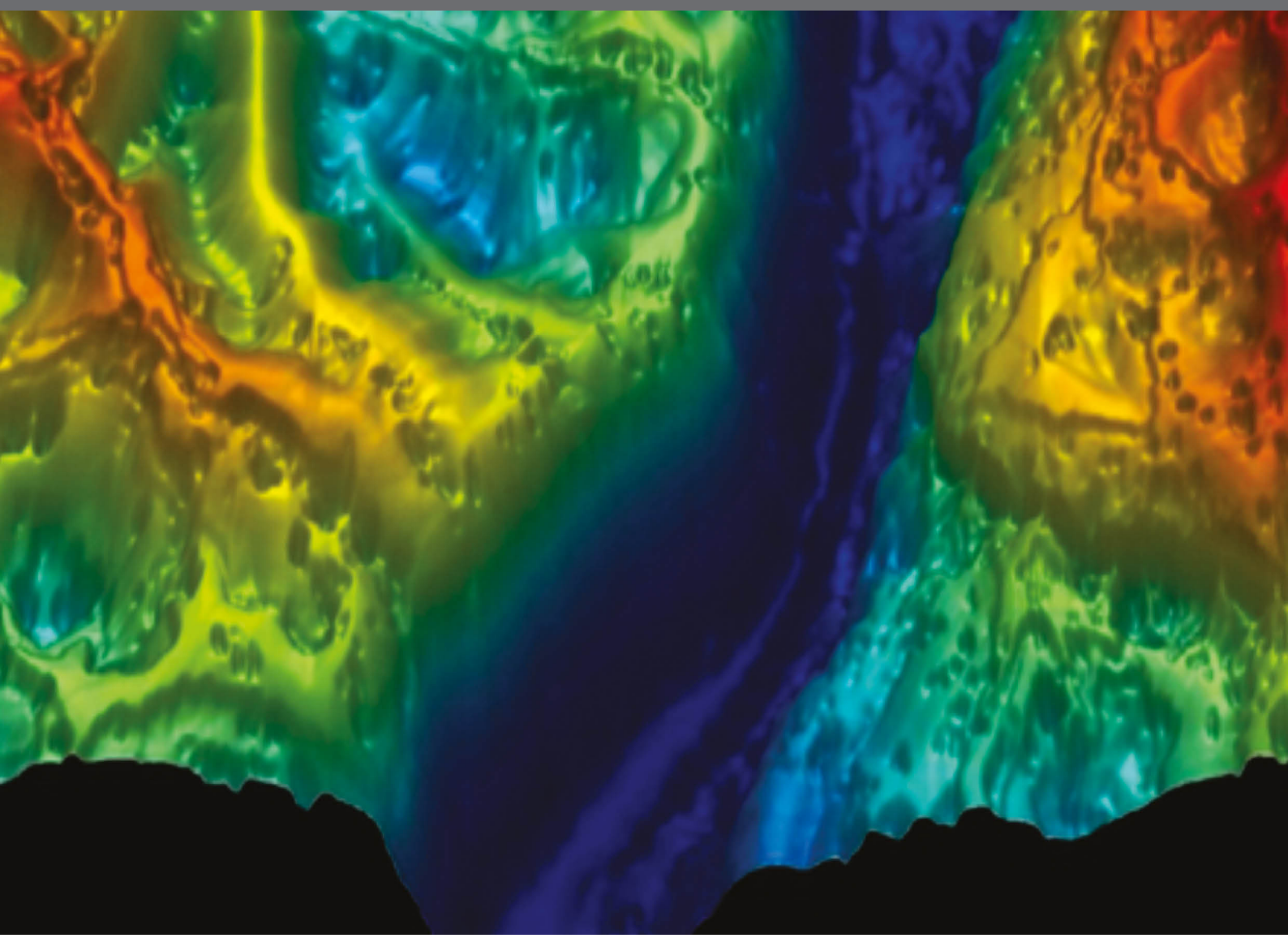
THE LEAST COST PATH FROM LANDSCAPE GENETICS TO LANDSCAPE GENOMICS

EDITED BY: Samuel A. Cushman, Andrew J. Shirk, Glenn T. Howe, Melanie A. Murphy, Rodney J. Dyer and Stéphane Joost

PUBLISHED IN: Frontiers in Genetics, Frontiers in Plant Science and Frontiers in Ecology and Evolution



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88945-548-5
DOI 10.3389/978-2-88945-548-5

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

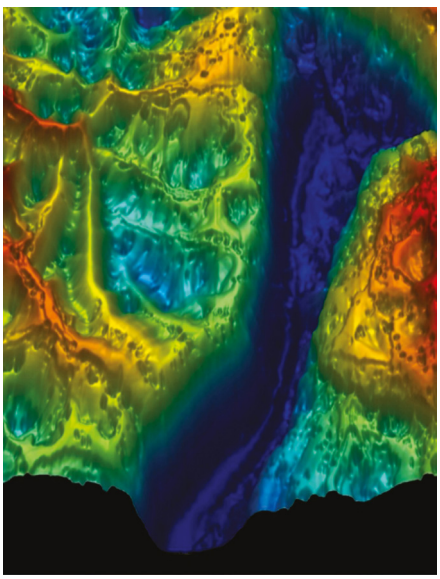
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

THE LEAST COST PATH FROM LANDSCAPE GENETICS TO LANDSCAPE GENOMICS

Topic Editors:
Samuel A. Cushman, USDA Forest Service, Rocky Mountain Research Station, United States
Andrew J. Shirk, Climate Impacts Group, University of Washington, United States
Glenn T. Howe, Oregon State University, United States
Melanie A. Murphy, University of Wyoming, United States
Rodney J. Dyer, Virginia Commonwealth University, United States
Stéphane Joost, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland



Cover image by Samuel A. Cushman

Ecosystems are the stage on which the play of evolution is acted, and ecosystems are complex, spatially structured and temporally varying. The purpose of this Research Topic is to explore critical challenges and opportunities for the transition from landscape genetics to landscape genomics. Landscape genetics has focused on the spatial analysis of small genetic datasets, typically comprised of less than 20 microsatellite markers, taken from clusters of individuals in putative populations or distributed individuals across landscapes. The recent emergence of large scale genomic datasets produced by next generation sequencing methods poses tremendous challenge and opportunity to the field. Perhaps the greatest is to produce, process, curate, archive and analyze spatially referenced genomic datasets in a way such that research is led by a priori hypotheses regarding how environmental heterogeneity and temporal dynamics interact to affect gene flow and selection. The papers in the Research Topic cover a broad range of topics under this area of focus, from reviews of the emergence of landscape genetics, to best practices in spatial analysis of genetic data. The compilation, like the emerging field itself, is eclectic and illustrates the scope of both the challenges and opportunities of this emerging field.

Citation: Cushman, S. A., Shirk, A. J., Howe, G. T., Murphy, M. A., Dyer, R. J., Joost, S., eds. (2018). The Least Cost Path From Landscape Genetics to Landscape Genomics. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-548-5

Table of Contents

frontiers in Genetics

EDITORIAL
published: 19 June 2018
doi: 10.3389/fgene.2018.00215

Editorial: The Least Cost Path From Landscape Genetics to Landscape Genomics: Challenges and Opportunities to Explore NGS Data in a Spatially Explicit Context

Samuel A. Cushman[1], Andrew J. Shirk[2], Glenn T. Howe[3], Melanie A. Murphy[4], Rodney J. Dyer[5] and Stéphane Joost[6]*

[1] USDA Forest Service, Rocky Mountain Research Station, Flagstaff, AZ, United States, [2] Climate Impacts Group, University of Washington, Seattle, WA, United States, [3] Forest Ecosystems and Society, Oregon State University, Corvallis, OR, United States, [4] Ecosystem Science and Management, University of Wyoming, Laramie, WY, United States, [5] Center for Environmental Science, Virginia Commonwealth University, Richmond, VA, United States, [6] Laboratory of Geographic Information Systems (LASIG), School of Architecture, Civil and Environmental Engineering (ENAC), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Keywords: landscape genetics, landscape genomics, next generation sequencing, spatial genetics, ecological genetics

The Editorial on the Research Topic

The Least Cost Path From Landscape Genetics to Landscape Genomics: Challenges and Opportunities to Explore NGS Data in a Spatially Explicit Context

Ecosystems are the stage on which the play of evolution is acted. Inferring evolutionary processes from the spatial and temporal genetic patterns they produce in populations is challenging because ecosystems are highly complex, spatially structured, and temporally varying. The field of landscape genetics has offered a means of navigating these challenges to make eco-evolutionary insights for many species. The emerging field of landscape genomics offers great promise to expand the potential of landscape genetic analysis even further. The purpose of this Research Topic for Evolutionary and Population Genetics is to explore a number of critical challenges and opportunities for the transition from landscape genetics to landscape genomics. To-date, landscape genetics has generally focused on spatial analyses of small genetic datasets, typically comprised of <20 microsatellite markers, taken from clusters of individuals in putative "populations" or distributed individuals across landscapes. The recent emergence of large-scale genomic datasets containing thousands of markers produced by next generation sequencing (NGS) methods poses tremendous opportunity and challenge to the field. Perhaps the greatest is to produce, process, curate, archive, and analyze spatially referenced genomic datasets in a way such that research is led by a priori hypotheses about how environmental heterogeneity and temporal dynamics interact to influence gene flow and selection. Effective progress in this transition to a robust field of landscape genomics will likely depend on integrating vast genomic datasets with powerful modeling and replicated and controlled experiments to test putative relationships between population processes and evolutionary and population genetic responses (Cushman, 2014). The recent availability of whole genome sequence (WGS) data offers incredible molecular resolution, but comes at great expense. This limits the spatial and temporal sample sizes for economic reasons, making it challenging to achieve spatial representativeness and temporal robustness.

OPEN ACCESS

Edited and reviewed by:
Norman A. Johnson, University of Massachusetts Amherst, United States

****Correspondence:***
Samuel A. Cushman
scushman@fs.fed.us

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *16 April 2018*
Accepted: *28 May 2018*
Published: *19 June 2018*

Citation:
Cushman SA, Shirk AJ, Howe GT, Murphy MA, Dyer RJ and Joost S (2018) Editorial: The Least Cost Path From Landscape Genetics to Landscape Genomics: Challenges and Opportunities to Explore NGS Data in a Spatially Explicit Context. Front. Genet. 9:215. doi: 10.3389/fgene.2018.00215

No single person has the expertise or the time to effectively bring these components together. More than ever, success in advancing our field will depend on collaborations across large multi-disciplinary groups (Cushman, 2014). Experts in the development of genomic, epigenomic, and transcriptomic data from high throughput technologies are needed to produce the genome-wide raw data for subsequent analysis. Bioinformatics specialists are needed to provide programming and computer science expertise to efficiently handle and analyze vast genomic datasets, and to effectively utilize high performance computing resources. Modelers will be needed to work with the bioinformaticians to explore the implications of hypotheses a priori, to refine hypotheses by optimizing fit to observed data, and predict how observed pattern-process relationships may propagate across scale through space and time. Experimenters should work closely with modelers to rigorously test hypotheses in controlled and replicated experiments. To be successful, this entire integration should be led by theoreticians who have a coherent vision for how each of these parts will synergize to address focused and falsifiable questions of importance in advancing the field.

In this research topic we recruited a number of leading experts in genomics, epigenetics, landscape genetics, and simulation modeling to explore the challenges and opportunities presented by the intersection of NGS data, spatial modeling, and replicated and controlled experimentation. Overall, this effort produced a series of 10 papers, not including this editorial. These papers covered a wide range of topics including (1) two reviews of recent developments and current status of landscape genomics, (2) one review of theory and mechanisms of epigenetics and their applications in a landscape genomic context, (3) two papers illustrating the cutting edge in individual-based, spatially-explicit simulation modeling applied to eco-evolutionary problems in landscape genomics, (4) one paper about using genetic rare variants, or singletons, to infer past demographic events over a species' history, (5) three empirical papers describing a range of analytical methods to explore the spatial and environmental drivers of selection and genetic differentiation in plants and animals, and (6) one paper focused on the landscape side of "landscape genomics" which provides a review and evaluation of best practices of using geographical information systems to compile, display and analyze environmental data in the most appropriate way for landscape genomic research.

Landscape genomics is at the exciting cutting edge of the recent spatial revolution that has led to the emergence of the field of landscape genetics. Given the recency of landscape genomics as a field of study, there are relatively few established research frameworks, analytical approaches or even conceptual models for what is meant by landscape genomics and how it is best conducted (see Balkenhol et al., 2017). The two reviews of recent landscape genomics literature in this Research Topic attempt to summarize the field as it stands now and identify its strengths, weaknesses and opportunities. In the first of these two review papers, Li et al. define landscape genomics as a new discipline that aims to reveal relationships between adaptive genetic variation and environmental heterogeneity, and note that there have been few formal landscape genomics papers published to date. Their review outlines the sampling strategies, molecular marker types and research categories in 37 articles published during the first 10 years of this field, and identifies major challenges and future directions for landscape genomics. The second review, by Storfer et al., emphasizes the role of emerging genomic technology in driving the emergence of landscape genomics as a field of study. In particular, they note that widely available next-generation sequencing data have resulted in immensely improved ability to detect candidate genes under selection and identify the environmental factors that drive that selection. However, they note that the transition between landscape genetics and landscape genomics is extremely challenging due to the difficulty of handling and interpreting vast genomic datasets. They also note the rapid emergence of a wide range of analysis methods and provide detailed discussion of outlier differentiation methods and genetic-environment association tests. They note that the key to choosing appropriate genome scan methods is an understanding of the underlying demographic structure of study populations, and such data can be obtained using neutral loci from the generated genome-wide data or prior knowledge of a species' phylogeographic history and summarize recent simulation studies that test the power and accuracy of genome scan methods under a variety of demographic scenarios and sampling designs. They conclude with a discussion of additional considerations for future method development, and a summary of methods that show promise for landscape genomics studies but are not yet widely used. These two reviews provide what is probably the most complete snap-shot of the field of landscape genomics produced to-date, and propose an excellent foundation for the more theoretical papers in the Research Topic as well as context for the papers that present empirical examples of current landscape genomic research.

Epigenetics has recently emerged as a topic of immense interest in evolutionary biology. Up to this time, landscape genetics and landscape genomics research has focused on sequence genetic variation in relation to natural gene flow and adaptive variation. However, it is appearing increasingly likely that a large portion of the variance in evolutionary responses is related not to variation in genomic sequences but to epigenetic regulation of the expression of those sequences. Fitness-related traits can be affected by heritable variation in epigenetic marks, resulting in transgenerational plasticity. Given the importance of epigenetics in evolutionary biology, it is critical to begin the integration of epigenetics with landscape genetics and landscape genomics (e.g., Paun et al., 2010). Whipple and Holeski take an exciting first step in this effort with their review of epigenetic theory and mechanisms and their relationships with landscape genomics and landscape genetics. In their paper they summarize the relevance of epigenetic inheritance to ecological and evolutionary processes, and review the literature on landscape-level patterns of epigenetic variation. They argue that landscape-level patterns of epigenomic variation in plants generally show greater levels of isolation by distance and isolation by environment than is found for the genome, suggesting a perhaps elevated role in the spatial population processes that are the focus of landscape genetics and genomics. They note that demonstrating transgenerational inheritance requires more

complex breeding and/or experimental designs, and argue that multi-generation common garden experiments conducted across multiple environments are required to understand epigenome inheritance and to separate the relative contributions of heritable epigenetic variation to the phenotype.

The two papers in the Research Topic that focus on individual-based, spatially-explicit simulation modeling of eco-evolutionary processes offer a tantalizing glimpse into the exciting emerging field of landscape genomic simulation modeling. In the first of these papers, Landguth et al. present the first application to a real-world ecological system of a new individual-based simulation model that incorporates spatially complex gene flow and spatially heterogeneous environmentally driven selection. They use the recent population declines to the high elevation western North America foundation species whitebark pine as a case study to illustrate the power of this modeling framework. Specifically, they present a simulation modeling framework to improve understanding of the long-term genetic consequences of the blister rust pathogen, the evolution of rust resistance, and scenarios of planting rust resistant genotypes of whitebark pine. By combining climatic niche modeling and eco-evolutionary landscape genetics modeling, they evaluate the effects of different scenarios of planting rust-resistant genotypes and impacts of wind field direction on patterns of gene flow. As such, Landguth et al. is the first paper to combine empirical data, experimentation, and large-scale population-wide simulation modeling of adaptive evolution in spatially-complex landscapes. The second simulation paper, by Cushman and Landguth uses the same individual-based, spatially-explicit modeling approach to explore the interactions of heterogeneous environmental selection with speciation driven by hybrid incompatibility. Within-species hybrid incompatibility arises when combinations of alleles at more than one locus have low fitness but where possession of one of those alleles has little or no fitness consequence for the carriers. In this paper, Cushman and Landguthuse simulation modeling to explore the effects of heterogeneous natural selection on the frequency, size and duration of reproductively isolated clusters of individuals in continuously distributed populations. They found that spatially heterogeneous selection produced clusters of reproductively isolated individuals that were much larger, longer lasting and spatially proximal. This pattern was strong across levels of gene flow and strength of selection, suggesting that even relatively weak selection acting in the context of strong gene flow may produce reproductively isolated clusters that are large and persistent, enabling incipient speciation in a continuous population without geographic isolation.

Another important topic in evolutionary theory and spatial genetics relates to the effects of past demographic events in species history on current patterns of genetic structure and differentiation. To address this issue, Cubry et al. argue that rare variants are important for drawing inference about past demographic events in a species' history, and specifically that singletons, which are variants for which genetic variation is carried by a unique chromosome in a sample, provide a particularly powerful lens to explore deep demographic history and its impacts on current population structure. They define the empirical distribution of singletons and then use computer simulations to evaluate the potential for the empirical distribution of singletons to provide a description of genetic diversity across geographic space. Using a Bayesian framework, they then show that this measure leads to accurate estimates of the geographic origin of range expansions and use this approach to estimate the origin of a cultivated plant species. Ultimately, this paper demonstrates that the empirical distribution of singletons is a useful measure to analyze results of sequencing projects based on large scale sampling of individuals across geographic space.

The three empirical case studies address two crop plants and one wild mammal species. In the first plant-based empirical example, Egea et al. explore the genomics of garlic. They use high-throughput genotyping-by-sequencing approaches to assess genetic diversity and structure of a large garlic-germplasm bank, relate genotypes to agronomical history and develop a cost-effective method to manage genetic diversity in germplasm banks. They identified three main garlic-groups and demonstrated that DArTseq is a cost-effective method to analyze species with large and expected complex genomes, like garlic. In the second plant-based empirical study, Abebe et al. focused on detecting adaptive loci in barley. They also used a genotyping by sequencing approach on a diverse population of barley landraces and compared genomic structure to climatic data. Partitioning the variance between climate variables and geographic distance indicated that climate variables accounted for most of the explainable genetic variation, and analysis of the associated SNPs revealed putative candidate genes for plant adaptation. This study highlights the utility of landscape genomic approaches to detect the presence of putative adaptive loci among barley landraces. The final empirical case study Zero et al.) focuses on how the persistence of small populations is influenced by genetic structure and functional connectivity. The authors used two network-based approaches to understand the persistence of the northern Idaho ground squirrel (Urocitellus brunneus) and the southern Idaho ground squirrel (U. endemicus), two rare species. They found that population graph analyses revealed that local extinction rapidly reduced connectivity for the southern species, while connectivity for the northern species could be maintained following local extinction. Results from gravity models complemented those of population graph analyses and indicated that potential productivity and large-scale topographic features drove connectivity in the northern species. The paper is one of the very first examples of using scenario analysis in landscape genetics to inform conservation strategies of other species exhibiting patchy distributions.

The final paper in the Research Topic addresses spatial analysis itself. There are two components of landscape genomics: landscape analysis and genetic data. However, a large majority of work has focused primarily on the genetic data component of the field, and much less on methods, theory and best practices in spatial analysis. Obtaining reliable knowledge about the pattern-process relationships that govern population demographics and evolution in complex environments requires rigorous approaches to link genetic, genomic, and epigenetic data to environmental and spatial drivers. To begin to address this critical need, Leempoel et al. explore the use of Geographic Information

Systems (GIS) in landscape genetics and landscape genomics. They note that GIS is a tool that is uniquely suited to overlaying genetic information with environmental data, which is the prerequisite to locate and analyze genetic boundaries of various plant and animal species or to study gene-environment associations (GEA). Their paper focuses on the power of free and open-source GIS approaches and provide essential information for their successful application in molecular ecology. The paper provides a useful introduction to the key concepts related to GIS and then presents an overview of open-source GIS-related software, file formats, major environmental databases. Then the authors focus on GIS applications in landscape genetics, such as sampling strategies for Next Generation Sequencing, data exploration and spatial statistics suited for the analysis of large genetic datasets, and provide suggestions to properly edit maps and to make them as comprehensive as possible.

The overall goal for this Research Topic was to produce a concentrated compilation of the current thinking, methods, and perspectives in the emerging field of landscape genomics. In that regard, the mixture of review papers, simulation modeling advances, empirical examples and methodological approaches, we hope, will serve the reader well as a broad, current overview of this field. We truly feel there are few subjects that can claim to have an equal degree of synergy and rapidity of development as landscape genomics. The collision of explosive advances in genomic data generation with powerful individual-based simulation modeling approaches, and their integration with experimental genetics studies, provides an incredibly powerful synergy that is transforming entire fields of genetics, ecology and conservation. We hope this Research Topic will serve in some small way to advance this exciting growth of knowledge.

Looking forward, we believe that advancing landscape genomics will depend on formally linking genomic datasets with modeling and experimentation (Cushman, 2014). The papers in this Research Topic provide some initial insight into the challenges of this integration and the current state of development in its several parts. Given that no single person has the expertise to effectively bring these components together, success in advancing our field will depend on collaborations across large multi-disciplinary groups. The broad range of topics and expertise represented in this Research Topic may be seen as the nucleus of such a cross-disciplinary effort at integration, but clearly there is a tremendous amount to be done and this initial step has, more than anything, revealed that. Experts genomic, epigenomic, and transcriptomic data must work with bioinformatics specialists to efficiently handle and analyze vast genomic datasets, and to effectively utilize high performance computing resources. Modelers and experimental geneticist must work collaboratively with the bioinformaticians and genomics experts to test hypotheses in controlled and replicated experiments and project the relationships identified into broad and complex landscaps in a rapidly changing world. Accelerating global change presents a tremendous threat to the biosphere and challenge to human civilization. Landscape genomics will provide extremely valuable tools and approaches to understand, predict and mitigate the negative effects of global change on biodiversity, but only if it progresses rapidly to integrate genomic data, spatial modeling and experimental genetics.

AUTHOR CONTRIBUTIONS

All authors co-edited this Research Topic. SC wrote the first draft of this editorial, and the other co-authors provided edits and revisions.

REFERENCES

Balkenhol, N., Dudaniec, R. Y., Krutovsky, K. V., Johnson, J. S., Cairns, D. M., Segelbacher, G., et al. (2017). *Landscape Genomics: Understanding Relationships between Environmental Heterogeneity and Genomic Characteristics of Populations*. Cham: Springer.

Cushman, S. A. (2014). Grand challenges in evolutionary and population genetics: the importance of integrating epigenetics, genomics, modeling, and experimentation. *Front. Genet.* 5:197. doi: 10.3389/fgene.2014.00197

Paun, O., Bateman, R. M., Fay, M. F., Hedrén, M., Civeyrel, L., and Chase, M. W. (2010). Stable epigenetic effects impact adaptation in allopolyploid orchids (Dactylorhiza: Orchidaceae). *Mol. Biol. Evol.* 27, 2465–2473. doi: 10.1093/molbev/msq150

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

REVIEW
published: 13 March 2018
doi: 10.3389/fgene.2018.00068

Navigating the Interface Between Landscape Genetics and Landscape Genomics

Andrew Storfer, Austin Patton and Alexandra K. Fraik*

School of Biological Sciences, Washington State University, Pullman, WA, United States

As next-generation sequencing data become increasingly available for non-model organisms, a shift has occurred in the focus of studies of the geographic distribution of genetic variation. Whereas landscape genetics studies primarily focus on testing the effects of landscape variables on gene flow and genetic population structure, landscape genomics studies focus on detecting candidate genes under selection that indicate possible local adaptation. Navigating the transition between landscape genomics and landscape genetics can be challenging. The number of molecular markers analyzed has shifted from what used to be a few dozen loci to thousands of loci and even full genomes. Although genome scale data can be separated into sets of neutral loci for analyses of gene flow and population structure and putative loci under selection for inference of local adaptation, there are inherent differences in the questions that are addressed in the two study frameworks. We discuss these differences and their implications for study design, marker choice and downstream analysis methods. Similar to the rapid proliferation of analysis methods in the early development of landscape genetics, new analytical methods for detection of selection in landscape genomics studies are burgeoning. We focus on genome scan methods for detection of selection, and in particular, outlier differentiation methods and genetic-environment association tests because they are the most widely used. Use of genome scan methods requires an understanding of the potential mismatches between the biology of a species and assumptions inherent in analytical methods used, which can lead to high false positive rates of detected loci under selection. Key to choosing appropriate genome scan methods is an understanding of the underlying demographic structure of study populations, and such data can be obtained using neutral loci from the generated genome-wide data or prior knowledge of a species' phylogeographic history. To this end, we summarize recent simulation studies that test the power and accuracy of genome scan methods under a variety of demographic scenarios and sampling designs. We conclude with a discussion of additional considerations for future method development, and a summary of methods that show promise for landscape genomics studies but are not yet widely used.

Keywords: landscape genomics, landscape genetics, local adaptation, selection, spatial analyses

OPEN ACCESS

Edited by:
Samuel A. Cushman, United States Forest Service (USDA), United States

Reviewed by:
Pablo Orozco-terWengel, Cardiff University, United Kingdom
Clinton Wakefield Epps, Oregon State University, United States
Paul F. Gugger, University of Maryland Center for Environmental Sciences, United States

****Correspondence:***
Andrew Storfer
astorfer@wsu.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *18 October 2017*
Accepted: *15 February 2018*
Published: *13 March 2018*

Citation:
Storfer A, Patton A and Fraik AK (2018) Navigating the Interface Between Landscape Genetics and Landscape Genomics. Front. Genet. 9:68. doi: 10.3389/fgene.2018.00068

INTRODUCTION

Understanding the spatial distribution of adaptive genetic variation is at the very core of evolutionary biology and population genetics. Recent advances in next-generation sequencing make studies of the genomic basis of local adaptation now possible for virtually any organism. Simultaneously, spatial data for nearly every corner of the Earth are available due to dramatic increases in GIS and mapping technologies. These technological developments have led to the rapid proliferation of studies that integrate geographic and genomic data to test for spatial patterns of genes under selection, collectively termed "landscape genomics" (Joost et al., 2007; Lowry, 2010; Manel et al., 2010).

Landscape genomics stems from landscape genetics, an explicitly spatial suite of analysis methods that focus on testing the influence of landscape features on genetic population structure (Manel et al., 2003; Storfer et al., 2007; Manel and Holderegger, 2013). The transition from landscape genetics to landscape genomics has come with the shift from utilizing a dozen or so loci (often microsatellites) to thousands and even millions of loci (often single nucleotide polymorphisms-SNPs)—and even complete transcriptomes or genomes—in studies of spatial genetic variation.

Is landscape genomics just landscape genetics with more loci? In the original article that coined the term "landscape genetics," Manel et al. (2003) state that, "*Dozens of markers are available for numerous taxa*" and that "*identification of loci under selection can help us understand the genetic basis of local adaptation…*" (p. 190). However, except for candidate gene approaches, where *a priori* information about the function of specific genes is known, dozens of markers are generally insufficient for tests of selection; such tests commonly rely on orders of magnitude more loci to have appropriate statistical power to conduct outlier analyses (Luikart et al., 2003; Pritchard and Di Rienzo, 2010) or genotype-environment associations (i.e., GEAs, Rellstab et al., 2015). As a result, the literature commonly refers to landscape genomics studies as those that (have the power to) focus on describing spatial patterns of selection and adaptation, whereas landscape genetics studies primarily focus on the influence of landscape variables on gene flow (Rellstab et al., 2015; Haasl and Payseur, 2016).

Semantics aside, scientists are now awash with data, and analytical methods have lagged behind our ability to generate massive data sets. The shift from analyzing dozens to thousands to millions of markers (and even whole genomes) brings about new computational challenges. Whereas landscape genetics relies upon a rich history of spatial statistics dating back to the 1950s and 1960s, genome-wide selection analyses have primarily been developed in the last decade. New methods are rapidly being developed, and embarking on a landscape genomics study may seem like a daunting task for some researchers. Here, we aim to disentangle some of the complexity involved in conducting a landscape genomics study and associated downstream analyses, and we hope to offer some perspective for novice and experienced researcher alike. We focus primarily on marker-based studies of non-model organisms, as it is in these systems that landscape genomics studies are most rapidly expanding. Additionally, inference in non-model organisms is particularly challenging as they lack the genomic tools such as reference genomes and transcriptomes, which are typically available for model systems (Manel et al., 2010; Storfer, 2015). We emphasize that this piece is not meant to be an exhaustive review of the subject, as many substantial articles have already been published to this effect (e.g., Haasl and Payseur, 2016; Hoban et al., 2016; Rellstab et al., 2016). Rather, we provide a brief guide to navigate this new and rapidly changing field and in the following sections, we focus on: (1) study design; (2) data generation; (3) analysis methods and associated challenges; (4) methods at the interface of landscape genetics and landscape genomics; and, (5) future directions.

STUDY DESIGN

Early work in landscape genetics went through an exploratory phase, where sampling was geographically widespread and involved testing the effects of various landscape variables on gene flow and population genetic structure (Storfer et al., 2007, 2010). Similarly, early landscape genomics studies lacked specific hypotheses and were designed to take an unbiased approach to search for candidate loci across populations that differed in key environmental variables (e.g., altitude; Haasl and Payseur, 2016). Instead of using candidate gene or QTL approaches (Stinchcombe and Hoekstra, 2008), tests for selection were conducted across a suite of loci spread throughout the genome without *a priori* information about putative function. High false positive rates are perhaps the most significant problem with landscape genomics studies that rely on genome scans (Lotterhos and Whitlock, 2014, 2015; Rellstab et al., 2015; Haasl and Payseur, 2016), and this is further exacerbated without *a priori* hypotheses. Studies that lack specific hypotheses are prone to choose candidate loci with the strongest associations with environmental variables, with a reasonable chance of detecting spurious result(s). One way to identify false positives is that loci in close proximity do not show a signature of selection. Even if loci detected in such analyses are "true" positives, the function of the candidate loci remains unknown, particularly when lacking a reference genome and thus the ability to map a candidate locus (Pavlidis et al., 2012). Even when a candidate is in linkage disequilibrium with a gene of known function, downstream functional verification may be necessary. Thus, landscape genomics studies should aim to be hypothesis-driven, because inference is stronger when there is documented variation in phenotypes or other specific information that provides evidence of spatial variation in local adaptation among populations.

It is also important to note that landscape genomics studies can test for candidate genes underlying local adaptation, as well as the effects of landscape variables on gene flow. That is, the large number of loci generated for landscape genomics studies can be partitioned into sets of loci that are putatively neutral and those that are putatively under selection, with the former being used to test spatial patterns of gene flow and population structure. Note, however, that sampling designs for assessing population genetic

structure and testing for loci under selection have important similarities and differences (**Table 1**).

For both landscape genetics and landscape genomics studies, choosing an appropriate spatial scale for a proposed study area is extremely important. In general, the extent of the study area and spacing of demes within that study area should match the spatial scale of dispersal and thereby the likely scale of environmentally-mediated selection for the study species (Anderson et al., 2010; Richardson et al., 2014; Rellstab et al., 2015; Hoban et al., 2016). Additionally, the resolution of the environmental data should be appropriate for the study species (e.g., sampling at a 2.5 km scale would be inappropriate for a slug species; Anderson et al., 2010). Also, GIS layers chosen for each study should be those deemed to be those most reasonable based on the ecology of the study species and what is known regarding habitat use. However, researchers should be aware that many environmental layers available for analysis in a GIS tend to be multicollinear (e.g., various temperature measures, such as seasonality and maximum temperature). Without some reduction of the number of variables (e.g., via ordination such as PCA), significant relationships between detected between environmental variables and allele frequencies may be spurious and/ or correlated with the true variables. Alternatively, problems with multicollinearity can be avoided by selecting one environmental variable as a representative of each correlated set (e.g., Trumbo et al., 2013). An overview of the use of GIS in landscape genomics studies is provided in Leempoel et al. (2017).

A key difference between landscape studies of gene flow and those designed to detect selection is regarding design of spatial sampling (**Table 1**). For example, in landscape genetics, when testing hypotheses about effects of a specific environmental variable such as precipitation on population genetic structure, a stratified random design is often preferred (Storfer et al., 2010). In contrast, landscape genomics simulations have repeatedly emphasized that replicated sampling of environmental extremes hypothesized to drive selection (e.g., high and low altitude) results in higher power to detect candidate loci under selection than random sampling or transect designs (De Mita et al., 2013; Lotterhos and Whitlock, 2014, 2015; Rellstab et al., 2015; Stucki et al., 2016; see also **Table 2**). Nonetheless, transect sampling can be appropriate when populations are expected to be maladapted to extremes, but locally adapted to intermediate conditions (Lotterhos and Whitlock, 2015). Sampling transects can also be useful when sampling across a zone of introgression or when geographic clinal analyses are to be employed (see Section Clinal Analyses). Thus, an important distinction to note between landscape genetics and landscape genomics studies is that the former involves study designs that tend to focus on sampling across environmental variation that should influence gene flow, whereas the latter should most often be designed

TABLE 1 | General differences between landscape genetics and landscape genomics studies.

	Questions	Scale of study	Sampling design	Analysis methods
Landscape genetics	Influence of landscape on gene flow	Among populations	**Stratified random**, opportunistic, clumped, individual-level	Mantel tests, *Assignment tests* (spatial and aspatial; e.g., Structure, Tess, Geneland), *Ordination (dbRDA, sPCA. MDS)*, Least cost paths (multiple regression, MLPE), Spatial autocorrelation, Spatial regression, EEMS*
	Influences of landscape on at-site variation	Within populations	**Across ecological gradients**, stratified	Graph models (e.g, Popgraph), GDMs, Structural equation models
	Barriers	Among populations	**Across hypothesized barrier(s)**	Wombling, Monmonier's maximum difference algorithm, spatial assignment tests (e.g., Geneland)
	Species' ecology	Within and among populations	**Across ecological gradients** (stratified)	Ordination, Least cost paths, Spatial autocorrelation, Spatial regression
	Source-sink dynamics	Among populations	**Across populations of different sizes or fragmentation levels**	Mantel tests, genetic diversity estimates (e.g., F-statistics, bottleneck tests)
Landscape genomics	Spatial patterns of selection	Among populations	**Paired sampling**, transect sampling	Outlier differentiation methods (eg., Bayescan, FLK, X^TX); Genotype-environment associations (e.g., Bayenv2, PC Adapt, LFMM, sGLMM, Samβada), *Ordination*, *Assignment tests* (e.g., FASTSTRUCTURE, Admixture, Tess3)
	Influence of landscape on local adaptation	Among populations	Transect sampling, paired sampling, stratified sampling	Outlier differentiation methods; Genotype-environment associations, *Ordination*, *Assignment tests*, Genomic cline analysis*, GDM*, EEMS*

*Note that, when conducting a landscape genomics study, that when loci under selection are removed and putatively neutral loci remain, that landscape genetics questions and analyses can then be conducted. Nonetheless, sampling designs generally differ between landscape genetics and landscape genomics studies, so some landscape genetics questions may not be addressable in studies with landscape genetics goals. Bolded sampling designs indicate preferred designs for that particular question. Not all analysis methods under each study type are listed, just those that are most commonly used or best suited to address the goals of the study. Note also that assignment test methods generally differ between landscape genetics and landscape genomics studies. Italicized words under analysis type indicate those commonly used in both landscape genetics studies of gene flow and landscape genomics studies of loci involved in adaptation. dbRDA, distance-based redundancy analyses; sPCA, spatial principal components analysis; MDS, multidimensional scaling; MLPE, maximum likelihood of population effects (Clarke et al., 2002); LFMM, latent factor mixed models; sGLMM, spatial generalized linear mixed models; EEMS, Estimated Effective Migration Surface (Petkova et al., 2016). Software names include: Geneland (Guillot et al., 2005), Structure (Pritchard et al., 2000), Tess (Durand et al., 2009), Popgraph (Dyer and Nason, 2004); Bayescan Foll and Gaggiotti, 2008, FLK (Bonhomme et al., 2010), Bayenv2 (Günther and Coop, 2013), PCadapt (Duforet-Frebourg et al., 2014) Faststructure (Raj et al., 2014), Admixture (Alexander et al., 2009), Tess3 (Caye et al., 2016). *indicates methods not yet widely used but show promise–see Sections Generalized Dissimilarity Modeling (GDM)–Clinal Analyses.*

to sample replicated pairs of populations that experience the same environmental extremes. Replication also helps reduce the chance that candidate loci under selection are false positives; loci detected repeatedly across different environments are less likely to result from confounding effects of population structure or environmental covariances (Rellstab et al., 2015).

With limited resources, researchers generally face a tradeoff between the total number of samples and the total number of localities that can be sampled in genetics studies of natural populations. Landscape genetics study designs often focus on maximizing the number of individuals per location to obtain accurate allele frequency estimates (Storfer et al., 2010; Manel and Holderegger, 2013). Most landscape genetics analyses are genetic distance-based, and inaccurate estimates of allele frequencies can bias gene flow estimates (Storfer et al., 2007, 2010). While replication of sites or transects is favored for reasons above in landscape genomics studies, the balance between sample size and number of sites depends on downstream analysis type. Power is generally limited by the total number of samples collected in landscape genomics studies (Lotterhos and Whitlock, 2015). Indeed, it is important to sample a sufficient number (e.g., > 10) of individuals per locality to generate accurate allele frequency estimates for analyses that rely on estimates of genetic differentiation among populations (i.e., differentiation outlier analyses below). However, optimizing the number of population pairs sampled (with smaller sample sizes per location) can be robust for detecting selection when sampling locations represent a range of environmental variable values across the study area (De Mita et al., 2013; **Table 2**).

DATA GENERATION

Initially, landscape genomics studies expanded from microsatellites commonly employed in landscape genetics studies to a few hundred AFLPs (amplified fragment-length polymorphisms; Joost et al., 2007). Currently, landscape genomics studies typically rely on genome-wide SNP marker sets generated using short-read next generation sequencing technologies (e.g., Illumina). Perhaps the most widely used of such reduced-representation approaches in the last few years is RAD-seq (restriction-associated digest DNA sequencing; Andrews et al., 2016; Lowry et al., 2017). RAD-seq is particularly appealing because it does not rely on availability of a reference genome. In short, whole genomic DNA is cut into fragments using a restriction enzyme, sequencing bar codes are ligated to restriction sites, individuals are bar-coded and fragments are sequenced using next-generation technology (Andrews et al., 2016). Homologous fragments among individuals are aligned (e.g., using Stacks Catchen et al., 2013 or other software), and thousands to millions of SNPs are identified. RAD-seq has been extremely beneficial for studies of population genetic structure, as well as pedigree and other analyses (Andrews et al., 2016; Catchen et al., 2017). Therefore, RAD-seq can be a powerful approach for landscape genetics studies. As with other genotyping-by-sequencing methods, RAD-seq, while beneficial for genotyping large numbers of individuals, suffers from marker attrition. That is, the more individuals sequenced, the fewer loci become available for robust analyses due to genotyping errors due low coverage or missing data. Additionally, a shortcoming of RAD-seq for landscape genomics studies is that generally only a small fraction of a genome is sampled, and thus loci involved in adaptation are often missed (Lowry et al., 2017). Further, without a reference genome, identified SNPs are anonymous, and downstream work is necessary to determine their function (Lowry et al., 2017).

As a potential solution, transcriptome sequencing and exome capture are reduced representation approaches that focus on genic (i.e., coding) regions. Genes will contain much of the functional genetic variation that underlies adaptation, and such regions are also in linkage with promoter regions also under selection (Hoekstra and Coyne, 2007; Stern and Orgogozo, 2008). RNA-seq is an approach to sequence total RNA or the mRNA transcriptome, which can be used to evaluate gene expression levels (in different environments) and, when multiple transcriptomes are sequenced, SNPs can be identified. A series of capture probes can then be designed to sequence the flanking region around identified SNPs in cDNA. Assembled transcriptomes, can then be used to annotate functional information for candidate SNPs since they are all found in coding DNA. Further, when SNP codon positions are identified, traditional sequence-based population genetic tests for selection can be applied (e.g., MK test; McDonald and Kreitman, 1991 or dN/dS ratios). Transcriptome sequencing, however, will only capture a subset of all coding genes, as gene expression is tissue-specific (Bishop et al., 1974). Exome capture sequencing will increase the number of coding loci (Jones and Good, 2016).

Another method used for genome-wide marker generation in non-model species is Pool-seq (reviewed Schlötterer et al., 2014), whereby a large number of individuals (dozens to hundreds) are pooled and sequenced together. Advantages include reduced cost, and genome-wide data generation that facilitates SNP identification and allele frequency generation for population genetic analyses. Disadvantages include lack of ability to identify individual samples, difficulties identifying rare variants, and potential alignment issues owing to non-homologous sequences (i.e., paralogs), and lower confidence in SNP assignment than other methods (Schlötterer et al., 2014). Software such as PoPoolation (Kofler et al., 2011) can help account for some of the bias introduced by pooling and sequencing errors. Nonetheless, pool-seq works much better when a reference genome is available and short-read sequences can be aligned and mapped to reduce alignment errors among pools. Even with a reference genome, structural variation (e.g., inversions, indels) between pooled resequenced samples and the reference can generate falsely identified SNPs (Tiffin and Ross-Ibarra, 2014).

ANALYSIS CONSIDERATIONS

Similar to landscape genetic studies, there is a wide array of analysis methods for landscape genomics analyses and new methods are continuously being developed (Hoban et al., 2016). The key difference between the two analytical frameworks is

TABLE 2 | Simulation studies of genome scan methods in landscape genomics.

Simulation study	Study aims	Methods tested	Demographic models	Simulated sampling strategies	Selection patterns	Major findings
De Mita et al., 2013	1. Compare methods evaluating differences in type I/II error rates and power 2. Evaluate impact of differences in selection, demography, and sampling strategy on inferences made by genome scans	Logistic Regression (LR; Joost et al., 2007) Generalized Estimated Equation (GEE; Poncet et al., 2010) Coop, Witonsky, Di Rienzo and Pritchard (CWDRP; Coop et al., 2010) Beaumont and Nichols test (FDIST2; Beaumont and Nichols, 1996) Foll and Gaggiotti (FG; Foll and Gaggiotti, 2008) Extended Lewontin and Krakauer (FLK; Bonhomme et al., 2010) Excoffier, Hofer and Foll (EHF; Excoffier et al., 2009) Vitalis, Dawson and Boursot (VDB; Vitalis et al., 2001)	Island Model (IM) Stepping Stone Model (SSM) Hierarchical Model (HM) Selfing + IM/SSM/HM Allogamy	S1-1 individual/population S2–4 individuals/population in 48 regularly sampled populations S3-6 random individuals/population in 12 populations S4-4 random individuals/population in 8 populations as two transects parallel to environmental gradient S5 - 4 random individuals/population in 4 populations sampled at extremes of gradient	None tested	LR and GEE have high FPR (false-postive rates), but fast run time Differentiation-based methods have low FPR, but slow run time Sampling fewer individuals in many populations (10/population for most methods) increases power Under allogamy and IM, all methods are comparable Under allogamy and HM or SSM, differentiation based methods have lower FPR Under selfing and IM, LR sampling using S1 is optimal. Under selfing SSM or HM, LR with S1, BN with S3, and FG with S2 perform best with respective sampling strategies
Frichot et al., 2013	1. Identify signatures of selection controlling for population structure 2. Introduce Latent Factor Mixed Models (LFMM) as a means to test for genetic-environment associations 3. Compare FPR and FDR between methods using spatially explicit neutral coalescent simulations	LFMM (Frichot et al., 2013) LRM (Storey and Tibshirani, 2003) Principle Component Regression (PCRM; Joost et al., 2007) Generalized Linear Models (GLMs; Joost et al., 2007) Standard Linear Mixed Models (GEMMA; Zhou and Stephens, 2012) Partial Mantel Test (PMT; Fumagalli et al., 2011) BayEnv (Coop et al., 2010)	Isolation by Distance (IBD)	None tested	P1 - Correlated with demographic history P2 - Along environmental gradient P3 - Low-intensity selection	LFMM has low FPR under IBD PMTs, LRMs and PCRMs have low power and high FPRs under IBD PMT, PCRM and GEMMA have high FNR when enviornement is strongly correlated with demography LFMM runs faster than BayEnv when analyzing large data sets LFMM performs better then BayEnv when genetic structure well characterized FDR (false-discovery rate) and FPR highly correlated
de Villemereuil et al., 2014	1. Individual-based simulation comparing power and error rates of genome scan methods 2. Characterize role of population structure and mode of selection on outlier detection	Allele frenquency-environmental linear regression (LRM; Storey and Tibshirani, 2003) Bayescan (Foll and Gaggiotti, 2008) BayEnv (Coop et al., 2010) Latent Factor Mixed Model (LFMM; Frichot et al., 2013)	Hierarchical Model (HM) Island Model (IM) Stepping Stone Model (SSM)	None tested	P1 - Correlated with demographic history P2 - Along environmental gradient P3 - Monogenic P4 - Polygenic	Decrease in power in methods under polygenic vs. monogenic selection Under polygenic selection LRM most powerful but has highest FDR BayEnv has low FDR under SSM, high under HM All methods have low power under P1 BayEnv and LRM have highest FPR, LFMM had the most true-positives under P1

(Continued)

TABLE 2 | Continued

Simulation study	Study aims	Methods tested	Demographic models	Simulated sampling strategies	Selection patterns	Major findings
Lotterhos and Whitlock, 2014	1. Test effects of IBD and range expansion to detect spatially divergent selection among methods 2. Compare effects of different parameterization and outlier differentiation/GEAs	Beaumont & Nichols test (FDIST2; Beaumont and Nichols, 1996) Bayescan (Foll and Gaggiotti, 2008) Extended Lewontin & Krakauer (FLK; Bonhomme et al., 2010) X^TX (Günther and Coop, 2013)	Island Model (IM) Isolation by Distance (IBD) Two Refugia (2R) One Refugium (1R)	None tested	Soft selection	Under IBD, FDIST2 and BayeScan have low power and high FPR FDIST2 and BayeScan have low FDR when assumptions of equilibrium are met FLK performs best when no neutral loci or null model is available BayEnv2 has highest power under IBD and non-eq demographic scenarios
Forester et al., 2015	1. Describe how variation in environment, strength of selection and dispersal affect strength of local adaptation 2. Determine which GEAs have the greatest power in competing scenarios	Principial components analysis (PCA) Principial coordinate analysis (PCoA; Bray and Curtis, 1957) Redundancy Analysis (RDA) Distance-based redundancy analysis (dbRDA) Latent Factor Mixed Model (LFMM; Frichot et al., 2013)	IBD with varying dispersal distances: 5% 10% 15% 25% 50%	None tested	P1 - Continuous (clinal) gradient P2 - Discrete spatial selection with habitat aggregation (10%) P3 - Discrete spatial selection with habitat aggregation (50%) P4 - Discrete spatial selection with habitat aggregation (90%)	RDA and dbRDA have highest power, low FPRs and strongest GEA indices under all scenarios PCA, PCoA & LFMM show stronger GEA indices at intermediate dispersal levels Ordination methods broadly control for population structure due to IBD better then other techniques Changes in habitat aggregation and selection have small effects on spatial structure at neutral sites
Lotterhos and Whitlock, 2015	1. Compare power of GEAs and outlier differentiation methods to detect loci involved in local adaptation based on: Sampling design and 2. Demography	X^TX (Günther and Coop, 2013) PCAdapt (Duforet-Frebourg et al., 2014) BayEnv2 (Günther and Coop, 2013) Latent Factor Mixed Model (LFMM; Frichot et al., 2013)	Island Model (IM) Isolation by D istance (IBD) Two Refugia (2R) One Refugium (1R)	S1 - Transect S2 - Paired sampling S3 - Random	Weak clinal selection	Pairwise sampling have high power for detecting genes under weak selection, transects better at detecting clines Total sample size influenced power more than distribution of populations LFMM has higher power then Bayenv2 with more samples, but higher FPR LFMM and Bayenv2 have high power because they explictly account for relatedness and environment

Summarized are questions, sampling methods, analysis methods and conclusions as to which methods lead to low false positive rates and high power to detect loci under selection.

that landscape genetics studies rely on use of putatively neutral markers to generate estimates of genetic population structure, whereas tests of selection in landscape genomic studies generally require the need to control for population structure (see **Table 1**). As above, note that genome-wide marker sets generated for landscape genomics tests of selection can also be parsed into neutral data and landscape genetics analyses can be employed (see Storfer et al., 2007, 2010; Guillot et al., 2009; Shirk et al., 2017). Landscape genomics studies employ tests for loci under selection using genome scans, candidate gene approaches, quantitative trait locus mapping and genome-wide association studies (see Stinchcombe and Hoekstra, 2008; Storfer, 2015). However, genome scans are the most widely used, as the latter analysis types tend to be used for model systems. It is important to note that numerous excellent reviews (e.g., Rellstab et al., 2015; Haasl and Payseur, 2016; Hoban et al., 2016) discuss in detail the benefits and limitations of the various genome scan methodologies and associated software. As such, we summarize the main considerations here.

Genome scans generally use two approaches to detect loci under selection: (1) differentiation outlier methods (which were previously called F_{ST}-outlier tests, but now include other methods of genetic differentiation among populations; Hoban et al., 2016); and, (2) genetic-environment association (GEA) tests (Schoville et al., 2012; Pardo-Diaz et al., 2015; Rellstab et al., 2015; Hoban et al., 2016). Differentiation outlier methods rely on the demonstration that, at migration-drift equilibrium under a neutral island model with spatially uniform migration and gene flow, population differentiation of allele frequencies (e.g., F_{ST}) across a large number of loci can be used to infer the process of selection acting on a subset of loci (Lewontin and Krakauer, 1973). Statistical outlier loci with significantly greater F_{ST} (or other genetic distance) values than the distribution of genome-wide F_{ST} values are presumed to be under diversifying or local selection or linked to those under selection (Black et al., 2001; Luikart et al., 2003). Similarly, loci with significantly lower F_{ST} values are inferred to be under stabilizing or purifying selection (Black et al., 2001; Luikart et al., 2003). Thus, unlike landscape genetics studies which generate genetic distance estimates among a small number of loci to elucidate effects of landscape variables on gene flow, landscape genomics studies rely on a very large number of loci to generate a frequency distribution of genetic distance values as a null against which to test for outliers under selection.

Early methods to conduct such outlier tests include FDIST (Beaumont and Nichols, 1996; implemented in LOSISTAN) to identify strong differences from the null distribution of F_{ST} values across loci. Later, the widely used BayeScan (Foll and Gaggiotti, 2008) was developed, which uses a Bayesian method to estimate the relative probability that each locus is under selection. PCAdapt is a recently developed popular method that uses a principal components analysis framework to detect candidate loci under local adaptation (Duforet-Frebourg et al., 2014). Methods that use genetic distance measures other than F_{ST} include FLK (Bonhomme et al., 2010), which uses a modified version of the Lewontin and Krakauer (1973) test for selection by comparing allele frequencies of different populations in a neighbor-joining tree constructed using a matrix of Reynold's genetic distance (Reynolds et al., 1983), and X^TX, which employs a Bayesian method to test individual SNPs against a null model generated by the covariance in allele frequencies between populations from the entire set of SNPs (utilized in Bayenv2; Coop et al., 2010; Günther and Coop, 2013). Summaries of differentiation outlier methods can be found in Hoban et al. (2016; Appendix 1). Notably, differentiation outlier methods are aspatial in nature.

GEAs (also referred to as EAAs or environmental association analyses; Rellstab et al., 2015) are spatial because they are designed to test for significant correlations between allele frequencies at particular loci with variation in environmental variable(s) (Joost et al., 2007; Hancock et al., 2011; Rellstab et al., 2015). Thus, unlike differentiation outlier approaches, GEAs require availability of environmental data from sources such as WorldClim data (http://www.worldclim.org, Hijmans et al., 2005). Widely used methods include Bayenv2, which tests for GEAs in addition to differentiation outliers, and latent factor mixed models (LFMM; Frichot et al., 2013). Bayenv2, tests for large allele frequency differences across environmental gradients by comparing observed allele frequency differences to transformed normal distribution of underlying population frequencies. Latent factor mixed models (LFMM; Frichot et al., 2013), include population structure as latent (or hidden) variables to limit false positive signals. Spatial generalized linear mixed models (SGLMMs; Guillot et al., 2014) are an extension to LFMMs and have proven to be computationally more efficient. Ordination approaches, such as redundancy analysis, can also be used in GEAs (Forester et al., 2015); ordination is also widely used in landscape genetics studies (Storfer et al., 2010). Another more recently developed GEA method is Samβada (Stucki et al., 2016), which is a multivariate analysis framework that accounts for underlying population structure with estimates of spatial autocorrelation in the data. To search for loci under selection, Samβada uses linear regressions to model the probability of observing a particular allele given the value of environmental variables at the location it was sampled for each locus independently (Stucki et al., 2016). A summary of GEAs and their assumptions can be found in Rellstab et al. (2015; **Table 1**).

Analysis Concerns

Fundamentally genome scan methods operate on the assumption that loci under selection can be differentiated from a null distribution of allele frequencies generated by neutral processes. Determining how much genetic differentiation can be expected in populations in the absence of selection, however, remains a great challenge (Lotterhos and Whitlock, 2014; Hoban et al., 2016). Thus, the primary concern with employing genome scan analyses is differentiating false positive signals from loci that are actually under selection.

Underlying population demographic structure, when not properly accounted for, can be a principal source of false positives. There are several demographic scenarios that can generate neutral allele frequency differentiation among populations that can falsely be interpreted as signals of selection

(Lotterhos and Whitlock, 2015; Rellstab et al., 2015; Haasl and Payseur, 2016). A straightforward example is illustrated by the case of allele surfing, whereby serial population bottlenecks that occur during founder effects of small populations migrating to new areas can result in fixed allelic differences among populations that are solely due to genetic drift (Excoffier et al., 2009; Waters et al., 2013). Similarly, recent population range expansions from refugia can generate correlations between allele frequencies and environmental variables that are not due to selection. In general, landscape genomics studies are challenging in small, patchy populations that are prone to genetic drift, which can result in the appearance of spatially distributed loci under selection. False signals of selection can also be generated by locus-specific hybridization or introgression from related taxa (Fraïsse et al., 2016; Hoban et al., 2016). Nonetheless, in cases where selection gradients follow the same spatial pattern as background genetic population structure, candidate loci under selection can be missed due to false negative signals.

In general, demographic structure can influence the null distribution of F_{ST} or other genetic differentiation measures and thereby bias significance testing (Lowry, 2010; Whitlock and Lotterhos, 2015). Each genome scan method utilizes a different way to account for underlying population demography. For example, FDIST assumes populations follow an island model (Beaumont and Nichols, 1996) to generate null F_{ST} distribution. The recently developed OutFLANK (Whitlock and Lotterhos, 2015), however, does not invoke a specific demographic model. Rather, OutFLANK infers the distribution of F_{ST} for loci unlikely to be strongly affected by spatially diversifying selection (Whitlock and Lotterhos, 2015). Specifically, OutFLANK uses a modified Lewinton-Krakauer method to infer a null F_{ST} distribution, which approximates a χ^2 distribution with adjusted degrees of freedom. Then, differentiation outliers are identified as those that fall outside this trimmed, putatively null F_{ST} distribution.

Approaches that use covariance matrices or linear models to account for population structure are also flexible because they have no explicit underlying population demographic model. For example, Bayenv2 is a GEA method that controls for genetic population structure in by generating a variance-covariance matrix of relatedness among samples; candidate loci are determined as those for which an environmental variable explains significantly more variation than the variance-covariance matrix of all other loci (Günther and Coop, 2013). Linear model approaches, such as LFMMs and SGLMMs, can limit false positives in both GEAs and outlier tests by including population structure as latent variables (Frichot et al., 2013; Lotterhos and Whitlock, 2015). Samβada uses estimates of underlying spatial autocorrelation in genetic data as a way to control for underlying population structure (Stucki et al., 2016).

A number of informative simulation studies that explore the power of the different methods under different demographic or other scenarios have recently been published (De Mita et al., 2013; Frichot et al., 2013; Jones et al., 2013; de Villemereuil et al., 2014; Lotterhos and Whitlock, 2014, 2015; Forester et al., 2015; See **Table 2** for a summary of the study conditions and their findings). The relative power of GEAs and differentiation outlier tests is dependent on the underlying demographic model. GEAs have higher power under an island model, whereas outlier tests have higher power under an isolation-by-distance model (Lotterhos and Whitlock, 2015). Within GEAs, the degree of patchiness in the landscape affects the power and false positive rates (Forester et al., 2015). With limited dispersal and strong isolation-by-distance, univariate GEAs had high false positive rates (FPRs; up to 55%) and constrained ordination procedures (e.g., redundancy analyses, or RDA) performed much better with lower FPRs (0–2%; Forester et al., 2015). Within outlier differentiation methods, Bayenv2 and FLK outperformed FDIST and Bayescan for systems experiencing IBD and recent range expansions (Lotterhos and Whitlock, 2014). Of all GEAs and outlier detection methods, LFMMs were generally found to have relatively low false positive rates (Type I error rates) than other methods (Jones et al., 2013; Joost et al., 2013).

Even after accounting for the underlying population structure, however, there are other important considerations that can affect the power of genome scan studies and their interpretation. To date, no methods have been developed to account explicitly for background selection (Hoban et al., 2016), which can result in population diversification due to purifying and not positive selection (Charlesworth et al., 1993). Background selection can thus cause errors in estimating the null distribution and thereby reduce power of genome scans (Tiffin and Ross-Ibarra, 2014; Haasl and Payseur, 2016). Signatures of local adaptation can also be incorrectly inferred as a result of spatially uniform positive selection. That is, across landscapes with limited gene flow, multiple beneficial mutations may arise to reach an optimal phenotype, resulting in a patchwork of allele frequencies. This can result in detectable genetic differentiation across the patches that produces false signals of selection by local environment (Hoban et al., 2016).

It is also important to note that genome scan analyses are biased to detect large effect loci, because power to detect small effect loci is generally low (Pritchard and Di Rienzo, 2010). Because most phenotypic traits are likely to be polygenic, and thus governed by many loci of small effect (Rockman, 2012), genome scan methods are prone to miss most loci involved in local adaptation (Stephan, 2015). Further, the polygenic nature of phenotypic traits means candidate loci explain a small proportion of phenotypic variation, which has been termed the "missing heritability problem" (Hindorff et al., 2009; Visscher et al., 2010; Yang et al., 2010, 2012). Recently, multilocus approaches have been developed that quantify the strength of selection acting on correlated loci using Bayesian sparse linear mixed models (Gompert et al., 2017). However, these approaches necessitate large sample sizes and time-series sampling, thereby limiting their widespread applicability. In addition, for studies that employ anonymous SNP markers when no reference genome exists, such as RAD-seq, candidate genes are assumed to be in linkage disequilibrium (LD) with loci under selection and are most often not under selection themselves (Lowry et al., 2017). With a reference genome, estimates of LD decay can be used to determine the size of the window to search for possible genes linked to a candidate SNP detected in a genome scan when the SNP is not in a gene itself. However, we do not know the extent

of LD for most species, and the size of LD blocks is not constant throughout the genome (Tiffin and Ross-Ibarra, 2014; Lowry et al., 2017). These factors can make mapping and annotating candidate markers prone to error.

Combinatorics and Other Multivariate Approaches

An important consideration in landscape genomics studies is how to integrate data analyses across multiple genome scan methods. One fairly standard approach is to construct Venn diagrams and use combinatorics as a method of validation for candidate loci. That is, the larger the number of genome scan methods that detect a particular candidate locus under selection, the more confident researchers tend to be that the candidate is truly under selection. However, genome scan methods each have different assumptions and different power to detect loci under selection, depending on population demography, sampling design and nature of the selective sweep (Lotterhos et al., 2017). Thus, reliance on concordance of multiple univariate methods to prioritize loci for further research is prone to miss loci under weak selection (Lotterhos and Whitlock, 2015).

Recent proposed solutions have included multivariate methods that combine *P*-values and control for false discovery rates (FDR; Benjamini and Hochberg, 1995). For example, de-correlated composite of multiple signals (DCMS) controls for genome-wide correlations among statistics by weighting each locus depending how correlated a particular statistic that detected the locus is to other statistics (Ma et al., 2015). Thus, the less a test statistic is correlated to another statistic(s), the higher the locus is weighted. François et al. (2016) built on earlier methods to control for FDR (e.g., Benjamini and Hochberg, 1995) using a "genomic inflation factor" to adjust the distribution of *p*-values. In general, composite methods tend to perform better than univariate methods, but their performance has only been evaluated in a narrow set of circumstances (Lotterhos et al., 2017).

Even newer methods include analyses to filter, visualize and integrate multiple univariate analyses in multivariate space (Lotterhos et al., 2017; Verity et al., 2017). For example, MINOTAUR (Multivariate vIsualizatioN and OuTlier Analysis Using R) is a program that uses one of four different distance measures (Mahalanobis distance, harmonic mean distance, nearest neighbor distance and kernel density deviance) to test the significance of loci (Verity et al., 2017). An important future direction is to continue to evaluate the variety of methods for evaluating and prioritizing candidate loci for future research. As we learn more about the genomic architecture of different species, we can continue to test the performance of existing methods, or develop new methods as appropriate.

