MATERIALS AND METHODS

Species sampling and genomic library preparation

Genomic data were collected from individuals in populations sampled across the Alaskan and north-western Canadian ranges of each of five species: collared pikas (59 individuals from 9 populations), hoary marmots (55 individuals from 9 populations), singing voles (62 individuals from 9 populations), brown lemmings (60 individuals from 8 populations) and arctic ground squirrels (63 individuals from 9 populations; see Fig. 1 and Tables S1.1 & S1.2 in Appendix S1 in Supporting Information). See Appendix S1 for details about DNA extraction and library construction.

Processing of illumina data

Sequences for each species were demultiplexed and reads with an average Phred score > 30 and an unambiguous barcode and restriction cut site were retained (scripts are available on Dryad under doi: 10.5061/dryad.8jm51). The Stacks 1.07 pipeline (Catchen *et al.*, 2013) was used to identify SNPs in the processed genomic data from the species-specific genomic libraries constructed for each species (see Appendix S1 for details about the processing of genomic data).

Genetic diversity statistics

After genomic variation was identified within individuals, the Stacks output files were loaded into species-specific MySQL databases. Loci were exported from each species' MySQL database using the *export_sql.pl* script, allowing one to four SNPs per RADseq locus; only biallelic RADseq loci were considered in order to comply with the assumptions of the current methods for analysing SNP data. The populations program in Stacks was used to calculate population genetic statistics on the exported RADseq loci, including nucleotide diversity (π), major allele frequency, observed heterozygosity and Wright's inbreeding coefficient (F_{IS}). Only loci present in at least two populations ($P = 2$) and genotyped in at least 50% of the individuals of each population ($r = 0.50$) were used to create molecular summary statistics for each species; in instances where 50% did not result in a round number of individuals in a population, the number of individuals

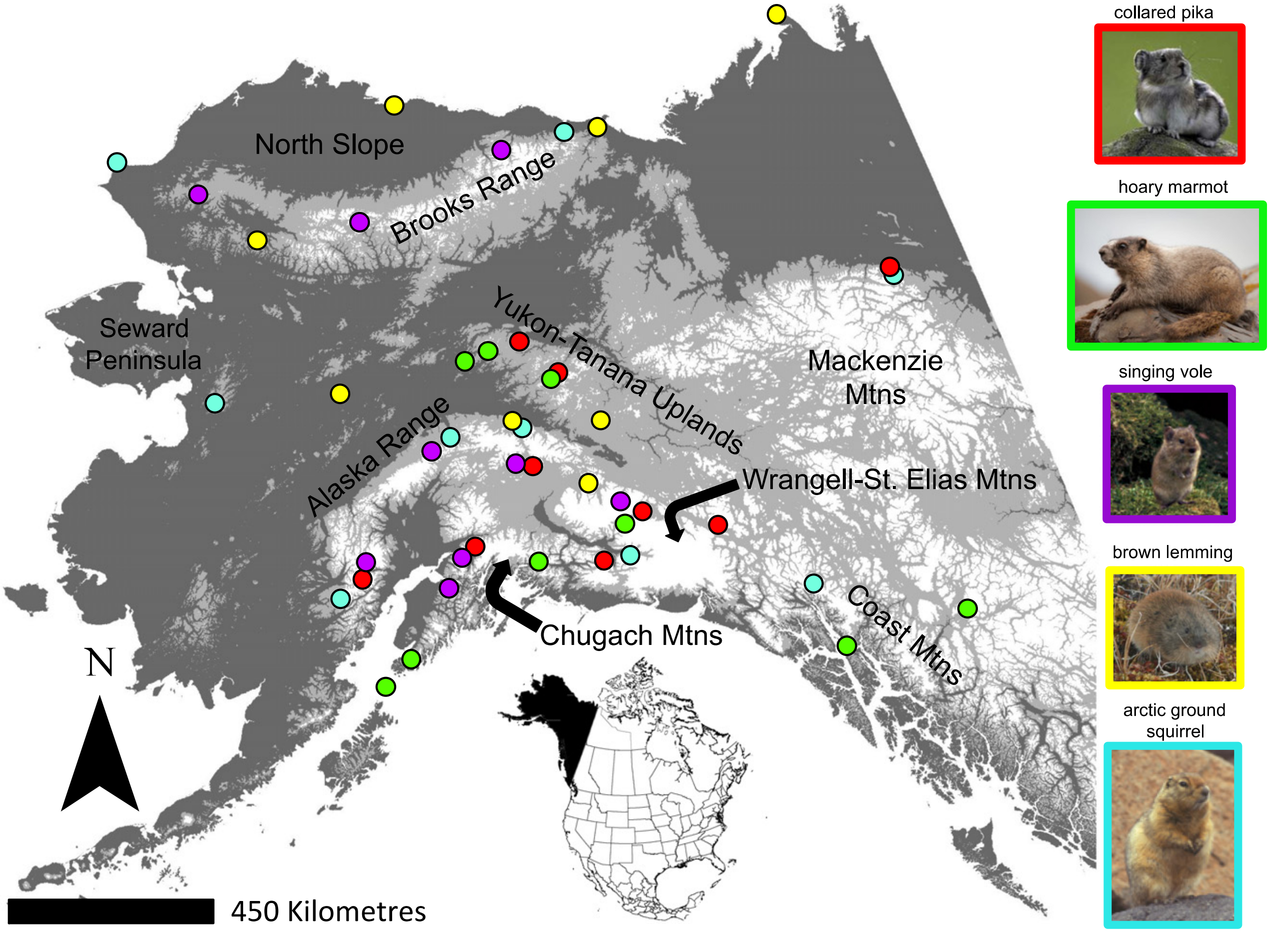

Figure 1 Sampled populations for the five mammal taxa in Alaska: collared pika (*Ochotona collaris*), hoary marmot (*Marmota caligata*), singing vole (*Microtus miurus*), brown lemming (*Lemmus trimucronatus*) and arctic ground squirrel (*Spermophilus parryii*). Also marked are the primary mountain ranges and uplands against a grey scale background. The sampling locations are representative of each species' range within the study area. The extent of overlap of their respective ranges differs; for example, collared pikas and hoary marmots do not occur in northern Alaska (e.g. the Brooks Range). Overlapping sampling points indicate species were collected from the same site. Photos credits: Moose Peterson (collared pika), Ivan Andrijevic (brown lemming), Link Olson (arctic ground squirrel), Jonathan L. Fiely (hoary marmot) and Creative Commons (singing vole).

required before the locus was processed was rounded up (e.g. a locus would need to be genotyped in three out of five individuals with $r = 0.50$).