Analysis Considerations-Summary

In general, researchers should avoid the temptation to analyze their data with as many genome scan methods as possible. Instead, several factors that should be considered when choosing genome scan method(s) to be employed. First, if attainable, knowledge of underlying demographic structure can be used to choose the most powerful methods that are least prone to Type I errors for that specific demographic history. For example, phylogeographic analyses can be used to assess whether there have been recent geographic range expansions from glacial refugia. To parameterize the number of latent factors (e.g., in LFMM or SGLMM), the number of genetic clusters (K) could be determined using a Bayesian clustering algorithm such as FastSTRUCTURE (Raj et al., 2014) or ADMIXTURE (Alexander et al., 2009). Note that incorrect assumptions about underlying demographic structure can increase both Type I and Type II error (Pérez-Figueroa et al., 2010; Jones et al., 2013; Lotterhos and Whitlock, 2014), and in such cases, model-free approaches may be preferred. Second, given the numerous additional concerns for which researchers have little ability to estimate (e.g., variation in genome-wide LD) or control for (e.g., the polygenic nature of most phenotypic traits), confidence in candidate loci as real targets of selection comes from their repeated detection across replicated transects or paired sampling locations. Similarly, candidate loci detected by multiple analysis methods also decreases the likelihood that they are false positives. Third, as stated above, inference of candidate loci is improved when selective agent(s) are known before embarking on a landscape genomics study. Candidate genes identified in genic pathways that influence particular phenotypes known to be under selection are less likely to be false positives than randomly detected loci or those without known function.

METHODS AT THE INTERFACE OF LANDSCAPE GENETICS AND LANDSCAPE GENOMICS

Generalized Dissimilarity Modeling (GDM)

Originally used to model species community turnover (Ferrier et al., 2007), GDMs have recently been adopted for use in landscape genetics studies. GDMs involve fitting I-splines that are monotonic, nonlinear functions that, when rescaled between 0 and 1, represent importance of environmental variables in explaining turnover of allele frequencies (Fitzpatrick and Keller, 2015). GDMs have been used to assess effects of at site environmental differences on gene flow (also called "isolation by environment"; Wang and Bradburd, 2014). I-splines can be non-linear, providing an advantage over linear approaches because they may be able to identify threshold values (i.e., the point along the environmental axis where the slope of the spline is greatest) for landscape variables. Similarly, GDMs can be applied to landscape genomics studies by fitting I-splines to the relationships of ecological variables on allele frequencies at putatively adaptive loci. Related to GDMs, which employ distance-based measures are gradient forests, an extension of random forests, which both employ machine-learning algorithms for model optimization (Breiman, 2001). Similar to GDM, gradient forests fit nonlinear monotonic functions to characterize allele-frequency turnover across environmental gradients for each locus independently (see Fitzpatrick and Keller, 2015). As such, both approaches can be used to identify a loci with high degree of allelic turnover associated with specific environmental variables, and thus yield candidate loci under selection.

Estimated Effective Migration Rate

Another recently developed method that can be applied to both landscape genetics and landscape genomics studies is the Estimated Effective Migration Surface (EEMS: Petkova et al., 2016). This method differs from other approaches that identify underlying population demographic structure (e.g., clustering and PCA-based approaches), because genetic differentiation is modeled as a function of estimated migration rates. EEMS uses a stepping stone model (Kimura and Weiss, 1964) that allows for migrations of variable rates to occur among a set of demes. This process is modeled by overlaying a dense regular grid over the study area and calculating an approximation of the expected genetic dissimilarity through the use of resistance distance, similar to "isolation-by-resistance" (McRae, 2006). Consequently, areas in which genetic dissimilarity decays more slowly will be assigned a greater value of Effective Migration Rate (EMR), than those for which genetic dissimilarity decays more rapidly.

EEMS offers two potential applications to landscape genomics studies. First, it can allow researchers to detect underlying demographic population structure, which can be used to help reduce false positive rates in genome scan methods. Second, EEMS analyses could be run separately on data sets containing only putatively neutral or putatively adaptive loci, and can then be used to visualize geographic features that impede gene-flow of neutral or adaptive loci, respectively.

Clinal Analyses

Clines have a rich history in population genetics and bridge both at-site and between-site analyses used in landscape genetics and genomics. To date, most clinal analyses on genome-scale data have focused on the study of hybrid zones and the detection of differential introgression (Gompert and Buerkle, 2010, 2011, 2012). While originally developed for use in identifying loci involved in adaptive divergence and reproductive isolation among hybridizing lineages, genomic cline models could be applied to identify candidate loci for population pairs for which a genome-wide admixture gradient (e.g., via ADMIXTURE or another assignment-based program) has been identified. Loci for which genomic clines possess outliers in one or both of these cline parameters may be subject to selective forces. Outlier loci with alleles introgressing most slowly can be interpreted as those involved in differential adaptation among populations, whereas loci introgressing most rapidly are likely to be uniformly advantageous.

Geographic cline models can explicitly measure the strength of selection on a locus, given the shape of a cline (Endler, 1977; Slatkin, 1987). Geographic cline analyses involve fitting a sigmoidal *tanh* cline model to allele frequencies and quantitative data such as environmental data or a measure of geographic distance (**Figure 1**; Szymura and Barton, 1986, 1991). Then, cline center, width and slope are estimated along a geographic transect (requiring transect sampling). GEAs are essentially clinal analyses but focus only on the slope of the cline between sampling locations. However, geographic cline analyses analyze the shape of the cline; selection tends to steepen the cline, gene flow widens and reduces the steepness of the cline, and genetic drift narrows the cline (**Figure 1**; Endler, 1977; Nagylaki, 1978). Researchers can then compare the shapes of observed allele frequency clines in putatively adaptive loci to the shape of clines for neutral loci, as well as those predicted by models of pure migration or drift (Nagylaki, 1978). Unfortunately, current implementations

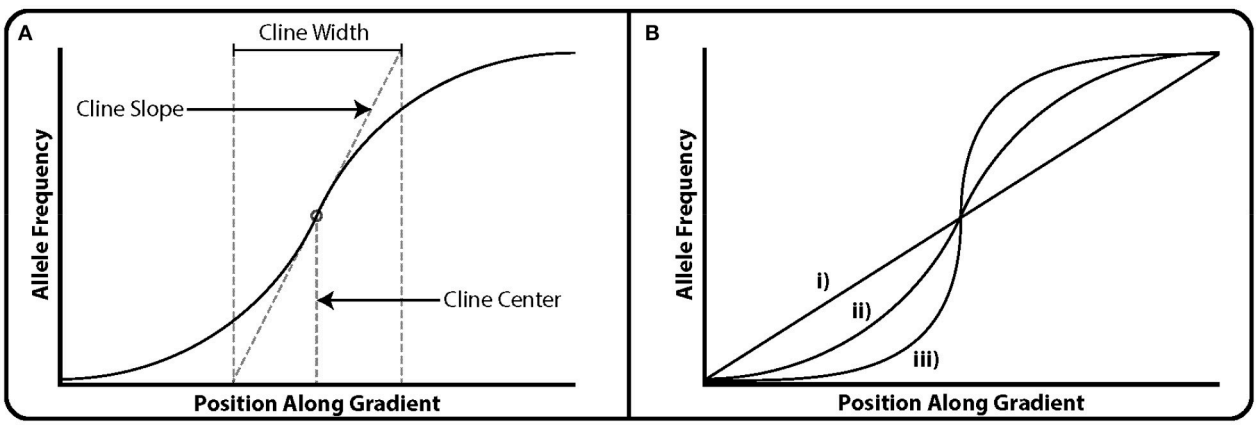


FIGURE 1 | An illustration of clines. X-axes correspond to position along geographic transects (ecological gradient) or hybrid indexes (genomic gradient) in the case of genomic cline analyses. **(A)** Illustration of the three parameters typically estimated in the use of geographic of genomic cline analysis. Cline slope is the estimate of the rate of allele frequency turnover at the steepest point in the cline. In genomic cline analysis this corresponds to the rate of introgression. Cline center corresponds to the point along the geographic transect or hybrid index at which allele frequency turnover is greatest. Cline width corresponds to the region along the gradient at which it's influence on allele frequency is greatest. **(B)** Three examples of clines. (i) A transect along which no selection appears to be acting, or the effects of gene flow are such that changes in allele frequency are purely a function of distance. In the case of genomic cline analyses, the loci under consideration appears to be favored equally in both parental taxa. (ii) A modest cline in which the allele favored by selection changes along the gradient. Given its shallower slope, selection may either be weaker, gene flow stronger (in the case of geographic transects) or the ecotone separating ends of the transect greater. (iii) A steep cline, often called a step cline. In the case of geographic clines, these are formed either by strong selection acting in favor of one allele along a sudden ecotone, or extremely limited gene flow along said ecotone. In the case of genomic clines, this may be due to heterozygote disadvantage, as in the case of reinforcement.

of geographic cline models (e.g., *Analyse*: Barton and Baird, 1995; *hzar*: Derryberry et al., 2014) are computationally burdensome, thus limiting cline fitting to datasets with small numbers of loci. Therefore, geographic cline analysis is currently best suited for use with a reduced set of candidate loci as identified by genome scans.

FUTURE DIRECTIONS

In the future, landscape genomics should integrate analyses on two scales—the landscape of the genome, and the ecological landscape. Specifically, the landscape of the genome refers to overall genomic architecture, such as the arrangement of loci on chromosomes, placement of inversions, deletions and copy number variants. All of these, ultimately, can affect gene expression, which is further modified by the environmental context in which an individual exists. However, the current state of landscape genomics studies is primarily to generate a list of candidate loci under selection, and, when possible annotate genes in LD with identified SNPs or other genetic variants. Nonetheless, scientists are increasingly aware that the genotype-phenotype relationship is influenced by far more of the genome than just genic sequences. For example, copy number variation and not sequence variation that determines how much human amylase, responsible for starch digestion, is expressed in saliva (Perry et al., 2007). Selection has acted on copy number variation in the amylase gene (*AMY1*) in the human populations; those with high starch diets have higher numbers of copies than populations with diets lower in starch (Perry et al., 2007). Similarly, camels have the highest number of copies known (11) of the *CYP2J* gene (related to salt homeostasis) likely due to selection for high salt tolerance necessary in desert environments (Wang et al., 2012). Transposable elements, which comprise over half the genome of many eukaryotes, were once thought of as parasitic or "junk" DNA (Federoff, 2012). However, evidence suggests that transposable elements are maintained in eukaryotic genomes due to their heritable role in epigenetic mechanisms, such as gene silencing (Federoff, 2012). DNA methylation patterns also influence gene expression and can also be heritable (Anway et al., 2005; Skinner et al., 2012). Promoters and other regulatory regions are also key determinants of gene expression levels and consequently phenotypes. Further, genes are expressed differently in different ecological environments, and selection varies spatially across the ecological landscape. In summary, genomic architecture plays a significant role in the genotype-phenotype relationship, as evidenced by the fact that "large effect SNPs" tend to explain a small fraction of phenotypic variation in natural populations (Hindorff et al., 2009; Rockman, 2012).

Given that technological advances continue to make whole genome sequencing more and more feasible in terms of cost and computational speed for genome assembly, a key challenge for the future of landscape genomics will be the development of methods that integrate multiple data types. Difficulties will include: (1) accounting for the effects of coding and non-coding regions of genomes and overall genomic architecture, combined with protein expression levels, on phenotypic variation; (2) coding for genomic features such as copy number, chromosome inversions or transposable element composition or location in our population genetic models (i.e., Can they be considered in the same way as alleles?); (3) constructing hierarchical models to integrate sources of error from different data types. Then, the challenge is compounded further with the necessity to integrate these complex genomic models with multiple types of spatial environmental data and habitat models in ways that optimize sampling while avoiding potential biases. Mapping the genotype-phenotype relationship has been a key challenge for evolutionary biology for over a century, and landscape genomics will provide the analytical framework to do so across spatially variable ecological environments. A long road may lie ahead, but it is certainly an exciting time for landscape genomics to unravel the complexity of the genomic architecture that underlies local adaptation.

CONCLUSIONS

Landscape genomics has emerged as a prominent framework for studying the genomic basis of local adaptation. Using large genomic data sets, researchers scan the genome for loci that exhibit signatures of selection across heterogeneous environments (Haasl and Payseur, 2016). These efforts have been highly successful, for example, in identifying genes underlying hypoxia adaptation in high-elevation human populations (Beall, 2007a,b; Simonson et al., 2010), environmental responses in Oak populations along climatic gradients (Sork et al., 2016), and differences in growth response amongst Salmon populations in response to geological conditions (Vincent et al., 2013). Studies of biotic factors, have also successfully in identified local adaptation to life history traits (Sun et al., 2015), community composition (Harrison et al., 2017), and disease prevalence (Leo et al., 2016; Mackinnon et al., 2016; Wenzel et al., 2016). Landscape genomics has already dramatically helped to further our understanding of the genomic basis of adaptation (Funk et al., 2012; Shryock et al., 2015). Here, we suggest the field can advance with a careful consideration of explicit hypotheses that, in turn, guide study design, and employment analysis methods that help control confounding factors such as underlying demographic structure. Future landscape genomic research will better integrate genomic architecture in assessments of candidate loci under selection.

AUTHOR CONTRIBUTIONS

AS conceived of, and wrote most of the paper. AP and AF contributed to the writing, as well as gathered information for, and assembled **Table 2**.

ACKNOWLEDGMENTS

This work was funded by NSF grant DEB-1316549 to AS. Additionally, we thank Mark Margres, Lauren Ricci, Matthew Lawrence, and Elisa Lopez-Contreras for insightful comments that helped improve the quality of the manuscript.

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Anderson, C. D., Epperson, B. K., Fortin, M. J., Holdregger, R., James, P. M. A., Rosenberg, M. S., et al. (2010). Considering spatial and temporal scale in landscape genetic studies of gene flow. *Mol. Ecol.* 19, 3565–3575. doi: 10.1111/j.1365-294X.2010.04757.x

Andrews, K. R, Good, J. M, Miller, M. R., Luikart, G., Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* 17, 81–92. doi: 10.1038/nrg.2015.28

Anway, M. D., Cupp, A. S., Uzumcu, M., and Skinner, M. K. (2005). Epigenetic transgenerational actions of endocrine disruptors and male fertility. *Science* 308, 1466–1469. doi: 10.1126/science.1108190

Barton, N. H., and Baird, S. J. E. (1995). *Analyse: An Application for Analysing Hybrid Zones*. Edinburgh: Freeware.

Beall, C. M. (2007a). Two routes to functional adaptation: Tibetan and Andean high-altitude natives. *Proc. Natl. Acad. Sci. U.S.A.* 104(Suppl. 1), 8655–8660. doi: 10.1073/pnas.0701985104

Beall, C. M. (2007b). Detecting natural selection in high-altitude human populations. *Respir. Physiol. Neurobiol.* 158, 161–171. doi: 10.1016/j.resp.2007.05.013

Beaumont, M. A., and Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. Lond. Ser. B Biol. Sci. Biol. Sci.* 263, 1619–1626. doi: 10.1098/rspb.1996.0237

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* 57, 289–300.

Bishop, J. O., Morton, J. G., Rosbash, M., and Richardson, M. (1974). Three abundance classes in HeLa cell messenger RNA. *Nature* 250, 199–204. doi: 10.1038/250199a0

Black, W. C., Baer, C. F., Antolin, M. F., and DuTeau, N. M. (2001). Population genomics: genome-wide sampling of insect populations. *Annu. Rev. Entomol.* 46, 441–469. doi: 10.1146/annurev.ento.46.1.441

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., et al. (2010). Detecting selection in population trees: the lewontin and krakauer test extended. *Genetics* 186, 241–262. doi: 10.1534/genetics.110.117275

Bray, J. R., and Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* 27, 325–349. doi: 10.2307/1942268

Breiman, L. (2001). Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat. Sci.* 16:3. doi: 10.1214/ss/1009213726

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., and Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Mol. Ecol.* 22, 3124–3140. doi: 10.1111/mec.12354

Catchen, J. M., Hohenlohe, P. A., Bernatchez, L., Funk, W. C., Andrews, K. R., and Allendorf, F. W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Mol. Ecol. Resou.* 22, 362–365. doi: 10.1111/1755-0998.12669

Caye, K., Deist, T. M., Martins, H., Michel, H., and François, O. (2016). TESS3: fast inference of spatial population structure and genome scans for selection. *Mol. Ecol. Res.* 16, 540–548. doi: 10.1111/1755-0998.12471

Charlesworth, B., Morgan, M. T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics.* 134, 1289–1303.

Clarke, R. T., Rothery, P., and Raybould, A. F. (2002). Confidence limits for regression relationships between distance matrices: estimating gene flow with distance. *J Agric. Biol. Environ. Stat.* 7, 361–372. doi: 10.1198/108571102320

Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J. K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics* 185. 1411–1423. doi: 10.1534/genetics.110.114819

Derryberry, E. P., Derryberry, G. E., Maley, J. M., and Brumfield, R. T. (2014). HZAR: hybrid zone analysis using an R software package. *Mol. Ecol. Resou.* 14, 652–663. doi: 10.1111/1755-0998.12209

De Mita, S., Thuillet, A. C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J., et al. (2013). Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Mol. Ecol.* 22, 1383–1399. doi: 10.1111/mec.12182

de Villemereuil, P., Frichot, É., Bazin, É., François, O., and Gaggiotti, O. E. (2014). Genome scan methods against more complex models: when and how much should we trust them?. *Mol. Ecol.* 23, 2006–2019. doi: 10.1111/mec.12705

Duforet-Frebourg, N., Bazin, E., and Blum, M. G. B. (2014). Genome scans for detecting footprints of local adaptation using a Bayesian factor model. *Mol. Biol. Evol.* 31, 2483–2495. doi: 10.1093/molbev/msu182

Durand, E., Jay, F., Gaggiotti, O. E., and François, O. (2009). Spatial inference of admixture proportions and secondary contact zones. *Mol. Biol. Evol.* 26, 1963–197310. doi: 10.1093/molbev/msp106

Dyer, R. J., and Nason, J. D. (2004). Population graphs: the graph theoretic shape of genetic structure. *Mol. Ecol.* 13, 1713–1727. doi: 10.1111/j.1365-294x.2004.02177.x

Endler, J. A. (1977). *Geographic Variation, Speciation, and Clines*. Princeton, NJ: Princeton University Press

Excoffier, L., Foll, M., and Petit, R. J. (2009). Genetic consequences of range expansions. *Annu. Rev. Ecol. Evol. Syst.* 40, 481–501. doi: 10.1146/annurev.ecolsys.39.110707.173414

Federoff, N. V. (2012). Transposable elements, epigenetics, and genome evolution. *Science* 338, 758–767. doi: 10.1126/science.338.6108.758

Ferrier, S., Manion, G., Elith, J., and Richardson, K. (2007). Using generalized dissimilarity modelling to analyze and predict patterns of beta diversity in regional biodiversity assessment. *Divers. Distrib.* 13, 252–264. doi: 10.1111/j.1472-4642.2007.00341.x

Fitzpatrick, M. C., and Keller, S. R. (2015). Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. *Ecol. Letts.* 18, 1–16. doi: 10.1111/ele.12376

Foll, M., and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180, 977–993. doi: 10.1534/genetics.108.092221

Forester, B. R., Jones, M. R., Joost, S., Landguth, E. L., and Lasky, J. R. (2015). Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. *Mol. Ecol.* 25, 104–120. doi: 10.1111/mec.13476

Fraïsse, C., Belkhir, K., Welch, J. J., and Bierne, N. (2016). Local interspecies introgression is the main cause of extreme levels of intraspecific differentiation in mussels. *Mol. Ecol.* 25, 269–286. doi: 10.1111/mec.13299

François, O., Martins, H., Caye, K., and Schoville, S. D. (2016). Controlling false discoveries in genome scans for selection. *Mol. Ecol.* 25, 454–469. doi: 10.1111/mec.13513

Frichot, E., Schoville, S. D., Bouchard, G., and François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol. Biol. Evol.* 30, 1687–1699. doi: 10.1093/molbev/mst063

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admettla, A., Pattini, L., and Nielsen, R. (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet.* 7:e1002355. doi: 10.1371/journal.pgen.1002355

Funk, W. C., McKay, J. K., Hohenlohe, P. A., and Allendorf, F. W. (2012). Harnessing genomics for delineating conservation units. *Trends Ecol. Evol.* 27, 489–496. doi: 10.1016/j.tree.2012.05.012

Gompert, Z., and Alex Buerkle, C. (2010). INTROGRESS: a software package for mapping components of isolation in hybrids. *Mol. Ecol. Resou.* 10, 378–384. doi: 10.1111/j.1755-0998.2009.02733.x

Gompert, Z., and Buerkle, C. (2011). Bayesian estimation of genomic clines. *Mol. Ecol.* 20, 2111–2127. doi: 10.1111/j.1365-294X.2011.05074.x

Gompert, Z., and Buerkle, C. A. (2012). bgc: software for Bayesian estimation of genomic clines. *Mol. Ecol. Resou.* 12, 1168–1176. doi: 10.1111/1755-0998.12009.x

Gompert, Z., Egan, S. P., Barrett, R. D., Feder, J. L., and Nosil, P. (2017). Multilocus approaches for the measurement of selection on correlated genetic loci. *Mol. Ecol.* 26, 365–382. doi: 10.1111/mec.13867

Guillot, G., Leblois, R., Coulon, A., and Frantz, A. C. (2009). Statistical methods in spatial genetics. *Mol. Ecol.* 18, 4734–4756. doi: 10.1111/j.1365-294X.2009.04410.x

Guillot, G., Mortier, F., and Estoup, A. (2005). Geneland: a program for landscape genetics. *Mol. Ecol. Notes* 5, 712–715. doi: 10.1111/j.1471-8286.2005.01031.x

Guillot, G., Vitalis, R., le Rouzic, A., and Gautier, M. (2014). Detecting correlation between allele frequencies and environmental variables as a signature of

selection. A fast computational approach for genome-wide studies. *Spat. Stat.* 8, 145–155. doi: 10.1016/j.spasta.2013.08.001

Günther, T., and Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics* 195, 205–220. doi: 10.1534/genetics.113.152462

Haasl, R. J., and Payseur, B. A. (2016). Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Mol. Ecol.* 25, 5–23. doi: 10.1111/mec.13339

Hancock, A. M., Brachi, B., Faure, N., Horton, M. W., Jarymowycz, L. B., Sperone, F. G., et al. (2011). Adaptation to climate across the *Arabidopsis thaliana* genome. *Science* 334, 83–86. doi: 10.1126/science.1209244

Harrison, T. L., Wood, C. W., Borges, I. L., and Stinchcombe, J. R. (2017). No evidence for adaptation to local rhizobial mutualists in the legume *Medicago lupulina*. *Ecol. Evol.* 7, 4367–4376. doi: 10.1002/ece3.3012

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* 25, 1965–1978. doi: 10.1002/joc.1276

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9362–9367 doi: 10.1073/pnas.0903103106

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., et al. (2016). Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *Am. Nat.* 188, 379–397. doi: 10.1086/688018

Hoekstra, H. E., and Coyne, J. A. (2007). The locus of evolution: evo devo and the genetics of adaptation. *Evolution* 61, 995–1016. doi: 10.1111/j.1558-5646.2007.00105.x

Jones, M. R., Forester, B. R., Teufel, A. I., Adams, R. V., Anstett, D. N., Goodrich, B. A., et al. (2013). Integrating landscape genomics and spatially explicit explicit approaches to detect loci under selection in clinal populations. *Evolution* 67, 3455–3468. doi: 10.1111/evo.12237

Jones, M. R., and Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Mol. Ecol.* 25, 185–202. doi: 10.1111/mec.13304

Joost, S., Bonin, A., Bruford, M. W., Després, L., Conord, C., Erhardt, G., et al. (2007). A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Mol. Ecol.* 16, 3955–3969. doi: 10.1111/j.1365-294X.2007.03442.x

Joost, S., Vuilleumier, S., Jensen, J. D., Schoville, S., Leempoel, K., Stucki, S., et al. (2013). Uncovering the genetic basis of adaptive change: on the intersection of landscape genomics and theoretical population genetics. *Mol. Ecol.* 22, 3659–3665. doi: 10.1111/mec.12352

Kimura, M., and Weiss, G. H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* 49, 561–576.

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., et al. (2011). PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS ONE* 6:e15925. doi: 10.1371/journal.pone.0015925

Leempoel, K., Duruz, S., Rochat, E., Widmer, I., Orozco-terWengel, P., and Joost, S. (2017). Simple rules for an efficient use of geographic information systems in molecular ecology. *Front. Ecol. Evol.* 5:33. doi: 10.3389/fevo.2017.00033

Leo, S. S., Gonzalez, A., and Millien, V. (2016). Multi-taxa integrated landscape genetics for zoonotic infectious diseases: deciphering variables influencing disease emergence. *Genome* 59, 349–361. doi: 10.1139/gen-2016-0039

Lewontin, R. C., and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74, 175–195.

Lotterhos, K. E., Card, D. C., Schaal, S. M., Wang, L., Collins, C., and Verity, B. (2017). Composite measures of selection can improve the signal-to-noise ratio in genome scans. *Methods Ecol. and Evol.* 8, 717–727. doi: 10.1111/2041-210X.12774

Lotterhos, K. E., and Whitlock, M. C. (2014). Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Mol. Ecol.* 23, 2178–2192. doi: 10.1111/mec.12725

Lotterhos, K. E., and Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Mol. Ecol.* 24, 1031–1046. doi: 10.1111/mec.13100

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., et al. (2017). Responsible RAD: striving for best practices in population genomic studies of adaptation. *Mol. Ecol. Res.* 17, 366–369. doi: 10.1111/1755-0998.12677

Lowry, D. B. (2010). Landscape evolutionary genomics. *Biol. Lett.* 6, 502–504.

Luikart, G., England, P. R., Tallmon, D., Jordan, S., and Taberlet, P. (2003). The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Genet.* 4, 981–994. doi: 10.1038/nrg1226

Ma, Y., Ding, X., Qanbari, S., Weigend, S., Zhang, Q., and Simianer, H. (2015). Properties of different selection signature statistics and a new strategy for combining them. *Heredity* 115:5. doi: 10.1038/hdy.2015.42

Mackinnon, M. J., Ndila, C., Uyoga, S., Macharia, A., Snow, R. W., Band, G., et al. (2016). Environmental correlation analysis for genes associated with protection against malaria. *Mol. Biol. Evol.* 33, 1188–1204. doi: 10.1093/molbev/msw004

Manel, S., and Holderegger, R. (2013). Ten years of landscape genetics. *Trends Ecol. Evol.* 28, 614–621. doi: 10.1016/j.tree.2013.05.012

Manel, S., Joost, S., Epperson, B. K., Holderegger, R., Storfer, A., Rosenberg, M. S.,et al. (2010). Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Mol. Ecol.* 19, 3760–3772. doi: 10.1111/j.1365-294X.2010.04717.x

Manel, S., Schwartz, M. K., Luikart, G., and Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends Ecol. Evol.* 18, 189–197. doi: 10.1016/S0169-5347(03)00008-9

McDonald, J. H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351, 652–654. doi: 10.1038/351652a0

McRae, B. H. (2006). Isolation by resistance. *Evolution* 60, 1551–1561. doi: 10.1111/j.0014-3820.2006.tb00500.x

Nagylaki, T. (1978). A diffusion model for geographically structured populations. *J. Math. Biol.* 64, 375–382. doi: 10.1007/BF02463002

Pardo-Diaz, C., Salazar, C., and Jiggins, C. D. (2015). Towards the identification of the loci of adaptive evolution. *Methods Ecol. Evol.* 6, 445–464. doi: 10.1111/2041-210X.12324

Pavlidis, P., Jensen, J. D., Stephan, W., and Stamatakis, A. (2012). A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Mol. Biol. Evol.* 29, 3237–3248. doi: 10.1093/molbev/mss136

Pérez-Figueroa, A., García-Pereira, M. J., Saura, M., Rolán-Alvarez, E., and Caballero, A. (2010). Comparing three different methods to detect selective loci using dominant markers. *J. Evol. Biol.* 23, 2267–2276. doi: 10.1111/j.1420-9101.2010.02093.x

Perry, L., Dickau, R., Zarrillo, S., Holst, I., Pearsall, D. M., Piperno, D. R., et al. (2007). Starch fossils and the domestication and dispersal of chili peppers (Capsicum spp. L.) in the Americas. *Science* 315, 986–988. doi: 10.1126/science.1136914

Petkova, D., Novembre, J., and Stephens, M. (2016). Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* 48:94. doi: 10.1038/ng.3464

Poncet, B. N., Herrmann, D., Gugerli, F., Taberlet, P., Holderegger, R., Gielly, L., et al. (2010). Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. *Mol. Ecol.* 19, 2896–2907. doi: 10.1111/j.1365-294x.2010.04696.x

Pritchard, J. K., and Di Rienzo, A. (2010). Adaptation–not by sweeps alone. *Nat. Rev. Gen.* 11:665. doi: 10.1038/nrg2880

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Raj, A., Stephens, M., and Pritchard, J. K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP datasets. *Genetics* 114:164350. doi: 10.1534/genetics.114.164350

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., and Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Mol. Ecol.* 24, 4348–4370. doi: 10.1111/mec.13322

Rellstab, C., Zoller, S., Walthert, L., Lesur, I., Pluess, A. R., Graf, R., et al. (2016). Signatures of local adaptation in candidate genes of oaks (Quercus spp.) with respect to present and future climatic conditions. *Mol. Ecol.* 25, 5907–5924. doi: 10.1111/mec.13889

Reynolds, J., Weir, B. S., and Cockerham, C. C. (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* 105, 767–779.

Richardson, J. L., Urban, M. C., Bolnick, D. I., and Skelly, D. K. (2014). Microgeographic adaptation and the spatial scale of evolution. *Trends Ecol. Evol.* 29, 165–176. doi: 10.1016/j.tree.2014.01.002

Rockman, M. V. (2012). The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evol. Int. J. Orgn. Evol.* 66, 1–17. doi: 10.1111/j.1558-5646.2011.01486.x

Schlötterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014). Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* 15, 749–763. doi: 10.1038/nrg3803

Schoville, S. D., Bonin, A., François, O., Lobreaux, S., Melodelima, C., and Manel, S. (2012). Adaptive genetic variation on the landscape: methods and cases. *Annu. Rev. Ecol. Evol. Syst.* 43, 23–43. doi: 10.1146/annurev-ecolsys-110411-160248

Shirk, A. J., Landguth, E. L., and Cushman, S. A. (2017). A comparison of regression methods for model selection in individual-based landscape genetic analysis. *Mol. Ecol. Res.* 8, 55–67. doi: 10.1111/1755-0998.12709

Shryock, D. F., Havrilla, C. A., DeFalco, L. A., Esque, T. C., Custer, N. A., and Wood, T. E. (2015). Landscape genomics of *Sphaeralcea ambigua* in the Mojave Desert: a multivariate, spatially-explicit approach to guide ecological restoration. *Conserv. Genet.* 16, 1303–1317. doi: 10.1007/s10592-015-0741-1

Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., Bai, Z., et al. (2010). Genetic evidence for high-altitude adaptation in Tibet. *Science* 329, 72–75. doi: 10.1126/science.1189406

Skinner, M. K., Mohan, M., Haque, M. M., Zhang, B., and Savenkova, M. I. (2012). Epigenetic transgenerational inheritance of somatic transcriptomes and epigenetic control regions. *Genome Biol.* 13:R91. doi: 10.1186/gb-2012-13-10-r91

Slatkin, M. (1987). Gene flow and the geographical structure of natural populations. *Science* 236, 787–792. doi: 10.1126/science.3576198

Sork, V. L., Squire, K., Gugger, P. F., Steele, S. E., Levy, E. D., and Eckert, A. J. (2016). Landscape genomic analysis of candidate genes for climate adaptation in a California endemic oak, *Quercus lobata*. *Am. J. Bot.* 103, 33–46. doi: 10.3732/ajb.1500162

Stephan, W. (2015). Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Mol. Evol.* 25, 76–88. doi: 10.1111/mec.13288

Stern, D. L., and Orgogozo, V. (2008). The loci of evolution: how predictable is genetic evolution? *Evolution* 62, 2155–2177. doi: 10.1111/j.1558-5646.2008.00450.x

Stinchcombe, J. R., and Hoekstra, H. E. (2008). Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* 100, 158–170. doi: 10.1038/sj.hdy.6800937

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9440–9445. doi: 10.1073/pnas.1530509100

Storfer, A. (2015). *Landscape Genetics*. *Oxford Bibliographies in Evolutionary Biology*. Oxford, UK: Oxford University Press.

Storfer, A., Murphy, M. A., Spear, S. F., Holderegger, R., and Waits, L. P. (2010). Landscape genetics: where are we now?. *Mol. Ecol.* 19, 3496–3514. doi: 10.1111/j.1365-294X.2010.04691.x

Storfer, A., Murphy, M. A., Evans, J. S., Goldberg, C. S., Robinson, S., Spear, S. F., et al. (2007). Putting the 'landscape' in landscape genetics. *Heredity*. 98:128. doi: 10.1038/sj.hdy.6800917

Stucki, S., Orozco-terWengel, P., Forester, B. R., Duruz, D., Colli, L., et al. (2016). High performance computation of landscape genomic models including local indicators of spatial simulation. *Mol. Ecol. Res.* 17, 1072–1089. doi: 10.1111/1755-0998.12629

Sun, Z. X., Zhai, Y. F., Zhang, J. Q., Kang, K., Cai, J. H., Fu, Y., et al. (2015). The genetic basis of population fecundity prediction across multiple field populations of *Nilaparvata lugens*. *Mol. Ecol.* 24, 771–784. doi: 10.1111/mec.13069

Szymura, J. M., and Barton, N. H. (1986). Genetic analysis of a hybrid zone between the fire-bellied toads, Bombina bombina and *B. variegata*, near Cracow in southern Poland. *Evolution* 40, 1141–1159.

Szymura, J. M., and Barton, N. H. (1991). The genetic structure of the hybrid zone between the fire-bellied toads *Bombina bombina* and *B. variegata*: comparisons between transects and between loci. *Evolution* 45, 237–261. doi: 10.1111/j.1558-5646.1991.tb04400.x

Tiffin, P., and Ross-Ibarra, J. (2014). Advances and limits of using population genetics to understand local adaptation. *Trends Ecol. Evol.* 29, 673–680. doi: 10.1016/j.tree.2014.10.004

Trumbo, D. R., Spear, S. F., Baumsteiger, J., and Storfer, A. (2013). Rangewide landscape genetics of an endemic Pacific northwestern salamander. *Mol. Ecol.* 22, 1250–1266. doi: 10.1111/mec.12168

Verity, R., Collins, C., Card, D. C., Schaal, S. M., Wang, L., and Lotterhos, K. E. (2017). minotaur: a platform for the analysis and visualization of multivariate results from genome scans with R Shiny. *Mol. Ecol. Res.* 17, 33–43. doi: 10.1111/1755-0998.12579

Vincent, B., Dionne, M., Kent, M. P., Lien, S., and Bernatchez, L. (2013). Landscape genomics in Atlantic salmon (*Salmo salar*): searching for gene–environment interactions driving local adaptation. *Evolution* 67, 3469–3487. doi: 10.1111/evo.12139

Visscher, P. M., Yang, J., and Goddard, M. E. (2010). A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). *Twin Res. Hum. Genet.* 13, 517–524. doi: 10.1375/twin.13.6.517

Vitalis, R., Dawson, K., and Boursot, P. (2001). Interpretation of variation across marker loci as evidence of selection. *Genetics* 158, 1811–1823.

Wang, I. J., and Bradburd, G. S. (2014). Isolation by environment. *Mol. Ecol.* 23, 5649–5662. doi: 10.1016/s0160-4120(97)00049-4

Wang, Z., Ding, G., Chen, G., Sun, Y., Sun, Z., Zhang, H., et al. (2012). Genome sequences of wild and domestic bactrian camels. *Nat. Comm.* 3:1202. doi: 10.1038/ncomms2192

Waters, J. M., Fraser, C. I., and Hewitt, G. M. (2013). Founder takes all: density-dependent processes structure biodiversity. *Trends. Ecol. Evol.* 28, 78–85. doi: 10.1016/j.tree.2012.08.024

Wenzel, M. A., Douglas, A., James, M. C., Redpath, S. M., and Piertney, S. B. (2016). The role of parasite-driven selection in shaping landscape genomic structure in red grouse (*Lagopus lagopus scotica*). *Mol. Ecol.* 25, 324–341. doi: 10.1111/mec.13473

Whitlock, M. C., and Lotterhos, K. E. (2015). Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of FST. *Am. Nat.* 186, S24–S36. doi: 10.1086/682949

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A., Heath, A. C., et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* 44, 369–375. doi: 10.1038/ng.2213

Yang, T. P., Beazley, C., Montgomery, S. B., Dimas, A. S., Gutierrez-Arcelus, M., Stranger, B. E., et al. (2010). Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics*. 26, 2474–2476. doi: 10.1093/bioinformatics/btq452

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824. doi: 10.1038/ng.2310

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

MINI REVIEW
published: 12 December 2017
doi: 10.3389/fpls.2017.02136

Ten Years of Landscape Genomics: Challenges and Opportunities

*Yong Li[1], Xue-Xia Zhang[1], Run-Li Mao[1], Jie Yang[1], Cai-Yun Miao[1], Zhuo Li[1] and Ying-Xiong Qiu[2]**

[1] College of Forestry, Henan Agricultural University, Zhengzhou, China, [2] Key Laboratory of Conservation Biology for Endangered Wildlife of the Ministry of Education and Laboratory of Systematic and Evolutionary Botany and Biodiversity, College of Life Sciences, Zhejiang University, Hangzhou, China

Landscape genomics is a relatively new discipline that aims to reveal the relationship between adaptive genetic imprints in genomes and environmental heterogeneity among natural populations. Although the interest in landscape genomics has increased since this term was coined, studies on this topic remain scarce. Landscape genomics has become a powerful method to scan and determine the genes responsible for the complex adaptive evolution of species at population (mostly) and individual (more rarely) level. This review outlines the sampling strategies, molecular marker types and research categories in 37 articles published during the first 10 years of this field (i.e., 2007–2016). We also address major challenges and future directions for landscape genomics. This review aims to promote interest in conducting additional studies in landscape genomics.

Keywords: adaptive evolution, genetic structure, landscape genomics, molecular ecology, population genetics

OPEN ACCESS

Edited by:
Renchao Zhou,
Sun Yat-sen University, China

Reviewed by:
Charles Masembe,
Makerere University, Uganda
Yanjun Zhang,
Wuhan Botanical Garden (CAS), China
Michael Benjamin Kantar,
Hawaii University, United States

****Correspondence:***
Ying-Xiong Qiu
qyxhero@zju.edu.cn

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *14 June 2017*
Accepted: *01 December 2017*
Published: *12 December 2017*

Citation:
Li Y, Zhang X-X, Mao R-L, Yang J, Miao C-Y, Li Z and Qiu Y-X (2017) Ten Years of Landscape Genomics: Challenges and Opportunities. Front. Plant Sci. 8:2136. doi: 10.3389/fpls.2017.02136

INTRODUCTION

Rapid global climate change is an important factor that affects biodiversity (Hoffmann and Sgrò, 2011). Adjusting their distribution range or local adaptation is the usual coping strategy of species toward rapid climate change (Aitken et al., 2008). Local adaptation requires the species to face long-term spatial environmental heterogeneity and eventually leads to adaptive differentiation of phenotypes. These changes might be due to phenotypic plasticity or heritable phenotypic variation. Exploring the adaptive evolution of species in response to spatial environmental heterogeneity will be useful in understanding initial adaptive divergence and evolutionary potential of a target species (Pluess et al., 2016). Landscape genomics is a powerful research field for investigating the adaptive evolution of species in response to spatial environmental heterogeneity (Vincent et al., 2013).

Joost et al. (2007) proposed landscape genomics as a relatively new discipline that aims to reveal the relationship between the adaptive genetic imprints in genomes and the environmental heterogeneity. Different from landscape genetics, landscape genomics requires a sufficient number of molecular markers to cover the entire genome. Emphasis is placed on adaptive evolution at the genome level (Miao et al., 2017). Landscape genetics, however, is biased toward using a relatively small number of molecular markers to reveal the relationship between environmental factors and the spatial genetic structure of populations (Dionne et al., 2008; Poelchau and Hamrick, 2012; Manel and Holderegger, 2013). Landscape genomic studies on many plant and animal species have been recently conducted (Berg et al., 2015; Manthey and Moyle, 2015; Leamy et al., 2016; Vangestel et al., 2016). These studies have

achieved considerable progress on understanding of the relative roles of adaptive and non-adaptive processes in shaping patterns of genomic variation and the effects of environmental variables on adaptive differentiation at the genomic level. Although landscape genomics has been pursued for a decade, the studies, basic theoretical frameworks, and universal hypotheses in this field are still scarce. Thus, additional landscape genomic studies are needed to assist the construction of basic theoretical frameworks and formulation of universal hypotheses. This review summarizes the progress of landscape genomic studies such as conceptual and methodological developments as well as applied contributions during the previous decade. We then outline expected future directions in the field and encourage researchers to participate in this field.

METHODS

By searching the theme "landscape genomics" in the database of web of science[1], and further looking through the related papers carefully, 37 articles focused on adaptive genetic imprints in genomes driven by environmental factors were finally selected. Supplementary Table S1 lists the molecular markers, sampling strategies, statistical methods, and research categories addressed in these articles.

SAMPLING STRATEGIES IN LANDSCAPE GENOMICS

Sampling strategies in landscape genomics are divided into two major categories: random and stratified. The random sampling design includes scattered and clustered sampling. In the scattered sampling design, samples are randomly collected from across the species distribution range, while in the clustered sampling design, populations are divided into clusters according to environmental or genetic factors and samples are randomly taken from each cluster. A stratified sampling design can be performed to capture the range of variability across landscape variable(s) of interest. Thus, this sampling design requires a large amount of biological and environmental information of a target species. The optimal sampling scheme will be obtained by model calculation (Manel et al., 2012). The two sampling strategies mentioned above can be implemented at the individual or population level. The advantage of population sampling is more conducive to detect variation in gene frequency among populations than individual sampling. The most controversial topic in population sampling is multiple samples in fewer populations versus fewer samples in multiple populations. The former strategy is more representative in landscape genomic studies, but its accuracy in estimating genetic parameters is often questioned. In population genetics studies, the minimum sample size of a population should not be less than 20 individuals, 25–30 individuals are considered to be more reasonable (Hale et al., 2012). Therefore, it is necessary to ensure a minimum population sample size in the landscape genomic studies. Compared to population-based sampling, the application of individual sampling in landscape genomic studies is relatively scarce, but nonetheless suitable for clinal populations or those with unclear population structure (Jones et al., 2013).

MOLECULAR MARKERS IN LANDSCAPE GENOMICS

Landscape genomic studies require molecular markers that are sufficiently spread throughout the genome (Balkenhol et al., 2009). However, most non-model species do not have established genomic information to appropriately place sufficient markers across the genome. Therefore, two characteristics, i.e., no requirement for *a priori* genome knowledge and a high covering density in genomes, are indispensable for the use of these molecular markers in landscape genomics (Yang et al., 2017).

Two types of molecular markers are suitable for landscape genomic studies. Type-I markers have no DNA sequence information, such as amplified fragment length polymorphisms, inter-simple sequence repeats, and start codon targeted polymorphisms. Type-II markers contain DNA sequence information, such as single-nucleotide polymorphisms (SNPs). Type-I markers require low generation cost but have few defects. Although these type-I markers may allow detecting loci potentially responsible for adaptation using outlier locus detection and environmental association analysis (EAA), the gene function of those loci cannot be easily validated, and thus might be false-positives. Type-II markers usually display high scanning density but have higher generation cost than type-I markers. However, type-II markers exhibit several advantages because these markers contain DNA sequence information. This information can help us annotate and map these markers on the genome. Based on the landscape genomic studies that we have selected (see Supplementary Table S1), SNP genotyping was mainly achieved through DNA microarrays. However, the use of DNA microarrays requires a large amount of prior gene information (Teng and Xiao, 2009). The recently developed reduced-representation genome sequencing (RRGS) is based on next-generation sequencing (NGS), which includes genotyping by sequencing (Elshire et al., 2011), restricted site associated DNA (Miller et al., 2007), and specific-locus amplified fragment sequencing (Sun et al., 2013). RRGS reduces the cost of sequencing, maintains high coverage of the genome, and does not require *a priori* genomic information. Thus, the use of RRGS is beneficial in landscape genomic research (Brauer et al., 2016). Most of the RRGS methods are currently based on Illumina sequencing platforms, which have an advantage of high accuracy and throughput and a disadvantage of short reading lengths. Third generation sequencing (TGS), such as MinION device by Oxford Nanopores and PacBio Sequel by Pacific BioSciences, has been recently developed to compensate for the short reading length of NGS. Although TGS maintains the speed and flux advantages of NGS, this method still exhibits some problems, such as high cost and error rate, which must be addressed (Mikheyev and Tin, 2014). In summary, the use of type-II markers to conduct landscape genomic studies can

[1] http://www.isiknowledge.com/

facilitate the indirect validation of loci potentially responsible for adaptation.

MAJOR RESEARCH CATEGORIES IN LANDSCAPE GENOMICS

There are a wide variety of questions in ecology and evolution that can be addressed using a landscape genomic approach. We group these questions under two major research categories: (1) quantifying influence of spatial environmental variables on genomic divergence; (2) uncovering the environmental factors that shape adaptive genetic variation and the genetic basis of adaptive change.

QUANTIFYING INFLUENCE OF SPATIAL ENVIRONMENTAL VARIABLES ON GENOMIC DIVERGENCE

Isolation by distance (IBD) describes the local accumulation of genetic differences when dispersal between populations is geographically restricted (Slatkin, 1993). For IBD, gene flow path is assumed to be in a linear geographic distance. However, in natural landscapes, the paths of gene flow between populations are often non-linear and complex. In fact, populations that have identical habitats or small distances may also diverge when intervening landscape features inhibit dispersal between them (Isolation by Resistance, IBR; McRae, 2006; Ruiz-Gonzalez et al., 2015). Nevertheless, local genetic adaptation can also reduce gene flow among natural populations. This adaptive reduction in the effective rate of gene flow can contribute to a pattern of "isolation by environment" (IBE; Wang and Summers, 2010; Wang and Bradburd, 2014; Mosca et al., 2016). A strong pattern of IBE indicates that divergent selection is maintaining population differentiation in the face of possible dispersal (Schluter, 1998; Kawecki and Ebert, 2004). IBE can also arise from divergent habitat choice or other forms of biased dispersal (Armsworth and Roughgarden, 2008; Bolnick and Otto, 2013). Therefore, these different processes affect the spatial distribution of genetic variation and landscape genetic structure. Recently developed analytical methods can partition the often confounded patterns of IBD, IBE, and IBR when explaining genetic divergence across a landscape (Supplementary Table S1). The basic strategy is to use IBD as a null hypothesis against which IBE (or IBR) can be tested. Partial Mantel tests have been widely used in landscape genomics studies to evaluate the relative influence of different ecological and evolutionary factors on genetic differentiation. However, such tests have low statistical power and are prone to false positives (Guillot and Rousset, 2013). Recently, structural equation modelling (SEM) (Wang et al., 2013) and multiple matrix regression with randomization (MMRR) (Wang, 2013) have been used to quantitatively compare how much genetic divergence depends on IBD versus IBE (Zhang et al., 2016). In addition, the BEDASSLE package (Bradburd et al., 2013) is also used to estimate the relative contributions of IBD and IBE to genetic differentiation This Bayesian method models the allele frequencies in a set of populations at a set of unlinked loci as spatially correlated Gaussian processes, in which the covariance structure is a decreasing function of both geographic and ecological distance (Bradburd et al., 2013). In landscape genomic studies, multivariate statistical models are more appropriate when multidimensional niches are analyzed to identify ecological drivers of population genetic variation (Orsini et al., 2013). Redundancy analysis (RDA) (Legendre and Legendre, 2012) and canonical correlation analysis (CCA) (Parisod and Christin, 2008; Hecht et al., 2015) are commonly used to estimate the relative contribution of spatial and environmental variables. The CCA method can control for demographic effects if spatial autocorrelation is included in the model design, while RDA and partial RDA analyses are alternative and robust approaches that can control for spatial effects while analyzing others (Sork et al., 2013).

UNCOVERING THE ENVIRONMENTAL FACTORS THAT SHAPE ADAPTIVE GENETIC VARIATION AND THE GENETIC BASIS OF ADAPTIVE CHANGE

Polymorphic sites across species genomes will establish their adaptive differentiation to acclimatize to the heterogeneous environment. Landscape genomics attempts to detect these adaptive loci under selection and reveal potential environmental drivers of selection by using correlative methods. The detection of loci responsible for adaptation usually involves two steps. One is to detect the outlier loci; and the other is to associate the outlier loci with environment variables, referred to as EAA. The commonly used methods for detecting the outlier loci are ARLEQUIN (Excoffier et al., 2009), BAYESCAN (Foll and Gaggiotti, 2008), FLK (Bonhomme et al., 2010), and spatial ancestry analysis (SPA) (Yang et al., 2012). ARLEQUIN is applied to simulate a null distribution of F_{ST} values under a hierarchical island model, which is insensitive to the hierarchically subdivided population samples or those with a recently shared history. BAYESCAN is an F_{ST}-based model to identify outlier loci according to Bayesian posterior probability. FLK deals with variation in effective population size and historical branching of populations by incorporating a population kinship matrix into the Lewontin and Krakauer (LK) statistic (Lewontin and Krakauer, 1973). SPA is a probabilistic model for the spatial structure of genetic variation that is used to identify loci showing extreme patterns of spatial differentiation. Compared with the two F_{ST}-based approaches, SPA is particularly sensitive to strong spatial patterns in allele frequency and works at the individual level rather than at the population level. These methods are usually combined to distinguish the selected loci from the neutral loci and thus effectively reduce the false-positive rate (Wang et al., 2016). EAA, followed by outlier analysis, will be conducted to test whether these outlier loci are associated with particular environmental factors and under adaptive evolution.

The methods for conducting EAA can be divided into five broadly defined categories, including categorical tests,

logistic regressions, matrix correlations, general linear models, and mixed effects models (Rellstab et al., 2015). A first category contains categorical tests, which compares allele frequencies of individuals or populations from different types of environments. The different types of environment are introduced as categorical variables in parametric or non-parametric tests. A second category comprises the statistical methods of logistic regressions, such as SAM, Samβada. The spatial analysis method (SAM; Joost et al., 2007) is the first implementation of logistic regression in EAA. SAM can compute multiple simultaneous univariate logistic regressions to test for association between allelic frequencies and environmental variables. However, this approach ignores neutral genetic structure, possibly leading to high false-positive rates under various demographic scenarios (De Mita et al., 2013; Frichot et al., 2013). Nevertheless, an extended version of SAM, Samβada (Joost et al., 2007) improves the performance of this method by adding neutral genetic structure as an additional factor (Rellstab et al., 2015). A third category contains a linear approach, matrix correlations, in which the effects of environmental factors and neutral genetic structure on allele frequencies are simultaneously estimated. The most widely used methods include a simple Mantel test and the partial Mantel test (Mantel, 1967). However, variations of the (partial) Mantel test may circumvent certain bias and autocorrelation problems (Legendre, 1993; Legendre et al., 2002). A fourth important category of statistical methods is general linear models in which a response variable is modeled as a linear function of some set of explanatory variables. The general linear model framework can be extended to models with multivariate response variables to account for the polygenic architecture of adaptive traits (Rellstab et al., 2015). The statistical methods include multiple linear regressions and univariate general linear models (Carl and Kuhn, 2007; Eckert et al., 2009) and canonical correlations and multivariate linear regressions, e.g., CCA (ter Braak and Smilauer, 2002; Legendre and Legendre, 2012) and RDA (Legendre and Legendre, 2012; Hecht et al., 2015). A fifth important category of statistical methods comprises the mixed effects models, such as BAYENV (Coop et al., 2010), LFMMs (Frichot et al., 2013), TASSEL (Bradbury et al., 2007), and EMMA (Kang et al., 2008). These approaches provide a unified statistical framework for controlling for the effects of neutral genetic structure (Rellstab et al., 2015). For example, BAYENV, based on a Bayesian generalized linear mixed model, is applied to test the correlation between allelic frequencies and environmental variables after correcting for population structure and size (Günther and Coop, 2013). Latent factor mixed models (LFMMs) implemented fast algorithms using a hierarchical Bayesian mixed model based on a variant of principal component analysis (PCA), in which the residual population structure is introduced via unobserved or latent factors (Frichot et al., 2013; Caye et al., 2016). In addition, a linear mixed-model method implemented in TASSEL (Bradbury et al., 2007) is used to identify candidate loci responsible for adaptation according to the association between the genotypes and climate variables (Yoder et al., 2014). Based on linear mixed models, Kang et al. (2008) developed an efficient mixed-model association (EMMA) method. As previously mentioned, in order to reduce the false-positive rate, it is desirable to combine more than two statistical methods to identify the environment-associated loci (Yang et al., 2017).

MAJOR CHALLENGES

Although great progress in landscape genomics has been achieved in the past decade, two major challenges remain to be solved in the future. One is the presence of false positives, which have been a major problem in landscape genomics because of the lack of validation for adaptive loci. Three solutions will help solve this major challenge. First, robust detection methods must be developed, and multiple detection methods must be used to reduce the false-positive rates. Second, type-II markers that contain DNA sequence information must be selected. Although type-I markers may allow detecting loci potentially responsible for adaptation, the gene function of these detected loci are difficult to be validated. Type-II markers have DNA sequence information, which can be indirectly validated through the annotation of gene function. Third, a part of the loci responsible for adaptation must be validated using gene transfer and gene knockout technologies. Since most of previous landscape genomics studies have focused on non-model species, the detected loci responsible for adaptation do not have functional verification. Thus, in future, more experiments are needed to validate the function and adaptive generality of the detecting loci responsible for adaptation. In addition, most previous studies have showed great concern on gene differentiation rather than phenotypic differentiation (Manthey and Moyle, 2015; Di Pierro et al., 2016). The acquisition of adaptive phenotypic data has been conducted in a few recent landscape genomic studies (De Kort et al., 2014; Roschanski et al., 2016). Thus, obtaining the phenotypic data through common garden experiments and reciprocal transplant experiments should be considered in future.

RECOMMENDATIONS FOR FUTURE RESEARCH

The present landscape genomics mainly addresses two issues, i.e., influence of spatial environmental variables on genomic divergence and effects of the environmental factors on adaptive genetic variation. The following concerns need to be addressed in landscape genomic studies. (1) Previous studies have determined the specific genes that undergo adaptive changes and the environmental factors that contribute to these changes. However, the specific reason why these particular genes or environmental variables exhibit these functions remains unknown. (2) Type-II markers can help us reveal these specific genes. However, the metabolic pathways of the involved genes and the adaptive phenotypes controlled by these genes need to be identified. (3) Regional species in extreme environments usually establish some convergent adaptive changes in their genes or

phenotypes. However, most regional species not living in extreme environments have various adaptive differentiations. Thus, the commonalities behind these diverse adaptive differentiations must be determined. (4) The distribution range of species and their ability to respond to climate change largely depend on their landscape adaptability, which is usually determined by the potential adaptive differentiation of the genome and the gene dispersal ability of the species. Thus, a landscape adaptation index must be established to measure the adaptability of species. In summary, landscape genomics is an efficient method to study the adaptive evolution of species. We hope that this review of studies on landscape genomics over the past 10 years will assist in promoting future research in this field.

AUTHOR CONTRIBUTIONS

X-XZ and R-LM wrote the original draft of article; JY, C-YM, and ZL revised this article; Y-XQ and YL conceived the ideas and contributed to substantial revisions; all authors read and approved the final version of the manuscript.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (31770225), the Henan Agricultural University Science and Technology Innovation Fund (KJCX2016A2), the Funding Scheme of Young Backbone Teachers of Higher Education Institutions in Henan Province (2015GGJS-081), and the Key Scientific Research Projects of Henan Higher School (16A220002).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2017.02136/full#supplementary-material

REFERENCES

Aitken, S. N., Yeaman, S., Holliday, J. A., Wang, T. L., and Curtis-McLane, S. (2008). Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evol. Appl.* 1, 95–111. doi: 10.1111/j.1752-4571.2007.00013.x

Armsworth, P. R., and Roughgarden, J. E. (2008). The structure of clines with fitness-dependent dispersal. *Am. Nat.* 172, 648–657. doi: 10.1086/591685

Balkenhol, N., Waits, L. P., and Dezzani, R. J. (2009). Statistical approaches in landscape genetics: an evaluation of methods for linking landscape and genetic data. *Ecography* 32, 818–830. doi: 10.1111/j.1600-0587.2009.05807.x

Berg, P. R., Jentoft, S., Star, B., Ring, K. H., Knutsen, H., Lien, S., et al. (2015). Adaptation to low salinity promotes genomic divergence in Atlantic Cod (*Gadus morhua* L.). *Genome Biol. Evol.* 7, 1644–1663. doi: 10.1093/gbe/evv093

Bolnick, D. I., and Otto, S. P. (2013). The magnitude of local adaptation under genotype-dependent dispersal. *Ecol. Evol.* 3, 4722–4735. doi: 10.1002/ece3.850

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., et al. (2010). Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics* 186, 241–262. doi: 10.1534/genetics.104.117275

Bradburd, G. S., Ralph, P. L., and Coop, G. M. (2013). Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution* 67, 3258–3273. doi: 10.1111/evo.12193

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Brauer, C. J., Hammer, M. P., and Beheregaray, L. B. (2016). Riverscape genomics of a threatened fish across a hydroclimatically heterogeneous river basin. *Mol. Ecol.* 25, 5093–5113. doi: 10.1111/mec.13830

Carl, G., and Kuhn, I. (2007). Analyzing spatial autocorrelation in species distributions using Gaussian and logit models. *Ecol. Model.* 207, 159–170. doi: 10.1016/j.ecolmodel.2007.04.024

Caye, K., Deist, T. M., Martins, H., Michel, O., and François, O. (2016). TESS3: fast inference of spatial population structure and genome scans for selection. *Mol. Ecol. Resour.* 16, 540–548. doi: 10.1111/1755-0998.12471

Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J. K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics* 185, 1411–1423. doi: 10.1534/genetics.110.114819

De Kort, H., Vandepitte, K., Bruun, H. H., Closset-Kopp, D., Honnay, O., and Mergeay, J. (2014). Landscape genomics and a common garden trial reveal adaptive differentiation to temperature across Europe in the tree species *Alnus glutinosa*. *Mol. Ecol.* 23, 4709–4721. doi: 10.1111/mec.12813

De Mita, S., Thuillet, A. C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J., et al. (2013). Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Mol. Ecol.* 22, 1383–1399. doi: 10.1111/mec.12182

Di Pierro, E. A., Mosca, E., Rocchini, D., Binelli, G., Neale, D. B., and La Porta, N. (2016). Climate-related adaptive genetic variation and population structure in natural stands of Norway spruce in the South-Eastern Alps. *Tree Genet. Genomes* 12:16. doi: 10.1007/s11295-016-0972-4

Dionne, M., Caron, F., Dodson, J. J., and Bernatchez, L. (2008). Landscape genetics and hierarchical genetic structure in Atlantic salmon: the interaction of gene flow and local adaptation. *Mol. Ecol.* 17, 2382–2396. doi: 10.1111/j.1365-294X.2008.03771.x

Eckert, A. J., Pande, B., Ersoz, E. S., Wright, M. H., Rashbrook, V. K., Nicolet, C. M., et al. (2009). High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (*Pinus taeda* L.). *Tree Genet. Genomes* 5, 225–234. doi: 10.1007/s11295-008-0183-8

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple Genotyping-By-Sequencing (GBS) approach for high diversity species. *PLOS ONE* 6:e19379. doi: 10.1371/journal.pone.0019379

Excoffier, L., Hofer, T., and Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity* 103, 285–298. doi: 10.1038/hdy.2009.74

Foll, M., and Gaggiotti, O. (2008). A genome scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180, 977–993. doi: 10.1534/genetics.108.092221

Frichot, E., Schoville, S. D., Bouchard, G., and François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol. Biol. Evol.* 30, 1687–1699. doi: 10.1093/molbev/mst063

Guillot, G., and Rousset, F. (2013). Dismantling the Mantel tests. *Methods Ecol. Evol.* 4, 336–344. doi: 10.1111/2041-210x.12018

Günther, T., and Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics* 195, 205–220. doi: 10.1534/genetics.113.152462

Hale, M. L., Burg, T. M., and Steeves, T. E. (2012). Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. *PLOS ONE* 7:e45170. doi: 10.1371/journal.pone.0045170

Hecht, B. C., Matala, A. P., Hess, J. E., and Narum, S. R. (2015). Environmental adaptation in Chinook salmon (*Oncorhynchus tshawytscha*) throughout their North American range. *Mol. Ecol.* 24, 5573–5595. doi: 10.1111/mec.13409

Hoffmann, A. A., and Sgrò, C. M. (2011). Climate change and evolutionary adaptation. *Nature* 470, 479–485. doi: 10.1038/nature09670

Jones, M. R., Forester, B. R., Teufel, A. I., Adams, R. V., Anstett, D. N., Goodrich, B. A., et al. (2013). Integrating landscape genomics and spatially explicit approaches to detect loci under selection in clinal populations. *Evolution* 67, 3455–3468. doi: 10.1111/evo.12237

Joost, S., Bonin, A., Bruford, M. W., Després, L., Conord, C., Erhardt, G., et al. (2007). A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Mol. Ecol.* 16, 3955–3969. doi: 10.1111/j.1365-294X.2007.03442.x

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–1723. doi: 10.1534/genetics.107.080101

Kawecki, T. J., and Ebert, D. (2004). Conceptual issues in local adaptation. *Ecol. Lett.* 7, 1225–1241. doi: 10.1111/j.1461-0248.2004.00684.x

Leamy, L. J., Lee, C. R., Song, Q. J., Mujacic, I., Luo, Y., Chen, C. Y., et al. (2016). Environmental versus geographical effects on genomic variation in wild soybean (*Glycine soja*) across its native range in northeast Asia. *Ecol. Evol.* 6, 6332–6344. doi: 10.1002/ece3.2351

Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? *Ecology* 74, 1659–1673. doi: 10.2307/1939924

Legendre, P., Dale, M. R. T., Fortin, M. J., Gurevitch, J., Hohn, M., and Myers, D. (2002). The consequences of spatial structure for the design and analysis of ecological field surveys. *Ecography* 25, 601–615. doi: 10.1034/j.1600-0587.2002.250508.x

Legendre, P., and Legendre, L. (2012). *Numerical Ecology*. Oxford: Elsevier.

Lewontin, R. C., and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74, 175–195.

Manel, S., Albert, C. H., and Yoccoz, N. G. (2012). "Sampling in landscape genomics," in *Data Production and Analysis in Population Genomics*, eds F. Pompanon and A. Bonin (New York, NY: Humana Press), 3–12.

Manel, S., and Holderegger, R. (2013). Ten years of landscape genetics. *Trends Ecol. Evol.* 28, 614–621. doi: 10.1016/j.tree.2013.05.012

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27, 209–220.

Manthey, J. D., and Moyle, R. G. (2015). Isolation by environment in white-breasted Nuthatches (*Sitta carolinensis*) of the Madrean Archipelago sky islands: a landscape genomics approach. *Mol. Ecol.* 24, 3628–3638. doi: 10.1111/mec.13258

McRae, B. H. (2006). Isolation by resistance. *Evolution* 60, 1551–1561. doi: 10.1111/j.0014-3820.2006.tb00500.x

Miao, C. Y., Li, Y., Yang, J., and Mao, R. L. (2017). Landscape genomics reveal that ecological character determines adaptation: a case study in smoke tree (*Cotinus coggygria* Scop.). *BMC Evol. Biol.* 17:202. doi: 10.1186/s12862-017-1055-3

Mikheyev, A. S., and Tin, M. M. Y. (2014). A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.* 14, 1097–1102. doi: 10.1111/1755-0998.12324

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., and Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* 17, 240–248. doi: 10.1101/gr.5681207

Mosca, E., Gugerli, F., Eckert, A. J., and Neale, D. B. (2016). Signatures of natural selection on *Pinus cembra*, and P. mugo, along elevational gradients in the Alps. *Tree Genet. Genomes* 12:9. doi: 10.1007/s11295-015-0964-9

Orsini, L., Mergeay, J., Vanoverbeke, J., and De Meester, L. (2013). The role of selection in driving landscape genomic structure of the waterflea *Daphnia magna*. *Mol. Ecol.* 22, 583–601. doi: 10.1111/mec.12117

Parisod, C., and Christin, P. A. (2008). Genome-wide association to fine-scale ecological heterogeneity within a continuous population of *Biscutella laevigata* (Brassicaceae). *New Phytol.* 178, 436–447. doi: 10.1111/j.1469-8137.2007.02361.x

Pluess, A. R., Frank, A., Heiri, C., Lalagüe, H., Vendramin, G. G., and Oddou-Muratorio, S. (2016). Genome-environment association study suggests local adaptation to climate at the regional scale in *Fagus sylvatica*. *New Phytol.* 210, 589–601. doi: 10.1111/nph.13809

Poelchau, M. F., and Hamrick, J. L. (2012). Differential effects of landscape-level environmental features on genetic structure in three codistributed tree species in Central America. *Mol. Ecol.* 21, 4970–4982. doi: 10.1111/j.1365-294X.2012.05755.x

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., and Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Mol. Ecol.* 24, 4348–4370. doi: 10.1111/mec.13322

Roschanski, A. M., Csilléry, K., Liepelt, S., Oddou-Muratorio, S., Ziegenhagen, B., Huard, F., et al. (2016). Evidence of divergent selection for drought and cold tolerance at landscape and local scales in *Abies alba* Mill. in the French Mediterranean Alps. *Mol. Ecol.* 25, 776–794. doi: 10.1111/mec.13516

Ruiz-Gonzalez, A., Cushman, S. A., Madeira, M. J., Randi, E., and Gómez-Moliner, B. J. (2015). Isolation by distance, resistance and/ or clusters? lessons learned from a forest-dwelling carnivore inhabiting a heterogeneous landscape. *Mol. Ecol.* 24, 5110–5129. doi: 10.1111/mec.13392

Schluter, D. (1998). "Ecological causes of speciation," in *Endless Forms: Species and Speciation*, eds D. J. Howard and S. H. Berlochers (Oxford: Oxford University Press), 114–129.

Slatkin, M. (1993). Isolation by distance in equilibrium and nonequilibrium populations. *Evolution* 47, 264–279. doi: 10.1111/j.1558-5646.1993.tb01215.x

Sork, V. L., Aitken, S. N., Dyer, R. J., Eckert, A. J., Legendre, P., and Neale, D. B. (2013). Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. *Tree Genet. Genomes* 9, 901–911. doi: 10.1007/s11295-013-0596-x

Sun, X., Liu, D., Zhang, X., Li, W., Liu, H., Hong, W., et al. (2013). SLAF-seq: an efficient method of large-scale de novo SNP discovery and genotyping using high-throughput sequencing. *PLOS ONE* 8:e58700. doi: 10.1371/journal.pone.0058700

Teng, X. K., and Xiao, H. S. (2009). Perspectives of DNA microarray and next-generation DNA sequencing technologies. *Sci. China Life Sci.* 52, 7–16. doi: 10.1007/s11427-009-0012-9

ter Braak, C. J. F., and Smilauer, P. (2002). *CANOCO Reference Manual and CanoDraw for Windows User's Guide: Software for Canonical Community Ordination (version 4.5)*. New York, NY: Microcomputer Power.

Vangestel, C., Vázquez-Lobo, A., Martínez-García, P. J., Calic, I., Wegrzyn, J. L., and Neale, D. B. (2016). Patterns of neutral and adaptive genetic diversity across the natural range of sugar pine (*Pinus lambertiana*, Dougl.). *Tree Genet.Genomes* 12:15. doi: 10.1007/s11295-016-0998-7

Vincent, B., Dionne, M., Kent, M. P., Lien, S., and Bernatchez, L. (2013). Landscape genomics in Atlantic salmon (*Salmo salar*): searching for gene-environment interactions driving local adaptation. *Evolution* 67, 3469–3487. doi: 10.1111/evo.12139

Wang, I. J. (2013). Examining the full effects of landscape heterogeneity on spatial genetic variation: a multiple matrix regression approach for quantifying geographic and ecological isolation. *Evolution* 67, 3403–3411. doi: 10.1111/evo.12134

Wang, I. J., and Bradburd, G. S. (2014). Isolation by environment. *Mol. Ecol.* 23, 5649–5662. doi: 10.1111/mec.12938

Wang, I. J., Glor, R. E., and Losos, J. B. (2013). Quantifying the roles of ecology and geography in spatial genetic divergence. *Ecol. Lett.* 16, 175–182. doi: 10.1111/ele.12025

Wang, I. J., and Summers, K. (2010). Genetic structure is correlated with phenotypic divergence rather than geographic isolation in the highly polymorphic strawberry poison-dart frog. *Mol. Ecol.* 19, 447–458. doi: 10.1111/j.1365-294X.2009.04465.x

Wang, T., Wang, Z., Xia, F., and Su, Y. J. (2016). Local adaptation to temperature and precipitation in naturally fragmented populations of *Cephalotaxus oliveri*, an endangered conifer endemic to China. *Sci. Rep.* 6:25031. doi: 10.1038/srep25031

Yang, J., Miao, C. Y., Mao, R. L., and Li, Y. (2017). Landscape population genomics of forsythia (*Forsythia suspensa*) reveal that ecological habitats determine the adaptive evolution of species. *Front. Plant Sci.* 8:481. doi: 10.3389/fpls.2017.00481

Yang, W. Y., Novembre, J., Eskin, E., and Halperin, E. (2012). A model-based approach for analysis of spatial structure in genetic data. *Nat. Genet.* 44, 725–731. doi: 10.1038/ng.2285

Yoder, J. B., Stanton-Geddes, J., Zhou, P., Briskine, R., Young, N. D., and Tiffin, P. (2014). Genomic signature of adaptation to climate in *Medicago truncatula*. *Genetics* 196, 1263–1275. doi: 10.1534/genetics.113.159319

Zhang, Y. H., Wang, I. J., Comes, H. P., Peng, H., and Qiu, Y. X. (2016). Contributions of historical and contemporary geographic and environmental factors to phylogeographic structure in a Tertiary relict species, *Emmenopterys henryi* (Rubiaceae). *Sci. Rep.* 6:24041. doi: 10.1038/srep24041

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Ecology and Evolution

REVIEW
published: 28 April 2017
doi: 10.3389/fevo.2017.00033

Simple Rules for an Efficient Use of Geographic Information Systems in Molecular Ecology

Kevin Leempoel[1], Solange Duruz[1], Estelle Rochat[1], Ivo Widmer[1], Pablo Orozco-terWengel[2] and Stéphane Joost[1]*

[1] Laboratory of Geographic Information Systems (LASIG), School of Architecture, Civil and Environmental Engineering (ENAC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, [2] Biomedical Science, Cardiff University, Cardiff, UK

Geographic Information Systems (GIS) are becoming increasingly popular in the context of molecular ecology and conservation biology thanks to their display options efficiency, flexibility and management of geodata. Indeed, spatial data for wildlife and livestock species is becoming a trend with many researchers publishing genomic data that is specifically suitable for landscape studies. GIS uniquely reveal the possibility to overlay genetic information with environmental data and, as such, allow us to locate and analyze genetic boundaries of various plant and animal species or to study gene-environment associations (GEA). This means that, using GIS, we can potentially identify the genetic bases of species adaptation to particular geographic conditions or to climate change. However, many biologists are not familiar with the use of GIS and underlying concepts and thus experience difficulties in finding relevant information and instructions on how to use them. In this paper, we illustrate the power of free and open source GIS approaches and provide essential information for their successful application in molecular ecology. First, we introduce key concepts related to GIS that are too often overlooked in the literature, for example coordinate systems, GPS accuracy and scale. We then provide an overview of the most employed open-source GIS-related software, file formats and refer to major environmental databases. We also reconsider sampling strategies as high costs of Next Generation Sequencing (NGS) data currently diminish the number of samples that can be sequenced per location. Thereafter, we detail methods of data exploration and spatial statistics suited for the analysis of large genetic datasets. Finally, we provide suggestions to properly edit maps and to make them as comprehensive as possible, either manually or trough programming languages.

Keywords: Geographic Information Systems, spatial analysis, landscape genetics, gene-environment associations, open-source software, geographic map

OPEN ACCESS

Edited by:
Samuel A. Cushman, United States Forest Service Rocky Mountain Research Station, USA

Reviewed by:
Yessica Rico, Institute of Ecology, Mexico
Rodolfo Jaffé, Vale Institute of Technology, Brazil

****Correspondence:***
Kevin Leempoel
k.leempoel@gmail.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution

Received: *11 October 2016*
Accepted: *31 March 2017*
Published: *28 April 2017*

Citation:
Leempoel K, Duruz S, Rochat E, Widmer I, Orozco-terWengel P and Joost S (2017) Simple Rules for an Efficient Use of Geographic Information Systems in Molecular Ecology. Front. Ecol. Evol. 5:33. doi: 10.3389/fevo.2017.00033

INTRODUCTION

Geographic Information Systems (GIS) are powerful tools to be used in the context of evolutionary studies. They are designed to store, handle, display, and analyze any kind of data representing objects (individuals, populations, areas, etc.) characterized by geographic coordinates (X = longitude and Y = latitude). With the help of GIS, geographic information can be combined with for example phenotype, genotype, or environmental data to display the spatial

distribution of genetic variants and to visualize factors influencing spatial evolutionary processes. The main advantage of GIS in evolutionary biology is to easily explore and display genetic information (neutral and adaptive genetic variation, gene flow) at multiple scales, and to overlay this information with physical barriers, land cover or topographic maps in order to generate subsequent analyses regarding the location and causes of genetic boundaries (Epperson, 2003; Manel and Holderegger, 2013). GIS have also proven useful in adaptive landscape genomics or gene-environment associations (GEA) studies, in the context of which they enable the retrieval of environmental variables at sampling locations. The integration of GIS with approaches from landscape ecology and population genetics, defined as landscape genetics by Manel et al. (2003), also has important implications for conservation biology (Petren, 2013).

However, we faced a paradigm change a few years ago with the advent of Next Generation Sequencing (NGS) data, whose use requires rethinking study pipelines. First, NGS currently presents an economic constraint as it is costly compared with genetic markers produced so far (e.g., microsatellites and AFLP). Consequently, we are unable to fully sequence several hundreds of individuals, requiring a careful selection of the samples to analyze. Appropriate sampling is thus key to achieve a precise and continuous evaluation of environment-driven selection on the genome (see **Box 1**; Manel et al., 2012; Hand et al., 2015; Rellstab et al., 2015). In addition, NGS datasets are large and must be treated differently to be efficient and to avoid computer memory overload. Finally, NGS data require new tools to display and analyze spatial patterns that are more computationally demanding.

The successful application of GIS tools is not intuitive for many biologists who are not familiar with the concepts relating GIS and the use of GIS software. Indeed, a large diversity of GIS tools is available and the difficulty of finding relevant information and instructions is an obstacle for non-expert users. To date, few scientific articles have defined the role of GIS in molecular ecology. For instance, Kozak et al. (2008) review the fast development of GIS-based environmental data and advocate for their usage as an alternative to unprecise proxies such as latitude of distance between populations. Another review by Joost et al. (2010) provided guidelines for GIS use in livestock genetics and enumerate the advantages of integrating data in a GIS environment. More recently, Rogers and Staub (2013) outlined spatial analyses and GIS methods in honey bees research. Their review is not specific to bees but instead aim to intensify the exploration of the spatial component of studies in ecology and related disciplines. Lastly, Balkenhol et al. (2015) published a book detailing the concepts and analytical steps of landscape genetics studies, such as sampling design, spatial analysis and environmental datasets. Also, GIS are exploited in many unrelated domains and it is thus difficult to find resources specifically targeted at biologists. The bases of GIS practices are readily found in freely available Massive Open Online Courses (MOOCs), such as the Coursera platform (Coursera, 2012) currently offering six courses on GIS. Yet, these reviews do not tackle the challenges brought by large genetic and environmental datasets, and fail to review the recurrent caveats related to spatial research. In this paper, we highlight the usefulness of GIS in population and landscape genomics and provide key information for their successful application to these fields.

GEOGRAPHIC COORDINATES

Geographic coordinates of samples constitute an invaluable source of information, ranging from the display of their spatial distribution to the retrieval of environmental variables. Whenever doing fieldwork, using a GPS is the best way to record the coordinates of samples. As such, we strongly recommend recording the location of each sample, instead of the location of the centroid of a population for instance. Firstly, it allows for a more precise retrieval of environmental values. Secondly, attributing the same location to several samples invokes pseudo-replication, a statistical bias that must be addressed in further analysis. Thirdly, coordinates of nearby individuals allow for a proper measurement of dispersal, using for example pairwise genetic relationship with distance. Regarding GPS devices, standard GPS, and to a lesser extent smartphones, are accurate enough in most cases. However, more precise devices, such as DGPS (differential GPS), are recommended for local scale studies in which samples are located less than a couple of meters apart: the precision of the location has to stay within the spatial resolution of the grain.

When GPS coordinates are not recorded, it is still possible to approximate sample locations with the help of satellite images or by encoding the address of the location (georeferencing or geocoding), although with a lower accuracy. In the former case, creating a new vector layer overlaid on a satellite image or an online map (see next Section) allows the recovering of samples coordinates from an approximately known location (e.g., a crossroad, a tree, a river; Docs.QGIS, 2014). For the latter case, plugins have been developed to read text delimited file containing addresses (e.g., your own house address) that you want to locate (for example the MMQGIS plugin in QGIS, Mangomap, 2012; MMQGIS Plugin, 2012). It must be noted that each line must contain the address, city, state and country.

Another essential consideration is choosing the relevant coordinate reference system. Indeed, GPS devices display the coordinates of a point in latitude and longitude values, usually in the World Geodetic System (WGS84). This is a global reference system in which the Earth is represented by an ellipsoid, and every position on the surface is defined by two angles at the center of the Earth: the latitude and longitude. However, projected systems for which a geographical location is converted from the ellipsoid (distances expressed in degrees) to a corresponding location on a two-dimensional surface (x and y expressed in meters) are preferred for analyses. It is important to note that, although global systems covering the whole planet exist, each country or region has its own coordinate system that is locally more accurate than the global system. Where no national projected system exists, it is still possible to use the Universal Transverse Mercator (UTM) coordinate system, a projected coordinate system covering the entire globe and dividing it into sixty 6°-wide longitudinal zones (Dmap, 1993). Even though

GIS software usually deal with different projection systems, the manual reprojection of all layers into the same local projection system is recommended to avoid potential incompatibilities (see next section). However, different GIS may not exactly use the same name for a coordinate system. Therefore, to facilitate the identification of coordinate systems across the diversity of GIS software, the EPSG (European Petroleum Survey Group) database (EPSG, 1985) is a widely used database referencing all projected coordinate systems, implemented in every GIS and providing them with a unique ID (Maling, 1992), e.g., EPSG: 4326 correspond to the WGS84 reference system.

SOFTWARE

There are many GIS software, with different functions and aimed at various audiences. Universal GIS software do not exist and, therefore, the choice is difficult for a beginner. Today, one of the most user-friendly GIS is QGIS (QGIS Development Team, 2015). It is ideal to explore geodata, able to read and convert a wide variety of input formats and suitable to produce high-quality maps. Note that QGIS, and all other GIS mentioned in this paper, is free and open source. Open source GIS can indeed perform the same tasks as their commercial counterparts, and include the opportunity to understand and improve GIS algorithms or enable a better collaboration as there is no problem related to license access (Ertz et al., 2014). In addition, a large community exists to support development efforts of open source GIS, and regularly creates extensions to add functions and improvements to the software. Forums and tutorial websites are also flourishing for newcomers (http://gis.stackexchange.com/, http://www.qgistutorials.com/, Sutton et al., 2009).

On the other hand, most analysis in GIS are not easily replicable and, therefore, programming languages such as R can be more efficient. R has been successfully used as a GIS for a long time and several packages and reviews have been published (Rodriguez-Sanchez, 2013; Brunsdon and Comber, 2015). Among them, we can mention *rgdal* for the importation of geodata (Bivand et al., 2016), *GISTools* for general GIS operations (Brunsdon and Chen, 2014), *rasters* for their display (Hijmans and van Etten, 2015), *spdep*, and *spatstat* for spatial statistics and analysis (Baddeley and Turner, 2005; Bivand and Piras, 2015). While these packages are relatively efficient to import, display large rasters and vectors, customization options are more limited than in dedicated GIS.

MAIN DATASET

The first step in a GIS project is usually to import a vector file containing samples coordinates. QGIS has a plugin to easily import GPS coordinates, either directly from a GPS device or through vector files, such as .kml or .gpx (Docs.QGIS, 2013). These formats are usually converted to shapefiles (.shp) due to the easier management of their attributes and projection system associated with vector units. Delimited text files (e.g., tabulator—tab—or space delimited) can be easily opened in QGIS as well and then be transformed into shapefiles. When opening a text file using "Add delimited text layer," QGIS should recognize automatically the delimiter used and the columns of coordinates (X Y, Latitude Longitude) (QGIStutorials, 2014). However, such delimited text files cannot be transformed to polygons or lines. In this case, one should already have a shapefile incorporating lines or polygons to which the text table can be joined. To do so, the shapefile and the text table should have the same column of unique IDs. When clicking on the properties of the shapefile, an option is proposed to join additional tables of attributes (QGIStutorial, 2014). As mentioned in the previous section, it is recommended to project all layers in the same coordinate system. In QGIS, this is done by right-clicking on the layer and by changing the coordinate system in the "save as" option. The newly projected layer will then be automatically loaded to the project. See Rogers and Staub (2013) for a more extensive review of the basic tasks in QGIS.

BACKGROUND DATASET

The second step is to add one or more background layer(s) to constitute the geographic context, either from raster data (see next section) or from an online map (Google, Bing, Open Street Map). The OpenLayer plugin in QGIS allows the addition of a background base-map to the QGIS interface (QGIS workshop, 2013). When using raster layers such as Elevation data or climatic variables, adding a semi-transparent shaded relief will enhance the contrast and reveal the topography. To this end, QGIS has a Terrain analysis module in which a hill-shade layer can be computed from a Digital Elevation Model (DEM, i.e., a matrix of elevation data). Then, the transparency of the layer can be adjusted in its properties. In addition, it is advisable to cut rasters and vectors to the size of the study area using the clipper tool to facilitate their display and reduce computation time. Note that the succession of layers in the main frame depends on the order of layers shown on the left panel of the application.

ENVIRONMENTAL AND LANDSCAPE VARIABLES

Environmental datasets have considerably evolved and represent new opportunities for the identification of environmental drivers of adaptation. One of the main applications of GIS software in landscape genomics is to extract values of environmental variables at the exact location where samples have been collected, or from the surrounding area by means of polygons representing a buffer, a forest, or a specific land cover class for instance. As databases containing georeferenced environmental variables are numerous, we propose a list of the 10 most important publicly accessible databases in **Table 1** (A more extensive list is proposed in Appendix 1, Supplementary Material). Raster environmental data are often delivered in geotiff (.tif) or Band Interleaved by Line (.bil) formats, similar to satellite images but containing only one layer of information (i.e., Temperature, Precipitation etc.). Regarding climate datasets, many studies rely on variables interpolated at large geographical scales on the basis of data provided by weather stations and distributed across territories,

TABLE 1 | Ten main public sources of environmental data (URLs: consulted on June 10, 2016).

Name	URL	Format	Observation
USGS Earth Explorer	http://earthexplorer.usgs.gov/	Raster	Remote sensing data (Aerial images, DEMs, infrared images)
WorldClim	http://www.worldclim.org/	Raster	Climate data (past, current and future)
Diva GIS	http://www.diva-gis.org/Data	Raster, Vector	Global climate data, biodiversity and crop collection data
Sentinel Satellite Data	https://scihub.copernicus.eu/dhus	Raster	10-m resolution satellite data Sentinel 2 data with 11 spectral bands, Synthetic aperture radar
Open Street Map	https://www.openstreetmap.org	Vector	Crowd sourced vector data. (Road network, land use, buildings etc.)
Global Biodiversity Information Facility	http://www.gbif.org/	Vector	Information on biodiversity of 1.6 million species, collected over three centuries
Map of Life	http://mol.org/	Vector	Species range map
UNEP	http://geodata.grid.unep.ch/	Vector	Data on environment, climate, emissions
FAO	http://www.fao.org/geonetwork/srv/en/main.home	Vector	Database containing inter-disciplinary information about biodiversity
Worldwide Global Forest Change	https://earthenginepartners.appspot.com/science-2013-global-forest	Raster	Time-series analysis of Landsat images characterizing forest extent and change

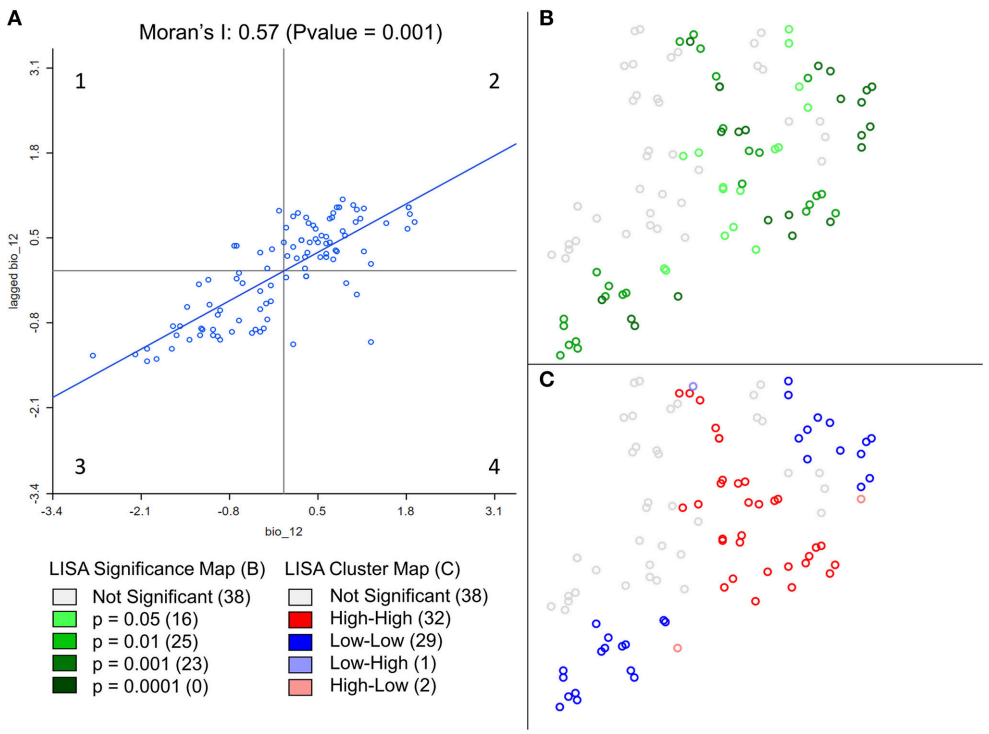


FIGURE 1 | Example of spatial statistic measurement in GeoDa. Results from global and local spatial autocorrelation (SA) were computed on Annual Precipitation at sampling locations of Ugandan cattle (Stucki, 2014). Annual Precipitation was extracted from the WorldClim dataset. In GeoDa, a weight file was created using the 10 nearest neighbors before computing spatial autocorrelation. Nine hundred ninety-nine permutations were performed to assess the significance of both SA measurements. The scatter plot of Global SA **(A)**, measured by the slope of the regression (0.57) displays the standardized precipitation values of each point on the X axis and standardized mean precipitation values of their 10 nearest neighbors on the Y axis. The scatterplot shows a positive correlation between most individuals and their neighbors. In other words, when precipitation is high (low) at a given location, close surrounding locations are more likely to experience high (low) precipitation as well. This positive correlation between neighboring locations is the translation of a clustering of values. On the other hand, significant local SA coefficients **(B)** are categorized **(C)** according to the 4 quadrants of the Moran's I plot **(A)**. In contrary to global SA, local SA indicates the location of positive SA or clustering (High-High–A2, Low-Low–A3), and negative SA or spatial outliers (High-Low–A1, Low-High–A4). Non-significant local SA coefficients are displayed in white.

such as the WorldClim dataset (Hijmans et al., 2005). These data are often delivered as continuous grids and their spatial resolution (i.e., area covered by a pixel) typically varies between 1 and 10 km^2. For more local or regional databases, however, national agencies are the most valuable sources (**Box 1**).

Alternatively or additionally, environmental variables can be computed from DEMs, and used as proxies to relevant ecological features (Kozak et al., 2008; Manel et al., 2010; Leempoel et al., 2015). DEMs are available on Earth Explorer (Earth Explorer, 2016) and come in formats such as geotiff or SAGA Grids

BOX 1 | Sampling design and scale.

Sampling design must be carefully chosen depending on the ecological scale of study, i.e., the spatial resolution and the extent of the area under study, and economic constraints. One important way of evaluating sampling strategies is to design them in a GIS environment to guarantee spatial randomness, representativeness or a constrained stratification of the sampling along environmental gradients among others.

Various optimized sampling strategies have been proposed and reviewed in the literature (Schwartz et al., 2009; Manel et al., 2012; Balkenhol et al., 2015; Rellstab et al., 2015). However, and regardless of the design chosen, one must consider sampling density and decide how many individuals will be sampled and then sequenced per location (or population). Indeed, the recent availability of NGS data implies to consider sub-sampling strategies for economic reasons. For example, a sub-sampling procedure using a hierarchical clustering can be applied in order to ensure a regular cover of both environmental and physical spaces (Stucki, 2014). For the former, stratified sampling techniques should be used over a range of climatic variables, previously filtered using a PCA. For the latter, a clustering index is minimized to ensure spatial spread. To ensure the representativeness of the entire area, the sampling can also be achieved using grid cells (see **Figure 2**). On the other hand, it is important to understand that landscape and population genomic sampling designs are difficult to reconcile (Joost et al., 2013). Indeed, sampling a small number of populations does not necessarily allow estimating changes in frequency along an environmental gradient. Conversely, sampling regularly along an environmental gradient may turn the assignment of individuals to populations more difficult. However, as pointed out by De Mita et al. (2013), for population genetics studies it is preferable to sample a high number of populations with few samples rather than a small number of populations with many samples. In addition, it is better to concentrate the sampling in a smaller area in order to obtain a greater density and higher statistical power (Joost et al., 2010).

Defining a scale of study also raises important questions regarding the relevance of environmental variables used. Indeed, when integrating different datasets (e.g., environmental, topographic, genetic), one must be aware that the spatial resolution of the raster data has to match the sampling density, and this is often not the case. Recently developed satellite imagery or DEMs show a fine resolution and a high accuracy, but the advantage of using high resolution data compared to data at coarser resolution remains under-studied (Levin, 1992; Marceau and Hay, 1999; Wilson and Gallant, 2000; Cavazzi et al., 2013). For example, while intuitively a fine resolution may be ideal, it may hold an excess of details and generate too much noise. Contrastingly, a too coarse resolution only shows generalized properties of the landscape and can have little explanatory power (Cavazzi et al., 2013). On the other hand, when the spatial resolution of the variable is too coarse, nearby samples will retrieve their environmental values from the same pixels (i.e., pseudo-replicates), thus inflating autocorrelation. One solution to this problem is to compute variables at multiple resolutions (Pradervand et al., 2014; Leempoel et al., 2015).

(.sgrd). We recommend not using text formats for grids (such as .asc or .xyz) since DEMs resolution has dramatically increased over the years, making these formats slow and heavy. The most common use of DEMs in ecology consists in retrieving altitude or computing primary terrain attributes (i.e., slope, orientation and curvature; Guisan and Zimmermann, 2000; Wilson and Gallant, 2000; Kozak et al., 2008; Manel et al., 2010). However, we recommend going beyond the traditional use of DEMs as a diversity of variables can be computed, like e.g., solar radiation, morphometric indices or hydrological variables (Leempoel et al., 2015). The treatment of DEMs and the production of topographic variables can be processed in software like SAGA GIS (SAGA GIS, 2004; Conrad et al., 2015) or GRASS GIS, now included in QGIS. SAGA GIS is the most DEM-oriented GIS to date and can compute a large panel of derived variables. It is also easily scriptable both using the command line or the R package RSAGA (Brenning, 2008), although the former is faster.

Satellite images covering the whole surface of the globe are also available through Earth Explorer and can be used e.g., to produce land cover maps. Most satellite sensors provide images with more than the 3 visible "colors," or bands, and it is thus the choice of the user to decide which bands to attribute to color channels (Red, Green and Blue). For example, by assigning the infrared band to the red channel and green and blue bands to their respective channels, one can easily identify trees or forests against water, fields or naked soils because plants reflect infrared wavelengths more than other land cover types. This process, the supervised classification of remote sensing data (satellite and aerial images, radar, etc.), can be operated in Opticks (Opticks, 2001) or in SAGA GIS.

Finally, vector databases, such as Openstreetmap (OSM) (OpenStreetMap, 2004), are precious to recover road networks, rivers, watershed boundaries, or landuse. OSM data can also be easily accessed through GeoFabrik (GeoFabrik, 2011) where cities or countries are already extracted. Note however that OSM data and most tiled web-maps are provided in Pseudo-Mercator projection (EPSG: 3857).

It is worth mentioning that in GEA studies, using a wide range of environmental variables often implies redundancy between these variables. However, statistical analyses require independent variables and, for this reason, it is important to perform multi-collinearity analysis (e.g.,) on the set of environmental variables, to understand which variables are highly correlated (Dobson and Barnett, 2008; Fischer et al., 2013). Such collinearity can be detected by performing a PCA, by using Variance Inflation Factor (VIF) or calculating pairwise correlation coefficients between pairs of variables, and then removing randomly one of the two variables from a pair that shows high correlation. See Rellstab et al. (2015) for a review of these methods. However, bear in mind that environmental variables, in particular DEM-derived ones, may not have a normal distribution. Variables should thus be transformed or non-parametric tests should be used to test for correlations (for example Spearman ranks instead of Pearson correlation coefficients).

SPATIAL ANALYSIS

Numerous spatial analysis techniques have been developed to address issues related to spatial data (Fortin and Dale, 2005). Here, we focus on exploratory spatial data analysis (ESDA) and spatial statistics given their central role in molecular ecology. For other spatial analysis methods, we suggest to have a look at the Geospatial analysis guide (Smith et al., 2005) and at the spatial analysis guide for ecologists (Fortin and Dale, 2005).

Evolutionary biology can benefit from ESDA (Joost, 2006), an interactive approach allowing the user to explore and analyze a dataset dynamically and in real-time through a combination

of various tools for data representation (Anselin, 1994). For example, maps can be used to display the position of samples, histograms and boxplots to evaluate the distribution of attribute values and Moran's scatter plot or conditional plots to analyze the relationship between the various variables. ESDA can also be useful for example to localize samples in areas showing extreme climatic conditions (outliers), to highlight regions where samples are highly correlated (clusters), or pinpoint populations with a low genetic diversity. A powerful ESDA tool is the open-source software GeoDa (GeoDa, 2005) that allows the exploration and spatial analysis of vector data (Anselin et al., 2006). GeoDa notably offers the possibility to create various maps (quantile, equal intervals, etc.) and to simultaneously analyze attributes with the help of other graphs.

Spatial autocorrelation (i.e., the degree of dependence among observations in a geographic space; SA) is often overlooked in ecological and evolutionary studies despite the fact that many environmental or biological characteristics show spatial dependence among observations, due to intrinsic process of dispersal and mating (Anselin, 1998; Hall and Beissinger, 2014). It is measured by comparing individual values of a defined variable with the mean of that variable in a defined neighborhood. By doing so for each sample, SA measures the degree of values similarities with location similarity. It is thus essential to measure SA in studies involving spatial data, not only because it is a natural phenomenon but also because it violates the assumption of independence required by standard statistical tests, such as student tests or regressions (Legendre, 1993; Wagner and Fortin, 2005). For example, Moran's I, a classic spatial autocorrelation statistic, can be used to estimate the scale/distance of gene flow in the landscape (Hall and Beissinger, 2014). In addition, Local Indicators of Spatial Association (LISA; Anselin, 1995) allows to identify and localize spatial autocorrelation patterns and study the spatial relationship between genetic markers and environmental features (Colli et al., 2014; Stucki, 2014). See **Figure 1** for an example. While GeoDa is better to visualize the SA of one variable, it cannot be automated to calculate it for many. For a fast computation of both global and local SA on genetic data, Sambada is handful (Stucki et al., 2016). It can be easily programmed to compute SA on millions of genetic markers and so with different neighborhood sizes and weighting schemes. The decrease of SA with distance can thus be measured using different lags and comparisons can be made between neutral and selected loci. The R package *spdep* can perform similar analyses (Bivand and Piras, 2015).

SPATIAL DATA REPRESENTATION

Maps illustrating the results of an analysis are often more powerful than tables to transmit a result or an idea. However, the creation of efficient maps requires a reflection phase about the graphical representation of the results. Indeed, maps can be too complex to read when too detailed or may be uninformative when too simple. Creating a map first requires choosing an appropriate display type. Traditional choropleth map, in which the entities are colored according to a scale based on the value of the attribute of interest, can be used in many situations. For example, to represent the membership coefficient of individuals to two populations distributed over a landscape (e.g., as frequently done for population genetic analyses of admixture), one can use a gradient passing through a neutral hue to contrast the two parts of the distribution (i.e., the membership of each individual to one or the other population) (**Figure 2**). Although most GIS provide colored gradients, it can be useful to understand how to obtain an appropriate color scheme using Color Brewer (Color Brewer 2, 2001). Alternatively, if individuals are grouped into more than two populations, bar charts can be more appropriate. Proportional circles can also be used, for example to indicate absolute numbers of individuals sampled in each population.

Background layers can then be added to provide more information on the geographic context, such as an aerial image or a DEM to situate the samples. This can potentially be combined with contour lines to compare the elevation from one location to another. One can also highlight the study area by darkening or de-saturating the rest of the map. Regarding points representing individuals or populations, simple shapes should be preferred. Labels should be readable and discarded if not.

Each map must then be edited before being published. Some elements must go along with a map: a legend (to identify the geographical units, or the different statistical classes used) and a scale. In the legend, the message should be simplified by regrouping categories, reducing the number of decimal places and removing unnecessary layers. Furthermore, a frame in a corner of the map, representing the region at a broader geographic scale (zoom out), is useful to situate the study area. Maps should be exported preferentially in .pdf format to keep the vector properties for potential future editions. We provide an example in **Figure 2**.

Lastly, GIS software are not particularly easy to use when it comes to producing maps iteratively. For example, creating maps of genetic markers under selection used to be feasible manually when the number of genetic markers tested was small. On the contrary, most GEA studies today use hundreds to millions of markers deriving from genomic analyses, and with many of them showing signatures of selection. In such cases, manually producing maps is neither smart nor informative. Computing software such as R should thus be favored with packages such as *Rgdal* and *Rasters* being very useful and sufficient to import genetic and geodata and to produce basic maps (Hijmans and van Etten, 2015; Bivand et al., 2016).

PERSPECTIVE

GIS are powerful tools for molecular ecologists but remain too often underexploited and misused, mainly because of the multitude of GIS software and databases available. We have presented in this paper useful guidelines making it possible for any GIS beginner to appropriate basic functions, to find specific learning resources for biologists, and we proposed a brief state of the art for the use of GIS in biology. However, it is intriguing that, in the big data era, geodatabases are not more frequently used to store and access genetic datasets. They would also speed up queries and reduce disk usage. There are in fact few examples of transformation of NGS data in spatial

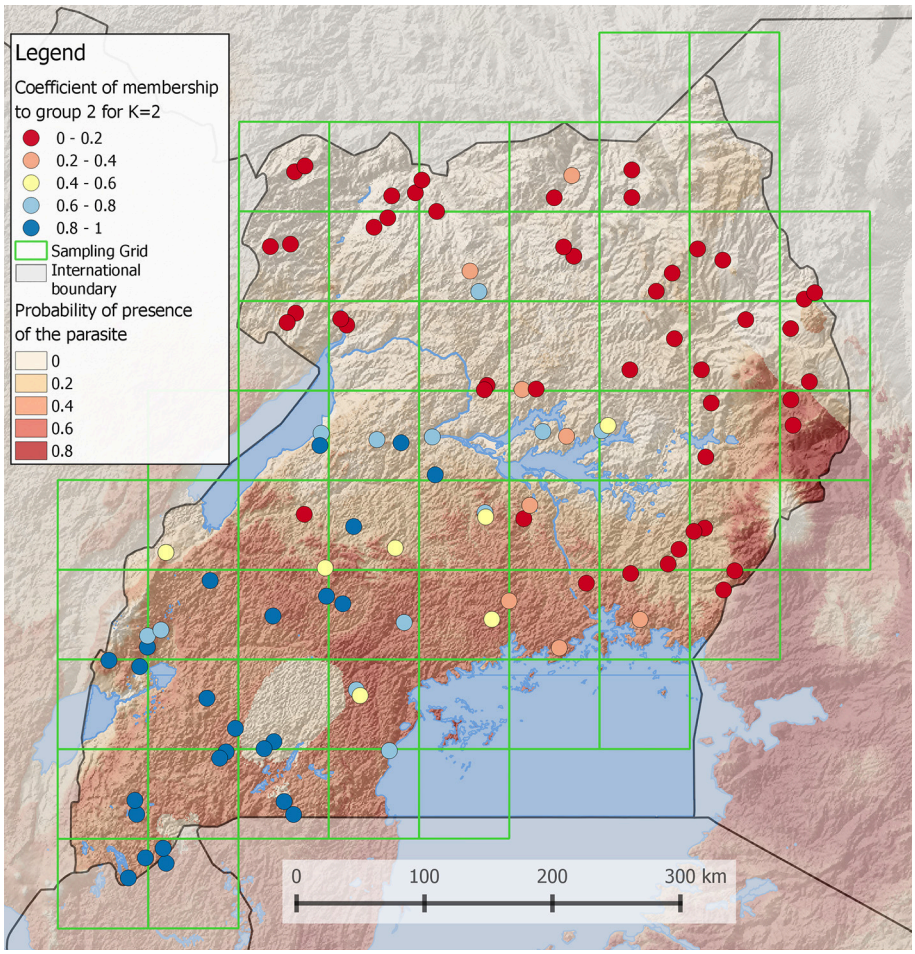


FIGURE 2 | Coefficient of membership to a genetic group of Ugandan Cattles (Stucki, 2014). Using the software admixture (Alexander et al., 2009), the most likely number of populations was found to be 2. In this case, it is possible to display the membership coefficient of each individual to one of the two populations. To do so, a gradient obtained from http://colorbrewer2.org is passing through a neutral hue to contrast the two populations. The order of layers in the legend is the same as in the map. In the background, a grid layer of probability of presence of a parasite is shown. A semi-transparent shaded relief is also displayed to reveal the topography. Lakes and international boundaries are overlaid on these raster layers. Ugandan boundaries are highlighted by darkening surrounding countries.

databases because of the high technicality of such task (Holl and Plum, 2009; Joost and Kalbermatten, 2010; Paila et al., 2013; Nandal et al., 2016; Piry et al., 2016). So far, the most compelling tool is the recently developed open source system TheSNPpit (Groeneveld and Lichtenberg, 2016). It allows for an integration of large genetic datasets in a PostgreSQL environment, which is also the backend of most GIS databases. Interestingly, this tool was mainly developed for breeding programs that already deal with thousands of individuals and millions of SNPs. A game changer that will most likely hit molecular biology in the future.

AUTHOR CONTRIBUTIONS

KL: structured and wrote the paper. SD and ER: wrote two paragraphs and reviewed the paper. IW: reviewed the paper. PO: reviewed the paper. SJ: supervised and reviewed the paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fevo.2017.00033/full#supplementary-material

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Anselin, L. (1994). Exploratory spatial data analysis and geographic information systems. *New Tools Spat. Anal.* 17, 45–54. doi: 10.1088/0957-4484/17/5/014

Anselin, L. (1995). Local indicators of spatial association — LISA. *Geogr. Anal.* 27, 93–115. doi: 10.1111/j.1538-4632.1995.tb00338.x

Anselin, L. (1998). "Exploratory spatial data analysis in a geocomputational environment," in *Geocomputation: A Primer*, eds P. A. Longley, S. M. Brooks, R. McDonnell, and W. Macmillian (New York, NY: Wiley and Sons), 77–94.

Anselin, L., Syabri, I., and Kho, Y. (2006). GeoDa: an introduction to spatial data analysis. *Geogr. Anal.* 38, 5–22. doi: 10.1111/j.0016-7363.2005.00671.x

Baddeley, A., and Turner, R. (2005). spatstat: an R package for analyzing spatial point patterns. *J. Stat. Softw.* 12, 1–42. doi: 10.1.1.126.8464

Balkenhol, N., Cushman, S. A., Storfer, A., and Waits, L. P. (2015). *Landscape Genetics : Concepts, Methods, Applications*.

Bivand, R., Keitt, T., and Rowlingson, B. (2016). *Package "rgdal."* R Package.

Bivand, R., and Piras, G. (2015). Comparing implementations of estimation methods for spatial econometrics. *J. Stat. Softw.* 63, 1–36. doi: 10.18637/jss.v063.i18

Brenning, A. (2008). Statistical Geocomputing combining R and SAGA: the example of landslide susceptibility analysis with generalized additive models. *SAGA Second. Out* 19, 23–32.

Brunsdon, C., and Chen, H. (2014). *GISTools: Some Further GIS Capabilities for R.* Available online at: https://cran.r-project.org/package=GISTools

Brunsdon, C., and Comber, L. (2015). *An Introduction to R for Spatial Analysis and Mapping*. SAGE Publications Available online at: https://books.google.com/books?id=zsf-AwAAQBAJ

Cavazzi, S., Corstanje, R., Mayr, T., Hannam, J., and Fealy, R. (2013). Are fine resolution digital elevation models always the best choice in digital soil mapping? *Geoderma* 195–196, 111–121. doi: 10.1016/j.geoderma.2012.11.020

Colli, L., Joost, S., Negrini, R., Nicoloso, L., Crepaldi, P., and Ajmone-Marsan, P. (2014). Assessing the spatial dependence of adaptive loci in 43 European and Western Asian goat breeds using AFLP markers. *PLoS ONE* 9:e86668. doi: 10.1371/journal.pone.0086668

Color Brewer, 2 (2001). Available online at: http://colorbrewer2.org/ (Accessed: October 5, 2016).

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., et al. (2015). System for Automated Geoscientific Analyses (SAGA) v. 2.1.4, *Geosci. Model Dev.* 8, 1991–2007. doi: 10.5194/gmd-8-1991-2015

Coursera (2012). Available online at: https://www.coursera.org (Accessed: October 5, 2016).

De Mita, S., Thuillet, A.-C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J., et al. (2013). Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Mol. Ecol.* 22, 1383–1399. doi: 10.1111/mec.12182

Dmap (1993). *UTM Grid Zones of the World.* Available online at: http://www.dmap.co.uk/utmworld.htm (Accessed on: October 5, 2016).

Dobson, A. J., and Barnett, A. (2008). *An Introduction to Generalized Linear Models, 3rd Edn*. Taylor & Francis. Available online at: http://books.google.ch/books?id=KodvPwAACAAJ

Docs.QGIS (2013). *GPS Plugin.* Available online at: http://docs.qgis.org/2.8/en/docs/user_manual/working_with_gps/plugins_gps.html (Accessed: October 5, 2016).

Docs.QGIS (2014). *Create New Vector in QGIS.* Available online at: https://docs.qgis.org/2.2/en/docs/training_manual/create_vector_data/create_new_vector.html (Accessed: October 5, 2016).

Earth Explorer (2016). Available online at: earthexplorer.usgs.gov/ (Accessed: October 5, 2016).

Epperson, B. (2003). *Geographical Genetics*. Available online at: http://press.princeton.edu/titles/7689.html

EPSG (1985). Available online at: http://www.epsg.org/ (Accessed: October 5, 2016).

Ertz, O., Rey, S. J., and Joost, S. (2014). The open source dynamics in geospatial research and education. *J. Spat. Inf. Sci.* 8, 67–71. doi: 10.5311/josis.2014.8.182

Fischer, M. C., Rellstab, C., Tedder, A., Zoller, S., Gugerli, F., Shimizu, K. K., et al. (2013). Population genomic footprints of selection and associations with climate in natural populations of *Arabidopsis halleri* from the Alps. *Mol. Ecol.* 22, 5594–5607. doi: 10.1111/mec.12521

Fortin, M. J., and Dale, M. R. T. (2005). *Spatial Analysis: A Guide for Ecologists.* Cambridge: Cambridge University Press.

GeoDa (2005). Available online at: http://geodacenter.github.io/ (Accessed October 5, 2016).

GeoFabrik (2011). Available online at: http://www.geofabrik.de (Accessed on: October 5, 2016).

Groeneveld, E., and Lichtenberg, H. (2016). TheSNPpit—a high performance database system for managing large scale SNP data. *PLoS ONE* 11:e0164043. doi: 10.1371/journal.pone.0164043

Guisan, A., and Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecol. Modell.* 135, 147–186. doi: 10.1016/S0304-3800(00)00354-9

Hall, L., and Beissinger, S. (2014). A practical toolbox for design and analysis of landscape genetics studies. *Landsc. Ecol.* 29, 1487–1504. doi: 10.1007/s10980-014-0082-3

Hand, B. K., Lowe, W. H., Kovach, R. P., Muhlfeld, C. C., and Luikart, G. (2015). Landscape community genomics: understanding eco-evolutionary processes in complex environments. *Trends Ecol. Evol.* 30, 161–168. doi: 10.1016/j.tree.2015.01.005

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* 25, 1965–1978. doi: 10.1002/joc.1276

Hijmans, R. J., and van Etten, J. (2015). *raster: Geographic Analysis and Modeling with Raster Data.* R Package Version 2.5-2. Available online at: http://CRAN.R-project.org/package=raster

Holl, S., and Plum, H. (2009). "PostGIS," in *GeoInformatics 03/2009*, 34–36.

Joost S. (2006). *The Geographical Dimension of Genetic Diversity - A GIScience Contribution for the Conservation of Animal Genetic Resources [Internet]*. Thesis, EPFL. doi: 10.5075/epfl-thesis-3454

Joost, S., Colli, L., Baret, P. V., Garcia, J. F., Boettcher, P. J., Tixier-Boichard, M., et al. (2010). Integrating geo-referenced multiscale and multidisciplinary data for the management of biodiversity in livestock genetic resources. *Anim. Genet.* 41, 47–63. doi: 10.1111/j.1365-2052.2010.02037.x

Joost, S., and Kalbermatten, M. (2010). "GEOME: A web-based landscape genomics geocomputation platform," in *Proceedings of the 1st ISPRS International Workshop on Pervasive Web Mapping, Geoprocessing and Services-WebMGS 2010* (Como: International Society for Photogrammetry and Remote Sensing).

Joost, S., Vuilleumier, S., Jensen, J. D., Schoville, S., Leempoel, K., Stucki, S., et al. (2013). Meeting review. Uncovering the genetic basis of adaptive change: on the intersection of landscape genomics and theoretical population genetics. *Mol. Ecol.* 22, 3659–3665. doi: 10.1111/mec.12352

Kozak, K. H., Graham, C. H., and Wiens, J. J. (2008). Integrating GIS-based environmental data into evolutionary biology. *Trends Ecol. Evol.* 23, 141–148. doi: 10.1016/j.tree.2008.02.001

Leempoel, K., Geiser, C., Daprà, L., Vittoz, P., Parisod, C., and Joost, S. (2015). Very high resolution digital elevation models: are multi-scale derived variables ecologically relevant? *Methods Ecol. Evol.* 6, 1373–1383. doi: 10.1111/2041-210x.12427

Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? *Ecology* 74, 1659–1673. doi: 10.2307/1939924

Levin, S. A. (1992). The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology* 73, 1943–1967.

Maling, D. H. (1992). *Coordinate Systems and Map Projections, 2nd Edn.* Amsterdam: Pergamon. doi: 10.1016/B978-0-08-037233-4.50003-0

Manel, S., Albert, C., and Yoccoz, N. (2012). "Sampling in landscape genomics," in *Data Production and Analysis in Population Genomics SE - 1 Methods in Molecular Biology*, eds F. Pompanon and A. Bonin (New York, NY: Humana Press), 3–12.

Manel, S., and Holderegger, R. (2013). Ten years of landscape genetics. *Trends Ecol. Evol.* 28, 614–621. doi: 10.1016/j.tree.2013.05.012

Manel, S., Poncet, B. N., Legendre, P., Gugerli, F., and Holderegger, R. (2010). Common factors drive adaptive genetic variation at different spatial scales in *Arabis alpina*. *Mol. Ecol.* 19, 3824–3835. doi: 10.1111/j.1365-294X.2010.04716.x

Manel, S., Schwartz, M. K., Luikart, G., and Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends Ecol. Evol.* 18, 189–197. doi: 10.1016/s0169-5347(03)00008-9

Mangomap (2012). *Make a Web Map from a List of Addresses in a Spreadsheet.* Available online at: http://blog.mangomap.com/post/74368997570/how-to-make-a-web-map-from-a-list-of-addresses-in (Accessed: October 5, 2016).

Marceau, D. J., and Hay, G. J. (1999). Remote sensing contributions to the scale issue. *Can. J. Remote Sens.* 25, 357–366.

MMQGIS Plugin (2012). Available at: https://plugins.qgis.org/plugins/mmqgis/ (Accessed: October 5, 2016).

Nandal, U. K., van Kampen, A. H. C., and Moerland, P. D. (2016). compendiumdb: an R package for retrieval and storage of functional genomics data. *Bioinformatics* 32, 2856–2857. doi: 10.1093/bioinformatics/btw335

OpenStreetMap (2004). Available at: https://www.openstreetmap.org/ (Accessed: October 5, 2016).

Opticks (2001). Available at: https://opticks.org/ (Accessed: October 5, 2016).

Paila, U., Chapman, B. A., Kirchner, R., and Quinlan, A. R. (2013). GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS Comput. Biol.* 9:e1003153. doi: 10.1371/journal.pcbi.1003153

Petren, K. (2013). The evolution of landscape genetics. *Evolution* 67, 3383–3385. doi: 10.1111/evo.12278

Piry, S., Chapuis, M.-P., Gauffre, B., Papaïx, J., Cruaud, A., and Berthier, K. (2016). Mapping Averaged Pairwise Information (MAPI): a new exploratory tool to uncover spatial structure. *Methods Ecol. Evol.* 7, 1463–1475. doi: 10.1111/2041-210X.12616

Pradervand, J.-N., Dubuis, A., Pellissier, L., Guisan, A., and Randin, C. (2014). Very high resolution environmental predictors in species distribution models: moving beyond topography? *Prog. Phys. Geogr.* 38, 79–96. doi: 10.1177/0309133313512667

QGIS Development Team (2015). *QGIS Geographic Information System.* Open Source Geospatial Foundation Project. Available online at: http://www.qgis.org/

QGIStutorial (2014). *Performing Table Joins.* Available online at: http://www.qgistutorials.com/en/docs/performing_table_joins.html (Accessed: October 5, 2016).

QGIStutorials (2014). *Importing Spreadsheets or CSV files.* Available online at: http://www.qgistutorials.com/en/docs/importing_spreadsheets_csv.html (Accessed: October 5, 2016).

QGIS workshop (2013). *Adding Basemaps using OpenLayers Plugin.* Available online at: http://maps.cga.harvard.edu/qgis/wkshop/basemap.php (Accessed: October 5, 2016).

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., and Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Mol. Ecol.* 24, 4348–4370. doi: 10.1111/mec.13322

Rodriguez-Sanchez, F. (2013). *Spatial Data in R: Using R as a GIS.* Available online at: https://pakillo.github.io/R-GIS-tutorial/#contents (Accessed: November 12, 2016).

Rogers, S. R., and Staub, B. (2013). Standard use of Geographic Information System (GIS) techniques in honey bee research. *J. Apic. Res.* 52, 1–48. doi: 10.3896/IBRA.1.52.4.08

SAGA GIS (2004). Available at: http://www.saga-gis.org/en/index.html (Accessed: October 5, 2016).

Schwartz, M. K., Luikart, G., McKelvey, K. S., and Cushman, S. A. (2009). "Landscape genomics: a brief perspecitve," in *Spatial Complexity, Informatics, and Wildlife* Conservation (Tokyo: Springer), 165–175.

Smith, M., de Longley, P., and Goodchild, M. (2005). *Geospatial Analysis - A Comprehensive Guide.* Available online at: http://www.spatialanalysisonline.com/ (Accessed: October 6, 2016).

Stucki, S. (2014). *Développement d'outils de géo-calcul Haute Performance Pour l'identification de Régions du Génome Potentiellement Soumises à la Sélection Naturelle - Analyse Spatiale de la Diversité de Panels de Polymorphismes Nucléotidiques à Haute Densité (800k) Chez Bos Taurus et B. Indicus en Ouganda.* Thesis, EPFL.

Stucki, S., Orozco-terWengel, P., Forester, B. R., Duruz, S., Colli, L., Masembe, C., et al. (2016). High performance computation of landscape genomic models including local indicators of spatial association. *Mol. Ecol. Resour.* doi: 10.1111/1755-0998.12629

Sutton, T., Dassau, O., and Sutton, M. (2009). A gentle introduction to GIS: brought to you with Quantum GIS, a Free and Open Source Software GIS Application for everyone. *T. Chief Dir. Spat. Plan.* doi: 10.1038/sj.bdj.2011.132

Wagner, H. H., and Fortin, M. J. (2005). Spatial analysis of landscapes: concepts and statistics. *Ecology* 86, 1975–1987. doi: 10.1890/04-0914

Wilson, J. P., and Gallant, J. C. (2000). *Terrain Analysis: Principles and Applications.* Wiley Wiley. Available online at: http://www.loc.gov/catdir/bios/wiley047/99089635.html

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

LIST OF TERMS

Raster: Regular grids of pixels that describe continuous phenomena, retaining information such as color (for aerial images), elevation, temperature.

Vector: Points, lines or polygons whose nodes are defined by geographical coordinates and describe discrete phenomenon such as borders, rivers, catchment areas. Vectors are usually stored in Shapefiles (.shp and associated files).

Datum: The datum defines the 3-dimensional sphere used to approximate the earth. It provides a frame of reference to measure coordinates in both geographic and projected coordinate systems.

Geographic Coordinate System: A GCS gives the coordinates (i.e., latitude and longitude) of a point as measured from the angles to the center of a defined sphere and meridian.

Projected Coordinate system: A PCS is a projection of the sphere on a flat, two-dimensional surface. Its coordinates (X and Y) are thus consistent and equally spaced.

DEM: Digital Elevation Models are grids of elevation data. Each pixel of that grid is spaced at regular horizontal intervals and contains one value of elevation.

Grain: The grain is the size of a pixel, the smallest unit on a grid. A small grain corresponds to a high spatial resolution.

Extent: The extent is the size of the study area.

frontiers in Genetics

MINI REVIEW
published: 25 October 2016
doi: 10.3389/fgene.2016.00189

Epigenetic Inheritance across the Landscape

Amy V. Whipple and Liza M. Holeski*

Department of Biological Sciences and Merriam-Powell Center for Environmental Research, Northern Arizona University, Flagstaff, AZ, USA

The study of epigenomic variation at the landscape-level in plants may add important insight to studies of adaptive variation. A major goal of landscape genomic studies is to identify genomic regions contributing to adaptive variation across the landscape. Heritable variation in epigenetic marks, resulting in transgenerational plasticity, can influence fitness-related traits. Epigenetic marks are influenced by the genome, the environment, and their interaction, and can be inherited independently of the genome. Thus, epigenomic variation likely influences the heritability of many adaptive traits, but the extent of this influence remains largely unknown. Here, we summarize the relevance of epigenetic inheritance to ecological and evolutionary processes, and review the literature on landscape-level patterns of epigenetic variation. Landscape-level patterns of epigenomic variation in plants generally show greater levels of isolation by distance and isolation by environment then is found for the genome, but the causes of these patterns are not yet clear. Linkage between the environment and epigenomic variation has been clearly shown within a single generation, but demonstrating transgenerational inheritance requires more complex breeding and/or experimental designs. Transgenerational epigenetic variation may alter the interpretation of landscape genomic studies that rely upon phenotypic analyses, but should have less influence on landscape genomic approaches that rely upon outlier analyses or genome–environment associations. We suggest that multi-generation common garden experiments conducted across multiple environments will allow researchers to understand which parts of the epigenome are inherited, as well as to parse out the relative contribution of heritable epigenetic variation to the phenotype.

Keywords: epigenetics, transgenerational plasticity, landscape genomics, adaptation, epigenome, phenotype

OPEN ACCESS

Edited by:
Samuel A. Cushman, Service Rocky Mountain Research Station, USA

Reviewed by:
Gonzalo Gajardo, University of Los Lagos, Chile
Tariq Ezaz, University of Canberra, Australia

****Correspondence:***
Amy V. Whipple
amy.whipple@nau.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *18 July 2016*
Accepted: *10 October 2016*
Published: *25 October 2016*

Citation:
Whipple AV and Holeski LM (2016) Epigenetic Inheritance across the Landscape. Front. Genet. 7:189. doi: 10.3389/fgene.2016.00189

INTRODUCTION

Understanding the ecological and evolutionary processes governing landscape patterns of genetic diversity and adaptive variation is important to predicting and managing the impacts of climate change on plant species distributions and function (Sork et al., 2013). Genomic, phenotypic, and environmental data are used to disentangle genetic and environmental influences on the phenotype and understand the distribution of adaptive variation among natural populations (Barrett and Hoekstra, 2011; Korte and Farlow, 2013; Lepais and Bacles, 2014; Rellstab et al., 2015). Evidence of adaptive differences across environmental gradients is common, but not universal (Sexton et al., 2014). There is increasing recognition that the inclusion of epigenetic-based transgenerational plasticity is likely to improve our understanding of adaptive phenotypic variation across the

landscape (e.g., Cushman, 2014; Verhoeven et al., 2016). Here, we summarize the relevance of epigenetic inheritance to ecological and evolutionary processes, and review the literature on landscape-level patterns of epigenetic variation (both transgenerational and not). Then, we discuss the implications of transgenerational epigenetic variation for various approaches to landscape genomics. Finally, we discuss designs that can partition the landscape distribution of adaptive genetic and epigenetic variation.

Epigenetic modifications are changes in phenotype that are mediated by the regulation of gene expression rather than alterations in the DNA sequence. Epigenetic modifications may be reset within an organism's lifespan or during meiosis, or they may be passed to offspring (Richards, 2006). These modifications can be inherited both maternally and paternally, via mechanisms such as DNA methylation, histone modification, and RNAi (Rapp and Wendel, 2005). The evolutionary relevance of transgenerational plasticity rests upon whether responses are adaptive and whether there is heritable genetic or epigenetic variation for the epigenetic modification (Day and Bonduriansky, 2011; Herman et al., 2014).