Quantitative comparison of the similarity between genes and geography

Principal components analyses were performed on customized STRUCTURE files. To create these files, the loci names exported from the MySQL databases (see above) were used as a 'whitelist' by the POPULATIONS program in STACKS; in order to stop POPULATIONS from filtering out loci from our data set (we wanted all of the data irrespective of the number of populations a locus occurred in or the number of individuals a locus was present in within a population), we set all parameters to 0. POPULATIONS wrote the genomic data in a Variant Call Format (vcf) file, which we converted to a STRUCTURE file format using PGDSPIDER 2.0.7.2 (Lischer & Excoffier, 2012). The STRUCTURE file for each species was edited to exclude linked SNPs, as well as SNPs and individuals that contained a high proportion of missing data, which can disproportionately affect patterns in a PCA (Wang *et al.*, 2012). First, all SNPs with > 70% missing data were deleted. Next, the amount of missing data per individual was calculated, and individuals with prohibitively high amounts of missing data (such that the final data set would contain too few SNPs) were excluded (these individuals were obvious because they generally contained > 90% missing data). The final step was to maximize the number of SNPs and individuals such that each individual had < 15% missing data (Lischer & Excoffier, 2012); the final number of individuals used in subsequent analyses is presented in Table S1.1 in Appendix S1.

Principal components analysis were performed on species-specific matrices in R (R Core Team, 2014) using the ADEGENET R package (Jombart *et al.*, 2008). Missing data were replaced by the mean frequency of the corresponding allele, which is recommended for centred PCAs (Jombart et al., 2008). Major axes for genome-wide SNP data were identified using the R DUDI.PCA function (centre = T, scale = T). An association between genetic differentiation and geography was assessed considering divergence along both latitudinal and longitudinal axes across populations using a Procrustes transformation approach. Specifically, species-specific PC1 and PC2 scores and the projected latitude and longitude of sampling localities were inputs in a Procrustes analysis, which maximizes the similarity between PCA maps of genetic variation and geographical locations of sampled populations (see Wang *et al.*, 2010, 2012). Geographical coordinates were transformed to an Albers Equal Area Conic projection using the SPTRANSFORM function in the RGDAL R package (Bivand *et al.*, 2014). Analyses were performed using the PROTEST function in the VEGAN R package (Oksanen *et al.*, 2013). Because Procrustes analysis superimposes a PCA plot of genetic variation onto a geographical map by rotating the PC axes to achieve maximum similarity to the geographical distribution of sampled locations (i.e. the sum of squared differences between the two data sets are minimized), it is ideal for quantitative comparison of the similarity between genes and geography across taxa (as it is for comparing the association between regions; see Wang *et al.*, 2010). We report the angle of the PCA map (i.e. θ, the rotation measured in degrees) that optimally minimizes the sum of squared Euclidean distance between the PCA map from the SNP data and the geographical map. The significance of the association statistic between the first two PCs of genetic variation and the geographical coordinates of the populations (denoted as t_0) for each species was evaluated based on 10,000 permutations, where geographical locations were randomly permuted across the different sample localities (note that all individuals from the same locality were assigned to a single geographical location in the permuted data set, such that observed levels of population structure were maintained).

As the aim of the work was to evaluate the overall similarity (or lack thereof) in the association between genes and geography across taxa, we assessed the robustness of our results by excluding one population at a time and repeating the PCA and Procrustes analyses on the new data sets. Comparison of the PCA coordinates from the new data sets and the original geographical data sets were applied systematically to identify the maximum extent to which the association between genes and geography might increase or decrease as different populations were excluded, denoted by the similarity score t'' (following the notation of Wang *et al.*, 2012). In addition, a similarity score denoted by t' (following the notation of Wang *et al.*, 2012) was computed between the new PCA coordinates for the SNP data and the original PCA coordinates for the SNP data (i.e. before removing any population) to assess how robust the patterns among populations in PCA space are to individual populations.

Environmental niche modelling

Environmental niche models (ENMs) were generated from bioclimatic variables for the present and the LGM with MAXENT 3.3.3e (Phillips *et al.*, 2006). We performed *a priori* model testing to determine optimal combinations of the regularization and feature parameters for the construction of each species' present-day ENM (Warren & Seifert, 2011). Specifically, we used SDMTOOLBOX (Brown, 2014) to test models over combinations of regularization parameters from 0.25 to 3 in intervals of 0.25 and the Linear, Quadratic, Hinge, Product and Threshold features. Each model parameter class was replicated 25 times using cross-validation. Georeferenced distribution points from vetted occurrence data used in the modelling were representative of the entire ranges of the five species, respectively, throughout northwestern North America (Dryad doi:10.5061/dryad.8jm51). For each species, occurrence data were spatially rarefied using SDMTOOLBOX at a resolution of 10 km to reduce spatial autocorrelation.

We used 19 bioclimatically informative variables to model present-day distributions (WorldClim 1.4; Hijmans *et al.*, 2005) and LGM distributions (PMIP2-CCSM; Braconnot *et al.*, 2007) for each species. To avoid overfitting of the distribution models, the geographical extent of the environmental layers was reduced to an area *c.* 20% larger than the known distribution of each species (Anderson & Raza, 2010) and coupled with background sampling bias files (Phillips *et al.*, 2009; Merow *et al.*, 2013). Sampling bias files were constructed in SDMToolBox using a buffer distance of 100 km, which was reasonable given the geographical extent of Alaska and the distance among species' occurrence points. Subsequently, the following procedure was carried out for each species to guard against the inherent difficulties in extrapolating distributions into novel climates (reviewed in Alvarado-Serrano & Knowles, 2014). Specifically, an iterative approach was used to generate ENMs for the LGM in which multivariate environmental similarity surfaces (MESS maps) were used to identify bioclimatic variables that result in areas of low reliability because of predicted values that are outside of the range of present-day environmental values for any given taxon (Elith *et al.*, 2010). Maxent was rerun excluding these out-of-range variables, and this process of analysis with MESS maps was repeated until no LGM variables were out-of-range compared to present-day bioclimatic variables. Because MESS maps do not indicate changes in correlations among the environmental variables used for LGM reconstructions (Elith *et al.*, 2010), we checked our ENM for the LGM using only the most informative variable for each species to ensure that we were not reporting errant distributional patterns. In addition, a present-day ENM was generated using the subset of variables that were not out-of-range during the LGM and compared to an ENM constructed using the most important variable (as determined by Maxent) and the remaining variables that had Pearson's *r* correlations to this variable of < 90%, as determined by ENMTools (Warren *et al.*, 2010); while these models were not expected to be identical, we checked that both models reported similar distributional patterns. Details about species-specific environmental variables and parameters for the different models are reported in Table S2.1 in Appendix S2.

RESULTS

Sequence data and genetic diversity

More than 500-million reads were produced across the four lanes of Illumina sequencing (average of 1,821,116 ± 825,584 reads per analysed individual across species; for details see Table S1.2 in Appendix S1). After excluding SNPs that were linked and/or that had greater than 15% missing data, the number of independent SNPs per species was: collared pika, 7463; hoary marmot, 5524; singing vole, 3666; brown lemming, 4718; arctic ground squirrel, 3502 (note that variation in the number of SNPs primarily reflects differences in genome size and effective population size across taxa, given the similar quality of reads, and the number and distribution of reads across specimens, in each library). Summaries of genetic diversity per population are given for each of the five taxa in Table S3.1 in Appendix S3. Heterozygosity was generally consistent across populations (with the exception of the Sud Island population of the hoary marmot, which had a considerably lower observed heterozygosity compared to other populations), but differed among taxa.

Procrustes analyses and ENMs

We find significant similarity between genes and geography across taxa (see Table S3.2 in Appendix S3). However, the strength of similarity differs among taxa and across geographical regions (see Fig. S3.2 in Appendix S3). Below we describe these associations between genes and geography on a per-species basis, including the robustness of the association with the exclusion of populations, as well as how the results from the Procrustes analyses conform to the projections of the species' distributions in the past based on the ENMs.

Collared pika

Although the similarity score between the pika populations in PC space and their actual geographical locations is significant ($t_0 = 0.71$; $P < 1.0^{-5}$), t_0 is generally low compared to other taxa (only the brown lemming has a lower t_0). This is in part due to departures associated with Jawbone Lake and the Pika Camp populations. For example, given the distance from Jawbone Lake in the east to Lake Kenibuna in the west, we would expect a large distribution of genetic variation along the longitudinal axis. Instead, individuals from these populations cluster with individuals from more centrally located populations (Fig. 2 and Fig. S3.1 in Appendix S3). In contrast, the Pika Camp population is more divergent genetically than would be predicted by geography alone (i.e. the population occupies a more distant area of PC space relative to the other populations).

Although reconstructions of glacial margins during the LGM suggest a north-western refuge (as previously suggested; Lanier *et al.*, 2015b), this was not supported by the Procrustes analyses (Fig. 4 and see Fig. S2.1 in Appendix S2). Collared pikas (like hoary marmots) are not known from the Brooks Range or anywhere north of the Yukon River in Alaska (Gunderson *et al.*, 2009; Lanier & Olson, 2013). If there was a more northerly population in the past, as predicted by the LGM ENM (see Fig. S2.1 in Appendix S2), it did not contribute to the current standing genetic diversity (i.e. we would expect strong deviations of populations from the central and southern areas if these areas were indeed colonized from a distant geographical source, which are not observed; Fig. 2). Likewise, despite the proximity of other sampled populations (e.g. Rock Lake) and suitable habitat during the past and present (Fig. 4), individuals from Pika Camp have a distinct ancestry that may

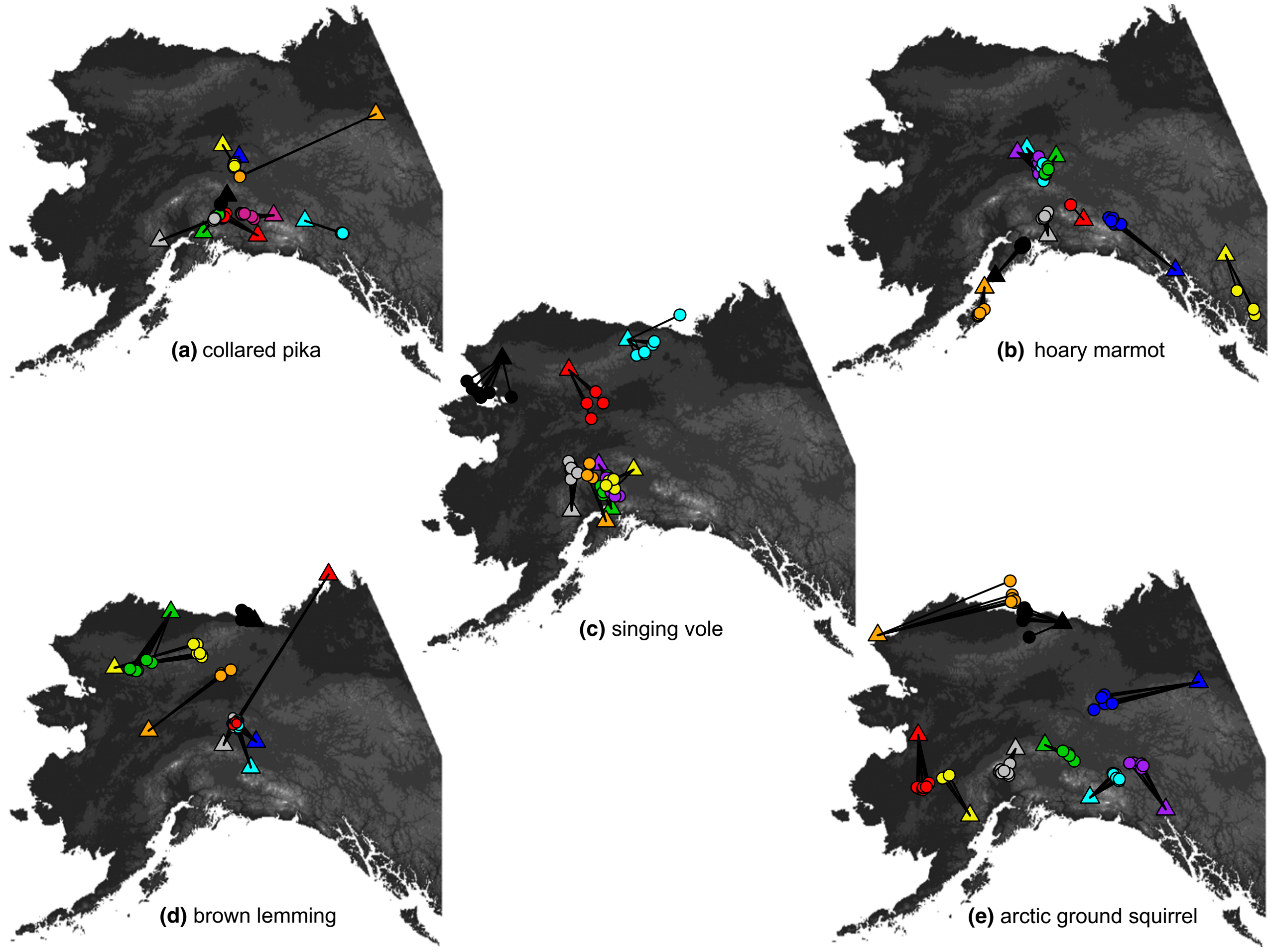

Figure 2 Procrustes-transformed PCA plot of genetic variation with each individual mapped in principal components (PC) space (small open circles) relative to the geographical location of populations (triangles) for each of the taxa (i.e., the plots for each taxa, a through e, as projected upon the map of Alaska). The length of the line connecting individuals in PC space to their geographical location represents the extent of the deviation from the expected pattern of genetic variation based on geography.

indicate that it was colonized from a different refugial source population (see below).