Approaches in Landscape-level Investigations of Genomic Variation

Landscape-level investigations of patterns of genomic variation in plants have examined both adaptive and neutral loci and have employed a wide variety of methods. Genomic regions influencing adaptive traits are routinely discovered via genome–phenotype associations, phenotype-free approaches, and common gardens (Sork et al., 2013; Rellstab et al., 2015). Genome-wide association studies (GWAS) identify genotype–phenotype associations and can be done using plants grown in common gardens or from natural populations (e.g., Ingvarsson and Street, 2011). Phenotype-free approaches use genomic information to detect signatures of selection. Examples of these include outlier and environmental association analyses (EAA) (Rellstab et al., 2015). Outlier analyses detect loci that show evidence of strong selection, relative to the bulk of assayed loci that show effects of only population structure and drift. EAA refers to a number of statistical methods for detecting association between environmental variables and particular loci (Rellstab et al., 2015). Finally, growing multiple genotypes within the same common garden environment allows the genetic basis of a phenotype to be identified. When the same genotype is grown in multiple common gardens, the approach can also be used to examine the effect of environment on phenotype (Clausen et al., 1948).

Landscape Genomic Patterns and Missing Heritability in Plants

Examination of gene flow within a species, based on neutral markers, typically shows isolation by distance (IBD) and/or isolation by environment (IBE). Genetic drift is the primary driver of IBD, where genetic differentiation increases with spatial distance (Wright, 1943; Charlesworth et al., 2003). IBE is influenced more by selection than is IBD, with genetic differentiation increasing with environmental distance (Bradburd et al., 2013; Wang and Bradburd, 2014). A review by Sexton et al. (2014) found both IBD and IBE were drivers of molecular genetic variation in natural plant populations, with IBD being more common than IBE.

Experiments combining an assessment of adaptive traits, molecular genetic loci underlying traits, genetic correlations, and gene flow barriers (distance, timing, and selection against immigrants) have thus far provided the most mechanistic understanding of landscape-level patterns of genetic variation (Lepais and Bacles, 2014). Often the realized heritability of populations detected in common garden and/or quantitative genetic designs cannot be fully explained by the loci detected in genomic approaches. The missing heritability problem can potentially be explained by a failure to detect loci of small effect or epistatic interactions among loci, but inherited (transgenerational) epigenetic variation is likely to be another source of the so-called missing heritability (Goldstein, 2009; Furrow et al., 2011). Epigenetic variation thus has potential implications for landscape-level adaptation.

Ecological and Evolutionary Relevance of Epigenetic Inheritance

Initial population-level work studying epigenetic inheritance has demonstrated the substantial impacts of epigenetic factors on phenotypic variation in traits such as floral symmetry and defense against herbivores and pathogens (Cubas et al., 1999; reviews in Kalisz and Purugganan, 2004; Herman and Sultan, 2011; Holeski et al., 2012). Most investigations of the adaptive role of epigenetic modification have focused on DNA methylation patterns (Cervera et al., 2002; Verhoeven et al., 2016). Use of epiRILs, recombinant inbred lines that differ primarily in epigenetic status, in *Arabidopsis thaliana,* revealed epigenetic quantitative trait loci that account for 60–90% of the heritability in two ecologically relevant traits, flowering time and primary root length (Cortijo et al., 2014). These lines of research have led to the suggestion that heritable epigenetic variation could be the source of "missing heritability" not identified by QTL and GWAS studies (Bonduriansky, 2012).

Epigenetic-based transgenerational inheritance is predicted to have particular relevance for evolution in scenarios in which genetic variation alone may not provide sufficient trait variation to result in a robust response to selection (Jablonka and Raz, 2009). These scenarios might include: rapidly changing environments, such as those predicted by climate change models; species with low genetic variation due to asexual reproduction or founder effects; and organisms with long generation times (Bossdorf et al., 2008; Bonduriansky and Day, 2009; Nicotra et al., 2010; Castonguay and Angers, 2012). Despite potentially greater importance in the evolution of long-lived and asexual species, most empirical work so far has been done in sexually reproducing annuals (but see Richards et al., 2012; Zas et al., 2013; Yakovlev et al., 2014; Preite et al., 2015). Thus, not only may epigenetic-based transgenerational inheritance be a source of adaptive variation across a variety of species, it may be particularly important to organisms such as clonal grasses

and long-lived trees, many of which are ecologically important foundation species.

EPIGENETIC PATTERNS ACROSS THE LANDSCAPE

The potential adaptive significance of the epigenome suggests its relevance to studies of adaptation across the landscape. A number of very recent landscape-level studies have investigated the role of epigenetics in intra-specific trait variation and adaptation (Medrano et al., 2014; Dubin et al., 2015; Preite et al., 2015; Foust et al., 2016; Gugger et al., 2016; Herrera et al., 2016; Keller et al., 2016). These studies focus on at least one of the following: (i) the relationship between genetic and epigenetic variation at the landscape level, (ii) correlations between environmental variables and epigenetic status, and (iii) correlations between epigenetic status and plant functional traits.

Genetic and epigenetic variations are spatially structured across a landscape. A positive relationship between geographic distance and epigenetic differences across eight studies was identified by Herrera et al. (2016). This pattern is compatible with IBD patterns that are often found for genetic differences among populations, which are the main determinant of spatial genetic structure in plants (Sexton et al., 2014). Herrera et al. (2016) also found evidence in their case study that nearby individuals were more similar in their epigenome than in their genome, especially at small spatial scales. This suggests the potential for environmental influences on the epigenome, rather than a direct genome–epigenome relationship. Epigenomic patterns due to the environment may change through a lifespan, be regenerated each generation, or be inherited across generations. The results of several additional landscape-level surveys of epigenetic variation suggest that environmental factors are more important than spatial distance or the genome in shaping epigenetic structure (Schulz et al., 2014; Huang et al., 2015; Herrera et al., 2016). While this may suggest IBE in the genomic context (Sexton et al., 2014), the interpretation in the case of the epigenome is more complicated. Greater epigenomic than genomic differentiation suggests additional factors other than simple genomic determination are involved, such as adaptation via a heritable epigenome or direct effects of the environment on the epigenome.

Numerous studies have found correlations between epigenetic variation and environmental factors across a landscape (Dubin et al., 2015; Foust et al., 2016; Gugger et al., 2016; Keller et al., 2016). This supports the prediction that epigenetic-based transgenerational inheritance might be particularly relevant for evolution in rapidly changing environments, as well as the relevance of IBE to epigenetic diversity. Both genome-wide genetic and epigenetic variation in *Arabidopsis* were correlated with climate and spatial variables across Sweden and Eurasia (Keller et al., 2016). However, such correlations are not always found, as was the case for in dandelion (*Taraxacum officinale*) across a north–south transect from Luxembourg to central Sweden, where no gradient in DNA methylation was found (Preite et al., 2015). Correlations between epigenetic variation and the environment may be inconsistent between species in the same environments. In a study of five populations (including four overlapping sites), of two perennial salt marsh species (*Spartina alterniflora* and *Borrichia frutescens*), significant correlations were found between epigenetic variation and habitat in *S. alterniflora*, but not *B. frutescens* (Foust et al., 2016).

While a number of studies have investigated correlations between epigenetic patterns and the environment, far fewer have identified fitness-relevant phenotypes that are putatively altered via epigenetic mechanisms. We know of only two studies, both of the perennial herb *Helleborus foetidus,* that have investigated the influence of epigenetic variation on plant functional traits (Alonso et al., 2014; Medrano et al., 2014). One study was done at a landscape level, and found that 8% of functional trait variation was explained by methylation sensitive amplified polymorphisms. Multivariate functional trait diversity was correlated with epigenetic diversity after genetic diversity was taken into account. The authors suggest that epigenetic influence on functional traits allows *H. foetidus* to adapt to or survive in an array of environmental conditions (Medrano et al., 2014). The second study was done at a small-scale landscape level. Using plants from three sites within the same region in Spain, Alonso et al. (2014) found a negative correlation between global methylation and plant reproductive output.

Research to date demonstrates linkage between the environment and epigenomic variation within a generation, but demonstrating transgenerational inheritance requires more complex breeding and/or experimental designs. In many cases, regeneration of the epigenotype by environmental exposure in the current generation is a strong alternative interpretation. The development of labor-intensive tools such as epiRILs, from crosses between plants containing epigenetic changes induced by natural levels of biotic or abiotic stress are one mechanistic way to determine the heritability of stress-related epigenomic variation (Johannes et al., 2009; Holeski et al., 2012). Rearing individuals for generations in alternate environments would also give additional insights into the number of generations environmental signals persist in the epigenome.

LANDSCAPE GENOMICS AND THE EPIGENOME

A critical next step is to determine the evolutionary relevance of the observed epigenetic patterns. Methods for detecting epigenetic variation at the landscape level are not designed to allow the researcher to differentiate between epigenetic variation that is reset within a generation and that which is inherited. In contrast to genetic or genomic patterns, the strength, and occurrence of transgenerational epigenetic inheritance at a landscape level, and thus its evolutionary implications, is poorly understood.

The evolutionary potential of transgenerational epigenetic variation is related to the degree to which it is inherited, as well as the extent to which it deviates from genetic variation. If genetic and epigenetic variations are strongly positively correlated, then the evolutionary trajectory of a population is not

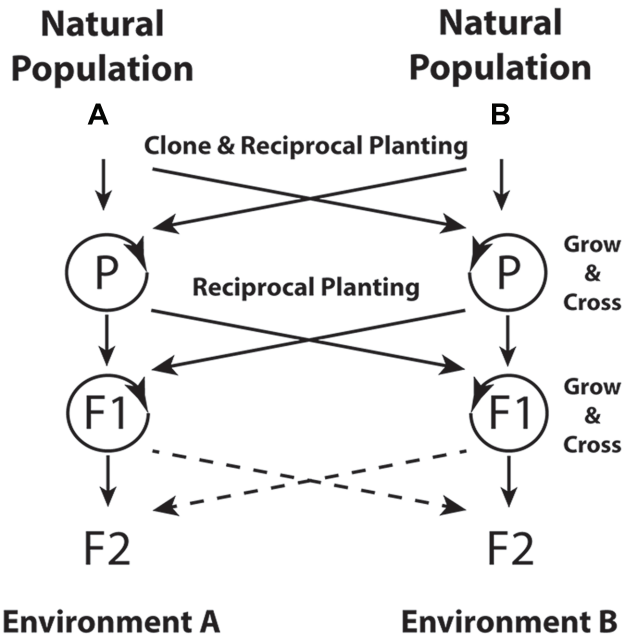


FIGURE 1 | Common garden design for distinguishing among genetic and epigenetic inheritance mechanisms for adaptive traits in a clonal plant species. The reciprocal environmental exposure (Environments A and B) in the parental (P) generation is the primary treatment for the question of whether the environment induces heritable epigenetic changes. A comparison of trait values of clonal replicates or full siblings planted in Environment A and B and derived from parent(s) in either environment A or B will expose within-generation phenotypic plasticity. In the F_2 generation, a comparison of trait values in offspring derived from the same parental genotypes grown in parental environments A versus B should expose phenotypic variation due to epigenetic inheritance. These offspring will share inherited genetic influences, and growth in the same environment eliminates the effects of within-generation plasticity. Additional generations or epigenomic assays may be needed to fully ascribe variation to epigenetic versus resource-based explanations. Dashed lines indicate optional reciprocal planting in the F_2 generation to further understand the time scale and reversibility of environmentally induced inheritance effects.

likely to deviate from that predicted by Mendelian patterns of inheritance (Day and Bonduriansky, 2011). In fact, in this case, the epigenome could be reset every generation and regenerated again from the genome with no influence of inheritance or current environment. In contrast, if genetic and epigenetic variation are weakly or not correlated, as has been demonstrated in *Arabidopsis* (Schmitz et al., 2013), then phenotypic change following selection could be decoupled from the genotype (Day and Bonduriansky, 2011; Liu, 2013). Evidence for both within-generation and transgenerational environmental influence on the epigenome (e.g., Saez-Laguna et al., 2014; Avramidou et al., 2015) suggests that complete correspondence between the genome and epigenome is unlikely. Thus, another expectation might be patterns of greater similarity of epigenomes in similar environments. This similarity could be further enhanced by inheritance of the epigenome.

Assessing the evolutionary relevance of epigenetic patterns across the landscape is a critical component in advancing the field of landscape-level studies of adaptation. How then, can this be done? Many of the methods currently used to detect adaptive genomic variation across the landscape (genome–phenotype associations, phenotype-free approaches, and/or common gardens) are not able to disentangle the effects of transgenerational epigenetic inheritance, relative to Mendelian genetic inheritance, on the phenotype.

Genome–phenotype association methods such as quantitative trait loci mapping and GWAS studies detect relationships between adaptive traits and genomic variation through the association of phenotypic and genomic data in individuals from crosses or natural populations (Rellstab et al., 2015). The inclusion of the phenotype in these analyses may complicate the interpretation of results because the influence of epigenetic effects on the phenotype remains unaccounted for. The results of phenotype-free approaches are relatively unaffected by the potential for transgenerational epigenetic inheritance because these analyses do not hinge on phenotypes that may integrate epigenetic influences. For example, outlier analyses use population genetic principles to detect loci that are likely to have experienced selection. Loci must be under sustained or strong selection to be detected, so this is a conservative technique that will miss many loci of small or fluctuating effect but will not be distorted by the occurrence of epigenetic variation. A second phenotype-free method, EAA, is based on genetic and environmental data and thus will also be unchanged by the occurrence of epigenetic variation. Environmental association analyses are being extended to analyze the association between the epigenome and the environment (Verhoeven et al., 2016), but the interpretation should include the possibilities of the genome and/or environment creating the epigenomic state without the involvement of transgenerational epigenetic inheritance.

CRITICAL NEED FOR COMMON GARDEN APPROACHES

Common garden studies are crucial for disentangling environmental and genetic influences on adaptive traits. Common garden and quantitative genetic designs rarely cover a landscape in as much detail as methods such as GWAS. However, in combination with genomic data, these studies can be used to more fully understand patterns of adaptive variation across the landscape (Sork et al., 2013; De Kort et al., 2014; Lepais and Bacles, 2014).

Multi-generation common garden experiments conducted across multiple environments will allow researchers to understand adaptive epigenomic inheritance (Robertson and Richards, 2015). In plants, transgenerational effects not explained by differences in seed size or mass (the primary visible indications of offspring provisioning) and persisting for multiple generations are hypothesized to occur via epigenetic mechanisms (Zas et al., 2013). Demonstrating that neither genetic loci nor the environment of the individual is the sole source of epigenetic expression, and that epigenomic variation influences adaptive traits, would provide strong evidence for the evolutionary relevance of epigenetic inheritance.

Controlling for genetic sources among environments, and especially the use of clones, would give insight into the extent of genetic determination of the epigenome. Use of multiple environments and of plant sources that span a landscape will allow testing of adaptive hypotheses. **Figure 1** shows a design for distinguishing between genetic and non-genetic inheritance mechanisms for adaptive traits. In this design, natural populations with contrasting environments are used in a reciprocal transplant context to test the adaptive nature of variation. The use of a clonal plant species allows for greater control of genetic background across treatments, and clonal replicates can help researchers minimize the effects of somatic mutation during the experiment. Multiple generations and reciprocal crosses between environments and sources in the parental generation enable separation of the various inheritance patterns (i.e., Day and Bonduriansky, 2011). Assays of the epigenome or additional generations of testing would strengthen an inference of epigenetic inheritance as a contributing causal agent in adaptation to the environment. Inclusion of environments spanning a species range, and modeling to interpolate across the landscape, could create a bridge between traditional common garden and landscape genomic scales (Cushman, 2014).

Thus far, only a few studies have taken advantage of common garden approaches for studying epigenetic inheritance. In two examples, clonal systems also helped narrow down potential sources of phenotypic variation. Studies in *Pinus pinaster* (Cendán et al., 2013; Zas et al., 2013) used seed orchards with cloned genotypes in contrasting common garden environments to assay for effects of maternal environment on offspring traits. They demonstrated effects of maternal environment on offspring traits that could be explained by resources (seed mass) and additional effects that could not be attributed to seed mass. Wilschut et al. (2016) made use of an unusual landscape genetic structure in the dandelion (*T. officinale*) where the same clone is distributed across a wide geographic area. The clonal identity controlled for genetic variation (excluding mutation in these clonal lines since divergence). Generations in different environments resulted in epigenetic differentiation among locations. When plants were grown in a common environment, traits of clonal replicates from different environments remained distinct, showing differentiation was not caused by the environmental exposure in the current generation.

CONCLUSION

Strong evidence exists for epigenetic inheritance and its potential to influence adaptive traits in plants. At the landscape level, studies have identified genomic variation that affects adaptation, but the genetic basis of additional phenotypic variation remains unaccounted for. A number of recent investigations of epigenetic variation across the landscape show patterns consistent with epigenetic inheritance contributing to adaptation. However, carefully designed common garden studies are needed to partition the contributions of genetic variation, phenotypic plasticity, and transgenerational epigenetic inheritance to adaptive phenotypes.

AUTHOR CONTRIBUTIONS

Both AW and LH contributed to the ideas, writing and editing of this manuscript.

FUNDING

This material is based upon work supported by the National Science Foundation under Grant Nos. (DBI-1126840 and EF-1442597) and by Northern Arizona University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

ACKNOWLEDGMENTS

The authors thank Ehren Moler for comments on the manuscript and Paul Heinrich for figure drafting. Two reviewers provided helpful comments which improved the manuscript.

REFERENCES

Alonso, C., Perez, R., Bazaga, P., Medrano, M., and Herrera, C. M. (2014). Individual variation in size and fecundity is correlated with differences in global DNA cytosine methylation in the perennial herb *Helleborus foetidus* (Ranunculaceae). *Am. J. Bot.* 101, 1309–1313. doi: 10.3732/ajb.1400126

Avramidou, E. V., Ganopoulos, I. V., Doulis, A. G., Tsaftaris, A. S., and Avranopoulos, F. A. (2015). Beyond population genetics: natural epigenetic variation in wild cherry (*Prunus avium*). *Tree Genet. Genomes* 11:95. doi: 10.1007/s11295-015-0921-7

Barrett, R. D., and Hoekstra, H. E. (2011). Molecular spandrels: tests of adaptation at the genetic level. *Nat. Rev. Genet.* 12, 767–780. doi: 10.1038/nrg3015

Bonduriansky, R. (2012). Rethinking heredity, again. *Trends Ecol. Evol.* 27, 330–336. doi: 10.1016/j.tree.2012.02.003

Bonduriansky, R., and Day, T. (2009). Nongenetic inheritance and its evolutionary implications. *Annu. Rev. Ecol. Evol. Syst.* 40, 103–125. doi: 10.1146/annurev.ecolsys.39.110707.173441

Bossdorf, O., Richards, C. L., and Pigliucci, M. (2008). Epigenetics for ecologists. *Ecol. Lett.* 11, 106–115.

Bradburd, G. S., Ralph, P. L., and Coop, G. M. (2013). Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution* 67, 3258–3273. doi: 10.1111/evo.12193

Castonguay, E., and Angers, B. (2012). The key role of epigenetics in the persistence of asexual lineages. *Genet. Res. Inter.* 4, 1–9. doi: 10.1155/2012/534289

Cendán, C., Sampedro, L., and Zas, R. (2013). The maternal environment determines the timing of germination in *Pinus pinaster*. *Env. Exp. Bot.* 94, 66–72. doi: 10.1016/j.envexpbot.2011.11.022

Cervera, M. T., Ruiz-Garcia, L., and Martinez-Zapater, J. (2002). Analysis of DNA methylation in *Arabidopsis thaliana* based on methylation- sensitive AFLP markers. *Mol. Genet. Genomics* 268, 543–552. doi: 10.1007/s00438-002-0772-4

Charlesworth, B., Charlesworth, D., and Barton, N. H. (2003). The effects of genetic and geographic structure on neutral variation. *Ann. Rev. Ecol. Evol. Syst.* 34, 99–125. doi: 10.1146/annurev.ecolsys.34.011802.132359

Clausen, J., Keck, D. D., and Hiesey, W. M. (1948). *Experimental Studies on the Nature of Species. III: Environmental Responses of Climatic Races of Achillea*. Washington, DC: Carnegie Institution of Washington.

Cortijo, S., Wardenaar, R., Colome-Tatche, M., Gilly, A., Etcheverry, M., Labadie, K., et al. (2014). Mapping the epigenetic basis of complex traits. *Science* 343, 1145–1148. doi: 10.1126/science.1248127

Cubas, P., Vincent, C., and Coen, E. (1999). An epigenetic mutation responsible for natural variation in floral symmetry. *Nature* 401, 157–161. doi: 10.1038/43657

Cushman, S. A. (2014). Grand Challenges in evolutionary and population genetics: the importance of integrating epigenetics, genomics, modeling, and experimentation. *Front. Genet.* 5:197. doi: 10.3389/fgene.2014.00197

Day, T., and Bonduriansky, R. (2011). A unified approach to the evolutionary consequences of genetic and nongenetic inheritance. *Am. Nat.* 178, E18–E36. doi: 10.1086/660911

De Kort, H., Vandepitte, K., Bruun, H. H., Closset-Kopp, D., Honnay, O., and Mergeay, J. (2014). Landscape genomics and a common garden trial reveal adaptive differentiation to temperature across Europe in the tree species *Alnus glutinosa*. *Mol. Ecol.* 23, 4709–4721. doi: 10.1111/mec.12813

Dubin, M. J., Zhang, P., Meng, D., Remigereau, M.-S., Osborne, E. J., Casale, F. P., et al. (2015). DNA methylation in *Arabidopsis* has a genetic basis and shows evidence of local adaptation. *eLife* 4:e05255. doi: 10.7554/eLife.05255

Foust, C. M., Preite, V., Schrey, A. W., Alvarez, M., Robertson, M. H., Verhoeven, K. J. F., et al. (2016). Genetic and epigenetic differences associated with environmental gradients in replicate populations of two salt marsh perennials. *Mol. Ecol.* 25, 1639–1652. doi: 10.1111/mec.13522

Furrow, R. E., Christiansen, F. B., and Feldman, M. W. (2011). Environment–sensitive epigenetics and the heritability of complex diseases. *Genetics* 189, 1377–1387. doi: 10.1534/genetics.111.131912

Goldstein, D. B. (2009). Common genetic variation and human traits. *N. Engl. J. Med.* 360, 1696–1698. doi: 10.1056/NEJMp0806284

Gugger, P., Fitz-Gibbon, S., Pellegrini, M., and Sork, V. L. (2016). Specieswide patterns of DNA methylation variation in *Quercus lobata* and its association with climate gradients. *Mol. Ecol.* 25, 1665–1680. doi: 10.1111/mec.13563

Herman, J. J., Spencer, S. G., Donohue, K., and Sultan, S. E. (2014). How stable "should" epigenetic modifications be? Insights from adaptive plasticity and bet hedging. *Evolution* 68, 632–643. doi: 10.1111/evo.12324

Herman, J. J., and Sultan, S. E. (2011). Adaptive transgenerational plasticity in plants: case studies, mechanisms, and implications for natural populations. *Front. Plant Sci.* 2:102. doi: 10.3389/fpls.2011.00102

Herrera, C., Medrano, M., and Bazaga, P. (2016). Comparative spatial genetics and epigenetics of plant populations: heuristic value and a proof of concept. *Mol. Ecol.* 25, 1653–1664. doi: 10.1111/mec.13576

Holeski, L. M., Jander, G., and Agrawal, A. A. (2012). Transgenerational defense induction and epigenetic inheritance in plants. *Trends Ecol. Evol.* 27, 618–626. doi: 10.1016/j.tree.2012.07.011

Huang, C. L., Chen, J. H., Tsang, M. H., Chung, J. D., Chang, C. T., and Hwang, S. Y. (2015). Influences of environmental and spatial factors on genetic and epigenetic variations in *Rhododendron oldhamii* (Ericaceae). *Tree Genet. Genomes* 11, 823. doi: 10.1007/s11295-014-0823-0

Ingvarsson, P. K., and Street, N. R. (2011). Association genetics of complex traits in plants. *New Phytol.* 189, 909–922. doi: 10.1111/j.1469-8137.2010.03593.x

Jablonka, E., and Raz, G. (2009). Trans–generational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *Q. Rev. Biol.* 84, 131–176. doi: 10.1086/598822

Johannes, F., Porcher, E., Teixeira, F. X., Saliba-Colombani, V., Simon, M., Agier, A., et al. (2009). Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet.* 5:e1000530. doi: 10.1371/journal.pgen.1000530

Kalisz, S., and Purugganan, M. D. (2004). Epialleles via DNA methylation: consequences for plant evolution. *Trends Ecol. Evol.* 19, 309–314. doi: 10.1016/j.tree.2004.03.034

Keller, T. E., Lasky, J. R., and Yi, S. V. (2016). The multivariate association between genome- wide DNA methylation and climate across the range of *Arabidopsis* thaliana. *Mol. Ecol.* 25, 1823–1837. doi: 10.1111/mec.13573

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9, 29. doi: 10.1186/1746-4811-9-29

Lepais, O., and Bacles, C. F. (2014). Two are better than one: combining landscape genomics and common gardens for detecting local adaptation in forest trees. *Mol. Ecol.* 23, 4671–4673. doi: 10.1111/mec.12906

Liu, Q. A. (2013). The impact of climate change on plant epigenomes. *Trends Genet.* 29, 503–505. doi: 10.1016/j.tig.2013.06.004

Medrano, M., Herrera, C., and Bazaga, P. (2014). Epigenetic variation predicts regional and local intraspecific functional diversity in a perennial herb. *Mol. Ecol.* 23, 4926–4938. doi: 10.1111/mec.12911

Nicotra, A. B., Atkin, O. K., Bonser, S. P., Davidson, A. M., Finnegan, E. J., Mathesius, U., et al. (2010). Plant phenotypic plasticity in a changing climate. *Trends Plant Sci.* 15, 684–692. doi: 10.1016/j.tplants.2010.09.008

Preite, V., Snoek, L. B., Oplaat, C., Biere, A., Van der Putten, W. H., and Verhoeven, K. J. F. (2015). The epigenetic footprint of poleward range-expanding plants in apomictic dandelions. *Mol. Ecol.* 24, 4406–4418. doi: 10.1111/mec.13329

Rapp, R. A., and Wendel, J. F. (2005). Epigenetics and plant evolution. *New Phytol.* 168, 81–91. doi: 10.1111/j.1469-8137.2005.01491.x

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., and Holderegger, R. (2015). A practical guide to environmental, association analysis in landscape genomics. *Mol. Ecol.* 24, 4348–4370. doi: 10.1111/mec.13322

Richards, C. L., Verhoeven, K. J., and Bossdorf, O. (2012). "Evolutionary significance of epigenetic variation," in *Plant Genome Diversity*, eds J. F. Wendel, J. Greilhuber, J. Dolezel, and I. J. Leitch (Viennna: Springer), 257–274.

Richards, E. J. (2006). Inherited epigenetic variation—revisiting soft inheritance. *Nat. Rev. Genet.* 7, 395–402. doi: 10.1038/nrg1834

Robertson, M., and Richards, C. (2015). Non-genetic inheritance in evolutionary theory-the importance of plant studies. *Non-Genetic Inheritance* 2, 3–11. doi: 10.1515/ngi-2015-0002

Saez-Laguna, E., Guevara, M. A., Diaz, L.-M., Sanchez-Gomez, D., Collada, C., Aranda, I., et al. (2014). Epigenetic variability in the genetically uniform forest tree species *Pinus pinea* L. *PLoS ONE* 9:e103145. doi: 10.1371/journal.pone.0103145

Schmitz, R. J., Schultz, M. D., Urich, M. A., Nery, J. R., Pelizzola, M., Libiger, O., et al. (2013). Patterns of population epigenomic diversity. *Nature* 495, 193–198. doi: 10.1038/nature11968

Schulz, B., Eckstein, R. L., and Durka, W. (2014). Epigenetic variation reflects dynamic habitat conditions in a rare floodplain herb. *Mol. Ecol.* 23, 3523–3537. doi: 10.1111/mec.12835

Sexton, J. P., Hangartner, S. B., and Hoffmann, A. A. (2014). Genetic isolation by environment or distance: which pattern of gene flow is most common? *Evolution* 68, 1–15. doi: 10.1111/evo.12258

Sork, V. L., Aitken, S. N., Dyer, R. J., Eckert, A. J., Legendre, P., and Neale, D. B. (2013). Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. *Tree Genet. Genomes* 9, 901–911. doi: 10.1007/s11295-013-0596-x

Verhoeven, K. J. F., Vonholdt, B. M., and Sork, V. L. (2016). Epigenetics in ecology and evolution: what we know and what we need to know. *Mol. Ecol.* 25, 1631–1638. doi: 10.1111/mec.13617

Wang, I. J., and Bradburd, G. S. (2014). Isolation by environment. *Mol. Ecol.* 23, 5649–5662. doi: 10.1111/mec.12938

Wilschut, R. A., Oplaat, C., Snoek, B., Kirschner, J., and Verhoeven, K. J. F. (2016). Natural epigenetic variation contributes to heritable flowering divergence in a widespread asexual dandelion lineage. *Mol. Ecol.* 8, 1759–1768. doi: 10.1111/mec.13502

Wright, S. (1943). Isolation by distance. *Genetics* 28, 114–138.

Yakovlev, I. A., Lee, Y., Rotter, B., Olsen, J. E., Skroppa, T., Johnsen, O., et al. (2014). Temperature-dependent differential transcriptomes during formation of an epigenetic memory in *Norway spruce* embryogenesis. *Tree Genet. Genomes* 10, 355–366. doi: 10.1007/s11295-013-0691-z

Zas, R., Cendan, C., and Sampedro, L. (2013). Mediation of seed provisioning in the transmission of environmental maternal effects in *Maritime pine* (*Pinus pinaster* Aiton). *Heredity* 111, 248–255. doi: 10.1038/hdy.2013.44

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 November 2016
doi: 10.3389/fgene.2016.00209

Spatially Heterogeneous Environmental Selection Strengthens Evolution of Reproductively Isolated Populations in a Dobzhansky–Muller System of Hybrid Incompatibility

Samuel A. Cushman[1]* *and Erin L. Landguth*[2]

[1] USDA Forest Service, Rocky Mountain Research Station, Flagstaff, AZ, USA, [2] Division of Biological Sciences, University of Montana, Missoula, MY, USA

Within-species hybrid incompatibility can arise when combinations of alleles at more than one locus have low fitness but where possession of one of those alleles has little or no fitness consequence for the carriers. Limited dispersal with small numbers of mate potentials alone can lead to the evolution of clusters of reproductively isolated genotypes despite the absence of any geographical barriers or heterogeneous selection. In this paper, we explore how adding heterogeneous natural selection on the genotypes (e.g., gene environment associations) that are involved in reproductive incompatibility affects the frequency, size and duration of evolution of reproductively isolated clusters. We conducted a simulation experiment that varied landscape heterogeneity, dispersal ability, and strength of selection in a continuously distributed population. In our simulations involving spatially heterogeneous selection, strong patterns of adjacency of mutually incompatible genotypes emerged such that these clusters were truly reproductively isolated from each other, with no reproductively compatible "bridge" individuals in the intervening landscape to allow gene flow between the clusters. This pattern was strong across levels of gene flow and strength of selection, suggesting that even relatively weak selection acting in the context of strong gene flow may produce reproductively isolated clusters that are large and persistent, enabling incipient speciation in a continuous population without geographic isolation.

Keywords: CDPOP, computer simulations, genotype-environment associations, hybrid-incompatability, landscape genomics

OPEN ACCESS

Edited by:
Guo-Bo Chen,
Evergreen Landscape&Architecture Studio, China

Reviewed by:
Yu-Ping Poh,
University of Massachusetts Boston, USA
Olivier Francois,
Grenoble Institute of Technology, France

***Correspondence:**
Samuel A. Cushman
scushman@fs.fed.us

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *13 August 2016*
Accepted: *10 November 2016*
Published: *24 November 2016*

Citation:
Cushman SA and Landguth EL (2016) Spatially Heterogeneous Environmental Selection Strengthens Evolution of Reproductively Isolated Populations in a Dobzhansky–Muller System of Hybrid Incompatibility. Front. Genet. 7:209. doi: 10.3389/fgene.2016.00209

INTRODUCTION

Hybrid incompatibility refers to when hybrids between species exhibit reduced viability, lower fertility, and/or phenotypic abnormalities, and is a form of postzygotic reproductive isolation. A number of researchers have argued that hybrid incompatibility is important to the speciation process (Coyne and Orr, 2004). Dobzhansky (1937) and Muller (1942) presented models arguing that hybrid incompatibility usually evolves due to changes in at least two different genetic loci. Genetic studies strongly support the Dobzhansky–Muller model (Coyne and Orr, 2004; Seehausen et al., 2014), and a growing number of these hybrid incompatibility genes have been identified (reviewed in Johnson, 2010; Presgraves, 2010a,b).

Hybrid incompatibility can also occur between different populations of the same species (e.g., in flour beetles, Demuth and Wade, 2007; in flies, Lachance and True, 2010; in nematodes, Seidel et al., 2008, 2011). Within-species hybrid incompatibility can arise given synthetic deleterious loci, sets of loci wherein individuals with combinations of alleles at more than one locus have low fitness but where possession of one of those alleles has little or no fitness consequence for the carriers (Phillips and Johnson, 1998). Analytical studies (Phillips and Johnson, 1998; Lachance et al., 2011) showed that these synthetic alleles could reach considerably high frequencies (roughly the quartic root of the mutation rate divided by the selection coefficient) in panmictic populations under mutation-selection balance (see also, Lachance et al., 2011). Indeed, synthetic lethality and sterility has been found at appreciable frequencies in populations of *Drosophila melanogaster* (e.g., Lachance and True, 2010).

Eppstein et al. (2009) showed that limited dispersal with small numbers of mate potentials alone can lead to the evolution of clusters of reproductively isolated genotypes despite the absence of any geographical barriers or heterogeneous selection. Such clusters evolved when several loci were underdominant (heterozygotes less fit than either homozygote). Non-additive fitness effects across loci (epistasis) enhanced the likelihood of clustering. Landguth et al. (2015) extended the work of Eppstein et al. (2009) to show that underdominance is not required for clustering of reproductively isolated genotypes. Landguth et al. (2015) simulated fitness determined by epistatic interactions, in form of the well-known Dobzhansky–Muller model, and unlike past simulation studies, which consider migration of individuals between demes (e.g., Gavrilets and Vose, 2007; Gavrilets et al., 2007), they modeled genetic divergence in an individual-based framework where gene flow, genetic drift, mutation, and selection were functions of individual-based movement and spatially-explicit interactions with environment (Landguth et al., 2012).

Landguth et al. (2015) showed that hybrid incompatibility can evolve within the same population when gene flow is strongly restricted in an isolation-by-distance model. They showed that under isolation-by-distance reproductively isolated clusters could arise and persist for many generations. Most of the models of sympatric speciation wherein reproductive isolation arises in the face of moderate or strong gene flow involve the counterbalancing force of relatively strong and heterogeneous natural selection. In these models, selection enables nascent species to evolve genetic differences that are incompatible with the evolved differences in the other nascent species (Gavrilets and Vose, 2007; Gavrilets et al., 2007; Nosil and Feder, 2012). In this paper, we expand upon the Landguth et al. (2015) work and explore how adding heterogeneous natural selection on the genotypes that are involved in reproductive incompatibility affects the frequency, size and duration of evolution of reproductively isolated clusters.

MATERIALS AND METHODS

Simulation Program

We used CDPOP v1.0 (Landguth et al., 2012), a landscape genetics tool for simulating the emergence of spatial genetic structure in populations resulting from specified landscape processes governing organism movement behavior. CDPOP models genetic exchange among spatially located individuals as a function of individual-based movement through mate selection and dispersal, incorporating vital dynamics (birth and death rates), and all the factors that affect the frequency of an allele in a population (mutation, gene flow, genetic drift, and selection). The landscape genetics framework of this program is such that individuals move as a probabilistic function of their environment (e.g., as habitat fragmentation increases, ability to disperse across gaps is reduced). These movement functions are scaled to a user-specified maximum dispersal and mate selection distance. This maximum movement value allows a user to control for short- and long-range movement of an organism by constraining all mate choices and dispersal distances to be within that limit, with probability specified by the user-defined movement function (e.g., inverse-square). The order of simulated events follow mate selection with given movement functions, birth and resulting Mendalian inheritance, mortality of adults, and offspring dispersal with given movement functions.

CDPOP v1.0 incorporates multi-locus selection, which is controlled via spatially-explicit fitness surfaces for each genotype under selection (Wright, 1932; Gavrilets, 2000). For example, in the case of a single two-allele locus, three relative fitness surfaces would be specified for the three genotypes (*AA*, *Aa*, and *aa*) from the two alleles, *A* and *a*. Selection is then implemented through differential survival of offspring as a function of the relative fitness of the offspring's genotype at the location on that surface where the dispersing individual settles (Landguth et al., 2012). CDPOP yields genetic patterns consistent with Wright–Fisher expectations when parameterized to match Wright–Fisher assumptions in simulations (Landguth and Cushman, 2010), as well as producing theoretical changes in allele frequency under selection for single and double diallelic locus (Landguth et al., 2012). For more details, see Landguth et al. (2012).

Our simulations consisted of 5000 diploid individuals with 100 biallelic loci; two of these loci were subject to selection. We initialized the 100 loci with a uniformly distributed random allele assignment (maximum allelic diversity). All loci experienced a 0.0005 mutation rate per generation (on the lower range of mammalian microsatellite rates) using the K allele model, a commonly used mutation model (Balloux, 2001; Haasl and Payseur, 2010), free recombination, and no physical linkage. Simulation parameters, other than for selection (described below), matched those in Landguth et al. (2015). Mating parameters represented a population of dioecious individuals with females and males mating with replacement. The number of offspring produced from mating was determined from a Poisson distribution (mean = 4), which produced an excess of individuals each generation to maintain a constant population size of 5000 individuals at every generation. Carrying capacity of the simulation surface was 5000 individuals. Excess individuals were discarded once all 5000 locations became occupied, which is equivalent to forcing out emigrants once all available home ranges are occupied (Balloux, 2001; Landguth and Cushman, 2010). We ran 10 Monte Carlo replicates of each simulation for 1250 generations, discarding the first 250 generations as

burn-in (no selection imposed) to establish a spatial genetic pattern prior to initiating the heterogeneous landscape selection configurations.

Simulation Scenarios

Our simulations combined dispersal in an isolation-by-distance (IBD) framework with heterogeneous natural selection for genotypes involved in reproductive incompatibility. The simulation modeling experiment involved all combinations of three factors (dispersal, landscape heterogeneity, and strength of selection; **Figure 1**).

The first factor is the degree of dispersal and we simulated six movement distances: 3, 5, 10, 15, 25, and 50% of the maximum extent of the landscape. These dispersal distances correspond to a broad range of possible dispersal destinations for a given offspring, as well as available mating partners for a given individual. Mating pairs of individuals and dispersal locations of offspring were chosen based on a random draw from the inverse-square probability function of distance, truncated with the specified maximum distance.

The second factor is the pattern of landscape heterogeneity of two habitat types providing differential selection for the genotypes involved in heterogeneous selection. Specifically, we used the neutral landscape model, QRULE (Gardner, 1999), to simulate binary landscape maps (1024 × 1024 pixels). Habitat fragmentation was controlled with the H parameter, which affects the aggregation of habitat pixels; higher values of H lead to higher levels of aggregation. The binary landscapes consisted of 50% of each of two habitat types and aggregation levels of H = 0.1 ("H1," **Figure 1A**), 0.5 ("H5," **Figure 1B**), and 0.9 ("H9," **Figure 1C**). Heterogeneous selection acted in a discrete fashion in which different homozygous genotypes (i.e., *AABB* and *aabb*; see below) were each favored by selection in one of the two habitat types. We produced 10 replicate landscapes for each *H*-value to assess stochastic variation among simulated landscapes.

Across these different heterogeneous landscapes and dispersal distances, we tested the third factor: strength of selection, defined as the difference of relative fitness of genotypes involved in hybrid incompatibility in the two habitat types and mediated in the simulations through density-independent (i.e., environment-driven) mortality (*s*) determined by genotypes at the selected loci. Selection strengths included $s = 0.02$ or "2%," $s = 0.04$ or "4%," $s = 0.08$ or "8%," $s = 0.16$ or "16%," $s = 0.32$ or "32%," and $s = 0.64$ or "64%" (see **Table 1**). Following the Dobzhansky–Muller model and the Landguth et al. (2015) simulations, we considered the two-locus (*A* and *B*), two-allele selection model (i.e., nine possible genotypes exist in the two-locus, two-allele selection model). We assumed that alleles *a* and *B* are incompatible and individuals that have these two alleles simultaneously have zero viability. This was implemented through relative fitness surfaces of 0.0 across the landscape for the genotypes *AaBB*, *AaBb*, *aaBB*, and *aaBb* as in Landguth et al. (2015). In this model, all offspring of matings between individuals *AABB* and *aabb* will have heterozygous genotype *AaBb* which will be inviable or sterile. The heterogeneous selection acting on the five remaining viable genotypes occurred relatively around

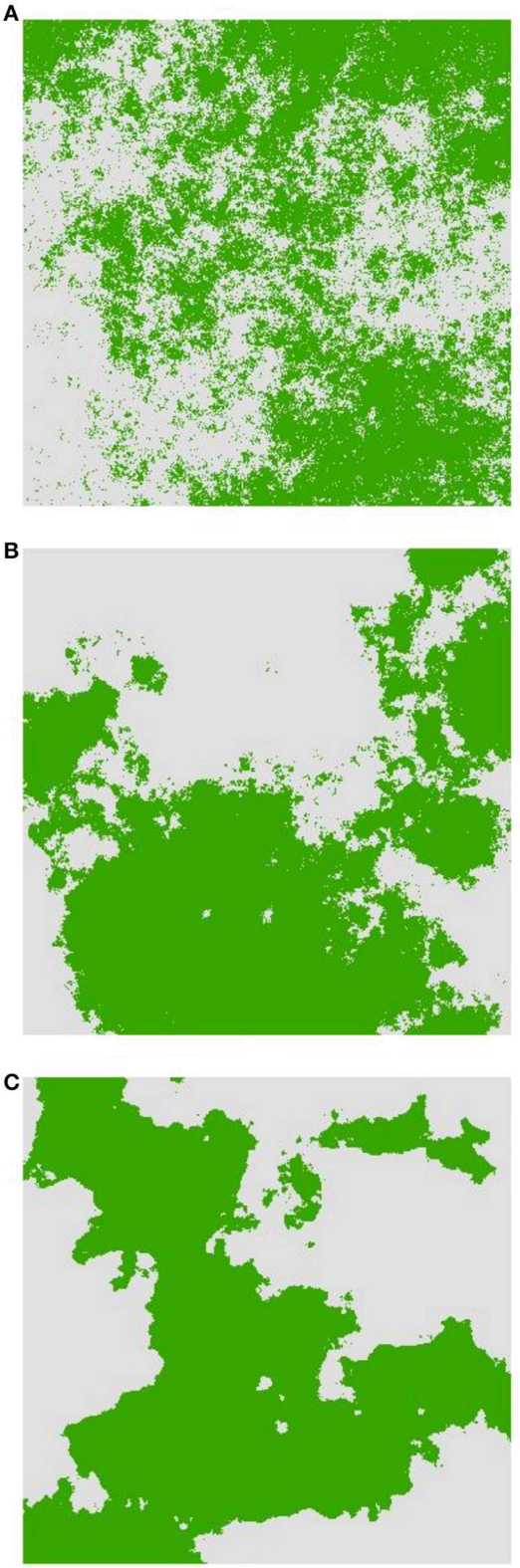


FIGURE 1 | Examples of landscape selection configurations used for simulations from least to most aggregated. (A) H1, **(B)** H5, and **(C)** H9. Dark areas represent *AABB* habitat and light areas represent *aabb* habitat.

TABLE 1 | The proportion of survival for each genotype in *AABB* habitat.

Selection scenario (%)	*AABB*	*AABb*	*AAbb*	*AaBB*	*AaBb*	*Aabb*	*aaBB*	*aaBb*	*aabb*
2	0.51	0.50	0.50	0.00	0.00	0.50	0.00	0.00	0.49
4	0.52	0.50	0.50	0.00	0.00	0.50	0.00	0.00	0.48
8	0.54	0.50	0.50	0.00	0.00	0.50	0.00	0.00	0.46
16	0.58	0.50	0.50	0.00	0.00	0.50	0.00	0.00	0.42
32	0.66	0.50	0.50	0.00	0.00	0.50	0.00	0.00	0.34
64	0.82	0.50	0.50	0.00	0.00	0.50	0.00	0.00	0.18

$s = 0.5$ or 50% mortality. *AABB* individuals had mortality less than 50% in "*AABB*" habitat patches and experienced high mortality (>50%) if they occurred in "*aabb*" habitat patches. *AABb* individuals had mortality less than 50% but greater than the favored *AABB* individuals. Individuals with *aabb* and *Aabb* genotypes experienced the opposite selection gradient from those of *AABB* and *AABb*, respectively. For example, in the $s = 0.02$ scenarios there would be a net 2% difference in fitness between *aabb* and *AABB* genotypes in the two habitat types, with *AABB* having 51% survival in its favored habitat type, and 49% survival in its disfavored type, while *aabb* would have 51% survival in its favored type and 49% survival in its disfavored type. The *AAbb* genotypes experienced a uniform selection of $s = 0.5$ or 50% mortality across the entire surface. **Table 1** lists the proportion of survival for each genotype corresponding to each relative selection strength scenario.

Evaluating Clusters of Reproductive Isolation

Following Landguth et al. (2015), we defined the occurrence of reproductive isolation in a continuously distributed population as the combination of two criteria: (1) an occurrence of a spatial cluster of individuals with genotype *AABB* that emerges simultaneously with another spatial cluster of individuals with genotype *aabb* (RI event) and (2) a RI event persisting in consecutive generations. To define an RI event, we used the density-based spatial clustering algorithm (DBSCAN; Ester et al., 1996), which finds spatial clusters if they contain sufficiently many points ($k = 4$) within a neighborhood ($\varepsilon = 2000\,\mu$; see Ester et al., 1996; Landguth et al., 2015). Then, the number of generations at which two separate clusters (*AABB* and *aabb*, respectively) emerged with the above criteria (RI events) was reported and averaged across the 10 Monte Carlo runs for each dispersal scenario. To assess persistence of RI events, we simply recorded the duration (in generations) of each RI event and reported the average time duration across each replicate and for each dispersal strategy. We also recorded the size of each RI event in terms of the number of individuals in the reproductively isolated cluster.

RESULTS

Mean Cluster Duration

Factorial analysis of variance found highly significant main effects for landscape heterogeneity, strength of environmental selection, and dispersal ability on the mean duration that reproductively isolated clusters of individuals persisted in the simulations (**Table 2**). The *F*-value was more than four times higher for selection and dispersal than for landscape heterogeneity, suggesting larger differences in cluster duration across levels of selection and dispersal than levels of habitat heterogeneity. There were significant interactions between landscape heterogeneity and selection and dispersal, and weaker interaction between landscape heterogeneity and selection.

To explore the main effects and the predominant interaction between landscape heterogeneity and dispersal we produced histograms in a dispersal × selection space, across the three levels of landscape heterogeneity (**Figures 2A–C**; **Supplementary Video S1** duration.avi). These charts illustrate two main patterns. First, reproductively isolated clusters persist for the entire simulation time when dispersal is low and environmental selection is high. Second, the duration of reproductively isolated clusters increases across levels of dispersal and selection as landscapes become less heterogeneous. For example, at H1, the most heterogeneous configuration, reproductively isolated clusters persist for the full simulation time at combinations of dispersal between 3 and 5% and selection levels of 32 or 64 (**Figure 2A**; **Supplementary Video S1** duration.avi). At the H5 level of heterogeneity, reproductively isolated clusters persist for the full simulation time for dispersal 3% when selection is 8 or above, at dispersal 5% when selection is 16 or above, at 10% dispersal when selection is 32 or above, and at dispersal 25% when selection is 64. The pattern continues at the highest level of aggregation, H9, when clusters have duration across the full extent of the simulation time or nearly the full extent for all combinations of dispersal and selection producing clusters (diagonal across dispersal-selection space from D3 to S64).

Mean Cluster Number

Factorial analysis of variance found highly significant main effects for landscape heterogeneity, strength of environmental selection, and dispersal ability on the mean number of reproductively isolated clusters of individuals (**Table 3**). The *F*-value was nearly ten times higher selection and dispersal than for landscape heterogeneity, suggesting larger differences in number of isolated clusters across levels of selection and dispersal than across levels of habitat

TABLE 2 | Analysis of variance table for factorial ANOVA of mean duration of reproductively isolated clusters (in generations) as function of dispersal ability (D: 3, 5, 10, 15, 25, 50% of breadth of landscape), selection (S: 2, 4, 8, 16, 32, 64% difference in relative fitness of genotypes *AABB* and *aabb* in habitat types 1 and 2 respectively), and landscape heterogeneity (Qrule H: 0.1, 0.5, 0.9) specifying the pattern of habitat types 1 and 2 in the landscape.

DF	SS	Mean square	*F*-value	Pr > F	DF
Heterogeneity	2	168,815	84,407	10.145	0.0002
Selection	5	1,832,356	366,471	44.045	2.00×10^{-16}
Dispersal	5	2,139,692	427,938	51.433	2.00×10^{-16}
Heterogeneity: Selection	10	282,283	28,228	3.393	0.00191
Heterogeneity: Dispersal	10	87,559	8756	1.052	0.41559
Selection:Dispersal	25	1,867,592	74,704	8.978	3.79×10^{-11}
Residuals	50	416,016	8320		

heterogeneity. There were significant interactions between landscape heterogeneity and selection and dispersal, and weaker interaction between landscape heterogeneity and selection.

The histograms (**Figures 2D–F**; **Supplementary Video S2** number.avi) illustrate three main patterns. First, as in the case of cluster duration, the number of reproductively isolated clusters is highest when dispersal is low and environmental selection is high. Second, and contrary to cluster duration, the number of clusters shows a wave pattern moving across the dispersal × selection space toward high dispersal and low selection as the landscape becomes less heterogeneous (e.g., from H1 to H5 to H9). For example, at H1 (the most heterogeneous scenario) there is a clear peak with the largest number of reproductively isolated clusters in scenarios with the shortest dispersal (3%) and strongest selection (64), with roughly linear decay along both selection and dispersal axes (**Figure 2D**). However, at H5, which is an intermediate level of landscape heterogeneity, the peak of number of isolated clusters turns into a ridge running diagonally across intermediate combinations of dispersal ability and selection (**Figure 2E**; e.g., D3S8, D4S16, D10S16, D15S32, D25S64). The pattern continues at the highest level of landscape aggregation (lowest heterogeneity; H9) with the ridge moving diagonally toward the foreground in (**Figure 2F**).

Mean Cluster Size

As with the other response variables (cluster duration and cluster number), factorial analysis of variance found highly significant main effects for landscape heterogeneity, strength of environmental selection, and dispersal ability on the size of reproductively isolated clusters of individuals (**Table 4**). The *F*-value was more twice as high for selection as for dispersal and four times higher than for landscape heterogeneity, suggesting larger differences in the size of isolated clusters across levels of selection, then dispersal, and weakest effect due to habitat heterogeneity. There were significant interactions between landscape heterogeneity and selection and dispersal, and weaker interaction between landscape heterogeneity and selection.

The histograms displaying size of reproductively isolated clusters across combinations of dispersal ability and strength of environmental selection (**Figures 2G–I**; **Supplementary Video S3** size.avi) show a pattern similar to those for cluster duration, except that in the case of cluster size selection seems to have a substantially larger effect than dispersal ability. Specifically, at all levels of habitat heterogeneity (H) the size of clusters of reproductively isolated individuals is highest at when selection is strong and dispersal is limited, but large clusters can persist at high levels of selection even when dispersal is relatively broad-scale (e.g., S32–S64 when D10–D15), while the converse is not true; clusters remain small when selection is weak even when dispersal is limited (e.g., S2–S8 when D3–D10). Second, there is a large effect of changing patterns of heterogeneity of the landscape features driving environmental selection of the genotypes involved in reproductive isolation (**Figures 2E–G**; **Supplementary Video S3** size.avi). For example, when H is 1 (highest level of heterogeneity) the largest cluster sizes are around 220 individuals (at D3S32). At H5 (intermediate heterogeneity) clusters of this size are found at levels of D3–D25 × S32–S64, and the largest cluster sizes exceed 450 individuals at combinations of dispersal and selection D3–D5 × S32–S64, and the largest clusters of over 500 individuals emerge at dispersal levels of between D5–D10 and selection level S64. The pattern continues at H9 (highest habitat aggregation) where clusters of over 630 reproductively isolated individuals emerge and clusters larger than 500 individuals are found at combinations of dispersal D3–D15 across selection levels of S32–S64 (**Figure 2G**).

DISCUSSION

Landguth et al. (2015) found that short-range dispersal strategies lead to the evolution of clusters of reproductively isolated genotypes despite the absence of any geographic barriers or heterogeneous selection. In addition, they found that clusters of genotypes that are reproductively isolated from other clusters can persist when migration distances are restricted such that the number of mating partners is below about 350 individuals. From these results they argued that under strong selection clusters of incompatible genotypes will readily evolve within continuously distributed populations when dispersal distances and potential mating choices are small relative to entire landscape extents and population size, respectively. Short mating distances reduce the rate at which genes moved through the population and reduce local effective population sizes such that local genetic structure would be maintained and not swamped by the homogenizing effects of high rates of gene flow. When mating and dispersal are very limited, reproductive isolation frequently evolves and reproductively isolated clusters may be highly persistent over time.

In this paper, we show that adding heterogeneous selection for the genotypes involved in reproductive isolation led to dramatic

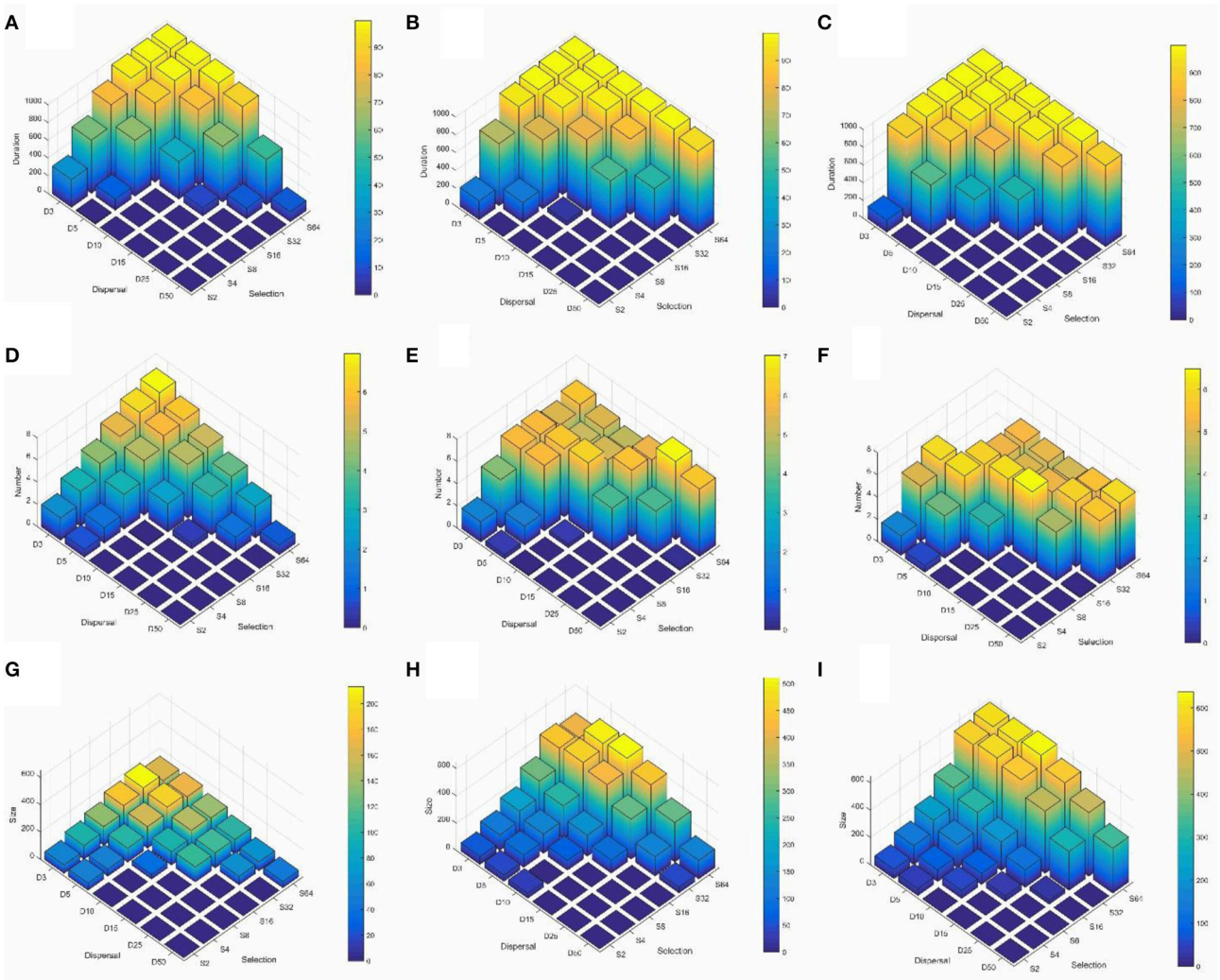


FIGURE 2 | Three-dimensional histograms of changes in the mean duration of reproductively isolated clusters of individuals (in generations; row 1, A–C), mean number of reproductively isolated clusters (row 2, **D–F**), and mean size of reproductively isolated clusters (individuals; row 3, **G–I**). Columns in the figure represent different levels of landscape aggregation of the two habitat types involved in environmental section of the genotypes contributing to reproductive isolation (column 1, **A,D,G** is H1, highly heterogeneous; column 2, **B,E,H** is H5, intermediate heterogeneity; column 3, **C,F,I** is H9, high aggregated patterns of the two habitat types). The 6 × 6 parameter space in each subfigure shows the combinations of six levels of dispersal (D3—3% of landscape extent, D5—5% of landscape extent, D10—10% of landscape extent, D15—15% of landscape extent, D25—25% of landscape extent, D50—50% of landscape extent) across six levels of selection (S2—2% difference in relative fitness of genotypes *aabb* and *AABB* in each of the two habitats, S4—4% difference in relative fitness, S8—8% difference in relative fitness, S16—16% difference in relative fitness, S32—32% difference in relative fitness, S64—64% difference in relative fitness). See **Supplementary Videos** for these histograms as they change through time.

increases in the duration, number, and size of reproductively isolated patches. Landguth et al. (2015) found that reproductively isolated clusters do not evolve when dispersal is >10% of the extent of the population, and that few clusters evolve and these only persist a short time when dispersal is >5% of the extent of the population. In strong contrast, we found that when there is spatially heterogeneous selection on genotypes involved in reproductive isolation, reproductively isolated clusters can evolve even at very high levels of dispersal, and these clusters can achieve very large size and very long duration, with number, size and duration increasing with the strength of selection.

We also found that strength of selection and dispersal ability affect the size, duration, and number of isolated clusters in roughly the same degree, and much more so than does the heterogeneity of the landscape. However, landscape heterogeneity does have substantial effects, such that when there is extremely high heterogeneity reproductively isolated clusters are less likely to emerge since there is a highly mixed pattern of selection that inhibits formation of large, aggregated clusters. This suggests that in evolutionary landscape genetics, as well as neutral differentiation (e.g., Cushman et al., 2012, 2013), there may be threshold effects where

TABLE 3 | Analysis of variance table for factorial ANOVA of mean number of reproductively isolated clusters as function of dispersal ability (D: 3, 5, 10, 15, 25, 50% of breadth of landscape), selection (S: 2, 4, 8, 16, 32, 64% difference in relative fitness of genotypes *AABB* and *aabb* in habitat types 1 and 2 respectively), and landscape heterogeneity (QRULE H: 0.1, 0.5, 0.9) specifying the pattern of habitat types 1 and 2 in the landscape.

	DF	SS	Mean square	*F*-value	Pr > F
Heterogeneity	2	35,035,209	17,517,605	9.759	0.000264
Selection	5	9.75×10^8	1.95×10^8	108.618	$<2 \times 10^{-16}$
Dispersal	5	6.38×10^8	1.28×10^8	71.09	$<2 \times 10^{-16}$
Heterogeneity: Selection	10	19,371,333	1,937,133	1.079	0.395622
Heterogeneity: Dispersal	10	6,113,877	611,388	0.341	0.965294
Selection:Dispersal	25	3.38×10^8	13,537,693	7.542	8.71×10^{-10}
Residuals	50	89,751,644	1,795,033		

TABLE 4 | Analysis of variance table for factorial ANOVA of size reproductively isolated clusters (individuals) as function of dispersal ability (D: 3, 5, 10, 15, 25, 50% of breadth of landscape), selection (S: 2, 4, 8, 16, 32, 64% difference in relative fitness of genotypes *AABB* and *aabb* in habitat types 1 and 2 respectively), and landscape heterogeneity (QRULE H: 0.1, 0.5, 0.9) specifying the pattern of habitat types 1 and 2 in the landscape.

	DF	SS	Mean square	*F*-value	Pr > F
Heterogeneity	2	46.7	23.33	22.998	8.28×10^{-08}
Selection	5	445.7	89.14	87.877	2.00×10^{-16}
Dispersal	5	217.6	43.53	42.91	2.00×10^{-16}
Heterogeneity: Selection	10	20.7	2.07	2.036	0.0489
Heterogeneity: Dispersal	10	2.8	0.28	0.272	0.9846
Selectoin:Dispersal	25	135.3	5.41	5.336	2.53×10^{-07}
Residuals	50	50.7	1.01		

landscape fragmentation limits emergence of reproductively isolated clusters. However, in contrast to the effect of habitat fragmentation on emergence of neutral genetic structure, in which genetic differentiation only occurs at high levels of landscape heterogeneity, evolution of reproductive isolation is facilitated by highly blocky landscapes with relatively low fragmentation.

In addition to the much larger total number, size, and duration of reproductively isolated patches when there is environmental selection, the pattern of cluster adjacency changes in critical ways that enable persistence of reproductively isolated clusters and therefore the potential for incipient speciation. Specifically in the Landguth et al. (2015) simulation, reproductively isolated clusters evolved only as a function of reproductive isolation and gene flow restriction by isolation-by-distance. This resulted in patterns of clusters in the landscape where putatively "reproductively isolated" clusters were rarely adjacent to clusters of individuals that were actually incompatible with them (**Figure 3**). They were most often adjacent to individuals that were not reproductively isolated from them, and clusters that were reproductively incompatible with them typically existed in other parts of the landscape with non-incompatible individuals in between. These non-incompatible individuals form a genetic "bridge" allowing gene flow between the putatively isolated clusters. While based on the criteria used by Landguth et al. (2015) this qualifies as evolution of reproductively isolated clusters, these clusters they were not isolated in the sense that individuals in these clusters could breed with the individuals that were adjacent to them, and could transfer genes between "isolated" clusters through the "bridge" of these compatible intervening individuals (**Figure 3**).

In contrast, when we added environmental selection on the genotypes involved in hybrid incompatibility very strong patterns of adjacency of mutually incompatible genotypes emerged such that these clusters were truly reproductively isolated from each other as there were no other reproductively compatible "bridge" individuals in the intervening landscape to allow gene flow between the clusters. This pattern was very strong across levels of gene flow and strength of selection, suggesting that even relatively weak selection acting in the context of strong gene flow may produce reproductively isolated clusters that are large and persistent, enabling incipient speciation in a continuous population without geographic isolation.

There are several lines of future work which should be explored to extend the scope of what was found in this paper. First, this paper used a simple two-locus model of hybrid incompatibility. While this is a model that is widely used in theoretical evolutionary ecology (Dobzhansky, 1937; Muller, 1942; Coyne and Orr, 2004) and applies to some real-world populations (Demuth and Wade, 2007, in flies, Lachance and True, 2010; in nematodes, Seidel et al., 2008, 2011), the majority of microevolutionary processes are likely mediated through polygenic selection in which many loci each contribute relatively small fitness effects. This paper serves as an initial analysis of a simple, classical model of two locus selection which provides clear theoretical insight. However, future work should explore how landscape heterogeneity, strength of selection, and dispersal ability interact within the context of multiple loci/allele selection (e.g., de Villemereuil et al., 2014) and how these factors influence the detection of local adaptation (e.g., genotype-environment associations; Bierne et al., 2011; Forester et al., 2016). In addition, future work should explore how underdominance, epistasis, and synonymous vs. nonsynonymous mutations interact in their influence on evolution of reproductively isolated clusters in continuous populations in heterogeneous landscapes. In addition, it will be important to combine simulation experiments with empirical studies and experiments (e.g., Cushman, 2014) to develop robust understanding of how landscape heterogeneity, patterns of gene flow and selection, and dispersal ability affect population differentiation and evolution. Simulation experiments such as presented here can describe the processes affecting populations and identify the conditions under which they have important influences. However, models without data are not compelling (Cushman, 2014). It is essential to confront these models with empirical data on the actual

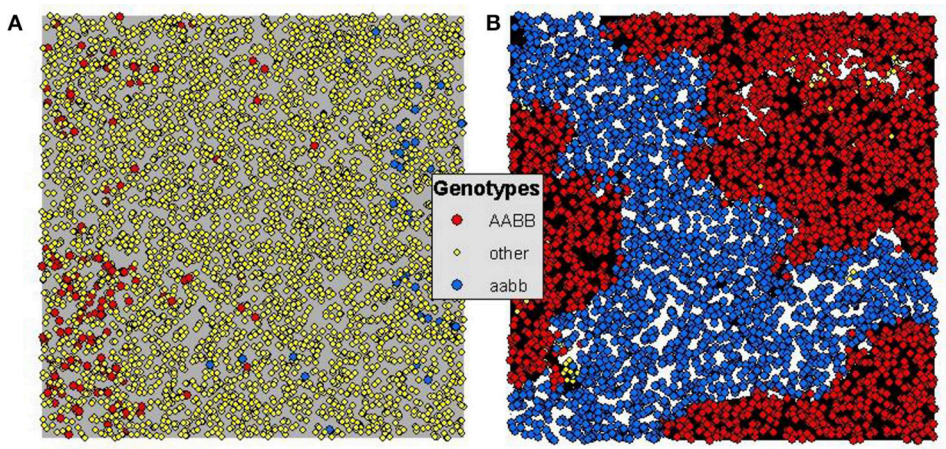


FIGURE 3 | Generation 1250 for 5% maximum dispersal scenarios of (A) uniform selection (i.e., Landguth et al., 2015) and **(B)** heterogeneous selection of $H = 0.9$ and $S = 64$. Orange dots indicate genotype *AABB*, yellow dots indicate genotype *aabb*, and all other genotypes as green dots. **(A)** Shows the pattern of genotypes (red and blue mutually reproductively isolated and yellow compatible with both) in the pure isolation-by-distance framework of Landguth et al. (2015) without heterogeneous selection. **(B)** Shows the pattern of genotypes for a heterogeneous selection scenario with dispersal limited to 5% of the extent of the population and selection set at 64. In **(A)** there are few and small reproductively isolated clusters and these are not truly isolated as the yellow genotypes provide a genetic bridge for gene flow between red and blue. In contrast in **(B)** there is nearly complete elimination of the yellow "bridge" genotypes, and extensive, large and immediately adjacent patches of mutually isolated genotypes (red next to blue).

patterns of genetic differentiation in complex landscapes, and to confirm the fitness relationships underlying these patterns in experimental studies such as common gardens (Cushman, 2014). Thus, we suggest future research that will combine simulation, experimentation, and large-scale population-wide empirical modeling of the influences of landscape heterogeneity, gene flow and strength of selection on the emergence of reproductive isolation.

AUTHOR CONTRIBUTIONS

SC and EL designed research. EL ran simulations. SC and EL performed the analyses, interpreted the data, and wrote the manuscript.

ACKNOWLEDGMENTS

This research was supported by National Science Foundation Grant No. EF-1442597 and the US Forest Service, Rocky Mountain Research Station.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2016.00209/full#supplementary-material

Supplementary Videos | AVIs of three-dimension histograms through time for (S1) change in mean duration of reproductively isolated clusters (S2), mean number of reproductively isolated clusters, and (S3) mean size of reproductively isolated clusters.

REFERENCES

Balloux, F. (2001). EASYPOP (Version 1.7): a computer program for population genetic simulations. *J. Heredity* 92, 301–302. doi: 10.1093/jhered/92.3.301

Bierne, N., Welch, J., Loire, E., Bonhomme, F., and David, P. (2011). The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Mol. Ecol.* 20, 2044–2072. doi: 10.1111/j.1365-294X.2011.05080.x

Coyne, J. A., and Orr, H. A. (2004). *Speciation*. Sunderland, MA: Sinauer Associates.

Cushman, S. A. (2014). Grand challenges in evolutionary and population genetics: the importance of integrating epigenetics, genomics, modeling, and experimentation. *Front. Genet.* 5:197. doi: 10.3389/fgene.2014.00197

Cushman, S. A., Shirk, A. J., and Landguth, E. L. (2012). Separating the effects of habitat area, fragmentation and matrix resistance on genetic differentiation in complex landscapes. *Landsc. Ecol.* 27, 369–380. doi: 10.1007/s10980-011-9693-0

Cushman, S. A., Shirk, A. J., and Landguth, E. L. (2013). Landscape genetics and limiting factors. *Conserv. Genet.* 14, 263–274. doi: 10.1007/s10592-012-0396-0

Demuth, J. P., and Wade, M. J. (2007). Population differentiation in the flour beetle *Tribolium castaneum* II. Haldane's rule and incipient speciation. *Evolution* 61, 694–699. doi: 10.1111/j.1558-5646.2007.00049.x

de Villemereuil, P., Frichot, É., Bazin, É., François, O., and Gaggiotti, O. E. (2014). Genome scan methods against more complex models: when and how much should we trust them? *Mol. Ecol.* 23, 2006–2019. doi: 10.1111/mec.12705

Dobzhansky, T. (1937). *Genetics and the Origin of Species*. New York, NY: Columbia University Press.

Eppstein, M. J., Payne, J. L., and Goodnight, C. J. (2009). Underdominance, multiscale interactions, and self-organizing barriers to gene flow. *J. Artif. Evol. Appl.* 2009:725049. doi: 10.1155/2009/725049

Ester, M., Kriegel, H. P., Sander, J., and Xu, X. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*

(KDD-96), eds E. Simoudis, J. Han, and U. M. Fayyad (Palo Alto, CA: AAAI Press), 226–231.

Forester, B. R., Jones, M. R., Joost, S., Landguth, E. L., and Lasky, J. R. (2016). Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. *Mol. Ecol*. 25, 104–120. doi: 10.1111/mec.13476

Gardner, R. H. (1999). "QRULE: map generation and a spatial analysis program," in *Landscape Ecological Analysis,* eds J. M. Klopatek and R. H. Gardner (New York, NY: Springer), 280–303.

Gavrilets, S. (2000). Waiting time to parapatric speciation. *Proc. R. Soc. Lond. B* 267, 2483–2492. doi: 10.1098/rspb.2000.1309

Gavrilets, S., and Vose, A. (2007). Case studies and mathematical models of ecological speciation. Palms on an oceanic island. *Mol. Ecol.* 16, 2910–2921. doi: 10.1111/j.1365-294X.2007.03304.x

Gavrilets, S., Vose, A., Barluenga, M., Salzburger, W., and Meyer, A. (2007). Case studies and mathematical models of ecological speciation. *Mol. Ecol.* 16, 2893–2909. doi: 10.1111/j.1365-294X.2007.03305.x

Haasl, R. J., and Payseur, B. A. (2010). The number of alleles at a microsatellite defines the allele frequency spectrum and facilities fast accurate estimates of θ. *Mol. Biol. Evol.* 27, 2702–2715. doi: 10.1093/molbev/msq164

Johnson, N. A. (2010). Hybrid incompatibility genes: remnants of a genomic battlefield? *Trends Genet*. 26, 317–325. doi: 10.1016/j.tig.2010.04.005

Lachance, J., Johnson, N. A., and True, J. R. (2011). The population genetics of X-autosome synthetic lethals and steriles. *Genetics* 189, 1011–1027. doi: 10.1534/genetics.111.131276

Lachance, J., and True, J. R. (2010). X-autosome incompatibilities in *Drosophila melanogaster*: tests of Haldane's rule and geographic patterns within species. *Evolution* 64, 3035–3046. doi: 10.1111/j.1558-5646.2010.01028.x

Landguth, E. L., and Cushman, S. A. (2010). CDPOP: a spatially-explicit cost distance population genetics program. *Mol. Ecol. Res.* 10, 156–161. doi: 10.1111/j.1755-0998.2009.02719.x

Landguth, E. L., Cushman, S. A., and Johnson, N. A. (2012). Simulating natural selection in landscape genetics. *Mol. Ecol. Res.* 12, 363–368. doi: 10.1111/j.1755-0998.2011.03075.x

Landguth, E. L., Johnson, N. A., and Cushman, S. A. (2015). Clusters of incompatible genotypes evolve with limited dispersal. *Front. Genet.* 6:151. doi: 10.3389/fgene.2015.00151

Muller, H. J. (1942). Isolating mechanisms. Evolution and temperature. *Biol. Symp.* 6, 71–125.

Nosil, P., and Feder, J. L. (2012). Genomic divergence during speciation: causes and consequences. *Philos. Trans. R. Soc. Land. B* 367, 332–342. doi: 10.1098/rstb.2011.0263

Phillips, P. C., and Johnson, N. A. (1998). The population genetics of synthetic lethals. *Genetics* 150, 449–458

Presgraves, D. C. (2010a). Darwin and the origin of interspecific genetic incompatibilities. *Am. Nat.* 176, S45–S60. doi: 10.1086/657058

Presgraves, D. C. (2010b). The molecular evolution of species formation. *Nat. Rev. Genet.* 11, 175–180. doi: 10.1038/nrg2718

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughmann, J. W., et al. (2014). Genomics and the origin of species. *Nat. Rev. Genet*. 15, 176–192. doi: 10.1038/nrg3644

Seidel, H. S., Allion, J. L., van Qudenaarden, A., Rockman, M. V., and Kruglyak, L., (2011). A novel sperm-delievered toxin causes late-stage embryonic lethality and transmission ratio distortion in C. *elegans. PLoS Biol.* 9:e1001115. doi: 10.1371/journal.pbio.1001115

Seidel, H. S., Rockman, M. V., and Kruglyak, L. (2008). Widespread genetic in C. elegans maintained by balancing selection. *Science* 318, 589–594. doi: 10.1126/science.1151107

Wright, S. (1932). "The roles of mutation, inbreeding, crossbreeding and selection in evolution," in *Proceedings of the VI International Congress of Genetrics,* Vol. 1 (Ithaca, NY), 356–366.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 10 February 2017
doi: 10.3389/fgene.2017.00009

Using Landscape Genetics Simulations for Planting Blister Rust Resistant Whitebark Pine in the US Northern Rocky Mountains

Erin L. Landguth[1], Zachary A. Holden[2], Mary F. Mahalovich[3] and Samuel A. Cushman[4]*

[1] Division of Biological Sciences, University of Montana, Missoula, MT, USA, [2] U.S. Department of Agriculture Forest Service, Missoula, MT, USA, [3] U.S. Department of Agriculture Forest Service, Northern, Rocky Mountain, Southwestern and Intermountain Regions, Moscow, ID, USA, [4] U.S. Department of Agriculture Forest Service, Rocky Mountain Research Station, Flagstaff, AZ, USA

Recent population declines to the high elevation western North America foundation species whitebark pine, have been driven by the synergistic effects of the invasive blister rust pathogen, mountain pine beetle (MPB), fire exclusion, and climate change. This has led to consideration for listing whitebark pine (WBP) as a threatened or endangered species under the Endangered Species Act, which has intensified interest in developing management strategies for maintaining and restoring the species. An important, but poorly studied, aspect of WBP restoration is the spatial variation in adaptive genetic variation and the potential of blister rust resistant strains to maintain viable populations in the future. Here, we present a simulation modeling framework to improve understanding of the long-term genetic consequences of the blister rust pathogen, the evolution of rust resistance, and scenarios of planting rust resistant genotypes of whitebark pine. We combine climate niche modeling and eco-evolutionary landscape genetics modeling to evaluate the effects of different scenarios of planting rust-resistant genotypes and impacts of wind field direction on patterns of gene flow. Planting scenarios showed different levels for local extirpation of WBP and increased population-wide blister rust resistance, suggesting that the spatial arrangement and choice of planting locations can greatly affect survival rates of whitebark pine. This study presents a preliminary, but potentially important, framework for facilitating the conservation of whitebark pine.

Keywords: assisted migration, CDMetaPOP, computer simulations, ecological niche modeling, genotype-environment associations, landscape genomics, wind resistance

OPEN ACCESS

Edited by:
Yann C. Klimentidis, University of Arizona, USA

Reviewed by:
Rodolfo Jaffé, Vale Institute of Technology, Brazil
Patrick M. A. James, Université de Montréal, Canada

****Correspondence:***
Erin L. Landguth
erin.landguth@mso.umt.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *13 August 2016*
Accepted: *18 January 2017*
Published: *10 February 2017*

Citation:
Landguth EL, Holden ZA, Mahalovich MF and Cushman SA (2017) Using Landscape Genetics Simulations for Planting Blister Rust Resistant Whitebark Pine in the US Northern Rocky Mountains. Front. Genet. 8:9. doi: 10.3389/fgene.2017.00009

INTRODUCTION

Whitebark pine (WBP; *Pinus albicaulis*) is one of the most intensively studied North American conifers, in part due to its unique relationship with the grizzly bear (*Ursus arctos horribilis*), Clark's nutcracker (*Nucifraga columbiana*), and over 20 other wildlife species (Lorenz et al., 2008), which depend on its seeds for food; thus it is considered a keystone and foundation species in high elevation forests within its range. Thus, recent declines associated with the spread of mountain pine beetle (MPB; *Dendroctonus ponderosae*), and the introduced invasive fungal pathogen white pine blister rust (WPBR; *Cronartium ribicola*) have led to consideration for listing the species

as threatened under the Endagered Species Act in 2010 (Federal Register 2010), intensifying interest in developing strategies for its conservation and management (see recent reviews by Keane et al., 2012, 2016).

One of the primary threats associated with WBP decline is WPBR-an invasive fungal pathogen introduced to the Pacific Northwest of North America around 1910 (Brar et al., 2015). WPBR affects the productivity and distribution of WBP by forming cankers, which girdle branches and boles, resulting in reduced cone production and increased tree mortality. It has since spread to five-needle pine species across the United States.

Genetic blister rust resistance was first identified in small samples of open-pollinated families by Bingham (1972) and Hoff et al. (1980). A larger trial of 110-seed sources later established the efficacy of identifying, propagating, and deploying blister rust resistant seedlings (Mahalovich et al., 2006). While major gene resistance has not been found in WBP, three resistance mechanisms exhibit as single-gene recessives. The no-spot and needle shed resistance mechanisms are present in very low frequencies (<1%), while the short shoot resistance mechanism is present in low frequency (5.2 percent, Mahalovich *in prep*). In the US Northern Rockies, offspring of over 1300 phenotypic selections are under evaluation in support of active restoration by planting proven, rust-resistant seedlings which have a combination of no-spot, needle-shed, bark reaction and shoot resistance mechanisms (Mahalovich and Dickerson, 2004; Greater Yellowstone Coordinating Committee whitebark pine Subcommittee, 2011; Keane et al., 2012, 2016).

Advances in landscape genetics and population genomics provide a robust means to predict the effects of landscape structure and climatic gradients on genetic structure, population connectivity, and adaptive genetic variation (Manel and Holderegger, 2013; e.g., Shryock et al., 2015). Recently developed simulation modeling tools provide effective means to link landscape patterns to gene flow and adaptive evolutionary processes to predict genetic characteristics of the population across its range under current and potential future conditions (Scribner et al., 2016). Simulation models offers several important benefits for landscape genomic research (Landguth et al., 2015). For example, simulation modeling can be used to predict how a system or its behavior will change if certain processes or parameters are altered. This is particularly relevant for predicting the effects of environmental change on a system, or for evaluating the likely outcomes of various management scenarios.

Our primary objective for this study was to develop a simulation modeling framework for assessing the connectivity of WBP across the US Northern Rocky Mountains and to assess the potential adaptive significance of genetic blister rust resistance. Specifically, we first developed climate niche models for WBP and WPBR distributions. Then, we used these models with an eco-evolutionary landscape genetics model to simulate demographic and genetic (i.e., demogenetic; Frank et al., 2011) responses with and without the presence of white pine blister rust. We conducted simulations that introduced a resistant gene for WPBR and simulated potential planting strategies with this genotype. We also tested the influence of wind field directionality on the ability of pollen to disperse rust-resistant genes through the landscape. Finally, future WBP landscape genetics studies are discussed, including planting strategies with WPBR resistant individuals in conjunction with adaptive simulation modeling experiments.

MATERIALS AND METHODS

Whitebark Pine Regeneration and White Pine Blister Rust Suitability Model

We developed correlative niche models (CNM; aka species distribution or habitat suitability models; Thuiller et al., 2005; Elith and Leathwick, 2009) for WBP and WPBR using occurrence records (presence and absence) to develop a probabilistic model of occurrence based on statistical relationships with climatic, topographic and biophysical variables. One criticism of CNM's applied to long-lived tree species is that they typically correlate adult occurrence records with climate data from relatively short time periods (i.e., 30–50 years). This means that at some locations, an adult tree >300 years old may have established under a very different climate than the one being used to represent its climatic suitability. Recent studies have suggested using juvenile rather than adult occurrences to provide a more realistic characterization of the relationship between a species and a suitable climatic period (Lenoir et al., 2009; Zhu et al., 2011; Bell et al., 2014; Dobrowski et al., 2015). In this study, we used juvenile (<130 mm diameter) occurrence records from Forest Inventory and Analysis (FIA) plot data on all public lands occurring within US Forest Service Northern Region. As predictors we developed a suite of high resolution (240 m) temperature, climatic water balance, and snow distribution models. Gridded data were extracted using the raster library in the R software environment using bilinear interpolation of the four nearest neighbor cells at each FIA plot location. Additional details on the development of the climatic water balance data are provided in Appendix 1. Details about the CNM for WBP and WPBR occurrence are provided in Appendixs 2, 3, respectively.

Whitebark Pine Simulation Model

We used CDMetaPOP (Landguth et al., 2016) to simulate how the presence of WPBR and individuals with resistance to WPBR influence WBP demogenetics. CDMetaPOP is a landscape-level, spatially-explicit, and individual-based eco-genetic model of meta-population processes. CDMetaPOP simulates demogenetic processes as interactions between individuals located across a number of "patches" (hereafter, stands) containing meta-populations. Individuals within a stand are assumed to share a common environment (e.g., carrying capacity, temperature). Within each stand, a class (age/stage/size) structure is used to simulate complex stochastic demographic processes, while movement of individuals (i.e., seeds and pollen) between stands is controlled as a function of spatially-explicit landscape resistance or permeability surfaces (e.g., directional wind resistance to movement). More simply stated, a landscape is populated with stands, which in turn are populated with individual trees. At the stand level, individuals undergo growth, reproduction, migration, and mortality, and the resulting genetic processes are simulated over time at the

individual-tree level. For more detailed information on the processes simulated in CDMetaPOP, see the user manual (https://github.com/ComputationalEcologyLab/CDMetaPOP).

Our WBP model required parameterization of a number of species-specific processes (see Appendix 4, Figure A4.1 and Table A4.1). After initialization of the model (e.g., stands, stage structure, and genetics), pollen dispersal (age 0) occurs during the summer. Then, cones from the current year's pollination/fertilization event emerged on each tree and seeds dispersed in the fall (age 1). Over winter, stage-structured density dependent mortality was implemented as a function of each stand's carrying capacity (K). Growth of all individuals and establishment of new mature individuals (age 20+) occurred by spring and the additional WPBR mortality on mature individuals was implemented at this time. More detailed methods with data sources used to parameterize the model are outlined below and in Appendix 4, Table A4.1.

Stands, Carrying Capacity, Age, and Size Classes

The WBP simulations were constrained to an extent in the US Northern Rockies that was delineated a priori by four zones (i.e., "seed zones"; Mahalovich and Hipkins, 2011; **Figure 1**). The extent contained 1059 initial spatially-delineated stand locations separated by at least 5 km. These WBP stands were designated by selecting all cells with >0.5 probability of WBP suitability, as predicted by the CNM described above (see Section Whitebark Pine Regeneration and White Pine Blister Rust Suitability Model and Appendix 2). For simplicity, we assumed a carrying capacity of 100 trees at each stand location.

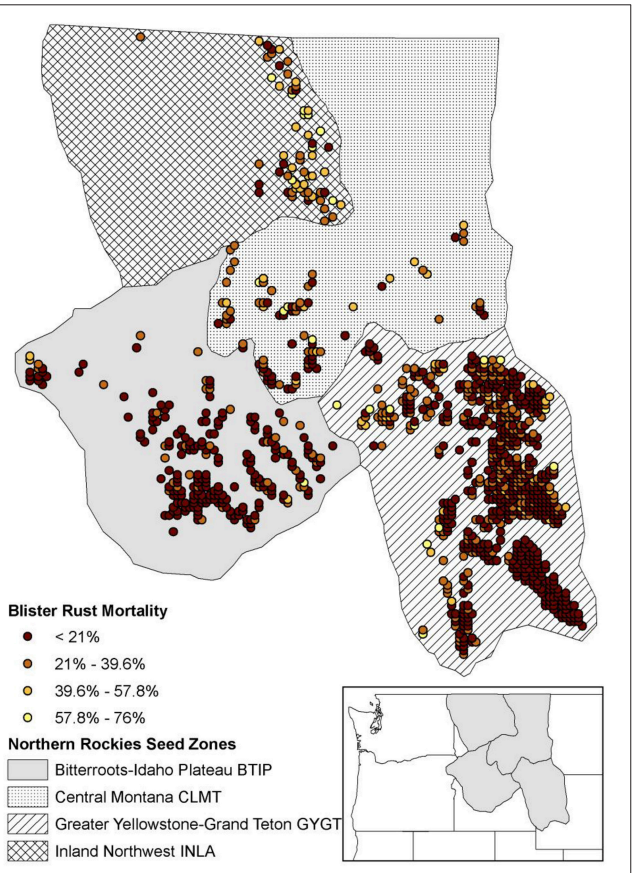


FIGURE 1 | Study area defined by the northern Rockies seed zones (Mahalovich and Hipkins, 2011) with initial 1059 simulated stand locations. WPBR relative spatial selection mortality shown for each stand.

We initialized the model at time = 0 with a random distribution of 500 age classes (Burns and Barbara, 1990). We ran the model without genetic exchange for an initial 25 years to allow the age distribution to stabilize, and then began genetic exchange (see next section). We defined age 0 "individuals" as fertilization events, which 12 months later emerged as age 1 cones producing seeds for dispersal. An annual increment of 0.2 cm diameter at breast height (DBH) (Keane et al., 2007) was used to grow each individual tree. As trees progressed through each size class, size-linked parameters (e.g., probability of mortality, probability of maturation, and fecundity) varied (Appendix 4).

Neutral and Adaptive Genetics

We initialized each individual's neutral genotypes with allele frequency files that match the frequency observed in each seed zone (Mahalovich and Hipkins, 2011), comprised of 16 loci with at most nine polymorphic alleles per locus. We did not consider mutation, which is reasonable considering the short simulation time period. In addition, we added a bi-allelic adaptive locus and assumed that only one gene confers resistance to WPBR (e.g., Kinloch et al., 1999; Lui et al., 2016). We initialized this selection-driven locus at time = 25 years with 0.01 and 0.99 frequency for the first and second allele, respectively. Any individual homozygous at the first allele (i.e., AA) in this selection-driven locus was assumed to have a selective advantage against blister rust infection.

This simple single-locus selection model was chosen because major gene resistance between the host species and pathogen has not been found in WBP (Bingham, 1983; Kinloch and Dupper, 2002), and much of our understanding of blister rust gene resistance comes from interior western white pine (*Pinus monticola*; Kinloch et al., 1999) and recently, Rocky Mountain white pine (*Pinus flexilis*; Lui et al., 2016). Thus, we assumed the blister rust resistance mechanisms acting in WBP are comparable to these species. Furthermore, the interior western white pine (*Pinus monticola*) blister rust screening program (Bingham, 1983; Mahalovich, 2010) serves as the basis for WBP blister rust screening trials (Bingham, 1983; McDonald and Hoff, 2001; Mahalovich et al., 2006). While there are other presumed single-gene recessive traits present in low frequency in blister rust screening trials (Mahalovich et al., 2006), the blister rust resistance trait chosen for modeling was the short shoot fungicidal reaction (Hoff and McDonald, 1971) due to the higher frequency of these genotypes in blister rust screening trials from 1999 to 2015 (Mahalovich, unpublished data). This resistance mechanism involves necrosis at the base of an infected needle fascicle bundle; thus, normal canker growth is halted and the branch and tree stem remains disease-free.

White Pine Blister Rust Resistance and Mortality

CDMetaPOP implements natural selection analogously to the adaptive-, or fitness-landscape of allele frequencies as originally envisioned by Wright (1932). This functionality enables extension of landscape genetic analyses to explicitly investigate the links between gene flow and selection in complex landscapes at the level of the individual (see Landguth et al., 2012a). We used WPBR occurrence (see Section Whitebark Pine Regeneration and White Pine Blister Rust Suitability Model) values at each stand as a proxy for differential mortality applied to mature trees only (e.g., WPBR occurrence of 0.5 would produce a 50% mortality at that stand; **Figure 1**). WPBR mortality rates in each stand were implemented based on the genotype of each individual and increased survival was associated with individuals that had *AA* in the selection-driven locus, which varied depending on the simulation scenario (see section Simulation Scenarios and Analysis). This allowed us to model evolution of WPBR resistance based on a single locus under selection with a single genotype being selected for.

Maturation and Fecundity

Mature individuals were defined as those of age 20 and greater. Although WBP may typically take longer to reach maturity when growing on poorer sites or at higher elevations (e.g., Krugman and Jenkinson, 1974; Mahalovich unpublished data), we used a lower bound of 20 years to allow for more generations in the model (Fire Effects Information System; http://www.fs.fed.us/database/feis/plants/tree/pinalb/all.html accessed September, 2015). We implemented a size-based fecundity model to determine the number of seeds produced at a given basal area per stand following the individual tree DBH conversion to basal area: Basal Area $= 0.00007854 * DBH^2$. To obtain a size-based seed production per individual tree, we used the value of 500 cones per 1 basal area (m^2/ha; Barringer et al., 2012) multiplied by 20 seeds per cone. Although cone and seed production varies spatially and temporally in our study area (Owens et al., 2008), no masting was considered and we assumed lower bound estimates (e.g., as low as 10 seeds per cone; Pigott, 2012) to reduce computational time.

Mortality

In order to isolate the effects of WPBR mortality, we only considered density-independent mortality based on class-based mortality probabilities. We applied a 99% probability of mortality to age 0 class to mimic 1% seed survival (DeMastus, 2013). We implemented a cumulative 35% probability of survival for age classes 1–15 (Izlar, 2002). Trees age 500 and older were assigned 25% probability of survival, which allowed for occasional long-lived trees (i.e., >500 years) given the length of the simulation time. If a stand reached K, then a random removal of excess individuals was conducted (e.g., Balloux, 2001).

Reproduction, Pollen Dispersal, and Wind Directionality

Reproduction within and across stands was monecious with selfing allowed. We considered two hypotheses for pollen movement in the summer months. Our first hypothesis assumed pollen moved according to a null model of isolation-by-distance: probability of pollen dispersal to a respective female cone locations was a function of the inverse-square Euclidean distance (Landguth and Cushman, 2010) with a 50% maximum study area distance threshold (450 km). Because pollen dispersal is governed by wind patterns, we also considered a second hypothesis that included directional movement with respect to prevailing wind direction (i.e., isolation-by-distance and wind). Thirty-year average (1979–2010) mean annual average wind direction was calculated from the North American Regional Reanalysis (NARR; Mesinger et al., 2006). Using the landscape connectivity program, UNICOR (Landguth et al., 2012b), we created asymmetrical costs for traversing with and against wind direction for all pairwise stand-to-stand locations. UNICOR creates a graph of a given resistance surface, which allows start and end node locations to find shortest paths on the resistance surface (i.e., Dykstra's algorithm). Given a wind direction map (and ignoring vector magnitude), a resultant vector was created in the 8-Moore neighborhood to weight direction in the graph creation. This produced an added cost resulting from the resultant vector calculation and when a path was traversing from a point and against wind direction, producing an asymmetrical cost distance matrix.

Cone/Seed Dispersal

Age 1 cones from the previous year were dispersed from individual trees (e.g., Clark's nutcracker, a bird which disperses and caches WBP seeds) following an isolation-by-distance movement pattern similar to pollen dispersal: probability of cone dispersal to a new stand location was a function of the inverse-square Euclidean distance with a 30 km maximum distance threshold (Lorenz et al., 2011). This produced the majority of cones staying in the same stand or nearest neighbor stands (i.e., dropping near parent tree) with occasional longer distance cone dispersal (e.g., Clark's Nutcracker). In addition to 1% seed survival (DeMastus, 2013), the ability for a seed to establish in a new stand location was determined based on resource availability (i.e., carrying capacity not exceeded in the destination stand).

Simulation Scenarios and Analysis

We conducted two blocks of simulation scenarios. The first block of simulations was used to help understand the added influence of WPBR mortality with and without an introduced gene that was resistant to WPBR. The second block of simulations was used to look at different spatial patterns for planting individuals with a resistance to WPBR. The spatially planting strategies we explored included planting in two regions (seed zones), as well as a broader distribution of planting across the entire extent outside of wilderness areas (**Figures 2A–C**). Each block compared pollen dispersal simulations for isolation-by-distance and directional pollen dispersal via wind. **Table 1** lists each block and respective scenario.

We ran simulations for 130 years, with the first 25 years considered "burn-in" for the population dynamics and age distributions to stabilize. We plotted mean population abundance, allelic diversity, and heterozygosity for all stands and for each block scenario. We used 10 replicate simulation runs

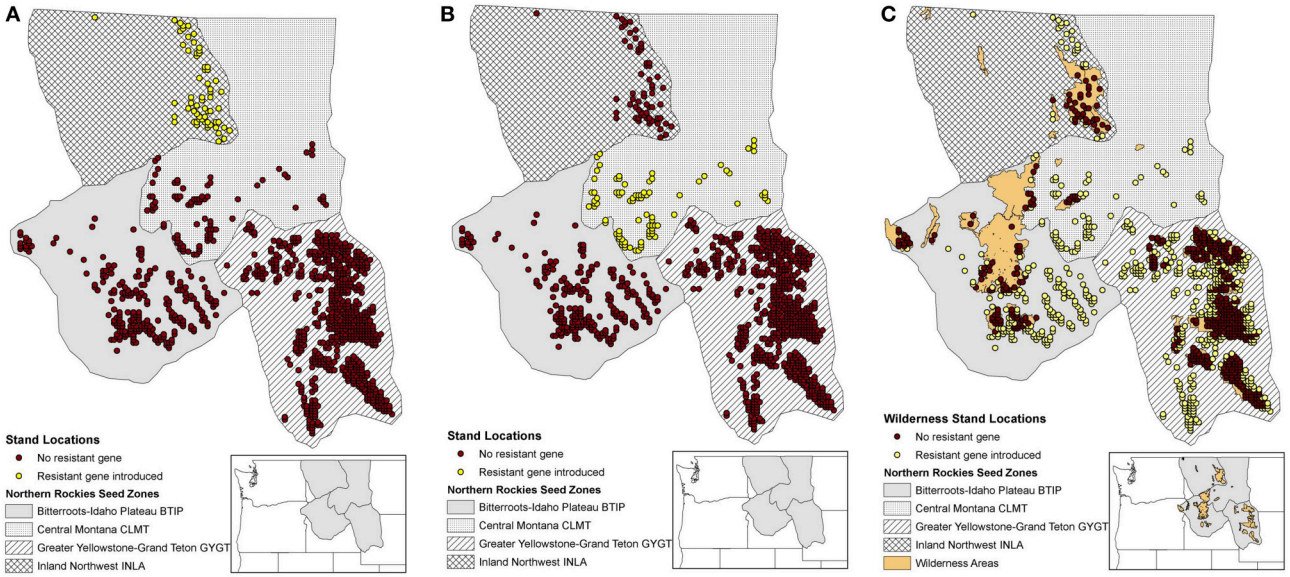


FIGURE 2 | Study area considered using the northern Rockies seed zones (Mahalovich and Hipkins, 2011) with initial 1059 stand locations. (A) WPBR resistant gene introduced in INLA zone stands (yellow dots). **(B)** WPBR resistant gene introduced in CLMT zone stands (yellow dots). **(C)** WPBR resistant gene introduced in stands (yellow dots) outside of wilderness areas (brown dots).

TABLE 1 | Simulation scenarios (WPBR–white pine blister rust).

Block Name	Scenario Name	Description
Block 1: WPBR mortality and resistance	No mortality	The null model in which no WPBR mortality considered
	All mortality	All stand locations applied the added WPBR mortality (**Figure 1**) regardless of genetic makeup.
	Resistant gene in all zones	All stand locations applied the added WPBR mortality (**Figure 1**). One genotype assumed to confer resistance to WPBR.
Block 2: WPBR resistance by planting strategy	Resistant gene in INLA zone	All stand locations applied the added WPBR mortality. One genotype assumed to confer resistance to WPBR only in the most northern zone (INLA; **Figure 2A**).
	Resistant gene in CLMT zone	All stand locations applied the added WPBR mortality. One genotype assumed to confer resistance to WPBR only in a central zone (CLMT; **Figure 2B**).
	Resistant gene in non-wilderness	All stand locations applied the added WPBR morality. One genotype assumed to confer resistance to WPBR only outside of wilderness areas (**Figure 2C**).

to assess variation in each metric. For a spatial representation of genetic differentiation, we calculated an overall pairwise genetic differentiation (G_{ST}) across all loci using the method of Nei (1973) and for each pair of zones at specified year $t = 100$.

RESULTS

Whitebark Pine and White Pine Blister Rust Maps

Results from the CNM for the presence or absence of juvenile WBP and WPBR within US Forest Service Northern Region are shown in **Figure 3**. See Appendix 2, 3 for supporting documentation on models. The distribution of juvenile WBP was reasonably well predicted by biophysical predictors, and presence or absences of juveniles was correctly classified at 92% of the forest inventory plots (Table A2.1). Mean maximum daytime temperature, followed by mean annual water balance deficit (unit of measure), were the strongest predictors in the WBP model. The model predicts that WBP occurs with highest probability at high elevation, cold sites with moderate to low water balance deficit. The distribution of WPBR was moderately well explained by climatic and biophysical predictors, with an overall classification accuracy of 81%.

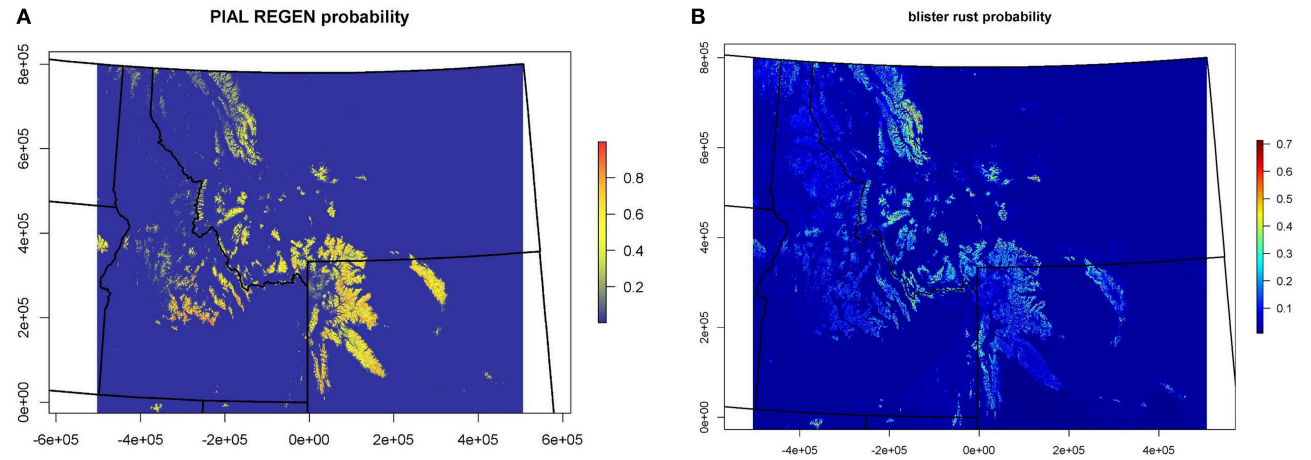


FIGURE 3 | Probability of occurrence maps for (A) whitebark pine and **(B)** white pine blister rust.

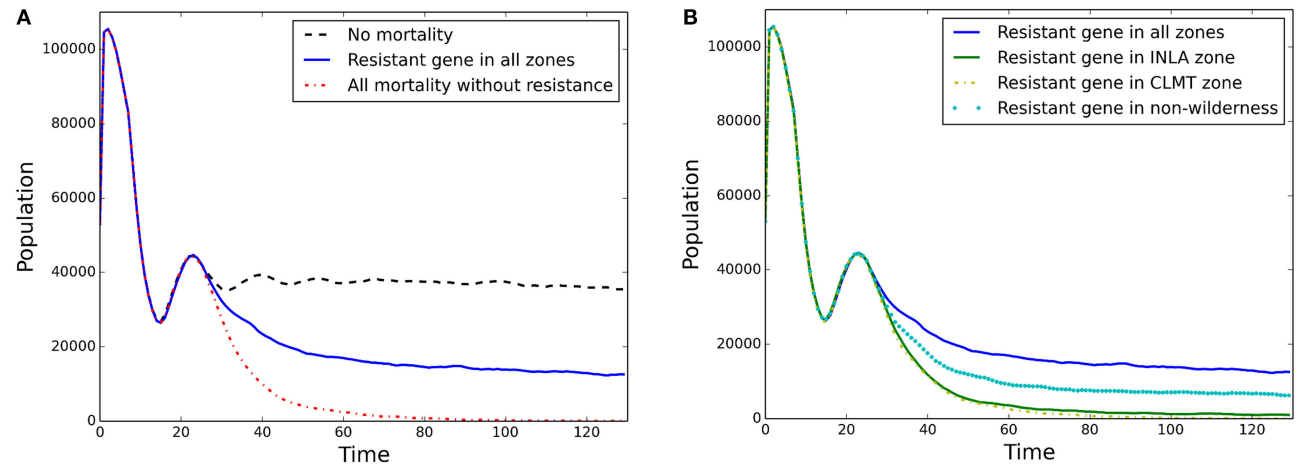


FIGURE 4 | Population abundance through time for each scenario in (A) Block 1 and **(B)** Block 2.

White Bark Pine Landscape Demogenetic Simulations

Overall population mean abundances (i.e., all stands) for each block of scenarios are shown in **Figure 4** for the simplest model of isolation-by-distance with no wind resistance included for pollen dispersal. Block 1 "No mortality" (**Figure 4A** black dashed line) shows stable population dynamics, while in the "All mortality" scenario the population declined smoothly to close to 0 by time 100 (**Figure 4A** red dash-dotted line). The introduction of a WPBR resistant gene for all individuals at every stand while still applying WPBR differential mortality led to stable population sizes of approximately 1/4th of the "No mortality" scenario (**Figure 4A** blue solid line).

Block 2 scenarios are shown in **Figure 4B**. Planting of individuals with resistance in the two different zones resulted in near extirpation of WBP (central CLMT zone; **Figure 4B** yellow dash-dotted line, and northern INLA zone; **Figure 4B** green line). **Figure 4B** also shows the scenario for the more widely distributed planting outside of Wilderness areas (**Figure 4B** cyan dotted line), which produced a stable population abundance at approximately $^1/_2$ of the "Resistant gene in all zones" (**Figure 4B** blue solid line). Similar results for mean stand growth rate are shown in Appendix 4 (Figures A4.2a,b).

Overall population mean allelic diversity is shown in **Figure 5** for the model of isolation-by-distance with no wind resistance included for pollen dispersal. The decline in allelic diversity revealed patterns similar to those of the population abundance graphs. The allelic diversity in the null model of no spatial differential mortality remained relatively constant at 0.39 (**Figure 5A** black dashed line). In the extreme scenario where WPBR was applied to every stand, allelic diversity steeply declined to 0.2 (**Figure 5A** red dashed-dotted line) and in the scenario in which WPBR resistant genotypes were planted in every stand, allelic diversity remained close to the null model

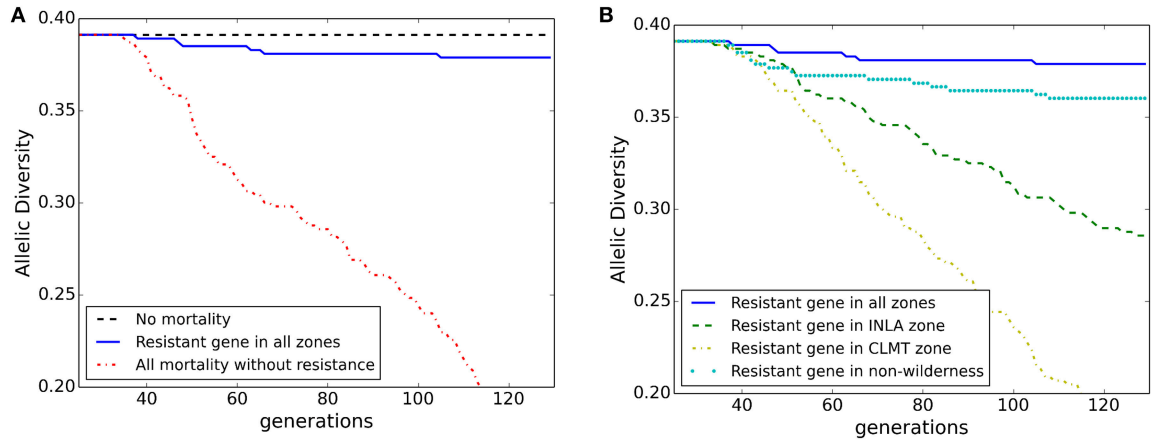


FIGURE 5 | Allelic diversity through time for each scenario in (A) Block 1 and **(B)** Block 2.

(0.38; **Figure 5A** blue solid line). However, there was a greater loss in allelic diversity with the central (CLMT) zone planting scenario (0.2; yellow dash-dotted line; **Figure 5B**) compared to the northern (INLA) zone planting scenario (0.3; green dashed line), despite equivalent population abundance, showing how genetic diversity may be more sensitive to spatial planting than overall abundance. Furthermore, planting of resistant WPBR individuals in a continuous distribution across the analysis extent produced higher allelic diversity numbers than the zone-specific planting (**Figure 5B** cyan dotted line). Similar results are shown for heterozygosity in Appendix 4 (Figures A4.3a,b).

Genetic differentiation for each zone is shown in **Figure 6** for time 100 for the model of isolation-by-distance with no wind resistance. The "No mortality" scenario (**Figure 6A**) shows little difference in genetic differentiation through time. However, as WPBR mortality is applied, genetic differentiation increases, with the largest differentiation in the "All mortality" scenario (**Figure 6C**). In fact, with the "All mortality" scenario, the CLMT zone becomes extirpated. The uniform introduction of a resistant WPBR gene produced patterns of genetic differentiation among zones similar to the "No mortality" scenario, with the exception of the INLA zone showing slightly higher differentiation (**Figure 6D**). The right panel in **Figures 6C–F** shows the Block 2 scenarios that varied spatial planting strategies for resistant genes to WPBR. Genetic differentiation increased under all planting strategies, with local extirpation occurring with the CLMT zone-specific scenario (**Figure 6E**). Genetic differentiation for non-Wilderness area planting of resistant genes only slightly increased (**Figure 6F**) from the null model of "No mortality" (**Figure 6A**).

When we included the effects of directional wind resistance on pollen dispersal we see an overall increase in genetic differentiation across all scenarios (**Figure 7**) with the exception of the "No mortality" scenario (**Figure 7A**), which remained at the same level of genetic differentiation as with the model of just isolation-by-distance. We also see more local extirpation, in particular in the scenario in which individuals with WPBR resistance are only planted in non-Wilderness areas (**Figure 7F**). These simulations show that incorporating more realistic effects of spatial processes, such as wind resistance, reduces pollen dispersal capability, thus reducing the ability of resistance genes to propagate through the landscape.

DISCUSSION

The goal of this paper is to provide an example of integrating species distribution modeling with landscape genetic simulation of neutral gene flow and adaptive evolution. Our specific focus was on exploring the effects of different levels of pathogen lethality and gene flow on the evolution of blister rust resistance in WBP and the effectiveness of several scenarios of planting rust resistant genotypes of WBP in different spatial configurations. This is the first simulation experiment to examine local and regional demogenetic patterns to the placement of resistant individuals, and the first to quantify differences in adaptive evolutionary processes as a function of directional and isotropic resistance to dispersal.

We first presented climate niche models for WBP and WPBR distributions. We used the climate niche models with a new eco-evolutionary landscape genetics model to simulate demogenetic responses with and without the presence of the disease agent, white pine blister rust. These models allowed us to produce baseline null models of a "healthy" disease-free system (e.g., **Figure 4A** black dashed line) with stable demographics and genetics and an extreme case of complete disease-ridden system (e.g., **Figure 4A** red dash-dotted line) with crashing demographics and genetics.

We then introduced individuals with a genotype that conferred resistance to WPBR and simulated potential planting strategies with this genotype in example zone-specific locations and across more broadly distributed areas across the study extent (i.e., outside of wilderness areas). This allowed us to quantify how much the introduction of disease resistant genotypes might mitigate the effects of WPBR and to evaluate this model systems sensitivity to the extent and pattern of introduction of disease

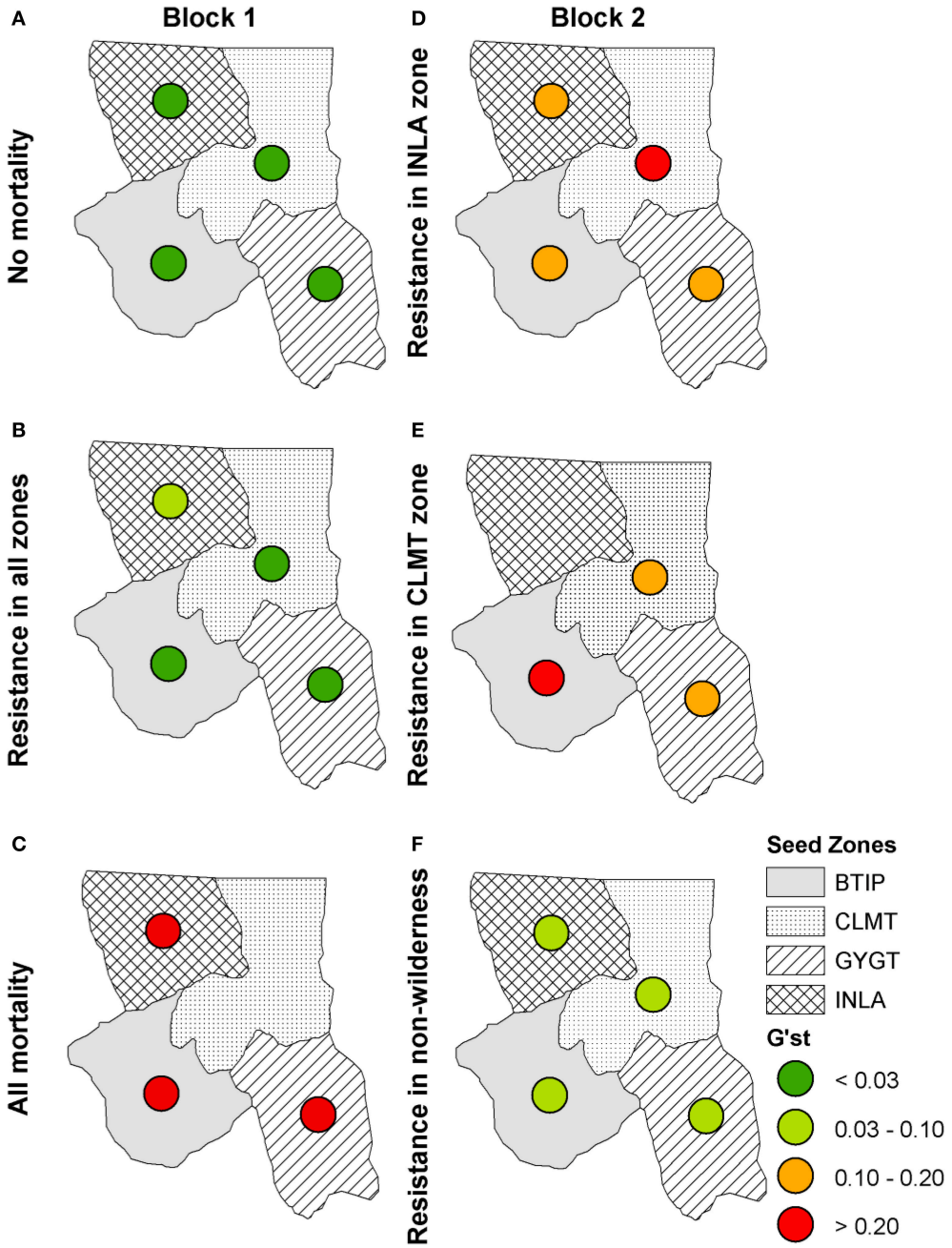


FIGURE 6 | Isolation-by-distance: G'_{ST} values for each seed zone at year 100 for Block 1 scenarios: (A) the null scenario "No mortality," **(B)** the scenario in which a resistant gene was introduced, **(C)** the scenario in which all stands receive WPBR mortality ("All mortality") and for Block 2 scenarios: **(D)** the scenario in which a resistant gene was introduced in only the INLA zone, **(E)** the scenario in which a resistant gene was introduced in the CLMT zone only, **(F)** the scenario in which a resistant gene was introduced outside of wilderness areas only.

resistant genotypes. Our results demonstrate that different patterns of planting resistant genotypes can influence genetic outcomes, and that genetic diversity and differentiation are more sensitive than population dynamics (**Figure 5B** compared to **Figure 4B**). Furthermore, planting of resistant WPBR individuals in a systematic distribution across the study area extent produced much higher allelic diversity numbers than more localized "clusters" (**Figure 5B**). A growing body of research has suggested that the loss of genetic diversity with increased disease may be a crucial mechanism driving population extinction risk (Whiteman et al., 2006). Thus, this finding could have additional important implications for management planning and

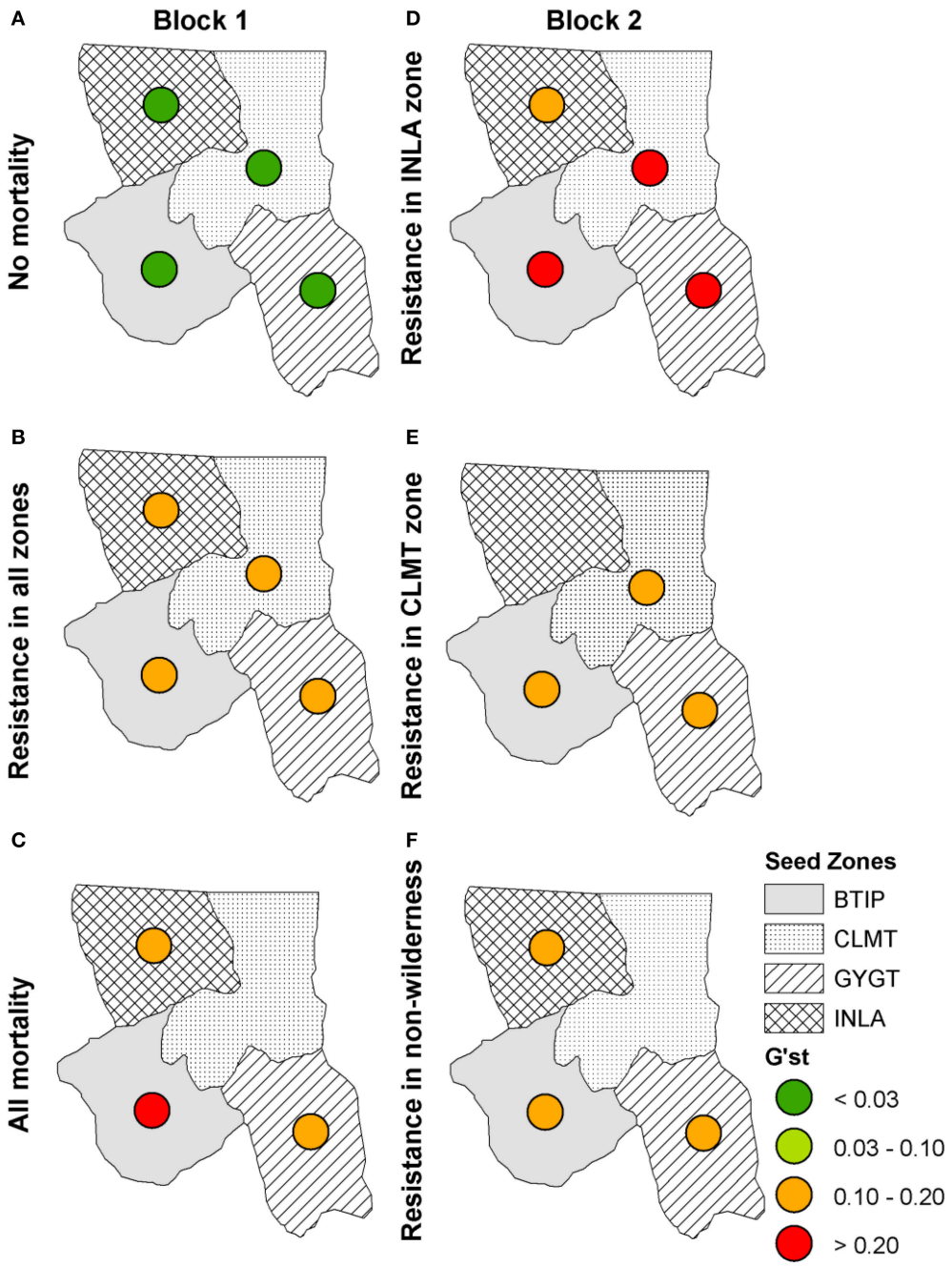


FIGURE 7 | Isolation-by-distance with directional wind resistance: G'_{ST} values for each seed zone at year 100 for Block 1 scenarios (left panel): (A) "No mortality," **(B)** "blister rust resistance in all zones," **(C)** "All mortality" and Block 2 scenarios (right panel): **(D)** "Resistance in INLA zone," **(E)** "Resistance in CLMT zone," **(F)** "Resistance in non-wilderness."

suggests that strategies should focus on implementing broad-scale, spatially continuous introductions rather than focusing on concentrating planting of disease resistant genotypes in particular nodal populations (e.g., Oyler-McCance et al., 2013). However, this also implies potential logistical limitations to effective management of WBPR through introduction of disease resistant genes across broad regions. Specifically, for rust resistance to spread in a local population it must be introduced with sufficiently high frequency to not be rapidly lost through drift before it can spread through selection. This is more easily achieved through concentrated introductions in patches or zones. However, our results show that broad-scale, continuous introductions are needed to effectively mitigate population and genetic effects of WPBR. It is not clear whether resources could

be sufficiently invested to implement such widely distributed planting at sufficient density to produce a lasting effect on the population.

Our results also show that large differences in predicted genetic differentiation are produced when models use simple isolation-by-distance assumptions as compared to when they implement more realistic spatial processes, such as isolation by resistance. Specifically, scenarios that incorporated the influence of wind resistance on the ability of pollen to disperse resistant genes through the landscape produced much higher rates of local extirpation along with higher genetic differentiation (**Figure 7**). These simulations showed that incorporating more realistic landscapes that control for movement, such as wind resistance, reduces pollen dispersal capability, thus reducing the ability of resistance genes to propagate through the landscape. This has important implications for spatial genomic and evolutionary modeling, most of which has to date utilized simple models of isolation-by-distance controlling gene flow (but see Forester et al., 2015). Our results show that it is essential to move this work into an explicitly landscape genomic framework in which gene flow is realistically driven by spatial patterns of landscape features that influence dispersal (such as wind fields in this case).

To produce reliable inferences about implications of adaptive variation, researchers must unambiguously determine whether markers for key adaptive traits, such as blister rust resistance, are under selection and identify the factors in the environment that drive that selection (Joost et al., 2013; Rellstab et al., 2015). This, however, remains a challenging task (see Vitalis et al., 2001; Luikart et al., 2003; Angeloni et al., 2011). For example, outlier detection methods will often detect signals of selection in markers that are not themselves under selection, but instead just linked to a gene that is (e.g., Jones et al., 2014). Moreover, when numerous regions of the genome are under divergent selection, outlier analyses can miss many regions that clearly are under selection (Michel et al., 2010). Further complications arise for the ability to detect adaptive loci when landscape configuration, dispersal ability, and selection strength intertwine (Forester et al., 2015), as well as the effects of sampling through design, replication, and resolution of markers (e.g., number of SNPs) (e.g., De Mita et al., 2013; Lotterhos and Whitlock, 2015). Developing methods for reliably identifying markers under selection is a major ongoing theme in landscape genomics research. Common garden experiments with reciprocal transplant of genotypes is a robust way to assess environmental selection (e.g., Whitham et al., 2006; Cushman, 2014) and can be readily extended to evaluate the interactions between environmental selection and pathogen resistance. For the simulation framework identified here to be truly useful to understand the potential of genetically mediated blister rust resistance to mitigate impacts on WBP populations, it will be important to identify the genetic mechanisms controlling resistance and how they may be linked to selection on other factors, such as drought and cold tolerance.

There are several lines of addition future work which should be explored to extend the scope of what we have presented here on the spatial dynamics of adaptation to the blister rust pathogen and the potential effectiveness of different strategies of planting resistant genotypes. First, this paper used a simple one-locus model of genotype-environment association. While this is a model that is widely used in theoretical evolutionary ecology (Coyne and Orr, 2004) and genotype-environment association testing (e.g., Jones et al., 2014; Forester et al., 2015) and applies to some proposed blister rust mechanisms (Kinloch et al., 1999), the majority of micro-evolutionary processes are likely mediated through polygenetic selection in which many loci each contribute relatively small fitness effects. This paper serves as an initial analysis of a simple classical model of one locus selection which provides insight. However, future modeling work should explore how the blister rust distribution and planting of resistant genotypes interacts within the context of multiple loci/allele selection, pleiotropy, and epistasis.

In addition, while this paper is the first to combine empirical data, experimentation, and large-scale population-wide simulation modeling, WBP and WPBR are complex systems that are still imperfectly understood and simulation models are a simplified representation of reality. Future studies should invest in improving how WBP and WPBR biology are represented in simulations (e.g., more realistic growth models or disease spread dynamics) and assess sensitivity and uncertainty in these systems. For example, simulations could explore the effects of habitat quality and density-dependent processes (Pfluger and Balkenhol, 2014) on the interaction between rust resistance and white bark pine population dynamics. In addition, simulation experiments, such as presented here can describe the processes affecting population and identify the conditions under which they have important influences. However, models without data are not compelling. It is essential to confront these models with empirical data on the actual patterns of genetic differentiation in complex landscapes, and to confirm the fitness relationships underlying these patterns in experimental studies, such as common gardens (Cushman, 2014).