Hoary marmot

This species showed the highest similarity between genes and geography ($t_0 = 0.90$, $P < 1.0^{-5}$) and it became very high with either the exclusion of the south-eastern Juneau ($t'' = 0.95$) or Northwestern British Columbia (NWBC) population ($t'' = 0.96$) (Fig. 3; see Table S3.2 in Appendix S3). The position in PC space of marmots from the Juneau population (Fig. 2) shows that they are genetically consistent with a population located much further to the west, whereas the NWBC population sampled at the same latitude, but just to the east (Fig. 2), is genetically consistent with populations farther south.

Geographical structuring of genetic patterns in hoary marmots in some ways mirrors that in collared pikas. The south-eastern-most marmot population is projected to a more southern location in the Procrustes analysis (Fig. 2), like the pikas. However, the inferred boundary between these putative refugia is discordant between the two species. Specifically, this west-versus-southern deviation occurs in the more eastern extent of the hoary marmot's sampled range compared to that for the collared pika. In both species, the central Alaskan populations show a strong correspondence between genes and geography (Fig. 2), suggesting historical stability (Fig. 4).

Singing vole

This species shows the second-highest similarity score between genes and geography ($t_0 = 0.89$; $P < 1.0^{-5}$) of all the taxa (see Table S3.2 in Appendix S3) as evidenced by the consistently small distortions of individuals from their expected genetic patterns based on the geographical location of sampled populations (Fig. 2). This pattern was generally robust to sequential population exclusion (Fig. 3).

The ENMs project suitable, stable habitat in both the northern and southern parts of the singing vole's current

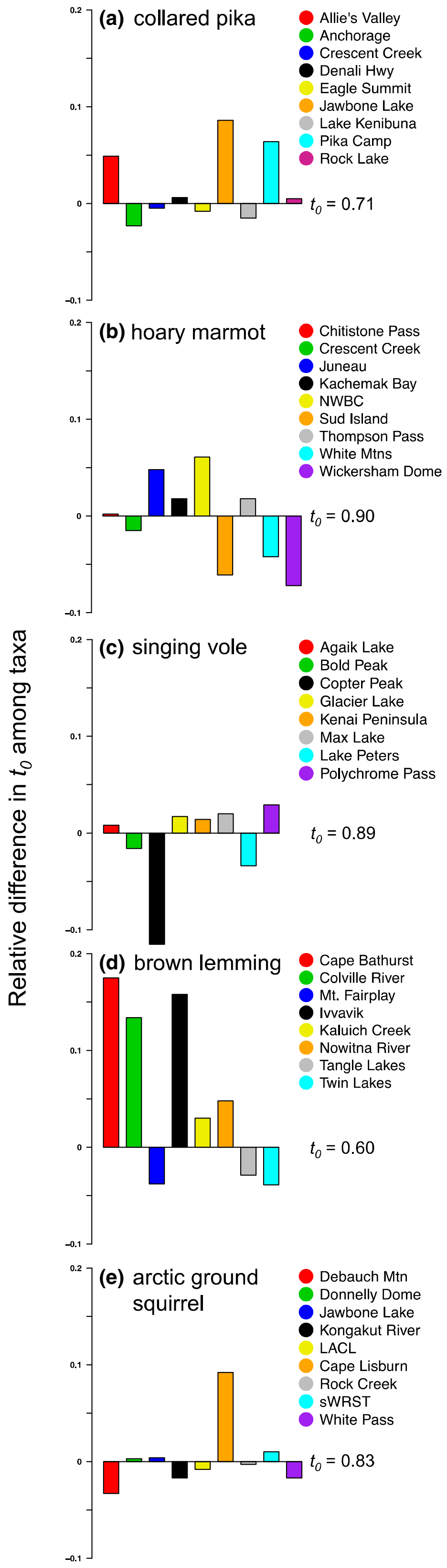

Figure 3 Comparison of the changes in the association between genes and geography with the exclusion of individual populations (i.e. t'') relative to when all populations are analysed (i.e. t_0) for: (a) collared pikas; (b) hoary marmots; (c) singing voles; (d) brown lemmings; and (e) arctic ground squirrels. Values for each species are standardized by t_0 (i.e. 0 on y-axis corresponds to t_0) such that positive values indicate a stronger association between genes and geography when a population is excluded, whereas negative values indicate a weaker association. Bar colours represent sampling populations; the same colours for each species' populations are used throughout all figures (Fig. 2 and Appendix SI).

range, but not in the central part of the range (Fig. 4). The northern populations remain genetically differentiated (see Fig. S3.2 in Appendix S3), and these differences generally correspond to the geographical distances separating the populations (Fig. 2), in contrast with patterns seen in the northern populations of the arctic ground squirrel and brown lemming (below), which tend to show less genetic distinctiveness. For example, northern arctic ground squirrel populations tend to be much more genetically similar to one another than similarly distributed singing vole populations (Fig. 2).

There is also no indication that singing voles were displaced to central Alaska based on the Procrustes analysis (Fig. 2), corresponding to the lack of suitable stable habitat in that region (Fig. 4). This is consistent with the rarity of reports of singing voles from central Alaska, despite intensive and repeated sampling efforts (Weksler *et al.*, 2010; Baltensperger & Huettmann, 2015).

Brown lemming

This species exhibits the lowest similarity between genes and geography of all five species (t_0 = 0.60, $P < 1.0\times10^{-5}$). Of the five focal species, the brown lemming is inferred to have the largest geographical extent of stable habitat (Fig. 4), based on projections of the species' present and past distributions (see Fig. S2.1 in Appendix S2). However, there is no obvious corridor of suitable habitat identified from the ENMs (Fig. 4) between the Cape Bathurst population in the north-east and the south-central Alaskan region that might explain the high genetic identity of individuals from this area with the central populations (Figs 1 & 2, and Fig. S2.1 in Appendix S2).