AUTHOR CONTRIBUTIONS

EL, ZH, MM, and SC designed research. EL ran simulations. EL and ZH performed the analyses. EL, ZH, MM, and SC interpreted the data and wrote the manuscript.

ACKNOWLEDGMENTS

Kay Izlar, Greg DeNitto, Blakey Lockman and Brytten Steed provided valuable feedback during the development of this project. This research was supported in part by funds provided by the Forest Health and Protection Group, Region 1, Forest Service, U.S. Department of Agriculture, Seattle City Light, and NASA grant NNX14AC91G.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2017.00009/full#supplementary-material

REFERENCES

Angeloni, F., Wagemaker, N., Vergeer, P., and Ouborg, J. (2011). Genomic toolboxes for conservation biologists. *Evol. Appl.* 5, 130–143. doi: 10.1111/j.1752-4571.2011.00217.x

Balloux, F. (2001). EASYPOP (version 1.7): a computer program for population genetic simulations. *J. Hered.* 92, 301–302. doi: 10.1093/jhered/92.3.301

Barringer, L. E., Tomback, D. F., Wunder, M. B., and McKinney, S. T. (2012). Whitebark pine stand condition, tree abundance, and cone production as predictors of visitation by Clark's Nutcracker. *PLoS ONE.* 7:e37663. doi: 10.1371/journal.pone.0037663

Bell, D. M., Bradford, J., and Lauenroth, W. K. (2014). Early indicators of change: divergent climate envelopes between tree life stages imply range shifts in western United States. *Glob. Ecol. Biogeogr.* 23, 168–180. doi: 10.1111/geb.12109

Bingham, R. T. (1972). "Taxonomy, crossability, and relative blister rust resistance of 5-needled pines," in *Biology of Rust Resistance in Forest Trees* (Washington, DC: USDA Forest Service Miscellaneous Publication), 271–278.

Bingham, R. T. (1983). *Blister Rust Resistant Western White Pine for the Inland Empire: The Story of the First 25 Years of the Research and Development Program*. General Technical Report INT-146, Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station, 45.

Brar, S., Tsui, C. K., Dhillon, B., Bergeron, M. J., Joly, D. L., Zambino, P. J., et al. (2015). Colonization history, host distribution, anthropogenic influence and landscape features shape populations of white pine blister rust, an invasive alien tree pathogen. *PLoS ONE* 10:e0127916. doi: 10.1371/journal.pone.0127916

Burns, R. M., and Barbara, H. H. (1990). *Silvics of North America: 1. Conifers; 2. Hardwoods*. Washington, DC: US Department of Agriculture, Forest Service.

Coyne, J. A., and Orr, H. A. (2004). *Speciation.* Sunderland, MA: Sinauer Associates.

Cushman, S. A. (2014). Grand Challenges in evolutionary and population genetics: the importance of integrating epigenetics, genomics, modeling, and experimentation. *Front. Genet.* 5:197. doi: 10.3389/fgene.2014.00197

DeMastus, C. R. (2013). *Effective Methods of Regenerating Whitebark Pine (Pinus albicaulis) Through Direct Seeding.* Doctoral dissertation, Montana State University-Bozeman, College of Letters & Science.

De Mita, S., Thuillet, A.-C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J. et al. (2013). Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Mol. Ecol.* 22, 1383–1399. doi: 10.1111/mec.12182

Dobrowski, S. Z., Swanson, A. K., Abatzoglou, J. T., Holden, Z. A., Safford, H. D., Schwartz, M. K., et al. (2015). Forest structure and species traits mediate projected recruitment declines in western US tree species. *Glob. Ecol. Biogeogr.* 24, 917–927. doi: 10.1111/geb.12302

Elith, J., and Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.* 40, 677–697. doi: 10.1146/annurev.ecolsys.110308.120159

Forester, B. R., Jones, M. R., Joost, S., Landguth, E. L., and Lasky, J. R. (2015). Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. *Mol. Ecol.* 25, 104–120. doi: 10.1111/mec.13476

Frank, B. M., Piccolo, J. J., and Baret, P. V. (2011). A review of ecological models for brown trout: towards a new demogenetic model. *Ecol. Freshw. Fish* 20, 167–198. doi: 10.1111/j.1600-0633.2011.00491.x

Greater Yellowstone Coordinating Committee whitebark pine Subcommittee (2011). *Whitebark Pine Strategy for the Greater Yellowstone Area*, 41

Hoff, R., Bingham, R. T., and McDonald, G. I. (1980). Relative blister rust resistance of white pines. *Forest Pathol.* 10, 307–316. doi: 10.1111/j.1439-0329.1980.tb00042.x

Hoff, R. J., and McDonald, G. I. (1971). Resistance to *Cronartium ribicola* in *Pinus monticola*: short shoot fungicidal reaction. *Can. J. Bot.* 49, 1235–1239. doi: 10.1139/b71-172

Izlar, D. K. (2002). *Assessment of Whitebark Pine Seedling Survival for Rocky Mountain Plantings*. M.S. Thesis Humboldt State Univeristy.

Jones, M. R., Forester, B. R., Teufel, A. I., Adams, R. V., Anstett, D. N., Goodrich, B. A., et al. (2014). Integrating spatially-explicit approaches to detect adaptive loci in a landscape genomics context. *Evolution* 67, 3455–3468. doi: 10.1111/evo.12237

Joost, S., Vuilleumier, S., Jensen, J. D., Schoville, S., Leempoel, K., Stucki, S., et al. (2013). Uncovering the genetic basis of adaptive change: on the intersection of landscape genomics and theoretical population genetics. *Mol. Ecol.* 22, 3659–3665. doi: 10.1111/mec.12352

Keane, R. E., Holsinger, L. M., Mahalovich, M. F., and Tomback, D. F. (2016). Evaluating future success of whitebark pine ecosystem restoration under climate change using simulation modeling. *Restor. Ecol.* doi: 10.1111/rec.12419. [Epub ahead of print].

Keane, R. E., Tomback, D. F., Aubry, C. A., Bower, A. D., Campbell, E. M., Cripps, C. L., et al. (2012). *A Range-Wide Restoration Strategy for Whitebark Pine (Pinus albicaulis)*, General Technical Report RMRS-GTR-279. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 108.

Keane, R., Gray, K., and Dickinson, L. (2007). *Whitebark pine Diameter Growth Response to Removal of Competition*. RMRS General Technical Report Research note RN-32.

Kinloch, B. B., and Dupper, G. E. (2002). Genetic specificity in the white pine-blister rust pathosystem. *Phytopathology* 92, 278–280. doi: 10.1094/PHYTO.2002.92.3.278

Kinloch, B. B., Sniezko, R. A., Barnes, G. D., and Greathouse, T. E. (1999). A major gene for resistance to white pine blister rust in western white pine from the western Cascade range. *Genet. Resist.* 89, 861–867. doi: 10.1094/phyto.1999.89.10.861

Krugman, S. L., and Jenkinson, J. L. (1974) "Pinus L. Pine," in *Seeds of Woody Plants in the United States*, ed C. S. Schopmeyer (Washington, DC: US Department of Agriculture, Agriculture Handbook 450), 598–638.

Landguth, E. L., Bearlin, A., Day, C., and Dunham, J. (2016). CDMetaPOP: an individual-based, eco-evolutionary model for spatially explicit simulation of landscape demogenetics. *Methods Ecol. Evol.* 8, 4–11. doi: 10.1111/2041-210X.12608

Landguth, E. L., and Cushman, S. A. (2010). CDPOP: a spatially explicit cost distance population genetics program. *Mol. Ecol. Resour.* 10, 156–161. doi: 10.1111/j.1755-0998.2009.02719.x

Landguth, E. L., Cushman, S. A., and Balkenhol, N. (2015). "Simulation modeling in landscape genetics," in *Landscape Genetics,* Vol. 6, eds N. Balkenhol, L. Waits and S. Cushman (London: Wiley). 99–116.

Landguth, E. L., Cushman, S. A., and Johnson, N. A. (2012a). Simulating natural selection in landscape genetics. *Mol. Ecol. Resour.* 12, 363–368. doi: 10.1111/j.1755-0998.2011.03075.x

Landguth, E. L., Hand, B. K., Glassy, J. M., Cushman, S. A., and Sawaya, M. (2012b). UNICOR: a species corridor and connectivity network simulator. *Ecography* 12, 9–14. doi: 10.1111/j.1600-0587.2011.07149.x

Lenoir, J., Gegout, J.-C., Pierrat, J.-C., Bontemps, J.-D., and Dhote, J.-F. (2009). Differences between tree species seedling and adult altitudinal distribution in mountain forests during the recent warm period (1986-2006). *Ecography* 32, 765–777. doi: 10.1111/j.1600-0587.2009.05791.x

Lorenz, T. J., Aubry, C., and Shoal, R. (2008). *A Review of the Literature on Seed Fate in Whitebark Pine and the Life History Traits of Clark's Nutcracker and Pine Squirrels*. General Technical Report PNW-GTR-742, USDA Forest Service, Pacific Northwest Research Station, Portland, OR.

Lorenz, T. J., Sullivan, K. A., Bakian, A. V., and Aubry, C. A. (2011). Cache-site selection in Clark's nutcracker (*Nucifraga columbiana*). *Auk* 128, 237—247. doi: 10.1525/auk.2011.10101

Lotterhos, K. E., and Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Mol. Ecol.* 24, 1031–1046. doi: 10.1111/mec.13100

Lui, J.-J., Schoettle, A. W., Sniezko, R. A., Sturrock, R. N., Zamany, A., Williams, H., et al. (2016). Genetic mapping of *Pinus flexilis* major gene (Cr4) for resistance to white pine blister rust using transciptome-based SNP genotyping. *BMC Genomics* 17:753. doi: 10.1186/s12864-016-3079-2

Luikart, G., England, P. R., Tallman, D., Jordan, S., and Taberlet, P. (2003). The power and promise of populations genomics: from genotyping to genome typing. *Nat. Rev. Genet.* 4, 1811–1832 doi: 10.1038/nrg1226

Mahalovich, M. F. (2010). *U.S.A. Inland Northwest Western White Pine Breeding and Restoration Program: History, Current and Future Directions*. Available online at: http://dnrc.mt.gov/divisions/forestry/docs/assistance/pests/miscellaneous-publications/mahalovich-2010.pdf/at_download/file

Mahalovich, M. F., Burr, K. E., and Foushee, D. L. (2006). "Whitebark pine germination, rust resistance and cold hardiness among seed sources in the Inland Northwest: Planting Strategies for Restoration," in *National Proceedings: Forest and Conservation Nursery Association; 2005 July 18-20*, (Park City, UT; Fort Collins, CO: Proceeding RMRS-P-43; US Department of Agriculture, Forest Service, Rocky Mountain Research Station), 91–101.

Mahalovich, M. F., and Dickerson, G. A. (2004). "Whitebark pine genetic restoration program for the Intermountain West (United States), in *Proceeding IUFRO Working Party 2.02.15 Breeding and Genetic Resources of Five-Needle Pines: Growth, Adaptability, and Pest Resistance, 23–27, July 2001*. (Medford, OR; Fort Collins, CO: USDA Forest Service, Rocky Mountain Research Station; Proceedings RMRS-P-32, 181-187).

Mahalovich, M. F., and Hipkins, V. D. (2011). "Molecular genetic variation in whitebark pine (*Pinus albicaulis Engelm.*) in the Inland West," in *High-Five Symposium: The Future of High-Elevation Five-Needle White Pines in Western North America. 2010 June 28–30*, ed R. E. Keane (Missoula, MT; For Collins, CO: Proceedings RMRS-P-63; USDA Forest Service, Rocky Mountain Research Station), 124–139.

Manel, S., and Holderegger, R. (2013). Ten years of landscape genetics. *Trends Ecol. Evol.* 28, 614–621. doi: 10.1016/j.tree.2013.05.012

McDonald, G. I., and Hoff, R. J. (2001). *Whitebark Pine Communities: Ecology and Restoration* (eds D. F. Tomback, S. F. Arno, and R. E. Keane), Washington, DC: Island Press, 193–220.

Mesinger, F., DiMego, G., Kalnay, E., and Mitchell, K. (2006). North american regional reanalysis. *Bull. Am. Meteorol. Soc.* 87, 343–360. doi: 10.1175/BAMS-87-3-343

Michel, A. P., Sim, S., Powell, T. H. Q., Taylor, M. S., Nosil, P., and Feder, J. L. (2010). Widespread genomic divergence during sympatric speciation. *Proc. Natl. Acad. Sci. U.S.A.* 107, 9724–9729 doi: 10.1073/pnas.1000939107

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U.S.A.* 70, 3321–3323. doi: 10.1073/pnas.70.12.3321

Owens, J. N., Kittirat, T., and Mahalovich, M. F. (2008). Whitebark pine (*Pinus albicaulis Engelm.*) seed production in natural stands. *For. Ecol. Manage.* 255, 803–809. doi: 10.1016/j.foreco.2007.09.067

Oyler-McCance, S. J., Fedy, B. C., and Landguth, E. L. (2013). Sample design effects in landscape genetics. *Conserv. Genet.* 14, 275–285. doi: 10.1007/s10592-012-0415-1

Pfluger, F. J., and Balkenhol, N. (2014). A plea for simultaneously considering matrix quality and local environmental conditions when analyzing landscape impacts on effective dispersal. *Mol. Ecol.* 23, 2146–2156. doi: 10.1111/mec.12712

Pigott, D. (2012). *Whitebark pine in British Columbia. Forest Genetics Council of British Columbia*. Available onile at: http://www.fgcouncil.bc.ca/Factsheet1-WhiteBarkPine_2011.pdf

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Mol. Ecol.* 24, 4348–4370. doi: 10.1111/mec.13322

Scribner, K., Lowe, W., Landguth, E. L., Luikart, G., Infante, D. M., Whelan, G., et al. (2016). Applications of genetic data to improve management and conservation of river fishes and their habitats. *Fisheries* 41, 174–188. doi: 10.1080/03632415.2016.1150838

Shryock, D. F., Havrilla, C., Defalco, L. A., and Wood, T. E. (2015). Landscape genomics of *Sphaeralcea ambigua* in the Mojave Desert: a multivariate, spatially-explicit approach to guide ecological resortation. *Conserv. Genet.* 16, 1303–1317. doi: 10.1007/s10592-015-0741-1

Thuiller, W., Lavorel, S., Araújo, M. B., Sykes, M. T., and Prentice, I. C. (2005). Climate change threats to plant diversity in Europe. *Proc. Natl. Acad. Sci. U. S. A.* 102, 8245–8250. doi: 10.1073/pnas.0409902102

Vitalis, R., Dawson, K., and Boursot, P. (2001). Interpretation of variation across marker loci as evidence of selection. *Genetics* 158, 1811–1823. Available online at: http://www.genetics.org/content/158/4/1811

Whiteman, N. K., Matson, K. D., Bollmer, J. L., and Parker, P. G. (2006). Disease ecology in the Galapagos Hawk (Buteo galapagoensis): host genetic diversity, parasite load and natural antibodies. *Proc. R. Soc.* 273, 797–804. doi: 10.1098/rspb.2005.3396

Whitham, T. G., Bailey, J. K., Schweitzer, J. A., Shuster, S. M., Bangert, R. K., and LeRoy, C. J. (2006). A framework for community and ecosystem genetics: from genes to ecosystems. *Nat. Rev. Genet.* 7, 510–523 doi: 10.1038/nrg1877

Wright, S. (1932). "The roles of mutation, inbreeding, crossbreeding and selection in evolution," *Proceedings of the Sixth International Congress of Genetics,* 1, 356–366.

Zhu, K., Woodall, C. W., and Clark, J. S. (2011). Failure to migrate: lack of tree range expansion in response to climate change. *Glob. Chang Biol.* 18, 1042–1052 doi: 10.1111/j.1365-2486.2011.02571.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 29 September 2017
doi: 10.3389/fgene.2017.00139

The Empirical Distribution of Singletons for Geographic Samples of DNA Sequences

Philippe Cubry[1], *Yves Vigouroux*[1] *and Olivier François*[2]*

[1] UMR DIADE, University of Montpellier, Montpellier, France, [2] TIMC-IMAG UMR 5525, Centre National de la Recherche Scientifique (CNRS), Université Grenoble-Alpes, Grenoble, France

Rare variants are important for drawing inference about past demographic events in a species history. A singleton is a rare variant for which genetic variation is carried by a unique chromosome in a sample. How singletons are distributed across geographic space provides a local measure of genetic diversity that can be measured at the individual level. Here, we define the empirical distribution of singletons in a sample of chromosomes as the proportion of the total number of singletons that each chromosome carries, and we present a theoretical background for studying this distribution. Next, we use computer simulations to evaluate the potential for the empirical distribution of singletons to provide a description of genetic diversity across geographic space. In a Bayesian framework, we show that the empirical distribution of singletons leads to accurate estimates of the geographic origin of range expansions. We apply the Bayesian approach to estimating the origin of the cultivated plant species *Pennisetum glaucum [L.] R. Br.* (pearl millet) in Africa, and find support for range expansion having started from Northern Mali. Overall, we report that the empirical distribution of singletons is a useful measure to analyze results of sequencing projects based on large scale sampling of individuals across geographic space.

Keywords: genetic diversity, singletons, geographic origin, range expansion, pearl millet

OPEN ACCESS

Edited by:
Samuel A. Cushman, United States Forest Service Rocky Mountain Research Station, United States

Reviewed by:
Pablo Orozco-terWengel, Cardiff University, United Kingdom
Ricardo T. Pereyra, University of Gothenburg, Sweden
Rita Rasteiro, University of Bristol, United Kingdom

***Correspondence:**
Olivier François
olivier.francois@univ-grenoble-alpes.fr

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *27 March 2017*
Accepted: *14 September 2017*
Published: *29 September 2017*

Citation:
Cubry P, Vigouroux Y and François O (2017) The Empirical Distribution of Singletons for Geographic Samples of DNA Sequences. Front. Genet. 8:139. doi: 10.3389/fgene.2017.00139

1. INTRODUCTION

High-throughput sequencing technologies have enabled studies of genomic diversity in model and non-model species at a dramatically increasing rate. Conducted at population and at individual levels, those studies have provided comprehensive surveys of common and rare variation in model species genomes (Weigel and Mott, 2009; 1000 Genomes Project Consortium et al., 2010; International HapMap 3 Consortium, 2010; 1000 Genomes Project Consortium, 2015). For example, the 1000 Genomes Project Consortium (2015) reported that the majority of variants in human genomes are rare. During the last decade, the role that rare variants play in shaping complex traits has been hotly debated (Pritchard, 2001; Schork et al., 2009; Tennessen et al., 2012), and accurately determining their distribution has become important for medical applications and association studies (Lee et al., 2014; Auer and Lettre, 2015). Beyond humans, rare variation has attracted considerable interest from genome sequencing projects for model organisms, including plants (Zhu et al., 2011; Weigel, 2012; Memon et al., 2016).

Rare variants are also important for drawing inference about past demographic events in a species history (Schraiber and Akey, 2015). Studies of human populations have shown that our

species has experienced a complex demographic history, and that a recent period of explosive growth has resulted in an excess of those variants (Coventry et al., 2010; Keinan and Clark, 2012). The analysis of private and rare variation has been used to reveal signals of differential demographic history among populations, and to refine models of human evolution (Marth et al., 2004; Gravel et al., 2011; Mathieson and McVean, 2014). In addition, estimating rare allele frequencies has enabled estimates of gene flow between populations, and has facilitated inference of fine-scale population structure (Slatkin, 1985; Novembre and Slatkin, 2009; O'Connor et al., 2015).

In this study, we define the empirical distribution of singletons in a sample of chromosomes as the proportion of the total number of singletons that each chromosome carries, where a singleton is a uniquely represented allele in the sample (Fu and Li, 1993). We provide theoretical and empirical analyses of the distribution of singletons in a sample of chromosomes, and we evaluate the potential for this distribution to provide an accurate description of genetic diversity at the individual level. Using spatial data, we use the distribution of singletons as an individual-based estimate of genetic diversity in geographic space.

The theoretical background for the analysis of the empirical distribution of singletons rely on the distribution of external branch lengths for coalescent genealogies (Blum and François, 2005; Caliebe et al., 2007). First, we use coalescent and spatially explicit simulations to evaluate individual contributions to genetic diversity in the sample based on singletons. Then we evaluate the use of the distribution of singletons in an approximate Bayesian Computation (ABC) framework to estimate the geographic origin of range expansions (Beaumont, 2010; Csilléry et al., 2010). We eventually provide an illustration of our theory by applying the ABC approach to the plant species *Pennisetum glaucum [L.] R. Br.* (pearl millet). Pearl millet is a cereal cultivated in semi-arid regions of Africa and the Indian subcontinent, and it is known to originate in Africa (Clotault et al., 2012). We evaluate the geographic origin of its range expansion by using 146 inbred lines from the whole African range.

2. THEORY

We consider a sample of n chromosomes from a population of N haploid organisms. We assume that there are L polymorphic loci, and that for each locus, 0 represents the ancestral or reference allele and 1 is the derived allele. A singleton is defined as a derived allele carried by a single chromosome in the sample. The total number of singletons, ξ_1, is the number of uniquely represented derived alleles in the sample, and it corresponds to the first component of the site frequency spectrum. We assume that the singletons are distributed over the n chromosomes in the sample. More specifically, the number of singletons decomposes as follows

$$\xi_1 = \sum_{i=1}^{n} \xi_1^{(i)},$$

where $\xi_1^{(i)}$ is the number of singletons carried by chromosome i. For each i, we denote by p_i the conditional probability that a singleton is carried by i. The n values $p_1, \ldots, p_n$ sum up to one, and those values define the *empirical distribution of singletons* in the sample (see below).

Next, we assume that the sample genealogies can be described by coalescent trees (Tavaré, 2004). For a particular locus, a tree is described by n tips and $n-1$ ancestral nodes. An external branch of the tree connects a tip to an ancestral node. For a given tree, we denote by $\tau^{(i)}$ the length of the external branch connecting chromosome i to its first ancestor node. The L coalescent trees exhibit complex patterns of statistical dependency along the chromosomes due to recombination among loci (Hudson, 1990). Measuring lengths in units of twice the total population size (N), and assuming a molecular clock model for mutations, the number of mutations falling on a particular branch of the tree has a Poisson distribution of rate $\theta/2$, where $\theta = 2\mu N$ and μ is the per generation mutation rate (Tavaré, 2004). Let ℓ be an arbitrary singleton locus. For all i, we write

$$\xi_1^{(i)} = \sum_{k=1}^{\xi_1} X_{i\ell},$$

where $X_{i\ell} = 1$ if singleton ℓ is carried by chromosome i, 0 otherwise. In the above formula, the summation runs over all singletons in the sample. Using mathematical properties of conditional distributions for the Poisson process, we have

$$p_i = \mathrm{P}(X_{i\ell} = 1) = \mathrm{E}\left[\frac{\tau_1^{(i)}}{\tau_1}\right],$$

where $\tau_1 = \sum_{i=1}^{n} \tau^{(i)}$. In this formula, the conditional probability that chromosome i carries a singleton at locus ℓ is given by the ratio of its external branch length to the total length of external branches in the sample genealogy at this locus. The distribution of singletons can be estimated by counting the number of singletons carried by each chromosome and normalizing as follows

$$\hat{p}_i = \xi_1^{(i)}/\xi_1, \quad i = 1, \ldots, n,$$

and the estimate is unbiased

$$\mathrm{E}\left[\hat{p}_i\right] = p_i\,.$$

In addition, the number of singletons carried by chromosome i, $\xi_1^{(i)}$, estimates the proportion of genetic diversity carried by chromosome i

$$\mathrm{E}[\xi_1^{(i)}] \approx \theta p_i, \quad i = 1, \ldots, n.$$

As a consequence of the theory presented in this section, the individual-based estimates of genetic diversity are unbiased

quantities regardless of demographic history, deviations from Hardy-Weinberg equilibrium and linkage disequilibrium. Limitations of the theory include the presence of closely related individuals, which should be removed from the sample prior to analysis. The approach is appropriate for modern sequencing data as soon as a few hundreds of DNA sequences are generated.

The rest of this study will evaluate the use of the empirical distribution of singletons in mapping genetic diversity in geographic space. To provide an elementary example, let us consider a sample of n chromosomes from a random mating population of size N. Using mathematical results for the neutral coalescent in a random mating population, the expected value of the number of singletons is an unbiased estimator of the genetic diversity in the sample (Fu and Li, 1993)

$$\mathrm{E}[\xi_1] = \theta.$$

For the lengths of external branch lengths, we have

$$\mathrm{E}[\tau^{(i)}] = 2/n, \quad i = 1, \ldots, n,$$

and $\mathrm{E}[\tau_1] = 2$ (Blum and François, 2005). Here, we expect that each chromosome contributes to genetic diversity equally. The above calculations show that, in a sample of size n from a random mating population, the distribution of singletons is uniform over the n chromosomes

$$p_i = 1/n, \quad i = 1, \ldots, n,$$

and we have

$$\mathrm{E}[\xi_1^{(i)}] = \mathrm{E}[\xi_1]\mathrm{P}(X_{ik} = 1) = \theta/n.$$

In other words, each individual contributes the same amount of genetic variation to the total sample diversity.

3. SIMULATION METHODS AND DATA SETS

3.1. Coalescent Simulations of Splitting Populations

We used the computer program *ms* to perform coalescent simulations for a two-population model (Hudson, 2002). In our simulations, we considered a population split model, in which two populations of sizes $N_1 = 50,000$ and $N_2 = sN_1$ ($s \in (0.01; 0.5)$, shrink rate) diverged t generations ago ($t \in (1,000; 10,000)$, split time). Population 1 expanded from an ancestral population of size $N_A = 5,000$, and the expansion started 10,000 generations ago. Samples of size $n = 100$ were considered and subdivided into subsamples of size 50 from each population. We simulated $L = 1,000$ unlinked haplotypes using the infinite-site model and an effective mutation rate $\theta \in (5; 10)$. The *ms* command line was written as follows: ./ms 100 1,000 -t theta -I 2 50 50 -g 1 46.05 -n 2 shrink.rate -eg 0.2 1 0.0 -ej split.time 2 1. The simulated data sets were processed by using the ".geno" format in the `R` package LEA (Frichot and François, 2015). We summarized the distribution of singletons by computing mean values and standard errors for each subsample. For all simulated samples, we used the `R` package *ape* to extract the coalescent trees generated by *ms*, and analyze the distribution of their external branch lengths (Paradis et al., 2004). We used the external branch length distribution to build a theoretical prediction for the distribution of singletons from each tree (see section 2), and summarized the theoretical distributions by computing mean values and standard errors for each subsample. The L coalescent simulations were replicated 200 times.

3.2. Range Expansions in Africa

Simulations of range expansions were performed by using the computer program SPLATCHE2 based on an array of 87 by 83 demes modeling the African continent (Currat et al., 2004). The demographic scenarios corresponded to range expansions from a single origin, simulated for a total duration of 1,600 generations. For each deme, the migration rate was equal to $m = 0.07$, and the growth rate was equal to $r = 0.1$. Additional parameters included an ancestral effective population size of 200 individuals, 200 generations before onset of expansion, and an effective mutation rate of 10^{-5} per base pair per generation.

Four types of demographic scenarios were considered. Two scenarios considered a "homogeneous" environment, for which the deme carrying capacities were set to a constant value $C = 100$ everywhere in Africa. Two other scenarios considered a heterogeneous environment linked to vegetation. In tropical semi-desert areas, the carrying capacities were set to $C = 60$, and in tropical extreme deserts and rain forests, the carrying capacities were set to $C = 30$. Demographic histories also differed by their geographic source of expansion. Range expansions were started either from an origin in West Africa (Mali, -4° E, 13° N) or from an origin in the Sahel area (Chad, 22° E, 20° N).

Ten haploid chromosomes were simulated for 30 population samples through the geographic range considered (300 chromosomes). Genetic variation was surveyed at 30,000 loci, and filtered out for monomorphic loci. From the resulting data sets, we computed the empirical distribution of singletons in each population sample, and compared this measure to expected heterozygosity for each population sample. Data files for running the SPLATCHE2 simulations are provided in Supplementary File 1. We reproduced the four scenarios by using individual sampling instead of population sampling. Here, individual genotypes were recorded at 300 distinct geographic sites, each obtained from a Gaussian perturbation of population centers with standard error of 2°. The Kriging method was used to interpolate the values of the expected heterozygosity and the empirical distribution of singletons on a geographic map of Africa (Cressie, 2015).

3.3. Pearl Millet Data

Whole genome sequencing data were obtained for 146 cultivated accessions of pearl millet (*Pennisetum glaucum [L.] R. Br.*) from the species range in Africa (International Pearl Millet Genome Sequencing Consortium, Varshney et al., 2017). A total of 169,095 SNPs were sampled after filtering out low quality

variants, and were used to estimate the distribution of singletons (Supplementary Material 1).

3.4. Approximate Bayesian Computation

We used Approximate Bayesian Computation (ABC) to evaluate the ability of the distribution of singletons to correctly estimate the onset of expansion in a range expanding species, and to estimate a posterior distribution for the location of this origin for cultivated pearl millet. We performed 20,000 range expansion simulations by considering a heterogeneous environment using the computer program SPLATCHE2. The deme carrying capacities were equal to $C = 100$ for tropical semi-desert areas, $C = 20$ for tropical extreme deserts and $C = 10$ for rain forests. Additional parameters included an ancestral effective population size of 200 individuals, 200 generations before onset of expansion, and an effective mutation rate of 10^{-5} per base pair per generation.

Prior distributions allowed the geographic coordinates of the origin of expansion to vary over the Sahel region. Longitude ranged between −16°E and 40°E, and latitude ranged between 5°N and 30°N. Lower prior probabilities were given to extreme latitudes and longitudes as a consequence of unsuitable habitats (water regions). Uninformative prior distributions were considered for the migration rate, the growth rate, the total duration of the demographic phase, the ancestral population size and the time before onset of expansion (Supplementary Table 1). In simulations, genetic variation was surveyed at 146 geographic sites corresponding to the exact sampling locations of pearl millet accessions. Ten thousands SNPs were simulated for each genotype. When evaluating summary statistics, a fraction of SNPs were removed from the simulated data in order to match with the amount of missing values observed in the original data set.

To define the summary statistics for ABC, we used a histogram for the distribution of singletons in the sample. The 146 accessions were grouped into spatial clusters according to a *k*-means algorithm and individual geographic information (Hartigan and Wong, 1979). The *k*-means algorithm resulted in 14 groups with more than 6 accessions in each group (**Figure 1**). To obtain a histogram, we computed the mean number of singletons in each group, and divided this value by the total number of singletons in the sample (Supplementary Table 2). Then ABC analysis was performed with the `R` package `abc` (Blum and François, 2010; Csilléry et al., 2012). Neural network models were used to estimate posterior distributions for the latitude and longitude of the geographic onset of expansion whereas the other parameters were considered as nuisance parameters without any interpretable unit. The tolerance rate was set to 0.05 and 250 neural networks were used in the `abc` function.

We first tested the accuracy of our estimates by using simulated data sets as inputs to the inference method. The sampling procedure and the ABC estimation were replicated 100 times, and we evaluated the correlation between coordinates of true origins and their estimated values. Then we considered the pearl millet data, and represented the prior and posterior densities of the geographic onset parameters by using

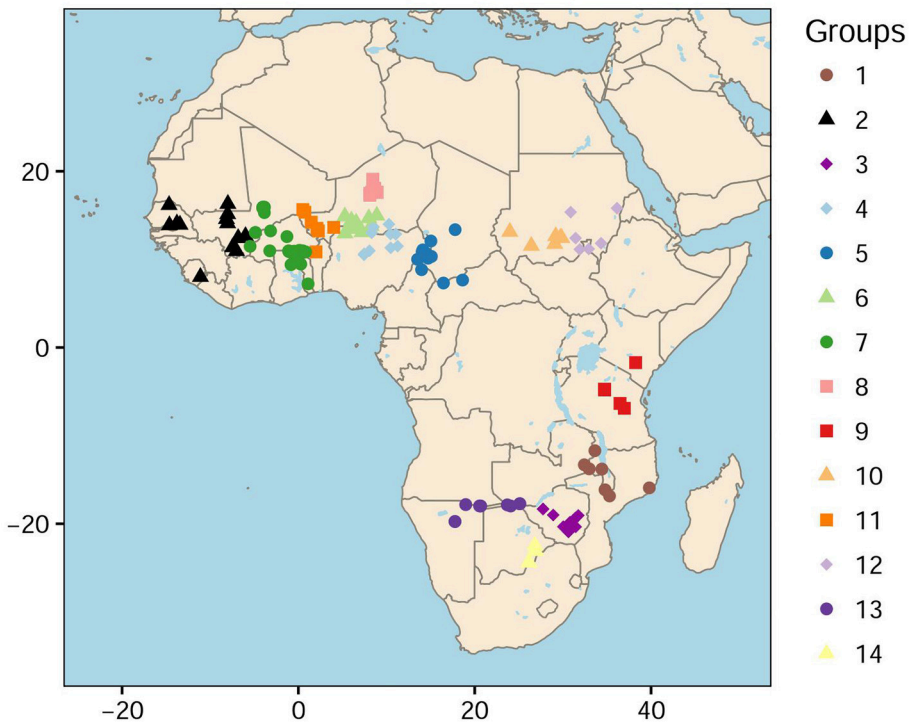


FIGURE 1 | Geographic distribution of 146 cultivated accessions of pearl millet. Fourteen geographic classes were defined as a result of a *k*-means procedure.

two-dimensional kernel density estimation with 100 grid points in each direction.

4. RESULTS

4.1. Coalescent Simulations of Splitting Populations

To evaluate statistical bias in the estimation of the distribution of singletons, we performed coalescent simulations of samples from two populations with unequal genetic diversity. The two populations diverged from an ancestral population t generations ago (*split time*), and at split time, the size of population 2 shrinked to s times the size of population 1 (*shrink rate*).

For each simulation, the number of polymorphic loci ranged between 7,883 and 39,761 (average value: 25,265 loci). For a value of the shrink rate $s \approx 1/3$, the average proportion of singletons in population 1 was about $\pi_1 = 0.0122$, and the average proportion of singletons in population 2 was about $\pi_2 = 0.0078$ ($\pi_1 + \pi_2 = 2/n$). This result reflected that genetic diversity in population 1 was higher than in population 2. The ratio was about $\pi_1/\pi_2 = 1.55$ (**Figure 2A**). The individual proportions were concentrated around their mean values with relatively small standard deviations ($SD_1 = 0.0010$, $SD_2 = 0.0008$).

The results from 200 replicates provided clear evidence that the empirical distribution of singletons is an unbiased estimate of its theoretical distribution based on coalescent trees (**Figure 2B**). The split time parameter had a weak influence on the distribution of singletons (Pearson correlation test, $P = 0.64$). The ratio π_1/π_2 reached values between 10 and 40 when the shrink rate was below 10%, and this parameter had a strong influence on the empirical distribution of singletons (**Figure S1**).

4.2. Range Expansions in Africa

For data sets generated under range expansion scenarios, the number of polymorphic loci ranged between 25,453 and 29,321 loci. The number of singletons ranged between 8,835 and 12,653, and the site frequency spectrum showed an excess of rare alleles as expected under explosive population growth. When the onset of expansion was set in Western Africa (cross in **Figure 3**), the maps of the empirical distribution of singletons and expected heterozygosity exhibited similar large-scale geographic patterns (**Figure 3**, Pearson's correlation coefficient 0.78). Because the computation of expected heterozygosities

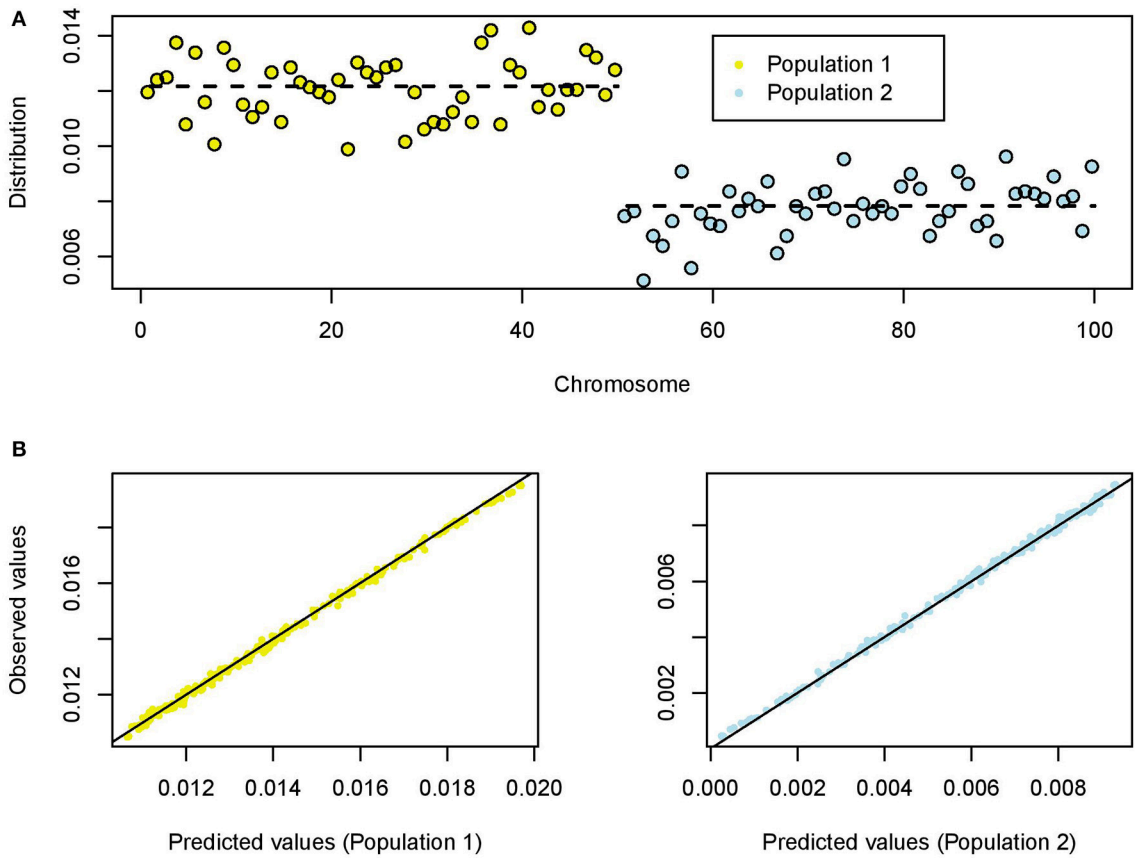


FIGURE 2 | Coalescent simulations of two splitting populations (100 chromosomes). **(A)** Empirical distribution of singletons for a value of the shrink rate $s = 0.33$. The dashed lines represent the averaged values for population 1 (expanding) and population 2 (skrinking). **(B)** Predicted and observed (empirical) values of the distribution of singletons for population 1 (left) and population 2 (right).

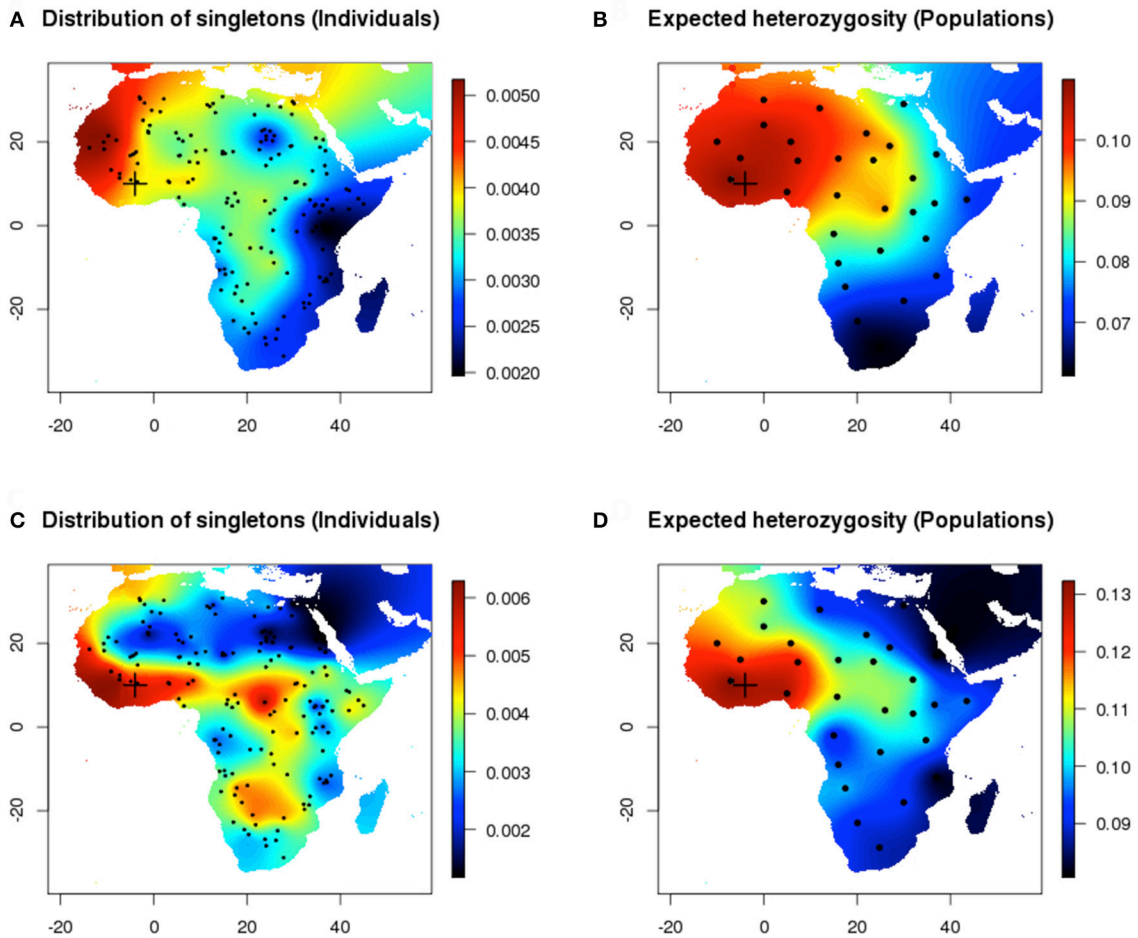


FIGURE 3 | Individual vs. population sampling after a range expansion simulation scenario (Western origin). **(A,B)** Homogeneous environment. Maps of the empirical distribution of singletons (individual sampling) and expected heterozygosity (true population sampling). **(C,D)** Inhomogeneous environment.

was based on a perfect assignment of samples to their true populations of origin, the interpolated maps corresponding to this measure (**Figures 3B,D**) contained less uncertainty than the maps of singletons (**Figures 3A,C**) that were based on random individual sampling. Considering environmental heterogeneity increased the variability of spatial estimates (**Figures 3C,D**).

Next, we compared estimates of heterozygosity for populations to the distribution of singletons in the same populations (**Figure 4**). Differences between maps produced with the empirical distribution of singletons and with expected heterozygosity decreased when the sampled chromosomes were perfectly assigned to their population of origin. The individual and population-based measures provided concordant estimates of genetic diversity in geographic space (Pearson's correlation coefficient 0.51). Similar results were observed when the onset of expansion was set in the Sahel area (20° E, 22° N) and were reported in **Figures S2**, **S3**.

4.3. Estimates of Expansion Onsets and Application to Pearl Millet

First, we used the distribution of singletons in ABC to infer origins of range expansion in 100 simulated data sets (**Figure 5**). The results provided evidence of the usefulness of the statistics to identify origins of range expansions. Estimated values for the longitude and latitude of the onset of expansion were highly correlated to the true values for these parameters. Pearson's squared correlation coefficients were equal to $R^2 = 0.950$ for the longitude and $R^2 = 0.948$ for the latitude (p-values < 0.01).

Next, we used the ABC approach to provide insights on the origin of range expansion of cultivated pearl millet in Africa. A total number of 41,032 singletons were found for 146 individuals, representing 24.27% of all variants. The posterior density for the longitude exhibited a mode around −7.52°E (CI:-11.26°E, 0.84°E) (**Figure 6**). For the latitude of origin, the posterior density exhibited a mode around 24.2°N and a large credible interval (CI: 11.03°N , 29.06°N) (**Figure 6**). The most probable

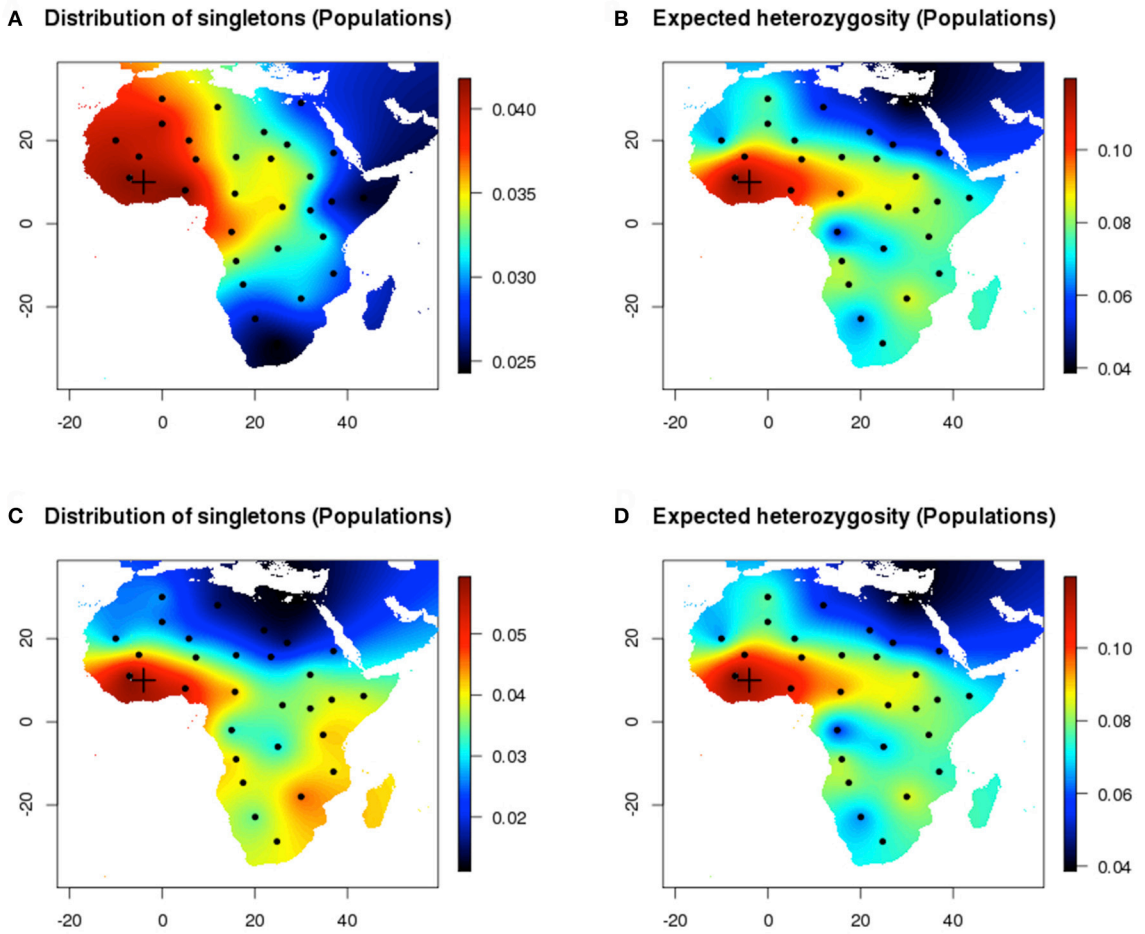


FIGURE 4 | Population sampling after a range expansion simulation scenario (Western origin). **(A,B)** Homogeneous environment. Maps of the empirical distribution of singletons (true population sampling) and expected heterozygosity (true population sampling). **(C,D)** Inhomogeneous environment.

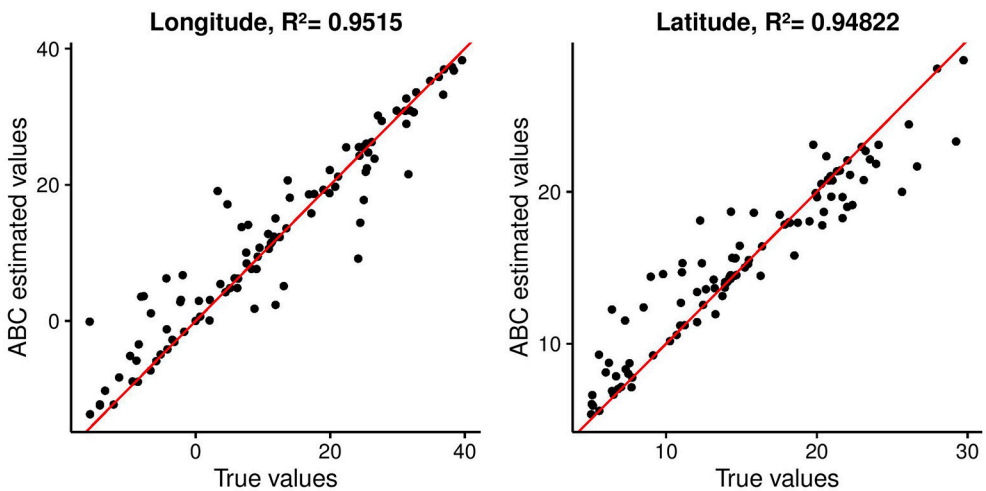


FIGURE 5 | Estimated coordinates of origin against their true values for 100 simulated data sets used as targets for ABC analysis. Pearson's correlation coefficients are reported.

location for the origin of expansion of pearl millet in Africa was found near the Mali-Mauritania border (**Figure 7**).

5. DISCUSSION

How singletons are distributed across geographic space provides a local measure of genetic diversity that can be measured at the individual level. In this study, we developed a theoretical background for the empirical distribution of singletons in a sample of chromosomes. We used simulations to provide evidence that the empirical distribution of singletons measures individual contributions to genetic diversity in the sample. The main advantage of this approach is to provide individual-based (local) estimates of genetic diversity that do not require the definition of populations.

Incorporated in an ABC framework, the empirical distribution of singletons led to accurate estimates of the geographic origin of range expansions in simulations. In ABC, the distribution of singletons was estimated by histograms obtained from clustering algorithms, and the histograms were used as summary statistics for Bayesian inference. Those statistics are appropriate to analyze the results of sequencing projects based on large scale sampling of individuals across geographic space. The method can be viewed as an interesting alternative to phylogenetic approaches when genomic sequences are used.

Potential factors that could bias our estimates of local genetic diversity includes missing data, genotyping errors, related individuals, and the use of a folded site frequency spectrum. Missing values or genotyping errors impacts individual data regardless of geography. By sharing genomic variation locally, related individuals reduce the number of unique variants drastically, and generate bias in global estimates of genetic diversity. Though those errors increase uncertainty in estimates, the biases on geographic estimates remain at small levels. Our ABC analysis took the potential biases into account by simulating the missing data, genotyping errors and the other issues. Alternative methods that could remove the biases would be based on genotype imputation and on the availability of genomic data from a closely related species.

We provided an illustration of the potential of singletons to inform demographic history by studying range expansion of pearl millet in Africa. Pearl millet is a widely grown staple

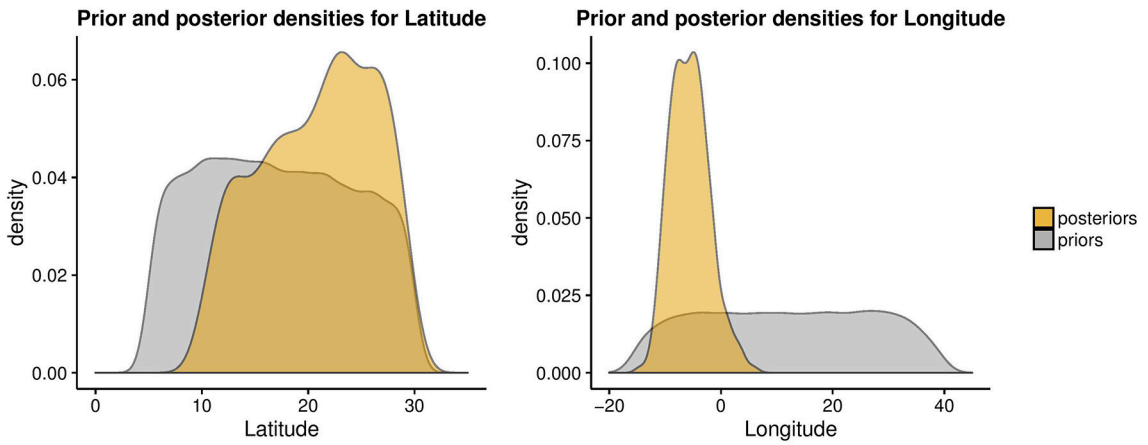


FIGURE 6 | Prior and posterior density estimates for the longitude and latitude of the expansion onset for cultivated pearl millet in Africa.

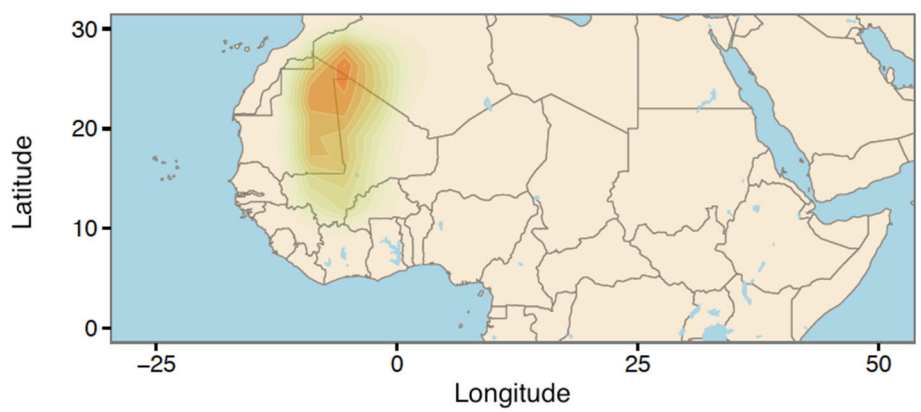


FIGURE 7 | Geographic origin of cultivated pearl millet expansion using kernel density estimation.

crop in Africa and India, but its precise origin is currently unknown (Tostain, 1992; Oumar et al., 2008; Clotault et al., 2012). When we applied an ABC approach to cultivated pearl millet genomes, we obtained a result supporting the Northern Mali region as the most probable geographic origin of expansion. Although the accuracy of the ABC approach was validated with extensive computer simulations of range expansion, the empirical results pointed out some limitations of our model for the data. The uncertainty around 18° reported for the latitude of origin was high, and improving our estimate would require supplementary information on past environmental conditions, carrying capacities and gene flow between pearl millet and related species. Interestingly, our results rejected an eastern origin for the expansion of the domesticated cereal. This result is consistent with recent archeological studies using both wild and cultivated samples, that pinpointed the Mali-Niger region as the most likely origin of domestication of pearl millet (Manning et al., 2011; Ozainne et al., 2014).

To conclude, singletons are a major component of the site frequency spectrum for many model and non-model species. The density of singletons in genomes has recently proven useful to detect selection in human genomes (Field et al., 2016). Here we showed that the density of singletons in geographic space is useful for providing local estimates of genetic diversity and key insights on the demographic history of a species.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

This work has been partially supported by the Agence Nationale de la Recherche, project AFRICROP, ANR-13-BSV7-0017, and by the LabEx PERSYVAL Lab, ANR-11-LABX-0025-01, funded by the French program Investissement d'Avenir.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2017.00139/full#supplementary-material

Figure S1 | Averaged proportion of singletons in population 1, and standard deviations in populations 1 and 2, as functions of the shrink rate.

Figure S2 | Individual vs. population sampling after a range expansion simulation scenario (Sahel origin). **(A,B)** Homogeneous environment. Maps of the empirical distribution of singletons (individual sampling) and expected heterozygosity (true population sampling). **(C,D)** Inhomogeneous environment.

Figure S3 | Population sampling after a range expansion simulation scenario (Sahel origin). **(A,B)** Homogeneous environment. Maps of the empirical distribution of singletons (true population sampling) and expected heterozygosity (true population sampling). **(C,D)** Inhomogeneous environment.

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073. doi: 10.1038/nature09534

Auer, P. L., and Lettre, G. (2015). Rare variant association studies: considerations, challenges and opportunities. *Genome Med.* 7:16. doi: 10.1186/s13073-015-0138-2

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Ann. Rev. Ecol. Evol. Syst.* 41, 379–406. doi: 10.1146/annurev-ecolsys-102209-144621

Blum, M. G. B., and François, O. (2005). Minimal clade size and external branch length under the neutral coalescent. *Adv. Appl. Probabil.* 37, 647–662. doi: 10.1017/S0001867800000409

Blum, M. G. B., and François, O. (2010). Non-linear regression models for approximate Bayesian computation. *Stat. Comput.* 20, 63–73. doi: 10.1007/s11222-009-9116-0

Caliebe, A., Neininger, R., Krawczak, M., and Rösler, U. (2007). On the length distribution of external branches in coalescence trees: genetic diversity within species. *Theor. Popul. Biol.* 72, 245–252. doi: 10.1016/j.tpb.2007.05.003

Clotault, J., Thuillet, A. C., Buiron, M., De Mita, S., Couderc, M., Haussmann, B. I., et al. (2012). Evolutionary history of pearl millet (*Pennisetum glaucum* [L.] R. Br.) and selection on flowering genes since its domestication. *Mol. Biol. Evol.* 29, 1199–1212. doi: 10.1093/molbev/msr287

Coventry, A., Bull-Otterson, L. M., Liu, X., Clark, A. G., Maxwell, T. J., Crosby, J., et al. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.* 1:131. doi: 10.1038/ncomms1130

Cressie, N. (2015). *Statistics for Spatial Data*. New-York, NY: John Wiley and Sons.

Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends Ecol. Evol.* 25, 410–418. doi: 10.1016/j.tree.2010.04.001

Csilléry, K., François, O., and Blum, M. G. B. (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* 3, 475–479. doi: 10.1111/j.2041-210X.2011.00179.x

Currat, M., Ray, N., and Excoffier, L. (2004). SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Mol. Ecol. Notes* 4, 139–142. doi: 10.1046/j.1471-8286.2003.00582.x

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., et al. (2016). Detection of human adaptation during the past 2000 years. *Science* 354, 760–764. doi: 10.1126/science.aag0776

Frichot, E., and François, O. (2015). LEA: an R package for landscape and ecological association studies. *Methods Ecol. Evol.* 6, 925–929. doi: 10.1111/2041-210X.12382

Fu, Y. X., and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics* 133, 693–709.

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., et al. (2011). Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U.S.A.* 108, 11983–11988. doi: 10.1073/pnas.1019276108

Hartigan, J. A., and Wong, M. A. (1979). A *K*-means clustering algorithm. *Appl. Stat.* 28, 100–108. doi: 10.2307/2346830

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surv. Evol. Biol.* 7:44.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18, 337–338. doi: 10.1093/bioinformatics/18.2.337

International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52–58. doi: 10.1038/nature09298

Keinan, A., and Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336, 740–743. doi: 10.1126/science.1217283

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* 95, 5–23. doi: 10.1016/j.ajhg.2014.06.009

Manning, K., Pelling, R., Higham, T., Schwenniger, J. L., and Fuller, D. Q. (2011). 4500-year old domesticated pearl millet (*Pennisetum glaucum*) from the Tilemsi Valley, Mali: new insights into an alternative cereal domestication pathway. *J. Archaeol. Sci.* 38, 312–322. doi: 10.1016/j.jas.2010.09.007

Marth, G. T., Czabarka, E., Murvai, J., and Sherry, S. T. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 166, 351–372. doi: 10.1534/genetics.166.1.351

Mathieson, I., and McVean, G. (2014). Demography and the age of rare variants. *PLoS Genet.* 10:e1004528. doi: 10.1371/journal.pgen.1004528

Memon, S., Jia, X., Gu, L., and Zhang, X. (2016). Genomic variations and distinct evolutionary rate of rare alleles in *Arabidopsis thaliana*. *BMC Evol. Biol.* 16:25. doi: 10.1186/s12862-016-0590-7

Novembre, J., and Slatkin, M. (2009). Likelihood-based inference in isolation-by-distance models using the spatial distribution of low-frequency alleles. *Evolution* 63:2914. doi: 10.1111/j.1558-5646.2009.00775.x

O'Connor, T. D., Fu, W., Turner, E., Mychaleckyj, J. C., Logsdon, B., Auer, P., et al. (2015). Rare variation facilitates inferences of fine-scale population structure in humans. *Mol. Biol. Evol.* 32, 653–660. doi: 10.1093/molbev/msu326

Ozainne, S., Lespez, L., Garnier, A., Ballouche, A., Neumann, K., Pays, O., et al. (2014). A question of timing: spatio-temporal structure and mechanisms of early agriculture expansion in West Africa. *J. Archaeol. Sci.* 50, 359–368. doi: 10.1016/j.jas.2014.07.025

Oumar, I., Mariac, C., Pham, J.-L., and Vigouroux, Y. (2008). Phylogeny and origin of pearl millet (*Pennisetum glaucum* [L.] R. Br) as revealed by microsatellite loci. *Theor. Appl. Genet.* 117, 489–497. doi: 10.1007/s00122-008-0793-4

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* 69, 124–137. doi: 10.1086/321272

Schork, N. J., Murray, S. S., Frazer, K. A., and Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* 19, 212–219. doi: 10.1016/j.gde.2009.04.010

Schraiber, J. G., and Akey, J. M. (2015). Methods and models for unravelling human evolutionary history. *Nat. Rev. Genet.* 16, 727–740. doi: 10.1038/nrg4005

Slatkin, M. (1985). Rare alleles as indicators of gene flow. *Evolution* 39, 53–65. doi: 10.1111/j.1558-5646.1985.tb04079.x

Tavaré, S. (2004). "Ancestral inference in population genetics," in *Lectures on Probability Theory and Statistics*, Lecture Notes Math. 1837, ed J. Picard (Berlin: Springer), 1–188.

Tennessen, J., Bigham, A., O'Connor, T., Fu, W., Kenny, E. E., Gravel, S., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337, 64–69. doi: 10.1126/science.1219240

Tostain, S. (1992). Enzyme diversity in pearl millet (*Pennisetum glaucum* L.). *Theor. Appl. Genet.* 83, 733–742. doi: 10.1007/BF00226692

Varshney, R. K., Shi, C., Thudi, M., Mariac, C., Wallace, J., Qi, P., et al. (2017). Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. *Nat. Biotechnol.* doi: 10.1038/nbt.3943. [Epub ahead of print]. Available online at: http://ceg.icrisat.org/ipmgsc/

Weigel, D. (2012). Natural variation in Arabidopsis: from molecular genetics to ecological genomics. *Plant Physiol.* 158, 2–22. doi: 10.1104/pp.111.189845

Weigel, D., and Mott, R. (2009). The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol.* 10:107. doi: 10.1186/gb-2009-10-5-107

Zhu, C., Li, X., and Yu, J. (2011). Integrating rare-variant testing, function prediction, and gene network in composite resequencing-based genome-wide association studies (CR-GWAS). *Genes Genomes Genet.* 1, 233–243. doi: 10.1534/g3.111.000364

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 July 2017
doi: 10.3389/fgene.2017.00098

Assessment of Genetic Diversity and Structure of Large Garlic (*Allium sativum*) Germplasm Bank, by Diversity Arrays Technology "Genotyping-by-Sequencing" Platform (DArTseq)

*Leticia A. Egea[1,2], Rosa Mérida-García[2], Andrzej Kilian[3], Pilar Hernandez[2] and Gabriel Dorado[1]**

[1] Departamento de Bioquímica y Biología Molecular, Campus Rabanales (C6-1-E17), Campus de Excelencia Internacional Agroalimentario (ceiA3), Universidad de Córdoba, Córdoba, Spain, [2] Instituto de Agricultura Sostenible (IAS-CSIC), Campus Alameda del Obispo, Córdoba, Spain, [3] Diversity Arrays Technology Pty. Ltd., Canberra, ACT, Australia

OPEN ACCESS

Edited by:
Samuel A. Cushman, United States Forest Service Rocky Mountain Research Station, United States

Reviewed by:
Turgay Unver, iBG-Izmir, International Biomedicine and Genome Institute, Turkey
Hikmet Budak, Montana State University, United States
Guillaume Besnard, UMR5174 Evolution Et Diversite Biologique (EDB), France

***Correspondence:**
Gabriel Dorado
bb1dopeg@uco.es

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *05 April 2017*
Accepted: *30 June 2017*
Published: *20 July 2017*

Citation:
Egea LA, Mérida-García R, Kilian A, Hernandez P and Dorado G (2017) Assessment of Genetic Diversity and Structure of Large Garlic (Allium sativum) Germplasm Bank, by Diversity Arrays Technology "Genotyping-by-Sequencing" Platform (DArTseq). Front. Genet. 8:98. doi: 10.3389/fgene.2017.00098

Garlic (*Allium sativum*) is used worldwide in cooking and industry, including pharmacology/medicine and cosmetics, for its interesting properties. Identifying redundancies in germplasm blanks to generate core collections is a major concern, mostly in large stocks, in order to reduce space and maintenance costs. Yet, similar appearance and phenotypic plasticity of garlic varieties hinder their morphological classification. Molecular studies are challenging, due to the large and expected complex genome of this species, with asexual reproduction. Classical molecular markers, like isozymes, RAPD, SSR, or AFLP, are not convenient to generate germplasm core-collections for this species. The recent emergence of high-throughput genotyping-by-sequencing (GBS) approaches, like DArTseq, allow to overcome such limitations to characterize and protect genetic diversity. Therefore, such technology was used in this work to: (i) assess genetic diversity and structure of a large garlic-germplasm bank (417 accessions); (ii) create a core collection; (iii) relate genotype to agronomical features; and (iv) describe a cost-effective method to manage genetic diversity in garlic-germplasm banks. Hierarchical-cluster analysis, principal-coordinates analysis and STRUCTURE showed general consistency, generating three main garlic-groups, mostly determined by variety and geographical origin. In addition, high-resolution genotyping identified 286 unique and 131 redundant accessions, used to select a reduced size germplasm-bank core collection. This demonstrates that DArTseq is a cost-effective method to analyze species with large and expected complex genomes, like garlic. To the best of our knowledge, this is the first report of high-throughput genotyping of a large garlic germplasm. This is particularly interesting for garlic adaptation and improvement, to fight biotic and abiotic stresses, in the current context of climate change and global warming.

Keywords: DNA fingerprinting, breeding, phenotype, somatic mutation, second-generation sequencing (SGS), third-generation sequencing (TGS), next-generation sequencing (NGS)

INTRODUCTION

Garlic (*Allium sativum*) is a plant producing an edible bulb, made of storage leaves known as cloves. It is of Asian origin, being *Allium longicuspis* considered its wild ancestor. It belongs to genus *Allium*, which includes almost 1,000 species, such as chive (*Allium schoenoprasum*), leek (*Allium ampeloprasum*), onion and shallot (*Allium cepa*) (Maab and Klaas, 1995; Kamenetsky et al., 2004; Meredith, 2008; Cardelle-Cobas et al., 2010; Pacurar and Krejci, 2010). Garlic has a large diploid genome ($2n = 2x = 16$), of an estimated haploid (1C) size of 15.9 gigabase pairs (Gbp); that is, 32 times larger than rice (*Oryza sativa*). Garlic is sterile (does not produce fertile botanical seeds by sexual reproduction), asexually propagating by its cloves, despite some progress in recent years to restore garlic fertility (Shemesh-Mayer et al., 2015). Besides, cloves must be reproduced every year, since they cannot be stored for longer periods and then germinated, as happens with standard botanical seeds. Such peculiarity adds extra cost and inconvenience to its maintenance, mainly for large germplasm collections. The peculiar garlic reproduction could lead to low genome diversity, since meiosis is not involved in its clonal reproduction by vegetative propagation (Kamenetsky et al., 2015). Yet, garlic shows a surprisingly high biodiversity, as well as environmental-adaptation capacity and phenotypic plasticity (Volk et al., 2004). All that leads to the large number of garlic varieties or cultivars available (traditionally classified by agromorphological characteristics). The reason for that is not fully understood, suggesting a complex genome (Green, 2001), due to its extremely large size containing many multicopy genes and other duplications, including non-coding sequences and tandem repeats (Arumuganathan and Earle, 1991; Jones et al., 2004; Ovesna et al., 2015), which should be better understood once sequenced. So far, partial and total genome duplications have been described (Supplementary Table S1). Additionally, somatic mutations have been also reported for this species, as well as somaclonal variation, differential gene-expression and alternative splicing (Al-Zahim et al., 1999; Rotem et al., 2007; Kamenetsky et al., 2015; Shemesh-Mayer et al., 2015). Probably, transposable elements are also involved in the evolution of this species.

Besides being appreciated in cooking as common seasoning for thousands of years (Cardelle-Cobas et al., 2010), garlic is also used in pharmacology and cosmetics. Indeed, it is known to have medical properties, protecting against different diseases, like, for instance, hypercholesterolemia, hypertension, atherosclerosis, and thrombosis, reducing the risk of developing cardiovascular disease (CVD). Other recognized bioactivities are antimicrobial (albeit being probiotic), antiasthmatic, antioxidant, anticarcinogenic, etc. (Corzo-Martínez et al., 2007; Pacurar and Krejci, 2010; Rana et al., 2011). Indeed, garlic contains bioactive compounds, including, among others: (i) lectins, which have wide applications in biomedicine and biotechnology (Smeets et al., 1997); (ii) peptides with angiotensin I-converting enzyme (ACE) inhibitory activity, being related to its antihypertensive activity (Suetsuna, 1998); and (iii) *N*-feruloyltyramine, which protects against CVD by suppressing platelet activation (Park, 2009). Besides, this species is rich in enzymes with industrial interest; for instance: (i) nucleases (DNase and RNase), with application in molecular biology (Carlsson and Frick, 1964); (ii) cellulases for biotechnological applications, like conversion of biomass into biofuel (Kim et al., 2010); (iii) superoxide dismutases (SOD), which represent a main defense against oxidative stress, being widely used in pharmacology/medicine, cosmetics, food, agriculture, and chemical industries (He et al., 2008; Liu et al., 2011); (iv) proteases/hemagglutinases, with application in medical tests (Parisi et al., 2008); and (v) alliinases (also known as alliinases), that catalyze conversion of alliin to allicin, which is the main therapeutic agent of garlic (Corzo-Martínez et al., 2007; Kim et al., 2010; Rathnasamy et al., 2014).

On the other hand, agricultural practices usually involve cultivation of a reduced number of species and varieties, which may lead to genetic erosion. That is especially relevant for monocultures, which on the other hand are required to feed an exponentially growing human population. It is therefore important to maintain germplasm banks as reservoirs of genetic variability for crop breeding. Thus, such collections may harbor genetic potential to improve productivity and adaptation/resistance to abiotic (drought, salinity, etc.) and biotic (diseases and plagues) stresses (Tanksley and McCouch, 1997). That is particularly relevant in the current frame of climatic change and global warming. Understanding this potential is critical for identification of biodiversity in biological resources and its efficient management, including conservation and selection of genetically divergent accessions to optimize breeding programs (Olukolu et al., 2012).

Yet, germplasm banks may be generated as mere raw collections of varieties over many years, being classified by criteria based on phenotypic/agronomic traits (passport data). That could lead to both homonymy (same name for genetically different cultivars) and duplications or synonymy (same cultivars with different names). That is especially problematic for species with similar appearance and significant phenotypic plasticity, like garlic. Thus, efficient identification of biodiversity is of paramount importance to manage and maintain such genetic-resources (Govindaraj et al., 2015). That is relevant not only to identify genuine variability for breeding purposes, but also to reduce space and maintenance costs, especially for large germplasm banks, generating reduced, albeit representative, core collections (Zhao et al., 2010).

The role of molecular markers as a tool for genetic analyses and crop improvement has gained importance through the years, as we have reviewed (Dorado et al., 2015c). Their use has become common in model species and important crops. Indeed, genetic diversity and polymorphism assessments are major priorities in plant and crop-breeding studies (Nybom and Bartish, 2000). Large-scale identification of molecular markers like single-nucleotide polymorphism (SNP) on genome and transcriptome represent interesting approaches (Ipek et al., 2016; Akpinar et al., 2017). Classical molecular-markers to assess genetic diversity and polymorphism in garlic have been described (Ovesná et al., 2014; Ipek et al., 2015). Among others, they include isozymes, random-amplified polymorphic DNA (RAPD) (Maab and Klaas, 1995), simple-sequence repeats (SSR) (DaCunha et al., 2014), amplified-fragment length polymorphism (AFLP) (Ipek et al.,

2005) and insertions-deletions (InDel) (Wang et al., 2016). Yet, such analyses of genetic diversity in this species are challenging (Kim et al., 2009).

Fortunately, recent technological developments overcome previous limitations. They include second-generation sequencing (SGS) and third-generation sequencing (TGS) approaches, sometimes known by the ambiguous next-generation sequencing (NGS) terminology, as we have reviewed (Dorado et al., 2015b). Thus, a high-throughput genotyping-by-sequencing (GBS) technology (DArTseq) has been developed. It combines diversity arrays technology (DArT) complexity reduction methods with SGS/TGS (Kilian et al., 2012; Courtois et al., 2013; Cruz et al., 2013; Raman et al., 2014), allowing to identify SNP. DArT markers are polymorphic segments of DNA that are found at specific genome sites, after complexity reduction, being detected by hybridization. Those markers may show dominant or codominant inheritance (Gupta et al., 2008). DArT markers exploit DNA-microarray platforms to analyze DNA polymorphisms, without requiring previous DNA-sequence knowledge. Their applications include genetic fingerprinting, like whole-genome profiling for molecular breeding, germplasm characterization and genetic mapping, among others (Jaccoud et al., 2001). DArTseq can be optimized for each organism and application, by selecting the most appropriate complexity-reduction method (both size of representation and fraction of selected genome for assays). This is particularly relevant for garlic, which has a large and expected complex genome, as previously described. Therefore, DArTseq has been used in the present work as a proof-of-concept, to analyze a large garlic-germplasm bank.

The main goals of this study are: (i) assess genetic diversity and structure of a large garlic-germplasm bank; (ii) create a core collection to reduce the number of original accessions, without losing genetic diversity; (iii) relate genotype to agronomical features; and (iv) describe a cost-effective method to manage genetic diversity that could be applied to germplasm banks and breeding projects of garlic and other species.

MATERIALS AND METHODS

Plant Material and DNA Isolation

A total of 417 *a priori* different garlic entries collected in Spain (some of them being originally derived from other countries) were used for DArTseq analyses: 408 from the main Garlic-Germplasm Bank at "Instituto Andaluz de Investigación y Formación Agraria, Pesquera, Alimentaria y de la Producción Ecológica" (IFAPA) of "Junta de Andalucía" in Cordoba; five from Cordoba University (C1 to C5); and four (G, K, L, and M) from "Centro de Ensayos de Evaluación de Variedades" at "Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria" (INIA) in Madrid (Supplementary Table S1). Garlic leaves were frozen in liquid nitrogen and stored at −80°C until needed.

DNA was isolated using cetyl trimethylammonium bromide (CTAB) protocol (Murray and Thompson, 1980), as we have optimized (Hernandez et al., 2001). It was dissolved in Tris-Na_2EDTA (TE; pH 8) and stored at 4°C. Isolated DNA was quantified by NanoDrop 2000c (Thermo Fisher Scientific, Waltham, MA, United States) and segregated by 1% (w/v) agarose [from United States Biological (Salem, MA, United States)] gel electrophoresis (AGE). Then it was stained with ethidium bromide from Sigma–Aldrich (St. Louis, MO, United States). Resulting DNA was visualized under ultraviolet (UV) light for quality evaluation, using a Molecular Imager VersaDoc MP 4000 System from Bio-Rad (Hercules, CA, United States). Additionally, DNA digestions with the frequent-cutter *Tru*1I restriction enzyme (RE; cutting at T$^{|}$ TA$_{|}$ A) from Thermo Fisher Scientific were performed, in order to check DNA quality and absence of contaminating nucleases.

DArTseq

DArTseq method from Diversity Arrays Technology (Canberra, ACT, Australia) is described elsewhere[1]. In short, the following steps were carried out: (i) complexity reduction, in which genomic DNA was digested with a combination of restriction enzymes. Then, adapters were ligated and only polymorphic fragments were selected. In this way, this technique allowed to exclusively focus in those sections of the genome which are interesting for genetic-diversity analyses, due to their polymorphism; (ii) polymorphic fragments were cloned into *Escherichia coli* bacteria to create a library. Each *E. coli* colony should contain one of those fragments; (iii) the generated library was amplified by polymerase chain-reaction (PCR), as we have reviewed (Dorado et al., 2015a); (iv) amplicons were cleaned and evaluated by capillary electrophoresis sizing; (v) fragments were sequenced; (vi) A FASTQ file was created with generated sequencing reads, including sequences from 30 to 60 base pairs (bp) of polymorphic fragments; (vii) an internal alignment was performed, using other reads from the library (this step is carried out in case of incomplete or absent reference genome, like in the present work); (viii) SNP and SilicoDArT markers were searched and filtered using algorithms; and (ix) resulting data were two presence/absence (1 and 0, respectively) matrices. One contained SNP and the other SilicoDArT markers, where each column represented an individual and each row a marker (Kilian et al., 2012).

In our case, four methods of complexity reduction were tested in garlic (data not shown), selecting the *Pst*I-*Nsp*I restriction enzymes (cutting at G| TGCA| G and R| CATG| Y, respectively). Briefly, DNA samples were processed in digestion/ligation reactions as previously described (Kilian et al., 2012), but replacing a single *Pst*I-compatible adaptor with two different adaptors, corresponding to two different RE overhangs. The *Pst*I-compatible adapter was designed to include flowcell-attachment sequence from Illumina (San Diego, CA, United States), sequencing-primer sequence and "staggered" barcode (varying-length region), similar to previously reported (Elshire et al., 2011). Reverse adapter contained flowcell-attachment region and *Nsp*I-compatible overhang sequence. Interestingly, an overrepresented sequence from cytoplasmic (chloroplastic) DNA, corresponding to >10% of total sequences, was identified (after initial optimization) in many *Pst*I-*Nsp*I garlic-library

[1]http://www.diversityarrays.com/dart-application-dartseq

samples. A cut site for *Alw*I (cutting at GGATCNNNN| N|) was identified within this overrepresented sequence, and thus such restriction enzyme was included in the digestion-ligation step of library construction. Only "mixed fragments" (*Pst*I-*Nsp*I) which did not have *Alw*I site were effectively amplified in 30 rounds of PCR, using the following reaction profile: (i) denaturation at 94°C for 1 min; (ii) 30 cycles [94°C for 20 s (denaturation), 58°C for 30 s (primer annealing) and 72°C for 45 s (primer extension)]; and (iii) final polymerization at 72°C for 7 min. Equimolar amounts of PCR amplicons from each sample reaction of 96-well microtiter plates were bulked and applied to c-Bot (Illumina) bridge PCR, followed by sequencing on HiSeq 2000 sequencing system from the same manufacturer. Single-read sequencing reactions were run for 77 cycles.

Sequences generated from each lane were processed using DArT analytical-pipelines. In the primary one, Fast-Alignment Sequence Tools Q (FASTQ) files were first processed. Thus, poor-quality sequences were filtered-away, applying more stringent selection criteria to the barcode region, as compared to the rest of the sequence. Assignments of sequences to specific samples in the "barcode split" step were very reliable. This way, approximately 2,000,000 sequences per barcode/sample were identified and used in marker calling. Finally, identical sequences were collapsed into "fastqcoll" files. These were "groomed" using the DArT PL's C++ algorithm, which corrects low-quality bases from singleton-tags into correct bases, using collapsed tags with multiple members as template.

Groomed fastqcoll files were used in the secondary pipeline (presence/absence of restriction fragments in representation), by DArT, PL, SNP, and SilicoDArT calling algorithms (DArTsoft version 14). In total, 33,423 presence/absence markers were generated. All tags from all libraries included in the DArTsoft analyses were clustered using the DArT PL's C++ algorithm (threshold distance of 3), for SNP calling. That was followed by cluster parsing into separate SNP loci, using a range of technical parameters; especially the balance of read counts for allelic pairs. Additional selection criteria were added to the algorithm, based on previous experience with analyses of approximately 1,000-controlled cross populations (data not shown). Testing for Mendelian distribution of alleles in these previous populations facilitated selection of technical parameters, discriminating well-true allelic variants from paralogous sequences. In addition, multiple samples were processed from DNA to allelic calls, as technical replicates and scoring consistency was used as the main selection criteria for high-quality/low error-rate markers. Calling quality was assured by high average-read-depth per locus (average across all markers was over 10 reads/locus).

Genetic Diversity and Structure Assessments

Three different analyses were performed, in order to study genetic diversity and structure of germplasm-bank accessions. After creating the SNP and SilicoDArT marker scoring matrices, a Gower's distance matrix was generated. Gower's distance is a coefficient that measures similarity between two samples, based on logical (absence/presence) information differing for several variables (Gower, 1971). These data were used to determine genetically redundant samples. Secondly, a hierarchical cluster-analysis was done with the "pvclust" R package (Suzuki and Shimodaira, 2015). The phylogenetic tree (dendrogram) was computed with a complete-linkage method. By doing complete-linkage clustering (agglomerative hierarchical clustering method), each element of a distance matrix was first individually clustered. Then, each sample was combined into a new cluster, according to the shortest distance (Defays, 1977). Besides previous tests, a principal-coordinates analysis (PCoA; also known as classical multidimensional scaling, Torgerson Scaling or Torgerson-Gower scaling) was also carried out, using R software version 3.2.2 (R-Development-Core-Team, 2015). Additionally, STRUCTURE software version 2.3.4 (Pritchard et al., 2000) was used to study genetic structure. The chosen parameters were five iterations, K ranging from 1 to 3, with a burnin length of 10,000 and 20,000 Markov Chain Monte Carlo (MCMC) repetitions after burnin.

RESULTS

DArTseq Analyses

A total of 417 garlic samples were analyzed using SilicoDArT markers (representing presence/absence of restriction fragments in DArT genomic representations) and SNP data. A total of 14,392 SNP were used for the analyses. DArTseq markers allowed identifying 286 unique (Supplementary Table S2) and 131 redundant samples. The latter were divided into 19 groups, showing a variable amount of individuals (two to 53; Supplementary Table S3). For instance, in group 1, samples 717 and 718 were from the same province (Jaen, Spain). Spanish White varieties were mainly associated in groups 2 and 3 (samples 238, 452, and 461, all from northern Spain). Additionally, for group 2, there was an internal structure between regions. Samples 335, 424, 433, 434, 457, 464, and 467 were from northern Spanish provinces; samples 360 and 368 came from Caceres (Spain) and samples 127, 130, and 553 from southern Spanish provinces. Groups 4 and 7 to 10 included Spanish Purple varieties. Particularly, samples in group 4 were all from Castilla-Leon (Spain). Group 7 was the most numerous, with a total amount of 53 redundant samples. Interestingly, some associations by province were found in this group. Thus, samples 2, 59, 486, and 489 were all from northern regions; samples 21, 37, and 366 from central provinces; and samples 3, 85, 107, 110, 125, 131, 139, 150, 171, 225, 344, 356, 715, and 720 were from southern provinces. Two samples (14 and 280) from Taiwan, were also included in group 7. On the other hand, no associations were found for groups 5, 6, and 11 to 19.

Germplasm-Diversity Assessments

The 417 garlic samples were further analyzed, in order to assess their genetic diversity and structure, to eliminate redundant accessions, and thus generate the germplasm-bank core collection. Two different analyses were performed: hierarchical cluster computed by complete-linkage method and PCoA. The dendrogram (**Supplementary Figure S1**) showed

three main clusters (I to III), besides a few samples diverging from them (A and B). Main branches were supported by high-bootstrap values (>90). Moreover, bootstrap values were mainly high as well inside the main three clusters. Only some final subgroups had statistically non-significant bootstrap values. The separation in the dendrogram of some well-characterized samples (C1 to C5) is of special interest. Thus, Spanish varieties (Purple C3 and White C4; highlighted in purple and pink, respectively, in **Supplementary Figure S1**) were more related between them than to Chinese varieties (White C1 and Purple C2; highlighted in brown in **Supplementary Figure S1**), which were closely related. Sample C5 is a Brazilian garlic (thought to be an old Spanish Purple variety exported to America during colonialism) brought back to Spain 5 years ago. Interestingly, it was nearer to Spanish samples (closer to C3 than to C4) than to other accessions (C1 and C2), being highlighted in purple (**Supplementary Figure S1**).

Agro-morphological information (Supplementary Table S1) showed data in agreement with the generated dendrogram. For instance, cluster A contained samples 167, 239, and 459, being hexaploid or giant varieties (**Supplementary Figure S1**; highlighted with orange dots). There was a fourth hexaploid individual (379), being located in cluster III. Another interesting case was made of samples grouped together and with similar geographical origins. Thus, accessions 511, 513, and 514 came from Egypt (**Supplementary Figure S1**; highlighted with brown dots). Additionally, there were clusters with samples from Castilla-Leon region like: (i) 380, 389, and 432; (ii) 376, 424, 425, and 431; and (iii) 54, 423, 434, and 438 in the case of cluster II (highlighted with pink dots). Samples 32, 123, 125, 136, 225, and 1390 in cluster III were from Andalusia region (Spain; highlighted with purple dots). Samples 265, 270, 272 to 274, 276, 300, and 373 from cluster B came from Japan.

In addition, most accessions were also grouped by garlic-variety color in the phylogenetic tree. Thus, samples 20, 54, 238, 335, 360, 368, 424, 452, and 467 were Spanish White varieties (cluster II, pink). Likewise, samples 2, 3, 16, 17, 19, 21, 27, 29, 30, 32, 33, 37, 38, 77, 85, 87, 110, 117, 120, 123 to 125, 131, 132, 136, 138 to 141, 149, 150, 158, 161, 166, 171 to 173, 296, 297, 342, 343, 349, 356, 366, 454, 489, 542, 543, 560, 566, 570, 572, 574, 577, 578, 694, 752, 774, 779, G and K were Spanish Purple, Red, Brown, or "Colorado" varieties (cluster III, purple). Conversely, some samples did not group as expected. Thus, accessions 176 and 353 (Brown and Spanish Purple, respectively) would belong to cluster III, in accordance to their available agro-morphological data, yet they were in cluster A. Likewise, samples 36, 43, 88, and 109 (being considered Red or Purple varieties) did not group in cluster III, but in cluster II instead. Additionally, sample 44 is described as Chinese and thus expected in cluster I, but showed in cluster II instead. Samples 28, 79, 101, 137, 268, 526, 753, 776, and L (described as White varieties) were expected in cluster II, but were in cluster III. Sample 51 (described as Spanish White) was conversely located in cluster I instead of II. Likewise for some Spanish Purple samples (7, 348, 363, 369, and 775). Finally, samples 263 and 300 (described as White varieties) were included in cluster B instead of II. All samples that were not assigned consistently with agro-morphological data were highlighted with red dots in **Supplementary Figure S1**.

Principal-coordinates analysis was performed to further evaluate dendrogram clusters (**Figure 1**). Variance (genetic diversity) explained by principal components (PC) (accounting for 0.99 of cumulative variance) was 0.93 for PC1, 0.04 for PC2, and 0.02 for PC3. The relationships for samples C1 to C5 were similar to the ones in the dendrogram. As expected, samples C1 and C2 were nearer among them (Chinese), as well as samples C3 to C5 (Spanish origin). In addition, samples C3 and C5 were also closer compared to C4, as displayed in dendrogram (Supplementary Table S4).

Germplasm Genetic-Structure

Genetic structure of the garlic germplasm-bank collection was evaluated with STRUCTURE software. Three groups were

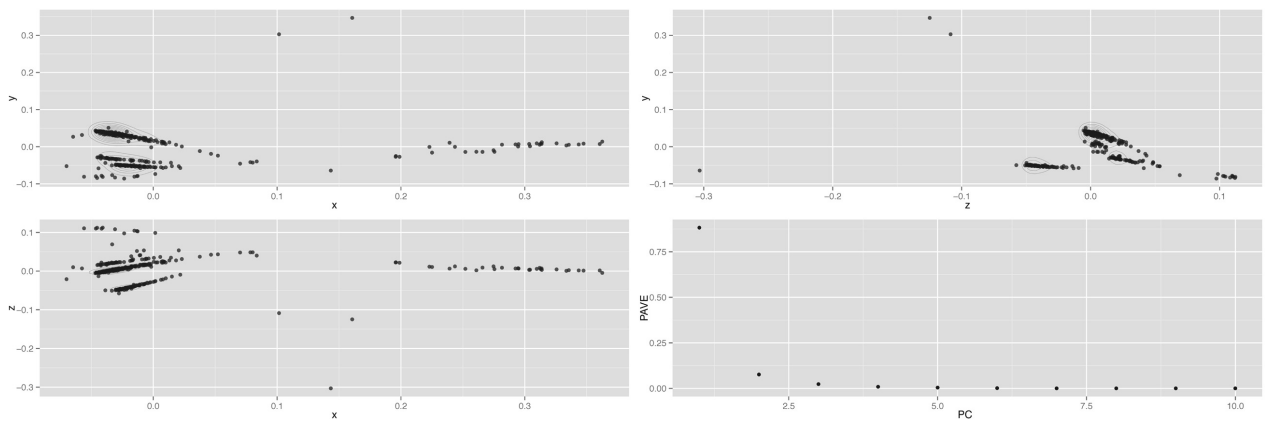


FIGURE 1 | Garlic principal-component plot. PCoA analysis was carried out to further analyze the garlic germplasm diversity. Upper-left represents principal coordinate (PC1; *x*-axis) with PC2 (*y*-axis); lower-left compares PC1 (*x*-axis) to PC3 (*z*-axis); and upper-right shows PC2 (*z*-axis) versus PC3 (*y*-axis). The lower-right graph corresponds to the Proportion of Analysis of Variance Explained (PAVE).

found, based on maximum likelihood and delta K (ΔK) values (**Supplementary Figure S2a**). As described above, this result is in agreement with cluster analysis and PCoA. Bar plot for $K = 3$ was also shown (**Supplementary Figure S2b**). In relation to the probability of membership of samples to clusters, Cluster I showed a score of 44.8%, being the group with the highest percentage. Clusters 2 and 3 had similar values (26.4 and 28.8%, respectively). When the probability of belonging to a group was high ($\leq$0.8 to 0.9), such individuals showed the same association found in hierarchical cluster-analysis. Well-known varieties (C1 to C5), also maintained the same relationships (Supplementary Table S5).

DISCUSSION

Garlic is known for multiple alimentary, medical and cosmetic uses worldwide. Yet, its classification and conservation in germplasm banks is challenging, due to homonymy and synonymy, being further complicated by its asexual life-cycle (Ipek et al., 2005). Previous information available allowed classifying the studied germplasm samples in this work by agro-morphological traits. Yet, such approach may be non-effective identifying true biodiversity, increasing redundancies and thus space and preservation costs in germplasm banks. In fact, it is known that the same garlic genotypes in different environmental conditions could exhibit diverse phenotypes (Volk et al., 2004). This is due to the high phenotypic plasticity of garlic, probably linked to its huge and expected complex genome, which somehow should compensate its lack of sexual reproduction.

Molecular markers have become an essential tool to identify, manage, and protect genetic diversity. Yet, developing them may be complicated, time-consuming and expensive for species like garlic, without sequenced reference genome, in which only scarce genomic-information is available (Ovesná et al., 2014). Additionally, classical molecular markers like isozymes, RAPD, SSR, or AFLP are not well suited to genotype garlic germplasm banks, due to its lack of resolution for such a peculiar genome in asexually reproducing accessions. Fortunately, technologies like DArT –and more recently, DArTseq– allow to reduce complexity and thus resolve complex genomic samples (Jaccoud et al., 2001).

Therefore, DArTseq was used in the present work to evaluate the genetic diversity and structure of 417 garlic samples (408 accessions from a garlic-germplasm bank). Data were analyzed by hierarchical-cluster computed by complete-linkage method, PCoA and genetic-structure approaches. Results showed a general consistency between accessions, geographic origins and groupings for expected/known garlic identities. All tests showed that individuals could be divided into three main groups (I, II, and III). Moreover, when the statistical probability of belonging to a group was high, the same association pattern of individuals was found in hierarchical-cluster analysis. Specifically, patterns for samples C1 to C5 (according to the previously known information) were maintained. Hence, DArTseq markers proved to be an effective and consistent genotyping approach to assess genetic diversity and structure.

Samples grouped by variety or geographical proximity were also found in non-redundant accessions, as described in the "Results" section. As expected, garlic samples of the same or near geographical regions grouped together. Indeed, cultivated varieties are usually selected by growers for several reasons, including being adapted to the climate in a specific region. In addition, the asexual garlic reproduction could lead to less genetic diversity and differentiation among varieties with similar geographical origins or different variants of the same variety. On the other hand, some samples were not grouped as expected, according to their agro-morphological information. Yet, such data is generated *de visu*, being therefore less accurate than molecular studies. In fact, it is known that morphological data are not always reliable to classify and detect genetic variation in germplasm collections (Jansky et al., 2015).

On the other hand, STRUCTURE assumes that markers are not in linkage disequilibrium (LD) within subpopulations. Yet, there are redundant lines in the data set, which could be against such assumption. But, there was a high consistency when comparing dendrogram clusters with those generated by STRUCTURE software. Thus, individuals assigned to the same cluster in the former, usually had higher probabilities to belong to the same group in the latter. Only three individuals were assigned differently in such analyses (4, 43, and 430) (Supplementary Table S5 and **Supplementary Figure S2**). This could be due to several reasons. In fact, criteria and calculations could lead to different results in each analysis. In the case of samples 4 and 430, they were located in an initial branch of cluster III, which indicates that they were genetically more different that the rest of assigned samples. Additionally, agro-morphological information was missing for samples 4 and 430.

The redundancy analysis showed that about one third of studied samples (131) could be considered as genetically redundant vs. 286 non-redundant (unique). This shows the higher resolution power and value of genomic analyses over agro-morphological ones. Thus, DArTseq results allowed to significantly reduce the analyzed garlic germplasm-bank size by 31.41%, generating a core collection, which was the main purpose of this research. Redundant accessions were divided into 19 groups (Supplementary Table S3). Samples included in each of them were in general related by variety (White, Purple, etc.) or location (same or near provinces). Interestingly, White varieties were more differentiated by location, whereas Purple ones were mainly associated in only one group. Samples 79 (Chinese White variety) and 526 (Spanish White variety) showed in group 7, in which Spanish Purple individuals were included. Curiously, this same lack of correlation was found in the hierarchical-cluster analysis, suggesting identities/differences not yet well understood. Further research is required to properly assess such results, including analyses of full genome sequences, once available in the future. That is now a possibility for large genomes like the garlic one, thanks to the throughput increase and cost reduction of TGS, which is expected to

become a mature technology in the next years (Dorado et al., 2015b).

As we have found, DArTseq is a cost-effective genotyping tool for creating and maintaining germplasm banks, allowing to properly ascertain, manage and maintain available biodiversity. Such technology has generated high-quality whole-genome profiles and genetic patterns, with dramatically increased resolution in relation to previous methodologies. Additionally, the high number of samples analyzed in this work, together with the large amount of marker data generated on lines with phenotypic information, should be useful for both genetic dissection of important traits and to help breeders improve this crop. Moreover, results obtained by DArTseq in any species can help to perform further analyses in germplasm collections without previous genetic information, even with high phenotypic-plasticity, complex genomes and asexual reproductive-systems that may hamper diversity analyses (Gebhardt, 2013). DArTseq sequences can be used to develop DArTseq markers and other molecular markers, such as SSR or SNP, which can be transferable to other germplasm banks (Belaj et al., 2011; Atienza et al., 2013). These tools can be associated to traits of interest, and thus used for marker-assisted breeding.

CONCLUSION

We have significantly reduced the analyzed garlic germplasm-bank size, identifying redundant accessions and thus generating a unique (non-redundant) core collection, with the consequent reduction in space and maintenance expenses. To our knowledge, this is the first work of high-throughput garlic genotyping. The obtained results show that DArTseq is a cost-effective method to perform genotyping-by-sequencing and genetic diversity analyses of such species with huge, expected complex and mostly unknown (without reference) genome, with clear applications for biodiversity conservation. This supports previous studies for characterizing and managing germplasm banks of other species. DArTseq has generated consistent results, in accordance with variety and geographical origin. They remark the relevance of genetic versus agro-morphological data, especially in the context of peculiar garlic-plasticity for environmental adaptation. Additionally, the high number of samples analyzed in this work and the amount of data generated should be useful for plant breeders in general, as well as for garlic adaptation and improvement in particular. This, along with other molecular markers and agro-morphological information represent useful tools to improve management strategies in germplasm-banks. In fact, having a core collection of characterized genotypes and phenotypes could help breeders to select plants with better adaptability. This is important for productivity and to face biotic and abiotic stresses, to fight the current climate change and global warming.

AUTHOR CONTRIBUTIONS

LE performed experiments, analyzed data, and wrote the manuscript; RM-G analyzed data and participated in manuscript writing; AK contributed to reagents, analysis tools, and manuscript writing; PH contributed to experimental design, materials, reagents, analysis tools, and manuscript writing; GD conceived and designed the experiments, contributed to materials, reagents, analysis tools, and manuscript writing; All authors read and approved the final version of the manuscript.

FUNDING

Supported by "Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria" (MINECO and INIA RF2012-00002-C02-02) and jointly funded by "Fondo Europeo de Desarrollo Regional" (FEDER); "Consejería de Agricultura y Pesca" (041/C/2007, 75/C/2009 and 56/C/2010), "Consejería de Economía, Innovación y Ciencia" (P11-AGR-7322) and "Grupo PAI" (AGR-248) of "Junta de Andalucía"; and "Universidad de Córdoba" ("Ayuda a Grupos"), Spain.

ACKNOWLEDGMENTS

We thank Francisco Mansilla (IFAPA, Córdoba, Spain) for germplasm samples and agro-morphological data. Likewise, Jesús Martín and Jaime Martín ("Universidad de Córdoba"; and "Innovolivo", Córdoba, Spain) and Antonio Escolano ("Centro de Ensayos de Evaluación de Variedades", INIA, Madrid) for additional garlic-samples. Teresa Hernández-Gutiérrez is acknowledged for support during sampling and other experimental work, and Jaroslava Ovesná (Crop Research Institute, Prague, Czechia) for comments on garlic genotyping.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2017.00098/full#supplementary-material

FIGURE S1 | Garlic dendrogram. Phylogenetic tree, with approximately unbiased (AU; red)/Bootstrap Probability (BP; green) percentage values and Euclidean distances, generated by complete-linkage method, to ascertain germplasm diversity. Cluster I includes C1 and C2 (Chinese varieties); Cluster II has C4 (Spanish White variety); and Cluster III shows C3 to C5 (Spanish Purple and Brazilian varieties). Samples C1 to C5, and others described in the text, are highlighted with colored dots. I corresponds to cluster II in STRUCTURE analysis, whereas II and III are equivalent to cluster I; and A and B correspond to cluster III using such software analysis.

FIGURE S2 | Garlic genetic structure. STRUCTURE software was used to analyze the studied garlic germplasm. **(a)** Diagram showing the three calculated clusters ($K = 3$); and **(b)** ΔK values.

REFERENCES

Akpinar, B., Lucas, S., and Budak, H. (2017). A large-scale chromosome-specific SNP discovery guideline. *Funct. Integr. Genom.* 17, 97–105. doi: 10.1007/s10142-016-0536-6

Al-Zahim, M., Ford-Lloyd, B., and Newbury, H. (1999). Detection of somaclonal variation in garlic (*Allium sativum* L.) using RAPD and cytological analysis. *Plant Cell Rep.* 18, 473–477. doi: 10.1007/s002990050606

Arumuganathan, K., and Earle, E. D. (1991). Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* 9, 208–218. doi: 10.1007/BF02672069

Atienza, S. G., de la Rosa, R., Domínguez-García, M. C., Martín, A., Kilian, A., and Belaj, A. (2013). Use of DArT markers as a means of better management of the diversity of olive cultivars. *Food Res. Int.* 54, 2045–2053. doi: 10.1016/j.foodres.2013.08.015

Belaj, A., Dominguez-García, M. D. C., Atienza, S. G., Urdíroz, N. M., Rosa, R. D., Satovic, Z., et al. (2011). Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. *Tree Genet. Genomes* 8, 365–378. doi: 10.1007/s11295-011-0447-6

Cardelle-Cobas, A., Soria, A. C., Corzo-Martinez, M., and Villamiel, M. (2010). "A comprehensive survey of garlic functionality," in *Garlic Consumption and Health*, eds M. Pacurar and G. Krejci (Hauppauge: Nova Science Publishers, Inc), 1–60.

Carlsson, K., and Frick, G. (1964). Partial purification of nucleases from germinating garlic. *Biochim. Biophys. Acta* 81, 301–310. doi: 10.1016/0926-6569(64)90046-x

Corzo-Martínez, M., Corzo, N., and Villamiel, M. (2007). Biological properties of onions and garlic. *Trends Food Sci. Technol.* 18, 609–625. doi: 10.1016/j.tifs.2007.07.011

Courtois, B., Audebert, A., Dardou, A., Roques, S., Ghneim-Herrera, T., Droc, G., et al. (2013). Genome-wide association mapping of root traits in a japonica rice panel. *PLoS ONE* 8:e78037. doi: 10.1371/journal.pone.0078037

Cruz, V. M. V., Kilian, A., and Dierig, D. A. (2013). Development of DArT marker platforms and genetic diversity assessment of the US collection of the new oilseed crop lesquerella and related species. *PLoS ONE* 8:e64062. doi: 10.1371/journal.pone.0064062

DaCunha, C. P., Resende, F. V., Zucchi, M. I., and Pinheiro, J. B. (2014). SSR-based genetic diversity and structure of garlic accessions from Brazil. *Genetica* 142, 419–431. doi: 10.1007/s10709-014-9786-1

Defays, D. (1977). Efficient algorithm for a complete link method. *Comput. J.* 20, 364–366. doi: 10.1093/comjnl/20.4.364

Dorado, G., Besnard, G., Unver, T., and Hernández, P. (2015a). "Polymerase chain reaction (PCR)," in *Reference Module in Biomedical Sciences*, ed. M. Caplan (Amsterdam: Elsevier).

Dorado, G., Gálvez, S., Budak, H., Unver, T., and Hernández, P. (2015b). "Nucleic-acid sequencing," in *Reference Module in Biomedical Sciences*, ed. M. Caplan (Amsterdam: Elsevier).

Dorado, G., Unver, T., Budak, H., and Hernández, P. (2015c). "Molecular markers," in *Reference Module in Biomedical Sciences*, ed. M. Caplan (Amsterdam: Elsevier).

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379. doi: 10.1371/journal.pone.0054603

Gebhardt, C. (2013). Bridging the gap between genome analysis and precision breeding in potato. *Trends Genet.* 29, 248–256. doi: 10.1016/j.tig.2012.11.006

Govindaraj, M., Vetriventhan, M., and Srinivasan, M. (2015). Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. *Genet. Res. Int.* 2015, 431487–431487. doi: 10.1155/2015/431487

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* 27, 857–871. doi: 10.2307/2528823

Green, E. (2001). Strategies for the systematic sequencing of complex genomes. *Nat. Rev. Genet.* 2, 573–583. doi: 10.1038/35084503

Gupta, P. K., Rustgi, S., and Mir, R. R. (2008). Array-based high-throughput DNA markers for crop improvement. *Heredity* 101, 5–18. doi: 10.1038/hdy.2008.35

He, N., Li, Q., Sun, D., and Ling, X. (2008). Isolation, purification and characterization of superoxide dismutase from garlic. *Biochem. Eng. J.* 38, 33–38. doi: 10.1016/j.bej.2007.06.005

Hernandez, P., de la Rosa, R., Rallo, L., Martin, A., and Dorado, G. (2001). First evidence of a retrotransposon-like element in olive (*Olea europaea*): implications in plant variety identification by SCAR-marker development. *Theor. Appl. Genet.* 102, 1082–1087. doi: 10.1007/s001220000515

Ipek, A., Yilmaz, K., Sikici, P., Tangu, N., Oz, A., Bayraktar, M., et al. (2016). SNP discovery by GBS in olive and the construction of a high-density genetic linkage map. *Biochem. Genet.* 54, 313–325. doi: 10.1007/s10528-016-9721-5

Ipek, M., Ipek, A., Almquist, S. G., and Simon, P. W. (2005). Demonstration of linkage and development of the first low-density genetic map of garlic, based on AFLP markers. *Theor. Appl. Genet.* 110, 228–236. doi: 10.1007/s00122-004-1815-5

Ipek, M., Sahin, N., Ipek, A., Cansev, A., and Simon, P. (2015). Development and validation of new SSR markers from expressed regions in the garlic genome. *Sci. Agric.* 72, 41–46. doi: 10.1590/0103-9016-2014-0138

Jaccoud, D., Peng, K., Feinstein, D., and Kilian, A. (2001). Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Res.* 29:E25. doi: 10.1093/nar/29.4.e25

Jansky, S. H., Dawson, J., and Spooner, D. M. (2015). How do we address the disconnect between genetic and morphological diversity in germplasm collections? *Am. J. Bot.* 102, 1213–1215. doi: 10.3732/ajb.1500203

Jones, M. G., Hughes, J., Tregova, A., Milne, J., Tomsett, A. B., and Collin, H. A. (2004). Biosynthesis of the flavour precursors of onion and garlic. *J. Exp. Bot.* 55, 1903–1918. doi: 10.1093/jxb/erh138

Kamenetsky, R., Faigenboim, A., Mayer, E., Ben Michael, T., Gershberg, C., Kimhi, S., et al. (2015). Integrated transcriptome catalogue and organ-specific profiling of gene expression in fertile garlic (*Allium sativum* L.). *BMC Genomics* 16:12. doi: 10.1186/s12864-015-1212-2

Kamenetsky, R., Shafir, I. L., Baizerman, M., Khassanov, F., Kik, C., and Rabinowitch, H. D. (2004). Garlic (*Allium sativum* L.) and its wild relatives from Central Asia: evaluation for fertility potential. *Adv. Vegetable Breed.* 83–91.

Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., et al. (2012). Diversity arrays technology: a generic genome profiling technology on open platforms. *Methods Mol. Biol. (Clifton, NJ)* 888, 67–89. doi: 10.1007/978-1-61779-870-2_5

Kim, A., Kim, R. N., Kim, D.-W., Choi, S-H., Kang, A., Nam, S.-H., et al. (2010). Identification of a novel garlic cellulase gene. *Plant Mol. Biol. Rep.* 28, 388–393. doi: 10.1007/s11105-009-0159-3

Kim, D.-W., Jung, T.-S., Nam, S.-H., Kwon, H.-R., Kim, A., Chae, S.-H., et al. (2009). GarlicESTdb: an online database and mining tool for garlic EST sequences. *BMC Plant Biol.* 9:61. doi: 10.1186/1471-2229-9-61

Liu, J., Wang, J., Yin, M., Zhu, H., Lu, J., and Cui, Z. (2011). Purification and characterization of superoxide dismutase from garlic. *Food Bioprod. Process.* 89, 294–299. doi: 10.1016/j.fbp.2010.07.003

Maab, H. I., and Klaas, M. (1995). Infraspecific differentiation of garlic (*Allium sativum* L.) by isozyme and RAPD markers. *Theor. Appl. Genet.* 91, 89–97.

Meredith, T. (2008). *The Complete Book of Garlic: A Guide for Gardeners, Growers, and Serious Cooks*. Portland: Timber Press.

Murray, M. G., and Thompson, W. F. (1980). Rapid isolation of high molecular-weight plant DNA. *Nucleic Acids Res.* 8, 4321–4325. doi: 10.1093/nar/8.19.4321

Nybom, H., and Bartish, I. (2000). Effects of life history traits and sampling strategies on genetic diversity estimates obtained with RAPD markers in plants. *Perspect. Plant Ecol. Evol. Syst.* 3, 93–114. doi: 10.1078/1433-8319-00006

Olukolu, B. A., Mayes, S., Stadler, F., Ng, N. Q., Fawole, I., Dominique, D., et al. (2012). Genetic diversity in *Bambara groundnut* (*Vigna subterranea* (L.) Verdc.) as revealed by phenotypic descriptors and DArT marker analysis. *Genet. Res. Crop Evol.* 59, 347–358. doi: 10.1007/s10722-011-9686-5

Ovesná, J., Leišová-Svobodová, L., and Kučera, L. (2014). Microsatellite analysis indicates the specific genetic basis of Czech bolting garlic. *Czech J. Genet. Plant Breed.* 50, 226–234.

Ovesna, J., Mitrova, K., and Kucera, L. (2015). Garlic (*A. sativum* L.) alliinase gene family polymorphism reflects bolting types and cysteine sulphoxides content. *BMC Genet.* 16:53. doi: 10.1186/s12863-015-0214-z

Pacurar, M., and Krejci, G. (eds). (2010). *Garlic Consumption and Health*. New York, NY: Nova Science Publishers.

Parisi, M., Moreno, S., and Fernandez, G. (2008). Isolation and characterization of a dual function protein from *Allium sativum* bulbs which exhibits proteolytic and hemagglutinating activities. *Plant Physiol. Biochem.* 46, 403–413. doi: 10.1016/j.plaphy.2007.11.003

Park, J. (2009). Isolation and characterization of N-Feruloyltyramine as the P-selectin expression suppressor from garlic (*Allium sativum*). *J. Agric. Food Chem.* 57, 8868–8872. doi: 10.1021/jf9018382

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

R-Development-Core-Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Raman, H., Raman, R., Kilian, A., Detering, F., Carling, J., Coombes, N., et al. (2014). Genome-wide delineation of natural variation for pod shatter resistance in *Brassica napus*. *PLoS ONE* 9:e101673. doi: 10.1371/journal.pone.0101673

Rana, S., Pal, R., Vaiphei, K., Sharma, S., and Ola, R. (2011). Garlic in health and disease. *Nutr. Res. Rev.* 24, 60–71. doi: 10.1017/S0954422410000338

Rathnasamy, S., Auxilia, L. R., and Purusothaman. (2014). Comparative studies on isolation and characterization of allinase from garlic and onion using PEGylation-a novel method. *Asian J. Chem.* 26, 3733–3735.

Rotem, N., Shemesh, E., Peretz, Y., Akad, F., Edelbaum, O., Rabinowitch, H., et al. (2007). Reproductive development and phenotypic differences in garlic are associated with expression and splicing of LEAFY homologue gaLFY. *J. Exp. Bot.* 58, 1133–1141. doi: 10.1093/jxb/erl272

Shemesh-Mayer, E., Ben-Michae, T., Rotem, N., Rabinowitch, H., Doron-Faigenboim, A., Kosmala, A., et al. (2015*). Garlic (*Allium sativum* L.) fertility: transcriptome and proteome analyses provide insight into flower and pollen development. *Front. Plant Sci.* 6:271. doi: 10.3389/fpls.2015.00271

Smeets, K., Van Damme, E., Van Leuven, F., and Peumans, W. (1997). Isolation and characterization of lectins and lectin-alliinase complexes from bulbs of garlic (*Allium sativum*) and ramsons (*Allium ursinum*). *Glycoconj. J.* 14, 331–343. doi: 10.1023/A:1018570628180

Suetsuna, K. (1998). Isolation and characterization of angiotensin I-converting enzyme inhibitor dipeptides derived from *Allium sativum* L (garlic). *J. Nutr. Biochem.* 9, 415–419. doi: 10.1016/S0955-2863(98)00036-9

Suzuki, R., and Shimodaira, H. (2015). *pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling*. Available at: http://stat.sys.i.kyoto-u.ac.jp/prog/pvclust/

Tanksley, S. D., and McCouch, S. R. (1997). Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* 277, 1063–1066. doi: 10.1126/science.277.5329.1063

Volk, G. M., Henk, A. D., and Richards, C. M. (2004). Genetic diversity among U.S. Garlic clones as detected using AFLP methods. *J. Am. Soc. Hortic. Sci.* 129, 559–569.

Wang, H., Li, X., Liu, X., Oiu, Y., Song, J., and Zhang, X. (2016). Genetic diversity of garlic (*Allium sativum* L.) germplasm from China by fluorescent-based AFLP, SSR and InDel markers. *Plant Breed.* 135, 743–750. doi: 10.1111/pbr.12424

Zhao, W. G., Chung, J. W., Lee, G. A., Ma, K. H., Kim, H. H., Kim, K. T., et al. (2010). Molecular genetic diversity and population structure of a selected core set in garlic and its relatives using novel SSR markers. *Plant Breed.* 130, 46–54. doi: 10.1111/j.1439-0523.2010.01805.x

Conflict of Interest Statement: AK works at Diversity Arrays Technology. This fact did not interfere with the objective, transparent and unbiased presentation of results, and does not alter the authors' adherence to all theoretical and applied genetics policies on data and material release.

The other authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 June 2017
doi: 10.3389/fgene.2017.00081

Complementary Network-Based Approaches for Exploring Genetic Structure and Functional Connectivity in Two Vulnerable, Endemic Ground Squirrels

Victoria H. Zero[1†], Adi Barocas[2,3†], Denim M. Jochimsen[4], Agnès Pelletier[5], Xavier Giroux-Bougard[6], Daryl R. Trumbo[7], Jessica A. Castillo[8], Diane Evans Mack[9], Mark A. Linnell[8], Rachel M. Pigg[10], Jessica Hoisington-Lopez[11], Stephen F. Spear[12], Melanie A. Murphy[3,13] and Lisette P. Waits[14]*

[1] Haub School of Environment and Natural Resources, University of Wyoming, Laramie, WY, United States, [2] Department of Zoology and Physiology, University of Wyoming, Laramie, WY, United States, [3] Program in Ecology, University of Wyoming, Laramie, WY, United States, [4] Department of Biological Sciences, University of Idaho, Moscow, ID, United States, [5] Department of Environmental Studies and Sciences, University of Winnipeg, Winnipeg, MB, Canada, [6] Department of Natural Resource Sciences, McGill University, Montreal, QC, Canada, [7] School of Biological Sciences, Washington State University, Pullman, WA, United States, [8] Department of Fisheries and Wildlife, Oregon State University, Corvallis, OR, United States, [9] Idaho Department of Fish and Game, McCall Subregion, McCall, ID, United States, [10] Division of Biology, Kansas State University, Manhattan, KS, United States, [11] The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO, United States, [12] The Wilds, Cumberland, OH, United States, [13] Department of Ecosystem Science and Management, University of Wyoming, Laramie, WY, United States, [14] Department of Fish and Wildlife Sciences, University of Idaho, Moscow, ID, United States

The persistence of small populations is influenced by genetic structure and functional connectivity. We used two network-based approaches to understand the persistence of the northern Idaho ground squirrel (*Urocitellus brunneus)* and the southern Idaho ground squirrel (*U. endemicus*), two congeners of conservation concern. These graph theoretic approaches are conventionally applied to social or transportation networks, but here are used to study population persistence and connectivity. Population graph analyses revealed that local extinction rapidly reduced connectivity for the southern species, while connectivity for the northern species could be maintained following local extinction. Results from gravity models complemented those of population graph analyses, and indicated that potential vegetation productivity and topography drove connectivity in the northern species. For the southern species, development (roads) and small-scale topography reduced connectivity, while greater potential vegetation productivity increased connectivity. Taken together, the results of the two network-based methods (population graph analyses and gravity models) suggest the need for increased conservation action for the southern species, and that management efforts have been effective at maintaining habitat quality throughout the current range of the northern species. To prevent further declines, we encourage the continuation of management efforts for the northern species, whereas conservation of the southern species requires active management and additional measures to curtail habitat fragmentation. Our combination of population graph analyses and gravity models can inform conservation strategies of other species exhibiting patchy distributions.

Keywords: functional connectivity, gene flow, graph theory, gravity model, landscape genetics, Sciuridae, *Urocitellus* [Spermophilus]

OPEN ACCESS

Edited by:
Marshall Abrams, University of Alabama at Birmingham, United States

Reviewed by:
Matthew Joseph Michalska-Smith, University of Chicago, United States
Jennifer Leonard, Consejo Superior de Investigaciones Científicas (CSIC), Spain

***Correspondence:**
Victoria H. Zero
vzero@uwyo.edu

†These authors have contributed equally to this work.

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *26 August 2016*
Accepted: *29 May 2017*
Published: *14 June 2017*

Citation:
Zero VH, Barocas A, Jochimsen DM, Pelletier A, Giroux-Bougard X, Trumbo DR, Castillo JA, Evans Mack D, Linnell MA, Pigg RM, Hoisington-Lopez J, Spear SF, Murphy MA and Waits LP (2017) Complementary Network-Based Approaches for Exploring Genetic Structure and Functional Connectivity in Two Vulnerable, Endemic Ground Squirrels. Front. Genet. 8:81. doi: 10.3389/fgene.2017.00081

INTRODUCTION

Habitat loss and fragmentation are threats to many species of conservation concern (Wilcox et al., 1985; Groombridge, 1992). These agents of landscape change decrease the size and structural connectivity of habitat patches, with consequences for long-term population viability and species distributions (Kareiva and Wennergren, 1995; Fahrig, 2002). Decreased animal movement, and subsequent reduction in gene flow, can lead to isolated populations and constricted species ranges (Andrews, 1990; Yahner and Mahan, 1997; Fahrig, 2002). Over time, reduced gene flow can decrease population size, alter population dynamics, and lower persistence probability (Meffe and Carroll, 1997; Ovaskainen and Hanski, 2003). Isolated populations typically have low levels of genetic variation (Frankham, 1997) that inhibit adaptation in the face of environmental change (Lande, 1988) and increase vulnerability to inbreeding depression (Frankham, 1995; Hedrick, 2005) and local extinction (Burkey, 1995; Frankham et al., 2002).

As the long-term persistence of populations in fragmented landscapes depends on functional connectivity, or how individuals respond to landscape composition (Tischendorf and Fahrig, 2000; Stevens et al., 2006), research that assesses the effects of landscape and ecological features on gene flow serves as a valuable conservation tool (McRae et al., 2008). The spatial context and composition of habitat patches generally have profound influences on animal movement beyond the effect of geographical distance alone (Ricketts, 2001). Landscape genetic methods are particularly suited to test how environmental context influences patterns of genetic variation and gene flow across temporal and spatial scales (Manel et al., 2003; Storfer et al., 2007; Holderegger and Wagner, 2008), and have recently been strengthened by the integration of graph theoretic approaches (Garroway et al., 2008; Murphy et al., 2016). These approaches provide a mathematical framework in which researchers can represent populations or sites as "nodes" and connections between them as "edges," and then evaluate patterns of connectivity to identify environmental factors underlying gene flow (Dyer and Nason, 2004; Garroway et al., 2008; McRae et al., 2008; Dyer et al., 2010; Murphy et al., 2010).

Graph theory can be used to assess functional connectivity, and may therefore provide important information for conservation planning. Network metrics such as degree centrality and betweenness (Everett and Borgatti, 2005) measure the relative contribution of sampled sites to overall population connectivity, and thus can pinpoint the best locations for conservation or management actions. Gravity models (Fotheringham and O'Kelly, 1989) can simultaneously evaluate the relative influence of geographic distance, local attributes of sampling locations (at-site characteristics), and the features that separate them (between-site characteristics) on gene flow (Murphy et al., 2010). Typical landscape genetic network models do not include the influence of local attributes. By including at-site characteristics in these models, we can incorporate additional factors contributing to gene flow by quantifying how habitat patches differ in quality (Ovaskainen and Hanski, 2003). Patches of higher quality habitat may produce more offspring and thereby contribute disproportionately to gene flow. Gravity models can help determine how landscapes should be managed to maintain connectivity and improve patch quality, and network metrics can identify where managers should focus conservation efforts.

The northern Idaho ground squirrel (*Urocetillus brunneus*; NIDGS) and the southern Idaho ground squirrel (*U. endemicus*; SIDGS) are two congeners of conservation concern. NIDGS and SIDGS are endemic to west-central Idaho and were originally classified as two subspecies (Yensen, 1991) but were recently elevated to distinct species based on genetic differences (U.S. Fish and Wildlife Service, 2015), morphology, behavior, and distinct geographic and ecological niches (Hoisington-Lopez et al., 2012). These species occur in small, discrete populations within a fragmented landscape (Van Horne et al., 2007; Yensen et al., 2008). Consequently, population graph analysis and gravity models can lend insight into factors affecting their population connectivity.

Their ranges are restricted and fragmented; both species have experienced population declines and reductions in the number and total area of sites occupied (Sherman and Runge, 2002; U.S. Fish and Wildlife Service, 2003; Yensen et al., 2008; Lohr et al., 2013). In recent years, the number of occupied locations and subpopulations has remained relatively stable, while the number of mature individuals appears to fluctuate according to several-year cycles (Evans Mack, personal communication). For example, between 2011 and 2016, overall population size ranged between just under 1,000 and over 2,500 individuals. Consequently, the United States Fish and Wildlife Service listed NIDGS as threatened in 2000 (Clark, 2000), while SIDGS was a candidate for listing until just recently (Federal Register, November 22, 2013 Vol. 68, No. 226:77 70103-7016). Primary threats to NIDGS include the loss of preferred habitat to ponderosa pine (*Pinus ponderosa*) encroachment due to fire suppression (Yensen and Sherman, 1997; Gavin et al., 1999; Sherman and Runge, 2002), and competition with the Columbian ground squirrel (*Urocitellus columbianus*; Dyni and Yensen, 1996). The latter species occurs throughout central Idaho, potentially overlapping populations of both Idaho ground squirrels. Declines in SIDGS are attributed to the invasion of non-native annual plants, including cheatgrass (*Bromus tectorum*) and medusahead (*Taeniatherum asperum*), which have increased fire frequency and intensity with subsequent shifts in vegetation composition (Yensen, 1991; Lohr et al., 2013).

The loss and degradation of preferred habitat have consequences for the long-term persistence of remaining NIDGS and SIDGS populations. Population divergence has been detected for NIDGS using allozymes (Gavin et al., 1999), and for both species using mitochondrial DNA (Yensen and Sherman, 1997; Garner et al., 2005; Hoisington-Lopez et al., 2012) and microsatellite data (NIDGS: $0.03 < F_{ST} < 0.46$; SIDGS: $0.04 < F_{ST} < 0.43$; Garner et al., 2005; Hoisington-Lopez et al., 2012). In addition, both species have low to moderate levels of genetic diversity (allelic richness, expected heterozygosity, and haplotype diversity; Garner et al., 2005; Hoisington-Lopez et al., 2012) that are likely a consequence of isolation and bottleneck events. The effects of landscape and environmental variables

on genetic diversity and connectivity of NIDGS and SIDGS have not been evaluated in depth. Understanding the ecological drivers underlying site productivity and factors facilitating gene flow among habitat patches is a critical conservation need for both species. Identifying sites that contribute the most to functional connectivity is also essential for making conservation and management decisions.