Arctic ground squirrel

This species exhibits several patterns unique among the sampled taxa. Individuals from the two northernmost populations (Figs 2 & 3) overlap genetically despite their geographical separation, unlike the patterns in singing voles and brown lemmings. In contrast to all other species with populations sampled in or near the northern Alaska Range (Fig. 1), arctic ground squirrels from this area do not overlap in PC space, instead remaining distinct (and even diverging from each other in opposite directions; Fig. 2). Furthermore, individuals from all of the remaining populations generally form distinct, non-overlapping clusters in the

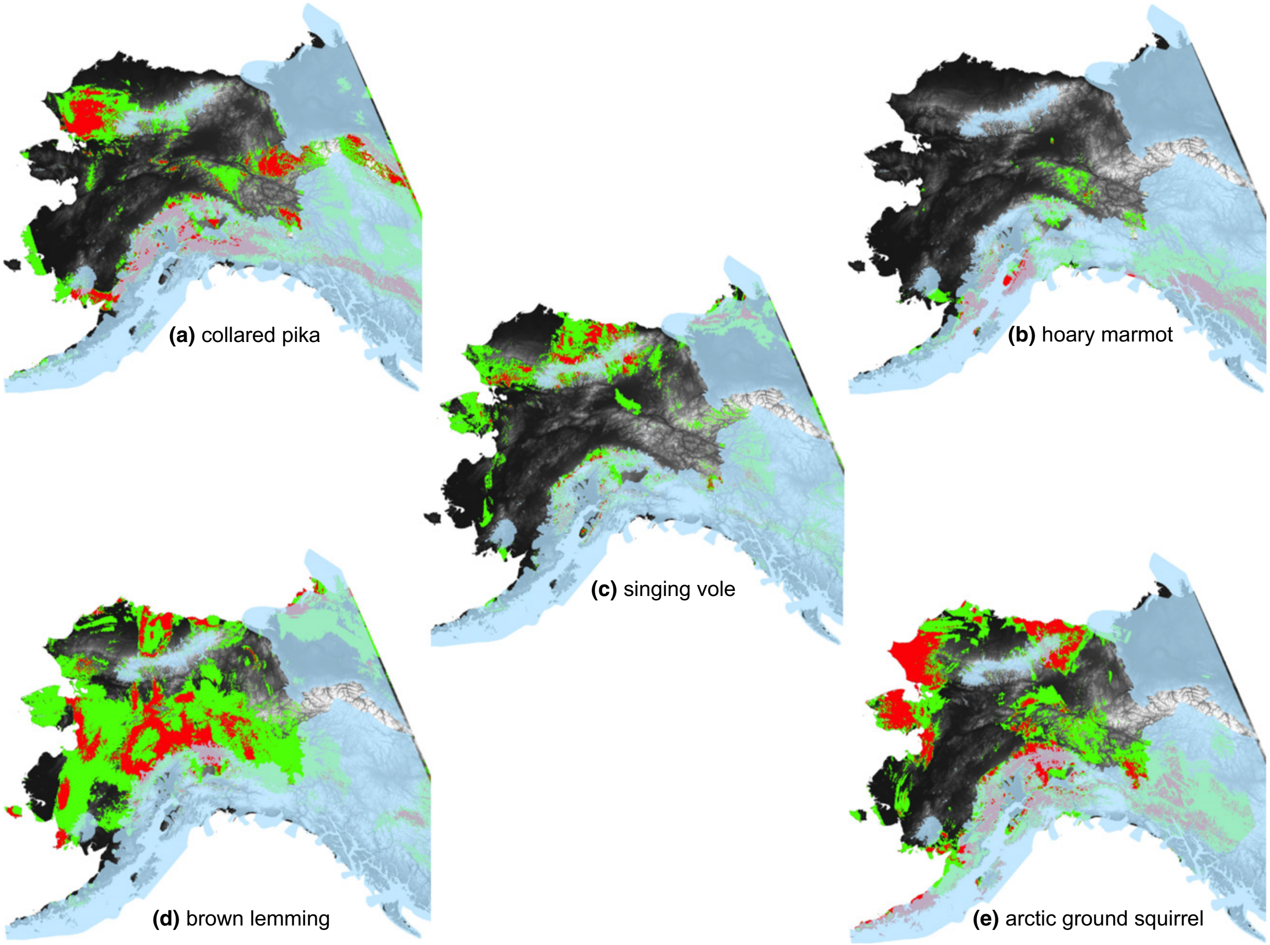

Figure 4 Maps of habitat predicted to be stable throughout Pleistocene glacial cycles. For each species (i.e., a through e), stable habitat (shown in red) is defined by the overlap of ENMs for the present and the Last Glacial Maximum (LGM), whereas unstable habitat (shown in green) is habitat predicted to be suitable in either the present or the LGM. The extent of glacial coverage at the LGM is shown in light blue. Note that the glacial reconstruction is based on independent geologic information from glacial moraines. Separate projections of current and past distributions are available in the supplement (see Fig. S2.1 in Appendix S2).

vicinity of their sampling localities (similar to the southern populations of hoary marmots). Despite these differences, individuals from Jawbone Lake still deviate longitudinally towards central Alaska, similar to the patterns seen in collared pikas and brown lemmings (from the more northern Cape Bathurst population).

While these patterns are unusual, the association between genes and geography in the arctic ground squirrel is significant and within the range of variation seen in the other taxa (t_0 = 0.83; $P < 1.0^{-5}$). Examination of stable habitat indicates that all of the individuals projected onto geographical space are near habitat expected to be stable in both the LGM and present (Fig. 4), except for the western and south-western populations of Debauch Mountain and LACL, respectively.

DISCUSSION

Species-specific analyses are useful for identifying a correspondence between genes and geography, but a comparison across taxa can also be used to generate hypotheses about shared versus taxon-specific biogeographical histories. In particular, the patterning of spatial variation differs among taxa, and the patterns of genetic variation in some areas more closely fit predictions based on where an individual was sampled compared to others. Below we highlight what our findings suggest about the history of arctic and subarctic alpine mammals, and in particular, specific hypotheses about their biogeographical and demographic histories. We also discuss the limitations of the approach, especially with respect to understanding the cause of deviations of genetic variation from expectations based on geography. Specifically, we focus on the utility of the approach for identifying hypotheses that might be tested with other approaches, rather than inferring process from the results of the Procrustes analyses themselves.

Comparison of Procrustes analyses across taxa

The similarity between geography and genes varied among taxa. For example, with all sampled populations included, t_0

ranged from a high of 0.90 in the hoary marmot to a low of 0.60 in the brown lemming (see Table S3.2 in Appendix S3). However, this variation reflects in part the disproportionate effect of individual populations (or combinations of several outlier populations) on decreasing $t_{0.}$ Indeed, in none of the species was the highest t_0-value observed when all populations were analysed. The highest similarity between geography and genes was achieved when a population was excluded (e.g. the similarity between geography and genes increased in all taxa, ranging from $t'' = 0.77$ in brown lemmings to $t'' = 0.96$ in hoary marmots). Interestingly, taxa differed with respect to which geographical regions, when excluded, maximized the association between genes and geography. For example, exclusion of the Cape Lisburn population of arctic ground squirrels maximized the similarity between genes and geography, but in the hoary marmot it was the south-eastern NWBC population that maximized t'' (Figs 2 & 3). This suggests there is no single and common cause to the departure from IBD across these taxa, which is relevant to forming comparative phylogenetic hypotheses for additional testing (see below).

Increases in the similarity between geography and genes when a particular population was excluded (i.e. t'') was not due to a disproportionate effect of the excluded population on the relative positions of individuals in PC space (see Table S3.2 in Appendix 3 for t'-values, which remained consistently very high). In all cases, the similarity between the PCA of genes with and without the population that maximized the similarity between genes and geography (i.e. t'') was 1.0 (the maximum value for t'), except for in hoary marmots where $t' = 0.96$. In contrast, for cases in which the exclusion of a population reduced the association between genes and geography (see Table S3.2 in Appendix S3), the large drop in t'' was also accompanied by a shift in the similarity of the PCAs of genes, t'. This suggests that inclusion of such populations is critical to characterizing the spatial structure in each taxon across the sampled region. In fact, in each species, the exclusion of several different populations results in t''-values that are lower than t_0-values, which highlights the importance of representative sampling across the species range when characterizing spatial structure (see DeGiorgio & Rosenberg, 2013). Again, the effect is not due solely to fewer data points when populations are excluded because in all species t_0, with all populations analysed, was lower than the maximum t''-value achieved when one population was excluded from the Procrustes analyses (as described above; see also Fig. 3).

Hypotheses motivated by results of the Procrustes analyses

The statistical association between genetic variation and geography in all species is an important finding. However, it is also noteworthy to consider what populations deviate from IBD expectations (especially when viewed in a comparative context and visualized geographically). In particular, these aspects of the Procrustes analyses can be useful for formulating hypotheses. To be clear, other approaches might be used to test for an association between genetic variation and geography (see Jombart *et al.*, 2008; Frichot *et al.*, 2012). However, with visualizations of distortions in genetic variation in relation to the geographical localities of sampled individuals, Procrustes analyses also provides a useful framework for generating hypotheses (see also Papadopoulou & Knowles, 2015a). As such, the output from Procrustes analyses can address one of the major challenges in statistical phylogeographical study – the identification of hypotheses (Knowles, 2009).

A notable departure (with regard to both the magnitude and geographical orientation of deviations) pertains to sampled populations along the periphery of the Alaskan mammal ranges relative to those from the interior. Specifically, the positioning of individuals in the Procrustes analyses span the entire latitudinal range of sampled populations in all the species. However, the full geographical extent of sampled populations along the longitudinal axis is underrepresented genetically, especially in collared pikas, brown lemmings and arctic ground squirrels (see Fig. S3.2 in Appendix S3). That is, individuals are clustered towards the Alaskan interior more than would be expected based on the longitudinal position of populations (Fig. 2). For example, any population sampled in the north-eastern portion of the ranges (e.g. in the area of the Mackenzie River Delta) shows patterns suggestive of a shared ancestor with other more centrally located populations, rather than an ancestral refugial source population in the north-east.

Other repeated patterns of deviations from IBD across taxa are suggestive of a shared biogeographical history in which populations within a region may have been colonized from multiple refugial source populations. For example, hoary marmots and collared pikas from the south-east are much more distant in genetic space from other geographically proximate populations (Fig. 2). Singing voles show a similar displacement (results not shown), but because of questions surrounding their taxonomic identity (Weksler *et al.*, 2010), these specimens were excluded from this study. However, not all taxa from this region show the same deviation. Arctic ground squirrels sampled in this south-eastern region (Fig. 1) are genetically most similar to populations to the north and west. Hence, although the genetic data in the taxa suggest the north-eastern region has not been consistently inhabited (i.e. these regions do not fit with general expectations under an equilibrium isolation-by-distance model), it seems unlikely that the deviations could be explained by one hypothesis regarding the geographical position of refugial source populations. For example, south-eastern populations of hoary marmots, collared pikas and possibly singing voles, but not arctic ground squirrels, may have been founded from an ancestral population further to the south and east than predicted by the current localities of sampled individuals (Fig. 2).

The results from the sequential exclusion of populations identifying regions (or populations) that have a disproportionate effect on the association between genes and geography can also be a source of information for developing hypotheses about region-specific processes. For example, a much higher association between genes and geography results when brown lemmings from northern coastal populations (Fig. 2) are excluded in Procrustes analyses (Fig. 3). This suggests that a possible hypothesis to explain the deviations between genes and geography in brown lemmings (Fig. 2) would have to accommodate the entire northern coastal region (not just one or two specific populations). Moreover, latitudinal differences in the genetic similarity of individuals suggest the region might have experienced fairly localized processes. These might include aspects of the demography of colonization and/or different ancestral source populations (i.e. individuals from Cape Bathurst and Colville River show genetic variation consistent with individuals sampled from more southern latitudes, in contrast to the Ivvavik population).

This more nuanced picture with concordance limited to specific taxa and certain geographical regions differs from more generalized hypotheses identified from mtDNA (Galbreath *et al.*, 2011; Lanier *et al.*, 2015a). Perhaps this is not entirely unexpected given that different markers provide differing degrees of resolution (Knowles, 2009). With the additional resolution of genomic markers it is increasingly clear that relying on mtDNA (or any single linkage partition) alone overlooks processes that may actually structure genomic variation. For example, unlike interior Alaska, which was part of ice-free Beringia during the LGM, formerly ice-covered localities within the hoary marmot's current distribution show the greatest discordance between genes and geography (i.e. Figs 2b & 3b). Likewise, a rapid expansion of hoary marmots from one or more south-central refugia (either nunataks or periglacial areas predicted as being suitable marmot habitat during the LGM; Fig. 4b) suggests a more dynamic history than suggested by past studies. Nevertheless, the genomic analyses also provide corroborative support for some species-specific hypotheses suggested by patterns of mtDNA differentiation. For example, a possible inland incursion from a coastal refugium (see Kerhoulas *et al.*, 2015) originating south of our sampling regime is suggested by the seemingly anomalous discordance in the NWBC marmots (Fig. 2).

Testing hypotheses developed from the findings of the Procrustes analyses

Some hypotheses suggested by the Procrustes analyses appear to be corroborated from independent data sources. For example, we have hypothesized that the Yukon-Tanana uplands are a potential refugium for hoary marmots and collared pikas based on deformations in the north-central parts of their range in the Procrustes analyses (Fig. 2). This area is also projected to be highly suitable and stable habitat by the ENMs (see Fig. S2.1 in Appendix S2), and it has been identified as a biodiversity hotspot for Alaskan small mammals (Baltensperger & Huettmann, 2015).

More generally, and as we advocate here, departures from IBD detected in the Procrustes analyses can be used to generate hypotheses for future study (as discussed above). However, the results from the Procrustes analyses, by themselves, are not sufficient for interpreting the processes underlying the lack of a correspondence between genes and geography (see below).

Not only might different processes leave similar signatures that can be difficult to distinguish, but the signal of a specific process may not be easily intuited from the pattern of deviations evident in the Procrustes plots, as with other summaries of genetic variation (see Knowles & Alvarado-Serrano, 2010; Brown & Knowles, 2012; He *et al.*, 2013; Wang & Bradburd, 2014). For example, it is difficult to identify one hypothesis that might have generated the deviations from IBD observed in the arctic ground squirrel (Fig. 2). Only the exclusion of the north-western population lead to an appreciable increase in this association (Fig. 3), leaving a fair amount of genetic variation that is not explained by IBD. A possible hypothesis that might be considered is isolation by colonization in which the populations were founded from a single centrally located ancestral source. However, this model alone wouldn't necessarily explain why the southern populations show latitudinal departures, but little deformation from longitudinal positions of populations (Fig. 2). Perhaps a non-equilibrium model in which the rate, or timing, of latitudinal spread differed from the longitudinal spread in the south could generate the observed deviations from IBD. Without further analysis, it is not possible to evaluate the likelihood of such a hypothesis. Such detailed demographic scenarios might be informed directly from the ENMs (see Fig. S2.1 in Appendix S2), including inferred areas of stability (Fig. 4), as with modelling approaches like the iDDC (He *et al.*, 2013). For example, changes in the suitability of habitats across the landscape, and changes in suitability over time, can be used to inform the colonization process associated with shifting distributions driven by glacial cycles (Brown & Knowles, 2012).

In addition to the multiple processes that might generate a departure from IBD, the magnitude and orientation of deformations in the Procrustes plots (i.e. the length of the arrows; see Fig. 2) may also be impacted by the timing of the events that cause a departure between genes and geography (e.g. Excoffier *et al.*, 2009). For example, for a recent expansion the direction of the deviations might be captured in a Procrustes analysis, but a population near the site of an expansion centre might show higher deviations relative to more geographicalally distant populations if the expansion has been recent (see simulation results in He *et al.*, 2013). Likewise, because PCA can be sensitive to the sampling of individuals over geographical space (e.g. over- or under-representative sampling for some regions; see DeGiorgio & Rosenberg, 2013), it is possible that such effects could influence some of the Procrustes analyses. We note that in general the patterns in the genetic PCs were not significantly impacted when we excluded one population at a time (see t'-values in Table S3.2 in Appendix S3). This suggests that results from analyses of the full Alaskan mammal data sets considered here are not being biased by geographical unevenness in the sampling of individuals. However, whether the results from Procrustes analyses are robust to different sample sizes across space is not known.

Does this mean that the results from Procrustes analyses have no utility for identifying the processes causing departures from IBD? Not at all – it just means that any interpretation will have to take into to account the uncertainty that would come with a single summary of genetic variation. For example, the statistical summaries from the Procrustes analyses (e.g. the t_0, t', and t''-values, as well as the angle of rotation to maximize the covariance between genes and geographical matrices) could provide valuable summary statistics for incorporation into procedures like Approximate Bayesian Computation (ABC) to test phylogeographical hypotheses. Likewise, integrated models of phylodemographic movements (e.g. iDCC; He *et al.*, 2013) may be useful in teasing apart these alternative hypotheses, especially if the differences among species discovered here are indicative of an interaction of species history and biology (e.g. Massatti & Knowles, 2014; Papadopoulou & Knowles, 2015b). In particular, our results hint at a possible distinction between more mesic species (such as brown lemmings) and more xeric species (such as collared pikas and hoary marmots). Brown lemmings show little geographical concordance in terms of direction of deformation relative to contemporary populations. For example, the projections onto geographical coordinates based on patterns of genetic variation do not overlap (i.e. individuals from populations form discrete clusters), and there is no concerted direction of movement as would be expected when previously glaciated habitat are colonized (Fig. 2). Other work has suggested that this region was a tundra mosaic (Elias *et al.*, 1996; Anderson *et al.*, 2004), which may have contributed to the lack of uniformity in the direction of deformation.

The Procrustes analyses are just the first step towards identifying future studies of genomic variation. With respect to this fascinating group of mammals, the lack of concordant genomic variation suggests there is no single geographical region in Alaska that has remained isolated geologically (i.e. a region that has remained independent of other regions) or ecologically (i.e. a barrier that prohibited historical gene flow among populations). However, some repeated patterns of variation across subsets of taxa in some parts of their ranges suggest a role for shared processes operating at more local geographical scales. Future tests will explore the hypotheses generated here, and evaluate the relative roles of taxon-specific versus regional processes in structuring genomic variation across these alpine small mammal communities.

ACKNOWLEDGEMENTS

This work was funded by NSF (DEB 1118815 to L.L.K., DDIG DEB 0808619 to L.E.O. and H.C.L.) and the Alaska Department of Fish and Game (State Wildlife Grants T-1-6, T-3-6, and T-9-1 to L.E.O.). We are also grateful to the more than 60 collectors without whose efforts this study would not have been possible.

REFERENCES

Alvarado-Serrano, D.F. & Knowles, L.L. (2014) Environmental niche models in phylogeographic studies: recent advances and precautions. *Molecular Ecology Resources*, **14**, 233–248.

Anderson, R.P. & Raza, A. (2010) The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela. *Journal of Biogeography*, **37**, 1378–1393.

Anderson, P.M., Edwards, M.E. & Brubaker, L.B. (2004) Results and paleoclimate implications of 35 years of paleoecological research in Alaska. *Developments in Quaternary Science* (ed. by A.R. Gillespie, S.C. Porter and B.F. Atwater), pp. 427–440. Elsevier, London.

Baltensperger, A.P. & Huettmann, F. (2015) Predictive spatial niche and biodiversity hotspot models for small mammal communities in Alaska: applying machine-learning to conservation planning. *Landscape Ecology*, **30**, 681–697.

Bivand, R., Keitt, T. & Rowlingson, B. (2014) rgdal: bindings for the geospatial data abstraction library. R package version 0.9-1. Available at: http://CRAN.R-project.org/package=rgdal.

Braconnot, P., Otto-Bliesner, B., Harrison, S. et al. (2007) Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum—part 1: experiments and large-scale features. *Climate of the Past*, **3**, 261–277.

Brown, J.L. (2014) SDMtoolbox: a python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses. *Methods in Ecology and Evolution*, **5**, 694–700.

Brown, J.L. & Knowles, L.L. (2012) Spatially explicit models of dynamic histories: examination of the genetic consequences of Pleistocene glaciation and recent climate change on the American Pika. *Molecular Ecology*, **21**, 3757–3775.

Catchen, J., Hohenlohe, P., Bassham, S., Amores, A. & Cresko, W.A. (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

DeGiorgio, M. & Rosenberg, N.A. (2013) Geographic sampling scheme as a determinant of the major axis of genetic variation in principal components analysis. *Molecular Biology and Evolution*, **30**, 480–488.

Elias, S.A., Short, S.K., Nelson, C.H. & Birks, H.H. (1996) Life and times of the Bering land bridge. *Nature*, **382**, 60–62.

Elith, J., Kearney, M. & Phillips, S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**, 330–342.

Excoffier, L., Foll, M. & Petit, R.J. (2009) Genetic consequences of range expansions. *Annual Review of Ecology Evolution and Systematics*, **40**, 481–501.

Francois, O., Currat, M., Ray, N., Han, E., Excoffier, L. & Novembre, J. (2010) Principal component analysis under population genetic models of range expansion and admixture. *Molecular Biology and Evolution*, **27**, 1257–1268.

Frichot, E., Schoville, S., Bouchard, G. & François, O. (2012) Correcting principal component maps for effects of spatial autocorrelation in population genetic data. *Frontiers in Genetics*, **3**, e254.

Galbreath, K.E., Cook, J.A., Eddingsaas, A.A. & DeChaine, E.G. (2011) Diversity and demography in Beringia: multilocus tests of paleodistribution models reveal the complex history of Arctic ground squirrels. *Evolution*, **65**, 1879–1896.

Gunderson, A.M., Jacobsen, B.K. & Olson, L.E. (2009) Revised distribution of the Alaska marmot, *Marmota broweri*, and confirmation of parapatry with hoary marmots. *Journal of Mammalogy*, **90**, 859–869.

He, Q., Edwards, D. & Knowles, L.L. (2013) Integrative testing of how environments from the past to the present shape genetic structure across landscapes. *Evolution*, **67**, 3386–3402.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Jombart, T., Devillard, S., Dufour, A. & Pontier, D. (2008) Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity*, **101**, 92–103.

Kerhoulas, N.J., Gunderson, A.M. & Olson, L.E. (2015) Complex history of isolation and gene flow in hoary, Olympic, and endangered Vancouver Island marmots. *Journal of Mammalogy*, **96**, 810–826.

Knowles, L.L. (2009) Statistical phylogeography. *Annual Review of Ecology Evolution and Systematics*, **40**, 593–612.

Knowles, L.L. & Alvarado-Serrano, D.F. (2010) Exploring the population genetic consequences of the colonization process with spatio-temporally explicit models: insights from coupled ecological, demographic, and genetic models in montane grasshoppers. *Molecular Ecology*, **19**, 3727–3745.

Lanier, H.C. & Olson, L.E. (2013) Deep barriers, shallow divergences: reduced phylogeographical structure in the collared pika (Mammalia: Lagomorpha: *Ochotona collaris*). *Journal of Biogeography*, **40**, 466–478.

Lanier, H.C., Gunderson, A.M., Weksler, M., Fedorov, V.B. & Olson, L.E. (2015a) Comparative phylogeography of eastern Beringian mammals highlights the double-edged sword of climate change faced by arctic- and alpine-adapted species. *PLoS ONE*, **10**, e0118396.

Lanier, H.C., Massatti, R., He, Q., Olson, L.E. & Knowles, L.L. (2015b) Colonization from divergent ancestors: glaciation signatures on contemporary patterns of genetic variation in Collared Pikas (*Ochotona collaris*). *Molecular Ecology*, **24**, 3688–3705.

Lischer, H.E.L. & Excoffier, L. (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298–299.

Massatti, R. & Knowles, L.L. (2014) Microhabitat differences impact phylogeographic concordance of co-distributed species: genomic evidence in montane sedges (*Carex* L.) from the Rocky Mountains. *Evolution*, **68**, 2833–2846.

McVean, G. (2009) A genealogical interpretation of principal components analysis. *PLoS Genetics*, **5**, e1000686.

Merow, C., Smith, M.J. & Silander, J.A. (2013) A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, **36**, 1058–1069.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., Stephens, M. & Bustamante, C.D. (2008) Genes mirror geography within Europe. *Nature*, **456**, 98–101.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Wagner, H. & Oksanen, J. (2013) Package Vegan. Community ecology package. Version 2.0–10.

Papadopoulou, A. & Knowles, L.L. (2015a) Species-specific responses to island connectivity cycles: refined models for testing phylogeographic concordance across a Mediterranean Pleistocene Aggregate Island Complex. *Molecular Ecology*, **24**, 4252–4268.

Papadopoulou, A. & Knowles, L.L. (2015b) Genomic tests of the species-pump hypothesis: recent island connectivity cycles drive divergence in Caribbean crickets across the Virgin Islands. *Evolution*, **69**, 1501–1517.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modeling*, **190**, 231–259.

Phillips, S.J., Dudik, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J. & Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181–197.

R Core Team (2014) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/. (accessed 3 January 2015)

Wang, I.J. & Bradburd, G.S. (2014) Isolation by environment. *Molecular Ecology*, **23**, 5649–5662.

Wang, C., Szpiech, Z.A., Degnan, J.H., Jakobsson, M., Pemberton, T.J. *et al.* (2010) Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Statical Applications in Genetics and Molecular Biology*, **9**, Article 13. doi: 10.2202/1544-6115.1493.

Wang, C., Zöllner, S. & Rosenberg, N.A. (2012) A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS Genetics*, **8**, e1002886. doi:10.1371/journal.pgen.1002886.

Warren, D.L. & Seifert, S.N. (2011) Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecological Applications*, **21**, 335–342.

Warren, D.L., Glor, R.E. & Turelli, M. (2010) ENMTools: a toolbox for comparative studies of environmental niche models. *Ecography*, **33**, 607–611.

Weksler, M., Lanier, H.C. & Olson, L.E. (2010) Eastern Beringian biogeography: historical and spatial genetic structure of singing voles in Alaska. *Journal of Biogeography*, **37**, 1414–1431.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Summaries of geographical information and genomic sampling.
Appendix S2 Summaries of ENM settings and projections of current and LGM distributions.
Appendix S3 PC maps of genetic variation and summaries of genetic variation.

BIOSKETCH

L. Lacey Knowles and her lab are interested in understanding the processes that structure patterns of genetic variation across geographical landscapes and among taxa. Her lab works on a diversity of empirical systems to discover how species-specific responses to past events, especially those caused by climate change, influence the connections among populations that shape divergence patterns over space and time. This work is also complemented by methodological study and development to identify approaches that are useful for making inferences about the processes that shape genetic patterns within and among taxa.

Author contributions: L.L.K., Q.H. and R.M. conceived the ideas; L.L.K. led the writing with input from all the other co-authors; L.E.O. and H.C.L. collected the specimens; H.C.L. and Q.H. collected the genomic data; Q.H. conducted bioinformatics processing of genomic data; R.M. did the genetic analyses and ENMs.

Editor: Brett Riddle