Our primary goal was to quantify functional connectivity among NIDGS and SIDGS populations and identify sites contributing the most to gene flow to help inform conservation and management efforts. We aimed to evaluate functional connectivity for each species using genetic patterns and identify at-site and between-site variables influencing gene flow. We hypothesized that the production of potential migrants from a site would be affected by forage availability as indicated by local climate measures. For NIDGS, availability of meadow (i.e., grassland) should be an important factor in population connectivity since this preferred habitat has been reduced by forest encroachment. For SIDGS, highly developed areas (measured by impervious surfaces) should reduce functional connectivity due to potential movement barriers and the likely increased incidence of non-native plants species. We also examined topographic complexity, waterways, soils, and competition from Columbian ground squirrels (*U. columbianus*) as potential drivers of functional connectivity.

MATERIALS AND METHODS

Study Area and Species

We examined the functional connectivity of northern and southern Idaho ground squirrel populations from 23 sites within 5 counties located in west-central Idaho (**Figure 1**, Table S1). No new field or genetic data were collected for this study. All procedures for initial data collection were approved by the University of Idaho Animal Care and Use Committee (2006-35), Idaho Fish and Game state permit (060308), and federal permit for *U. brunneus* (subpermit FWSSRBO-5). Extant, sampled NIDGS and SIDGS populations were previously determined by methods described by Yensen (1991). Mean sampling location area (±SE) was 0.44 ± 0.21 km^2 for NIDGS and 0.42 ± 0.11 km^2 for SIDGS. The study area includes the geographically discrete ranges of both species, extending between the Salmon and Payette Rivers. NIDGS inhabit mid to high elevations (1,150–2,300 m) in xeric, montane meadows, and grasslands surrounded by coniferous forests (Yensen, 1991; Yensen and Sherman, 1997). SIDGS occur at lower elevations (670–975 m) in sagebrush and bitterbrush habitats with interspersed perennial bunchgrasses and forbs (Hafner et al., 1998; IDFG unpublished data). The majority of habitat is under public ownership, with private land primarily at lower elevations (U.S. Fish and Wildlife Service, 2013). Land use includes logging, agriculture, grazing, and suburban developments (Yensen et al., 2008).

Genetic Data

We obtained multilocus, microsatellite genotypes from previous studies that, cumulatively, sampled most IDGS populations comprised of more than 10 individuals (**Figure 1**; Supplementary Data Sheet 1, Garner et al., 2005; Hoisington-Lopez et al., 2012). NIDGS sampling occurred in 2002 and 2006, and included 316 individuals from 13 locales (Table S1; Hoisington-Lopez et al., 2012). We excluded one NIDGS population from this study (Round Valley), as previous results indicate that it is both geographically and genetically isolated from all other NIDGS populations, and thus could lead to spurious correlations with landscape variables (Cushman and Landguth, 2010). SIDGS sampling consisted of 263 individuals in 2002 and 2006 from 11 locations (Table S1). When samples were collected at the location in multiple years, we tested for differences in allele frequency distributions before combining data (Hoisington, 2007). We used data from previously published microsatellite loci (n = 8) in Hardy–Weinberg equilibrium that showed no linkage disequilibrium (Hoisington-Lopez et al., 2012).

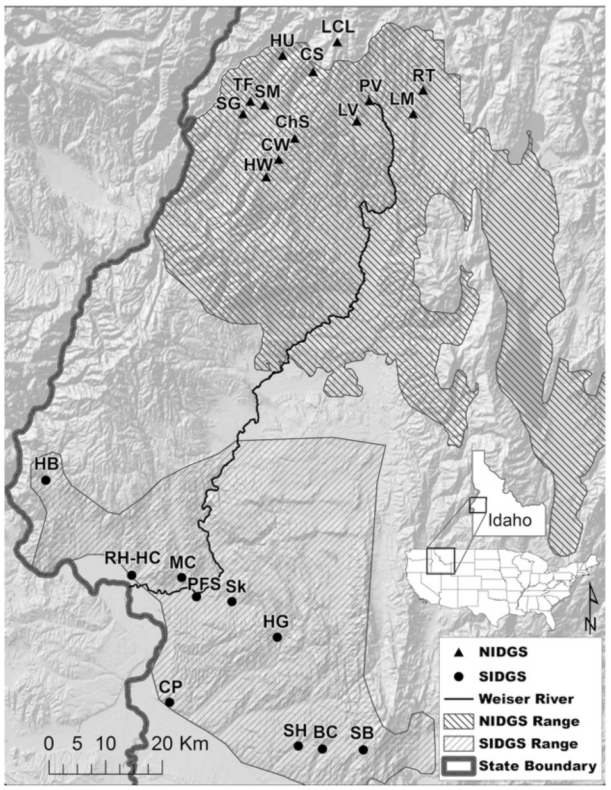


FIGURE 1 | Sampling locations of genetic data for northern Idaho ground squirrel (*Urocetillus brunneus*; NIDGS) and southern Idaho ground squirrel (*U.endemicus*; SIDGS). Also shown are the NIDGS probable historic distribution (U.S. Fish and Wildlife Service, 2003) and the current known range for SIDGS (Idaho Game and Fish Department). Individuals were sampled from 2002 to 2006 (Hoisington-Lopez et al., 2012). Background hillshade map was produced from the National Elevation Dataset (http://ned.usgs.gov). Full site names and sample sizes can be found in Table S1.

For each species, we calculated three measures of genetic distance between populations to serve as response variables in network models: (1) the proportion of shared alleles (D_{PS}; Bowcock et al., 1994), calculated using Microsatellite Analyzer 4.05 (Dieringer and Schlötterer, 2003); (2) conditional genetic

distance (cGD; Dyer et al., 2010), using Genetic Studio (Dyer, 2009) in R (gstudio 0.8, R Core Development Team, 2012); and (3) the fixation index (F_{ST}), a commonly used measure of population structure, calculated using Fstat (Goudet, 1995). D_{PS} is not subject to the equilibrium assumptions inherent in divergence (F_{ST}) measures, and thus may be more appropriate for measuring genetic connectivity among populations subject to recent disturbance, and cGD has been shown to outperform F_{ST} in some situations. Furthermore, cGD focuses only on population pairs that exhibit conditional dependence with one another and thus are likely to be directly exchanging migrants, and ignores population pairs that are conditionally independent and likely not directly exchanging migrants (Dyer et al., 2010). For each species, we additionally performed a Mantel test (Smouse et al., 1986) to examine the correlation of geographic distance with the two relevant metrics of genetic distance, D_{PS} and F_{ST} (Table S2).

Population Graph Analysis

To conduct the population graph analyses, we used cGD (Dyer et al., 2010), a metric that calculates the distance between each pair of nodes, thereby accounting for the genetic covariance in the whole network. The method examines pairwise correlations in inverted cGD values among sampling locations and draws an edge between two nodes if the partial correlation between them is significantly higher than expected by chance. The subsequent pruned graph contains the minimal number of edges which will sufficiently describe the total covariance structure among populations (Dyer et al., 2010). Because pruned networks are more information than saturated networks (Dyer and Nason, 2004), we kept them for subsequent analyses.

To help guide conservation actions, we determined the number of significant genetic units (genetic clusters) using two community detection methods, which identify "communities" of more highly connected nodes (Girvan and Newman, 2002). The first, Girvan-Newman uses an optimization procedure based on eigenvalues to calculate the support for different cluster numbers in terms of modularity (Q; the existence of non-overlapping groups of nodes in the network). The best-supported model of community division receives the highest modularity value (Newman, 2006). The second, the *Walktrap* algorithm, finds subgraphs of more densely connected nodes based on random walks and also calculates overall modularity (Pons and Latapy, 2006). To perform these analyses, we built a binary network for each species.

To determine the relative contribution of each sampling location to overall gene flow, we investigated the network topologies of both species by calculating four network metrics for each node: (1) degree centrality—the number of connections that each node has in the network (Everett and Borgatti, 2005), (2) strength centrality—the sum of all association indices (i.e., weighted connections among nodes) that each node has in the network (Garroway et al., 2008), (3) betweenness—the number of shortest paths that a particular node or edge lies on, which can identify bottlenecks (Everett and Borgatti, 2005), and (4) coreness—an algorithm that tests for the existence of a core/periphery structure in the network and calculates the location of each node in relation to the core. Based on the number of core nodes, we additionally calculated a concentration score (ranging from 0 to 1) which quantifies how close the network is to an idealized core-periphery model, in which all nodes in the core are connected within the core and to the periphery nodes and all nodes in the periphery are not connected (Borgatti and Everett, 1999). In the context of genetic networks, the coreness of a node can be interpreted as the extent to which it acts as a source for dispersing individuals. Sampling location abbreviations are presented in Table S1.

To examine the vulnerability of each species to local extinction, we assessed network sensitivity to node removal (**Figure 2**). Node removal simulates local patch extinction, a recurrent event in species that exhibit metapopulation structure (Hanski, 1998). We sequentially removed random nodes to generate up to 100 population graphs for each scenario (e.g., 1, 2, 3 nodes removed). For each of the simulated graphs, we assessed overall gene flow using two metrics: (1) Proportion of fully connected graphs, quantifying the extent to which the population graph will become fragmented as a result of node removal; (2) Size of the largest graph component, measuring the maximal number of nodes that retained connectivity among them. We calculated this metric proportional to the total network size. We built 95% confidence intervals, based on standard errors, around the proportional size of the largest component for each node removal scenario.

Gravity Models

We used gravity models (Fotheringham and O'Kelly, 1989; Murphy et al., 2010) to analyze the effects of abiotic and biotic variables on population connectivity. We modeled gene flow [1-genetic distance (D_{PS})] as a function of geographic (Euclidean) distance (w), attributes of nodes (v), and landscape resistance factors (c) that limit or facilitate movement of individuals between nodes (Murphy et al., 2010). We developed a set of *a priori* hypotheses to describe ecologically relevant processes affecting at-site production of migrants and between-site landscape resistance for both species (**Table 1**).

We used 30 m landcover data from the LANDFIRE Existing Vegetation Type dataset, and used our between-site calculations to assess habitat permeability (http://landfire.cr.usgs.gov/viewer). We extracted the landcover data for grassland, shrubland, agriculture, and impervious surfaces (i.e., roads and developed areas). We then calculated percent cover for each cover type within a 90 × 90 pixel moving window. We calculated surface relief ratio (*srr*; Evans, 1972) from 10 m Shuttle Radar Topography Mission digital elevation models using two neighborhood sizes (3 × 3 and 27 × 27 pixels), to assess topographic resistance to gene flow. We used the Geomorphometry and Gradient Metrics Toolbox (http://evansmurphy.wix.com/evansspatial#!arcgis-gradient-metrics-toolbox/crro) in ArcMap 10.2. We tested 6 biotic and abiotic variables hypothesized to affect at-site production/attraction (v) of IDGS migrants, such as climate, soil type, vegetation cover, and inter-specific competition. For landscape resistance between sites (c), we developed a set of 6 abiotic and biotic variables that relate to habitat permeability, topography, hydrologic complexity, and road density. For between-site variables, we calculated the

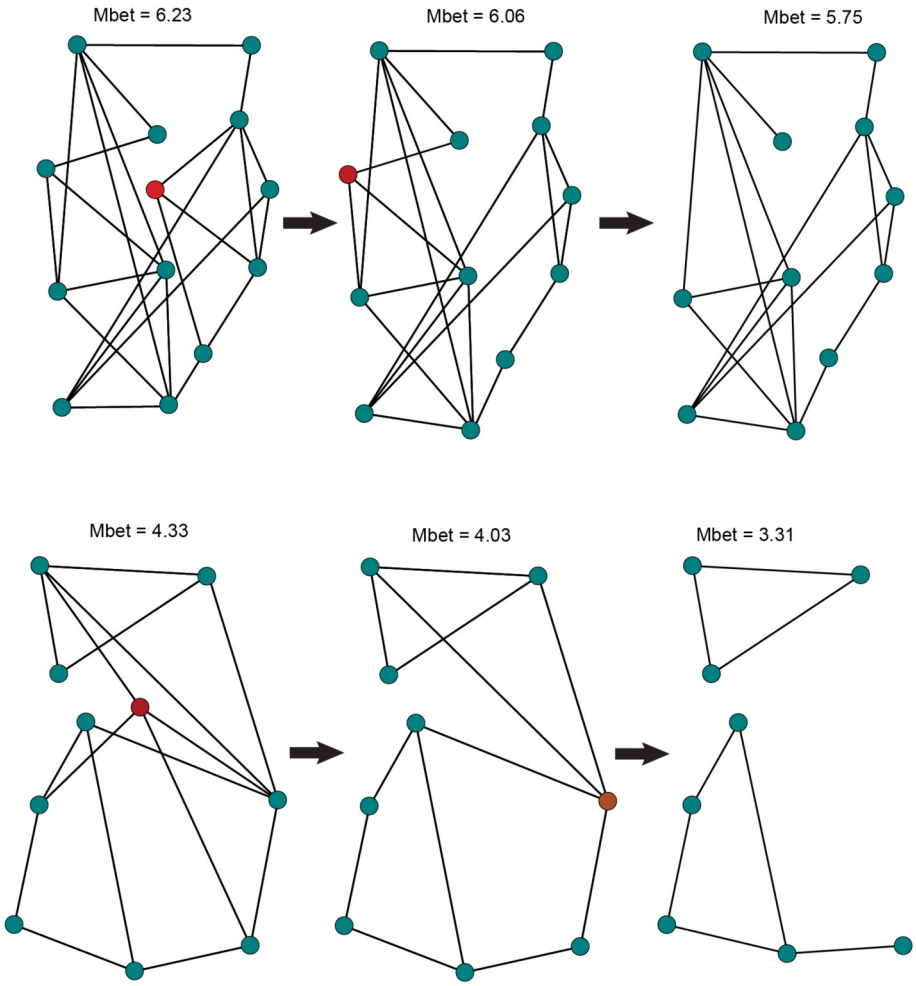


FIGURE 2 | Illustration of the node removal procedure used to simulate population extinction events in NIDGS (top) and SIDGS (bottom). In each step, a randomly selected node (in red), representing a sampling location, is removed from the network along with the edges connecting it to additional nodes. Network mean betweenness values are given on top. Following the removal of 2 nodes, the SIDGS network fails to create a single component, becoming fragmented.

average or variance along each edge (30 m width) connecting populations in the network. We also tested for the effect of spatial scale of each variable by building buffers along edges of 30, 150, and 300 m widths, and then calculating between-site values within each buffer (Murphy et al., 2010). Since each of these metrics was highly correlated with the along-line calculations ($R^2 > 0.8$), we used straight-line, 30 m width edge results for these metrics.

In spatially explicit genetic networks, incomplete sampling of nodes can lead to bias when using a pruned graph (Naujokaitis-Lewis et al., 2013). Given the small number of locations sampled for each species, we retained the fully connected networks for the gravity modeling procedure. Gravity models were run in R using the GeNetIt package. We used a hierarchical modeling approach to compare models that included one or more landscape variables with a distance-only (null) model. We used singly constrained models as they account for non-independence of pairwise comparisons. Gravity models were solved in mixed effects linear models using maximum likelihood (Zuur et al., 2009). We specified at-site and between-site variables as fixed effects and the identities of nodes as random effects (Murphy et al., 2010). We initially ran a null (distance) model and subsequently modeled at-site variables and between-site variables separately. We then built combined gravity models that included both classes of variables, via the inclusion of the best-supported, at-site and between-site variables identified during the first procedure. To avoid co-linearity, models did not include pairs of candidate variables correlated at 0.7 or higher (Table S5). We used Akaike information criterion scores adjusted for small sample size (AICc) to identify the best-supported models (Akaike, 1973; Burnham and Anderson, 2002). We additionally calculated conditional (including both fixed and random factors) R^2 values for each model (Nakagawa and Schielzeth, 2013). We plotted network flow against each variable identified as significant in the best-supported models to assess the direction of the effects of candidate variables. We subsequently calculated

TABLE 1 | Independent variables tested in candidate gravity models to explain functional connectivity of two species of Idaho ground squirrels.

Parameter	Process	Variable	Code	Predicted effect	Description	Ecological justification	Source	Species modeled
Geographic distance (w)	IBD	Geographic distance	*w*	–	Euclidean distance (m)	IBD previously detected in SIDGS (Hoisington, 2007).	SRTM	NIDGS* SIDGS*
At-site production (v)	Habitat	Vegetation	*shrub, grass*	+	Percent cover meadow (NIDGS) or shrubland (SIDGS)	An increased proportion of meadow habitat could yield larger populations (Hoisington-Lopez et al., 2012).	NLCD	NIDGS SIDGS
		Soil texture	*soil*	+	Percent loam	IDGS are associated with well-drained, loamy soils (Yensen and Sherman, 1997).	NRCS	SIDGS
	Productivity	Heat load index	*hli*	+	Measure of solar intercept as derived from slope aspect (McCune and Keon, 2002)	Slope and aspect influence forage production; data suggest inherent preferences (U.S. Fish and Wildlife Service, 2003; Lohr et al., 2013).	SRTM	NIDGS* SIDGS*
		Frost-free period	*ffp*	+	Length of frost-free period	Longer frost-free periods indicative of higher plant productivity (Hoisington-Lopez et al., 2012).	Spline	NIDGS SIDGS*
		Growing season precipitation	*gsp*	+	Precipitation received from April to September	Annual precipitation important in ecological niche models (Hoisington-Lopez et al., 2012). Growing season precipitation may be more directly related to plant productivity.	Spline	NIDGS SIDGS*
	Competition	Columbian ground squirrel distribution	*cogs*	–	Percent overlap of CGS core areas with NIDGS core areas	NIDGS may be competitively inferior in meadows where CGS occur (Yensen, 1991).	IDFG	NIDGS
Between-site resistance (c)	Topography	Topographic complexity	*srr3, srr27*	–	Elevation relief ratio (Evans, 1972)	Fine scale topographic complexity (3 × 3) may make movement energetically costly. Large-scale complexity (27 × 27) may represent barriers.	SRTM	NIDGS* SIDGS*
	Habitat	Land cover shrub, grass	*shrub, grass*	+	Proportion of intervening shrub and grassland	Meadows provide suitable habitat for NIDGS (Yensen and Sherman, 1997). Suitable burrowing and foraging habitat between populations could enhance dispersal opportunities.	NLCD	NIDGS SIDGS
		Land cover agriculture	*agri*	–	Proportion of intervening agricultural and developed areas	Cultivated areas could limit dispersal and are associated with human activity.	NLCD	SIDGS
		Ephemeral streams	*eph_strm*	+	Stream density	Ephemeral and intermittent streams may act as dispersal corridors when dry (Roach et al., 2001).	NHD	NIDGS SIDGS

(Continued)

TABLE 1 | Continued

Parameter	Process	Variable	Code	Predicted effect	Description	Ecological justification	Source	Species modeled
	Barriers	Impervious surfaces	*imperv*	–	Percent imperviousness	Road traffic may cause mortalities on all road types. Paved roads may represent habitat loss and adjacent areas may be altered.	NLCD	SIDGS*
		Perennial streams, rivers	*per_strm*	–	Stream density	Streams and rivers may represent absolute barriers to dispersal. The Weiser River was identified as a barrier to SIDGS gene flow (unpublished data).	NHD	SIDGS

*Parameter: the parameter estimated in the gravity equation [distance (w), production (v), and resistance (c)]. Process: the landscape process influencing gene flow: isolation-by-distance (IBD), habitat quality/permeability, site productivity, interspecific competition, topography, or geographic barriers. Predicted relationship: expectation of a positive (+) or negative (–) relationship between the independent variable and gene flow. Source: source of data containing the variable or used to derive the variable. NLCD, National Land Cover Database (2001); NRCS, Natural Resources Conservation Service; IDFG, Idaho Department of Fish and Game; SRTM, Shuttle Topographic Radar Mission digital elevation model; a climate spline model (Spline; Rehfeldt, 2006), and NHD: National Hydrography Dataset. Species modeled: NIDGS, northern Idaho ground squirrel; SIDGS, southern Idaho ground squirrel. Asterisks indicate variables that were found to be important in gravity models (**Table 2**).*

cumulative AIC weight for each variable by summing the weights of each model in which this variable was included (Burnham and Anderson, 2002).

RESULTS

Population Graph Analysis

In the cGD pruning procedure, population graphs retained a total of 24 (31% of saturated network) edges connecting 13 nodes for NIDGS and 16 (36% of saturated network) edges connecting 10 nodes for SIDGS (**Figure 3**). We identified support for 2 and 3 genetic clusters via Girvan–Newman and Walktrap [modularity scores: $Q(2)_{\text{Girvan}-\text{Newman}} = 0.33$; $Q(2)_{\text{Walktrap}} = 0.34$, $Q(3) = 0.33$] for NIDGS. The 2-cluster model included 1 cluster in the northwestern portion of NIDGS range and a second cluster in the southeastern portion. Both algorithms agreed on all sampling location cluster assignments except study site LCL. For SIDGS, the model with 3 clusters received the highest modularity score [modularity scores: $Q(2) = 0.26$, $Q(3)_{\text{Girvan}-\text{Newman}} = 0.29$; $Q(3)_{\text{Walktrap}} = 0.30$]. Modularity, reflecting compartmentalization within each network, was slightly lower in SIDGS compared to NIDGS. The 3-cluster model included a cluster in the northwestern portion of the species' distribution, separated by the Weiser River and the agricultural area surrounding it from 2 discrete clusters, located in the southern and central area of the range. There was no evident spatial segregation between the southern and central clusters.

In NIDGS, the node strength centrality and betweenness metrics suggested higher connectivity for the western populations (Table S3). The coreness analysis provided the best support for a model with 5 nodes at the core and 8 at the periphery. For the 5-node core model, the concentration score was 0.91. The 5 core nodes, corresponding to the CW (betweenness = 15.08), HU (13.08), SG (2.25), SM (9.41), and HW (11.08) populations were located in the western portion of the range, confirming the patterns suggested by the other network metrics (Table S3). Spatial patterns were less evident in the network topology analysis of SIDGS. Sampling locations RH-HC (17.50), HG (29.44), SB (10.32), and CP (14.17), representing distinct areas of the species' range and all 3 genetic clusters, had the highest betweenness (Table S3). The core/periphery model results revealed that the optimal model included 3 nodes at the core (corresponding to HG, SB, and Sk) with a concentration score of 0.84.

The node removal simulation analysis indicated that in the range of 2–5 removed nodes, SIDGS networks had higher probability of fragmentation by not creating a fully connected component (e.g., 3 nodes removed: NIDGS—97%, SIDGS—66% fully connected networks; **Figure 4A**). This larger fragmentation probability resulted in the largest components in SIDGS proportionally consisting of fewer nodes compared to NIDGS (e.g., 3 nodes removed: NIDGS—0.99, 95% CI = 0.98–1.0; SIDGS—92%, 95% CI = 0.9–0.94; **Figure 4B**). Taken together, both network connectivity metrics indicated higher resilience of the NIDGS network to node removal.

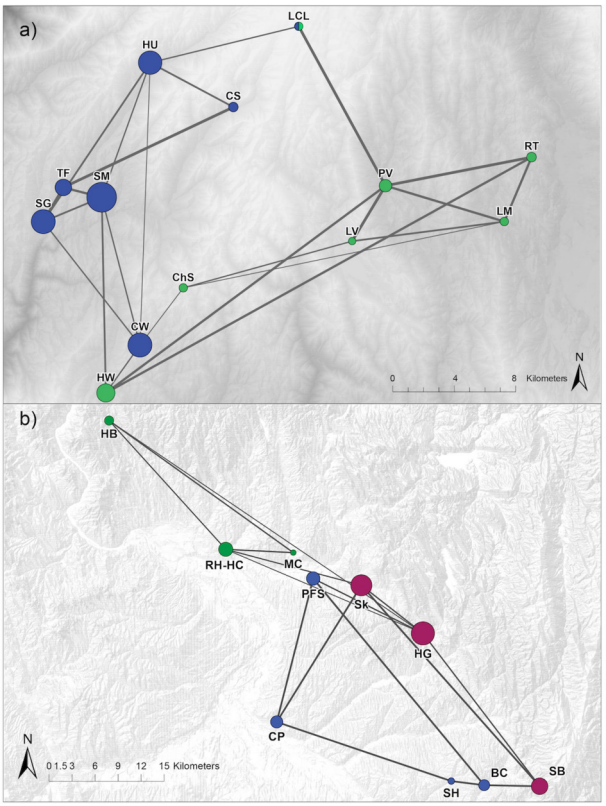


FIGURE 3 | Network diagrams representing the genetic relationships between northern **(a)** and southern **(b)** Idaho ground squirrel sampling locations. Networks are pruned using conditional genetic distance (cGD; Dyer and Nason, 2004). Individuals were sampled during 2002–2006. Node colors differ by cluster assignment with the Girvan-Newman algorithm. Nodes are placed according to geographic location and scaled to reflect coreness, a network metric that quantifies proximity to the core in a core/periphery model (Table S3). Edge width is proportional to the genetic flow between sampling locations.

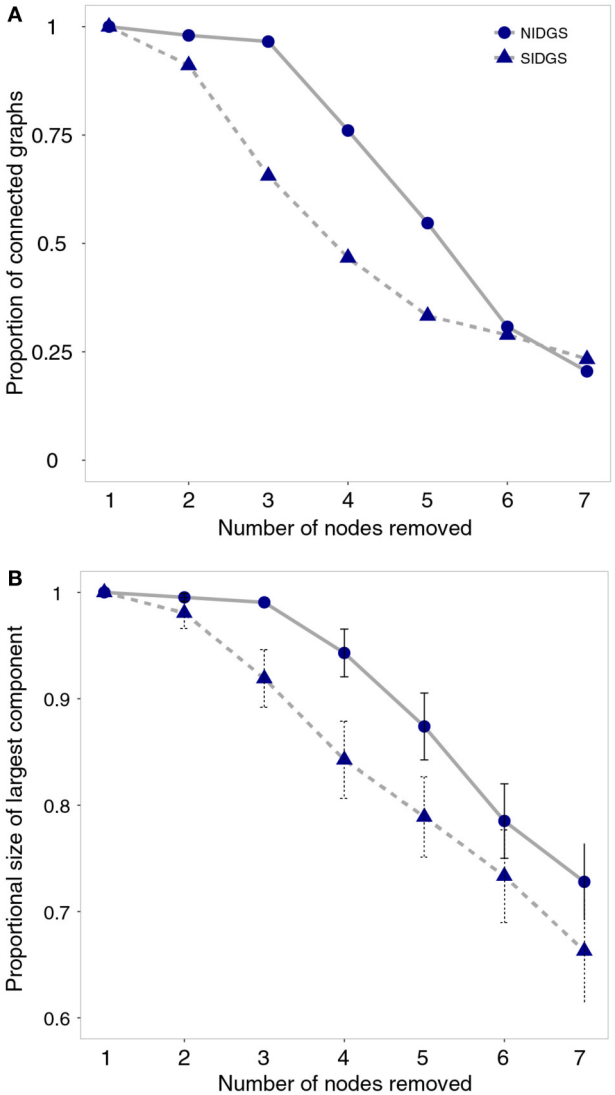


FIGURE 4 | The effects of random sequential removal of nodes on proportion of connected graphs **(A)** and the largest remaining network component **(B)** for northern Idaho ground squirrels (solid line) and southern Idaho ground squirrels (dashed line). Proportions, means and 95% confidence intervals were calculated for up to 100 simulated networks in each scenario. Idaho ground squirrels were sampled from 23 locations during 2002–2006 and genotyped using 8 microsatellite loci (Hoisington-Lopez et al., 2012).

Gravity Models

The mean geographic distance between sampled locations for NIDGS and SIDGS was 16.1 and 28.9 km, respectively. For NIDGS subpopulations, the pairwise genetic distance (D_{PS}) averaged (±SE) 0.41 ± 0.01, ranging from 0.23 to 0.56. SIDGS subpopulations had an average D_{PS} of 0.34 ± 0.02, ranging from 0.17 to 0.53. F_{ST} values were more similar among species, with means ±SE 0.19 ± 0.01 in NIDGS (range: 0.03–0.48) and 0.18 ± 0.01 for SIDGS (range: 0.04–0.41). For both species, pairwise genetic distance metrics were highly correlated (NIDGS $r = 0.89$; SIDGS $r = 0.96$). Mantel test results indicated a significant correlation between geographic distance and both metrics of genetic distance for NIDGS (D_{PS}: $r = 0.39$, $P = 0.001$; F_{ST}: $r = 0.38$, $P = 0.002$), and a stronger pattern in SIDGS (D_{PS}: $r = 0.64$, $P < 0.001$; F_{ST}: $r = 0.57$, $P < 0.001$).

The top variables for the NIDGS saturated network included those associated with potential site productivity (*v*: *hli, gsp*) and topography (*c*: *srr27, srr3*; **Table 2**, **Figure 5**). For NIDGS, geographic distance (*w*) was the sixth-ranked model, with a ΔAICc of 4.4. Heat load index (*hli*) positively correlated with gene flow and had the greatest support among at-site variables (variable weight: 0.63), while growing season precipitation had a weight of 0.06. One additional at-site variable, frost-free period (*ffp*: 0.04), received some variable weight but did not appear in those models that improved on the distance-only model. Measures of large-scale (*srr27*: 0.54) and small-scale (*srr3*: 0.08) topographic complexity negatively correlated to gene flow, and were the between-site variables with the greatest weights. Variables describing land cover, interspecific competition, and human disturbance received negligible support (Table S3). For individual parameter estimates, see Table S6.

TABLE 2 | Gravity model results of the best-supported models for northern and southern Idaho ground squirrels.

Species	Full model description	Type	ΔAICc	AIC weight	Conditional R^2
northern Idaho ground squirrel	*w + hli − srr27*	at + between	0.0	0.33	0.40
	w − srr27	between	1.3	0.17	0.41
	w + hli	at	2.5	0.09	0.36
	w + hli − srr3	at + between	3.9	0.05	0.37
	w + gsp + hli − srr27	at + between	4.2	0.04	0.39
	w	distance	4.4	0.04	0.38
southern Idaho ground squirrel	*w + ffp + hli − imperv − srr3*	at + between	0	0.42	0.47
	w + gsp + hli − imperv − srr3	at + between between	0.5	0.33	0.47
	w + gsp − imperv − srr3	at + between between	3.6	0.07	0.46
	w + hli − imperv − srr3	at + between	3.9	0.06	0.44
	w + ffp − imperv − srr3	at + between	5.3	0.03	0.46
	w + ffp + hli − agri − srr3	at + between	6.7	0.01	0.44
	w − imperv − srr3	between	6.9	0.01	0.44
	w + gsp + hli − agri − srr3	at + between	7.3	0.01	0.44
	w	distance	15	0.00	0.37

Type indicates whether the model includes at-site, between-site, or both categories of predictors. A full list of models in available in Tables S3, S4.

For SIDGS, gene flow was positively correlated with at-site productivity and negatively correlated with between-site factors associated with reduced landscape permeability including human disturbance (*imperv*, *agri*) and small-scale topographic complexity (*srr3*; **Table 2**). For SIDGS, heat load index (*hli*) at sites was positively correlated with gene flow (variable weight: 0.83). Growing season precipitation (*gsp*: 0.42) and frost-free period (*ffp*: 0.46) also positively related to gene flow. Small-scale topographic complexity (*srr3*: 0.95) appeared in all top models and negatively correlated with gene flow. Impervious surfaces appeared in six of eight top models, contributing 92% AIC weight, and was negatively correlated with gene flow. Agricultural areas impeded gene flow, but this land cover type received minimal weight (*agri*: 0.02). Variables describing land cover classes and stream densities received negligible support (Table S4).

DISCUSSION

We combined two graph theoretic approaches to enhance our understanding of the functional connectivity of two Idaho ground squirrel species and to inform conservation efforts. Population graph analysis revealed that the pattern and strength of network connectedness differed by species. Node removal simulations suggested that in the event of local patch extinction, SIDGS would likely lose connectivity rapidly, while NIDGS would maintain gene flow despite the removal of several patches or nodes. Gravity models revealed the influence of at-site productivity variables in both species, a finding that would not have been detected in traditional network approaches. These models also revealed effects of topographic complexity at two different spatial scales: fine-scale variation for SIDGS and broad-scale and fine-scale variation for NIDGS. Development, as measured by impervious surfaces, was a major hindrance to SIDGS gene flow.

Patterns of Genetic Structure

We found support for 2 or 3 genetic clusters in NIDGS and similar support for 3 genetic clusters in SIDGS using network community detection (Newman, 2006). Functional connectivity among habitat patches in NIDGS appears to be limited by a mountain ridge, with subpopulations clustered in the northwestern and southeastern portions of the range, and this result is similar to that obtained via STRUCTURE (Pritchard et al., 2000; Hoisington, 2007). However, one site (CW) located in the southwestern portion of the species' range (**Figure 3**), deviated from this pattern. Interestingly, the population graph links CW to 3 populations in the northwestern cluster, and the 2 populations spatially adjacent to CW (ChS and HW) in the southeastern genetic cluster. This pattern, which is consistent with mitochondrial DNA analyses (Hoisington, 2007), could be explained by repeated translocations of individuals from SM and surrounding sites into CW (Gavin et al., 1999) as well as natural recolonization from HW. For SIDGS, our population graph detected a unique genetic cluster in the northern portion of its range, composed of 3 populations. This suggests that the Weiser River acts as a barrier to gene flow as suggested in previous analyses (Garner et al., 2005; Hoisington, 2007). The optimal model for SIDGS had a modularity score lower than 0.3. Our results reveal genetic connectivity across the southern portion of the SIDGS range despite considerable fragmentation due to agriculture. In general, there was congruence between the community detection results and previous Bayesian clustering analyses (Hoisington, 2007). Population graph community detection algorithms base their calculations on genetic distances among nodes and thus have the advantage of including the contribution of ancestries from other genetic clusters. The

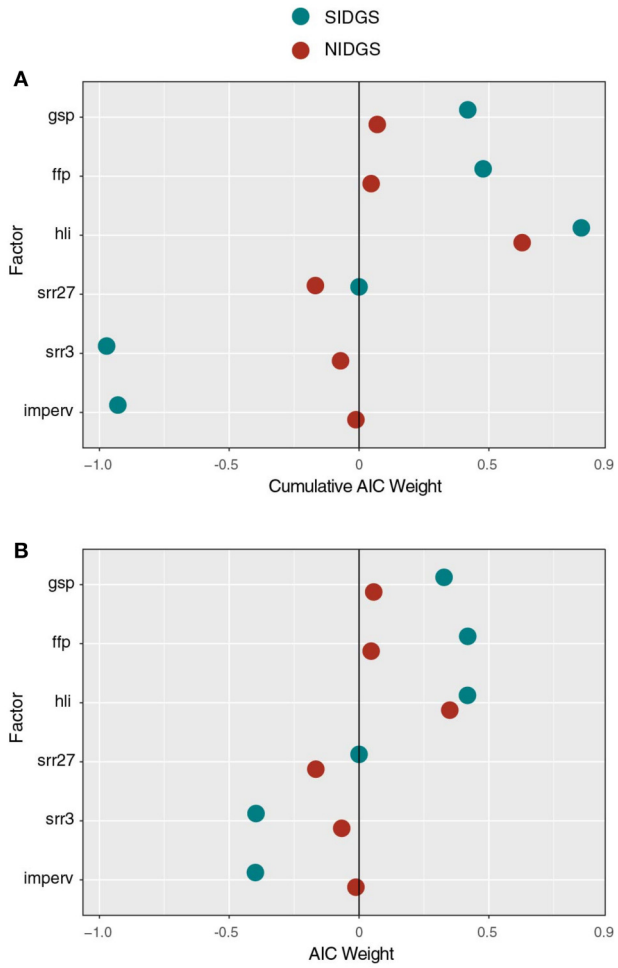


FIGURE 5 | Cumulative AICc weights of several factors examined with gravity models as candidates for driving NIDGS (red circles) and SIDGS (blue circles) gene flow. Panels include cumulative AICc weights for each factor in all examined models **(A)** and only the top-ranked model for each factor **(B)**. Positive and negative values represent the direction and magnitude of relative support for each factor's influence.

similarity in results suggests that these methods are well suited for our study system and applicable in additional systems where genetic data can be represented as population graphs.

Network topology metrics, calculated at the sampling location level, were consistent with these patterns. NIDGS nodes with the highest strength and betweenness were the ones that belonged to the core according to the core-periphery model. One exception was PV, which is spatially central and highly connected, but did not constitute a core node. All NIDGS core nodes were found in the western portion of the species' range. In addition, the majority of edges among these nodes were retained in the pruned network (**Figure 3A**). This may indicate that the western portion of the range is a source for dispersing individuals. With the exception of the CW population, the NIDGS population graph topology indicates a west-to-east gradient of connectivity. In SIDGS, nodes with high overall connectivity according to degree, strength, and betweenness, such as CP and PFS, were not included in the core and had relatively lower coreness (**Figure 3B**; Table S3). Core nodes (HG, SB, and Sk) all belonged to the same genetic cluster and were located in the southeastern part of the species' range. Interestingly, the most spatially central locations (PFS and MC) were not the most connected ones, suggesting that there are additional factors driving SIDGS gene flow beyond geographic distance.

The correlation of network structure to the idealized core/periphery model was slightly higher for NIDGS (0.91) compared to SIDGS (0.84), as was the proportion of core nodes. This slight difference may be explained by the lack of spatial organization in the SIDGS network (**Figure 3**). Overall, our population graph analyses indicate that gene flow among NIDGS locations is higher compared to SIDGS, which is consistent with the relatively large geographic distances found among SIDGS populations. Our use of core/periphery models to assess genetic data is a novel application of a methodology previously developed for social networks (Borgatti and Everett, 1999), and provides an additional metric to quantify node contribution, which may reflect the degree to which discrete sites are sources or sinks for dispersers.

Simulated node removals indicated an immediate decline in overall connectivity among SIDGS nodes, compared to the relative robustness to node removal in NIDGS (**Figure 4**), suggesting that the few connections retained in the SIDGS population graph have an increased conservation value for this species. In addition, these results imply that local extinction of 2 current subpopulations would drive a substantial decline in functional connectivity. SIDGS occur in areas prone to intense human activity and subpopulations are separated by large geographic distances. Our results highlight the susceptibility of this species to future habitat loss and fragmentation, and raise concern over further isolation of the remaining subpopulations. In contrast, simulated node removal in the NIDGS population graph suggests that this species is relatively robust to localized extinctions. The pruned population graph retained a similar proportion of edges, comprised of shorter distances, among subpopulations compared to the SIDGS graph. These results, in conjunction with lower levels of human disturbance across the NIDGS range, suggest that in the event of local extinctions the species may be better able to maintain population connectivity (Fahrig, 2002; Driscoll, 2004).

Functional Connectivity

The variables that were important in gravity models differed between species. We predicted that at-site variables associated with potential productivity would be positively correlated with functional connectivity for both species. Population size at each site would likely be an important predictor of gene flow, but these data were not available. However, population estimates are relevant to the conservation of both species and should be a priority for data collection. We also hypothesized that between-site variables indicative of high habitat quality would facilitate gene flow, while variables reflecting human activity would inhibit gene flow.

Model fit, as measured by conditional R^2, was moderate for both species ($R^2 \sim 0.4$). These results could be an artifact of our limited power to detect variation in habitat variables across the study areas (Short Bull et al., 2011), especially in light of the small number of extant populations occurring over restricted ranges (**Figure 1**).

Nevertheless, a number of at-site variables were identified as predictors of gene flow. Heat load index (*hli*), a surrogate for vegetation productivity, was one at-site variable that contributed to gene flow in both species. This metric had substantial cumulative AIC weight across models of NIDGS ($w = 0.63$) and SIDGS ($w = 0.83$) connectivity. Sites with a higher *hli* may yield a larger number of squirrels with improved body conditions due to increased forage availability and quality. The finding that NIDGS are primarily structured, apart from isolation by distance, by at-site productivity, would have been difficult to detect with other landscape genetic statistical approaches. Additionally, two other at-site variables associated with potential productivity facilitated gene flow for both species. Longer frost-free periods and increased growing-season precipitation were associated with higher connectivity. Lohr et al. (2013) reported that the greatest densities of SIDGS were associated with higher cover of perennial grasses, native perennial forbs, and higher plant species diversity. The combination of solar intercept (*hli*), long growing season (*ffp*), and greater rainfall (*gsp*) may result in high forage quality and quantity for ground squirrels. Therefore, at-site vegetation production is likely an important characteristic in maintaining viable populations for both species.

Landscape features that restricted gene flow differed for the two species. The population graph results for NIDGS revealed a division between the western and eastern sampling areas that are geographically separated by a mountain ridge. This is mirrored in the gravity model results, for which large-scale topographic complexity (*srr27*) received 54% weight across models. At this broad scale, *srr* is likely detecting ridges as a filter to movement, and this pattern is visually apparent when the graph of population structure is overlaid on topography (**Figure 3A**). Three landscape features were identified as barriers to gene flow for SIDGS: impervious surfaces, small-scale topographic complexity, and, to a minor extent, agriculture. Populations were less connected in highly developed areas as measured by imperviousness of surfaces along edges connecting nodes. Impervious surfaces primarily reflect the presence of roads. Gene flow could be disrupted across roads due to avoidance of high traffic areas or altered roadside habitat, increased mortality from vehicle collisions, or a combination of these factors. Although roads are often considered an important source of mortality for many wildlife species (Forman, 1998), small mammals may select these areas (Oxley et al., 1974), and the effects on small mammal behavior and movement may be contingent on road type and traffic volume (Brock and Kelt, 2004). Previous results indicate that dispersing Idaho ground squirrels repeatedly use dirt roads as corridors (Panek, 2005). The absence of support for road effects on NIDGS could be attributed to lower densities of high-volume traffic (paved) roads surrounding the sampling sites for this species. The negative impact of agricultural areas on gene flow may imply an avoidance of these areas, although the variable weight for this metric was small.

Restriction of gene flow in both species due to small-scale topographic complexity (*srr3*) likely reflects a preference for low-elevation, flat grasslands characteristic of the meadows. Gravity models failed to show any support for either ephemeral or perennial streams as drivers of gene flow (Tables S3, S4). However, our population graph analysis identified the Weiser River as a likely barrier to gene flow. Thus, our inability to detect an important barrier to gene flow with gravity models was supplemented by the results from our population graph analysis. These complementary results highlight the benefits of using multiple analytical methods for detecting patterns in genetic data.

Conservation Implications

Our findings of differences in functional connectivity and its drivers highlight the need for different conservation and management strategies for each species of Idaho ground squirrel. Results from the node removal analysis suggest that NIDGS populations are more connected and relatively resistant to metapopulation collapse from local population extinctions. Although, SIDGS are no longer a candidate for federal listing, their subpopulations may be more susceptible to future habitat loss and fragmentation than NIDGS (Hoisington-Lopez et al., 2012). Connectivity in NIDGS was driven mainly by potential site productivity and topographic characteristics, and not a lack of suitable habitat. These combined lines of evidence suggest that recent conservation efforts for NIDGS have been effective at maintaining this species' gene flow and diversity, and should therefore be continued.

Our results for southern Idaho ground squirrels suggest this species is extremely vulnerable. SIDGS sites are geographically distant from one another and highly sensitive to node removal (i.e., local extinction). Sites that are poorly connected, and thus unlikely to be recolonized following an extirpation event, may be good candidates for reintroduction. Additionally, sites that are highly connected might be examined for landscape characteristics that could be used as part of novel site reintroduction selection criteria. Translocations have been attempted with apparent success for SIDGS (Yensen and Tarifa, 2012), and these efforts, combined with supplementation from captive breeding, may become important for maintaining genetic connectivity and diversity in SIDGS populations (Hoisington-Lopez et al., 2012). Given the distances that separate SIDGS sites, we support the recommendation of Garner et al. (2005) that managers consider establishing additional populations to serve as stepping stones for connectivity. Our gravity model results suggest that factors relating to at-site vegetation productivity affect SIDGS genetic structure. A large amount of SIDGS habitat is located either in agricultural areas or sites dominated by invasive cheatgrass, both of which may be difficult to restore. While it appears that NIDGS have responded positively to habitat restoration, this strategy is less likely to successfully improve SIDGS habitat and functional connectivity due to the pervasive invasion of exotic weeds in their range (Yensen, 1991).

CONCLUSIONS

When working with species of conservation concern, it is important not only to assess genetic structure, but also to identify the factors that influence genetic connectivity. Here, we illustrate the value of using recently developed network-based approaches to examine functional connectivity for two vulnerable species of Idaho ground squirrels. Population graphs enhanced our understanding of each species' resistance to potential future loss of habitat patches or populations. Gravity models provided new insights into landscape-related processes that drive genetic structure of these imperiled species, particularly by identifying at-site influences on gene flow. We conclude that the combination of these methodologies allows stronger inference and a more complete assessment of genetic structure. Network models are especially advantageous for representing gene flow in species exhibiting patchy distributions. We encourage further exploration of these methodologies as a framework for hypothesis testing in future landscape genetics studies.

AUTHOR CONTRIBUTIONS

MM and LW conceived the research. JH collected data. VZ, AB, DJ, AP, XG, DT, RP, and JH analyzed data. All authors contributed to the writing of the manuscript.

FUNDING

Funding agencies included the American Genetics Association, the Idaho Department of Fish and Game, the US Fish and Wildlife Service, the University of Idaho Center for Research on Invasive Species and Small Populations, and Wyoming NASA Space Grant Consortium (NASA Grant #NNX10 AO95H).

ACKNOWLEDGMENTS

This project was conducted as part of the Landscape Genetics Distributed Graduate Course. We would like to thank HH. Wagner for her contributions to the course and synthesis meeting, the students and faculty that participated in the meeting for their constructive comments, and advisors and lab mates who contributed support and insightful discussion while working on the manuscript. In addition, we would like to thank M. Gould and K. Lohr for their collaboration on early stages of the project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2017.00081/full#supplementary-material

REFERENCES

Akaike, H. (1973). "Information theory as an extension of the maximum likelihood principle," in *Second International Symposium Information Theory*, eds B. N. Petrov and F. Csaki (New York, NY: Springer), 267–281.

Andrews, A. (1990). Fragmentation of habitat by roads and utility corridors: a review. *Aust. Zool.* 26, 130–141. doi: 10.7882/AZ.1990.005

Borgatti, S. P., and Everett, M. G. (1999). Models of core/periphery structures. *Soc. Netw.* 21, 375–395. doi: 10.1016/S0378-8733(99)00019-2

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R., and Cavalli-Sforza, L. L. (1994). High-resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368, 455–457. doi: 10.1038/368455a0

Brock, R. E., and Kelt, D. A. (2004). Influence of roads on the endangered Stephens' kangaroo rat (*Dipodomys stephensi*): are dirt and gravel roads different? *Biol. Conserv.* 118, 633–640. doi: 10.1016/j.biocon.2003.10.012

Burkey, T. V. (1995). Extinction rates in archipelagoes: implications for populations in fragmented habitats. *Conserv. Biol.* 9, 527–541. doi: 10.1046/j.1523-1739.1995.09030527.x

Burnham, K., and Anderson, A. (2002). *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*. New York, NY: Springer-Verlag.

Clark, J. R. (2000). Endangered and threatened wildlife and plants; determination of threatened status for the northern Idaho ground squirrel. *Fed. Regist.* 65, 17779–17786.

Cushman, S. A., and Landguth, E. L. (2010). Scale dependent inference in landscape genetics. *Landsc. Ecol.* 25, 967–979. doi: 10.1007/s10980-010-9467-0

Dieringer, D., and Schlötterer, C. (2003). Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Mol. Ecol. Notes* 3, 167–169. doi: 10.1046/j.1471-8286

Driscoll, D. A. (2004). Extinction and outbreaks accompany fragmentation of a reptile community. *Ecol. Appl.* 14, 220–240. doi: 10.1890/02-5248

Dyer, R. J. (2009). GeneticStudio: a suite of programs for spatial analysis of genetic-marker data. *Mol. Ecol. Resour.* 9, 110–113. doi: 10.1111/j.1755-0998.2008.02384.x

Dyer, R. J., and Nason, J. D. (2004). Population Graphs: the graph theoretic shape of genetic structure. *Mol. Ecol.* 13, 1713–1727. doi: 10.1111/j.1365-294X.2004.02177.x

Dyer, R. J., Nason, J. D., and Garrick, R. C. (2010). Landscape modelling of gene flow: improved power using conditional genetic distance derived from the topology of population networks. *Mol. Ecol.* 19, 3746–3759. doi: 10.1111/j.1365-294X.2010.04748.x

Dyni, E. J., and Yensen, E. (1996). Dietary similarity in sympatric Idaho and Columbian ground squirrels (*Spermophilus brunneus* and *S. columbianus*). *Northwest Sci.* 70, 99–108.

Evans, I. S. (1972). "General geomorphometry, derivatives of altitude, and descriptive statistics," in *Spatial Analysis in Geomorphology*, ed R. Chorley (New York, NY: Harper & Row), 17–90.

Everett, M. G., and Borgatti, S. P. (2005). "Extending centrality," in *Models and Methods in Social Network Analysis*, eds P. J. Carrington, J. Scott, and S. Wasserman (New York, NY: Cambridge University Press), 57–76.

Fahrig, L. (2002). Effects of habitat fragmentation on the extinction threshold : a synthesis. *Ecol. Appl.* 12, 346–353. doi: 10.2307/3060946

Forman, R. T. T. (1998). Road ecology: a solution for the giant embracing us. *Landsc. Ecol.* 13, 3–5. doi: 10.1023/A:1008036602639

Fotheringham, A., and O'Kelly, M. (1989). *Spatial Interaction Models: Formulation and Applications*. Dordrecht: Kluwer Academic.

Frankham, R. (1995). Inbreeding and extinction: a threshold effect. *Conserv. Biol.* 9, 792–799. doi: 10.1046/j.1523-1739.1995.09040792.x

Frankham, R. (1997). Do island populations have less genetic variation than mainland populations? *Heredity* 78, 311–327. doi: 10.1038/hdy.1997.46

Frankham, R., Ballou, J., and Briscoe, D. (2002). *Introduction to Conservation Genetics*. Cambridge: Cambridge University Press.

Garner, A., Rachlow, J. L., and Waits, L. P. (2005). Genetic diversity and population divergence in fragmented habitats: conservation of Idaho ground squirrels. *Conserv. Genet.* 6, 759–774. doi: 10.1007/s10592-005-9035-3

Garroway, C. J., Bowman, J., Carr, D., and Wilson, P. J. (2008). Applications of graph theory to landscape genetics. *Evol. Appl.* 1, 620–630. doi: 10.1111/j.1752-4571.2008.00047.x

Gavin, T. A., Sherman, P. W., Yensen, E., May, B., and Gavin, A. (1999). Population genetic structure of the northern Idaho ground squirrel (*Spermophilus brunneus brunneus*). *J. Mammal.* 80, 156–168. doi: 10.2307/1383216

Girvan, M., and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7821–7826. doi: 10.1073/pnas.122653799

Goudet, J. (1995). FSTAT (Version 1.2): a computer program to calculate F-statistics. *J. Hered.* 86, 485–486. doi: 10.1093/oxfordjournals.jhered.a111627

Groombridge, B. (1992). *Global Biodiversity: State of the Earth's Living Resources.* New York, NY: Chapman and Hall.

Hafner, D., Yensen, E., and Kirkland, G. (1998). *North American Rodents: Status Survey and Conservation Action Plan*. Gland; Cambridge: IUCN/SSC Rodent Specialist Group.

Hanski, I. (1998). Metapopulation dynamics. *Nature* 396, 41–49.

Hedrick, P. (2005). *Genetics of Populations*. London: Jones and Bartlett.

Hoisington, J. (2007). *Conservation Genetics, Landscape Genetics and Systematics of the Two Subspecies of the Endemic Idaho Groud Squirrel (Spermophilus brunneus).* Master's thesis, University of Idaho, Moscow, ID.

Hoisington-Lopez, J. L., Waits, L. P., and Sullivan, J. (2012). Species limits and integrated taxonomy of the Idaho ground squirrel (*Urocitellus brunneus*): genetic and ecological differentiation. *J. Mammal.* 93, 589–604. doi: 10.1644/11-MAMM-A-021.1

Holderegger, R., and Wagner, H. H. (2008). Landscape genetics. *Bioscience* 58:199. doi: 10.1641/B580306

Kareiva, P., and Wennergren, U. (1995). Connecting landscape patterns to ecosystem and population processes. *Nature* 373, 299–302. doi: 10.1038/373299a0

Lande, R. (1988). Genetics and demography in biological conservation. *Science* 241, 1455–1460. doi: 10.1126/science.3420403

Lohr, K., Yensen, E., Munger, J. C., and Novak, S. J. (2013). Relationship between habitat characteristics and densities of southern Idaho ground squirrels. *J. Wildl. Manage.* 77, 983–993. doi: 10.1002/jwmg.541

Manel, S., Schwartz, M. K., Luikart, G., and Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends Ecol. Evol.* 18, 189–197. doi: 10.1016/S0169-5347(03)00008-9

McCune, B., and Keon, D. (2002). Equations for potential annual direct incident radiation and heat load. *J. Veg. Sci.* 13, 603–606. doi: 10.1111/j.1654-1103.2002.tb02087.x

McRae, B. H., Dickson, B. G., Keitt, T. H., and Shah, V. B. (2008). Using circuit theory to model connectivity in ecology, evolution, and conservation. *Ecology* 89, 2712–2724. doi: 10.1890/07-1861.1

Meffe, G., and Carroll, C. (1997). *Principles of Conservation Biology*. Sunderland, MA: S. A. Inc.

Murphy, M., Dezzani, R., Pilliod, D., and Storfer, A. (2010). Landscape genetics of high mountain frog metapopulations. *Mol. Ecol.* 19, 3634–3649. doi: 10.1111/j.1365-294X.2010.04723.x

Murphy, M., Dyer, R. J., and Cushman, S. A. (2016). "Graph theory and network models in landscape genetics," in *Landscape Genetics: Concepts, Methods, Applications*, eds N. Balkenhol, S. A. Cushman, A. Storfer, and L. P. Waits (West Sussex: John Wiley & Sons, Ltd.), 165–180.

Nakagawa, S., and Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* 4, 133–142. doi: 10.1111/j.2041-210x.2012.00261.x

Naujokaitis-Lewis, I. R., Rico, Y., Lovell, J., Fortin, M. J., and Murphy, M. A. (2013). Implications of incomplete networks on estimation of landscape genetic connectivity. *Conserv. Genet.* 14, 287–298. doi: 10.1007/s10592-012-0385-3

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8577–8582. doi: 10.1073/pnas.0601602103

Ovaskainen, O., and Hanski, I. (2003). Extinction threshold in metapopulation models. *Ann. Zool. Fennici* 40, 81–97.

Oxley, D. J., Fenton, M. B., and Carmody, G. R. (1974). The effects of roads on populations of small mammals. *J. Appl. Ecol.* 11, 51–59. doi: 10.2307/2402004

Panek, K. (2005). *Dispersal, Translocation and Population Connectivity in Fragmented Populations of Southern Idaho Ground Squirrels*. Master's thesis, Boise State University, Boise, ID.

Pons, P., and Latapy, M. (2006). Computing communities in large networks using random walks. *J. Graph Algorithms Appl.* 10, 191–218. doi: 10.7155/jgaa.00124

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

R Core Development Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org

Rehfeldt, G. L. (2006). *A Spline Model of Climate for the Western United States*. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station.

Ricketts, T. H. (2001). The matrix matters: effective isolation in fragmented landscapes. *Am. Nat.* 158, 87–99. doi: 10.1086/320863

Roach, J. L., Stapp, P., Van Horne, B., and Antolin, M. F. (2001). Genetic structure of a metapopulation of black-tailed prairie dogs. *J. Mammal.* 82, 946–959. doi: 10.1644/1545-1542(2001)082<0946:GSOAMO>2.0.CO;2

Sherman, P. W., and Runge, M. C. (2002). Demography of a population collapse: the northern Idaho ground squirrel (*Spermophilus brunneus brunneus*). *Ecology* 83, 2816–2831. doi: 10.1890/0012-9658(2002)083[2816:DOAPCT]2.0.CO;2

Short Bull, R. A., Cushman, S. A., Mace, R., Chilton, T., Kendall, K. C., Landguth, E. L., et al. (2011). Why replication is important in landscape genetics: American black bear in the Rocky Mountains. *Mol. Ecol.* 20, 1092–1107. doi: 10.1111/j.1365-294X.2010.04944.x

Smouse, P. E., Long, J. C., and Sokal, R. R. (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst. Zool.* 35, 627–632. doi: 10.2307/2413122

Stevens, V. M., Verkenne, C., Vandewoestinjne, S., Wesselingh, R. A., and Baguette, M. (2006). Gene flow and functional connectivity in the natterjack toad. *Mol. Ecol.* 15, 2333–2344. doi: 10.1111/j.1365-294X.2006.02936.x

Storfer, A., Murphy, M. A., Evans, J. S., Goldberg, C. S., Robinson, S., Spear, S. F., et al. (2007). Putting the "landscape" in landscape genetics. *Heredity* 98, 128–142. doi: 10.1038/sj.hdy.6800917

Tischendorf, L., and Fahrig, L. (2000). On the usage and measurement of landscape connectivity. *Oikos* 90, 7–19. doi: 10.1034/j.1600-0706.2000.900102.x

U.S. Fish and Wildlife Service (2003). *Recovery Plan for the Northern Idaho Grund Squirrel.* Region 1, U.S. Fish and Wildlife Service. Portland, OR.

U.S. Fish and Wildlife Service (2013). *Review of Native Species that are Candidates for Listing as Endangered or Threatened; Annual Notice of Findings on Resubmitted Petitions; Annual Description of Progress on Listing Actions 77 FR 70103 70162,* Vol. 78, No. 226.

U.S. Fish and Wildlife Service (2015). *Federal Register,* Vol. 80, No. 120. Rules and Regulations.

Van Horne, B., Wolf, J. O., and Sherman, P. W. (2007). "Conservation of ground squirrels," in *Rodent Societies: An Ecological and Evolutionary Perspective*, eds J. O. Wolff and P. W. Sherman (Chicago, IL: University of Chicago Press), 463–471.

Wilcox, B. A., Murphy, D. D., and Jun, N. (1985). Conservation strategy : the effects of fragmentation on extinction. *Am. Nat.* 125, 879–887. doi: 10.1086/284386

Yahner, R. H., and Mahan, C. G. (1997). Behavioral considerations in fragmented landscapes. *Conserv. Biol.* 11, 569–570. doi: 10.1046/j.1523-1739.1997.96322.x

Yensen, E. (1991). Taxonomy and distribution of the Idaho ground squirrel, *Spermophilus brunneus. J. Mammal.* 72, 583–600. doi: 10.2307/1382142

Yensen, E., Hammerson, G., Jefferson, J., and Cannings, S. (2008). "*Spermophilus brunneus,*" in *IUCN Red List of Threatened Species. Version 2013.2* (IUCN). Available online at: www.iucnredlist.org

Yensen, E., and Sherman, P. W. (1997). *Spermophilus brunneus*. *Mamm. Species* 560, 1–5. doi: 10.2307/3504405

Yensen, E., and Tarifa, T. (2012). *Can Southern Idaho Ground Squirrels be Translocated Successfully?* Annual Report Zoo Boise Conserv. Fund, 1–37.

Zuur, A. F., Ieno, E. N., Walker, N., Savelieve, A. A., and Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. New York, NY: Spring Science and Business Media.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 02 October 2015
doi: 10.3389/fpls.2015.00813

Landscape genomics reveal signatures of local adaptation in barley (*Hordeum vulgare* L.)

Tiegist D. Abebe, Ali A. Naz and Jens Léon*

Department of Crop Genetics and Biotechnology, Institute of Crop Science and Resource Conservation, Rhenish Friedrich-Wilhelm University of Bonn, Bonn, Germany

Land plants are sessile organisms that cannot escape the adverse climatic conditions of a given environment. Hence, adaptation is one of the solutions to surviving in a challenging environment. This study was aimed at detecting adaptive loci in barley landraces that are affected by selection. To that end, a diverse population of barley landraces was analyzed using the genotyping by sequencing approach. Climatic data for altitude, rainfall and temperature were collected from 61 weather sites near the origin of selected landraces across Ethiopia. Population structure analysis revealed three groups whereas spatial analysis accounted significant similarities at shorter geographic distances (<40 Km) among barley landraces. Partitioning the variance between climate variables and geographic distances indicated that climate variables accounted for most of the explainable genetic variation. Markers by climatic variables association analysis resulted in altogether 18 and 62 putative adaptive loci using Bayenv and latent factor mixed model (LFMM), respectively. Subsequent analysis of the associated SNPs revealed putative candidate genes for plant adaptation. This study highlights the presence of putative adaptive loci among barley landraces representing original gene pool of the farming communities.

Keywords: landscape genomics, local adaptation, *Hordeum vulgare*, genotyping by sequencing, spatial genetic structure, adaptive loci

OPEN ACCESS

Edited by:
Stéphane Joost, Ecole Polytechnique Fédérale de Lausanne, Switzerland

Reviewed by:
Jacob A. Tennessen, Oregon State University, USA
Sevan Suni, Harvard University, USA
Torsten Günther, Uppsala University, Sweden

****Correspondence:***
Ali A. Naz, Department of Crop Genetics and Biotechnology, Institute of Crop Science and Resource Conservation, Rhenish Friedrich-Wilhelm University of Bonn, Katzenburgweg 5, D-53115 Bonn, Germany
a.naz@uni-bonn.de

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *29 June 2015*
Accepted: *17 September 2015*
Published: *02 October 2015*

Citation:
Abebe TD, Naz AA and Léon J (2015) Landscape genomics reveal signatures of local adaptation in barley (Hordeum vulgare L.). Front. Plant Sci. 6:813. doi: 10.3389/fpls.2015.00813

Introduction

Natural selection is the key evolutionary process that generates the adaptation of plants to their environments (Andrews, 2010). During this, the best fitted alleles to the specific environment become prevalent through positive selection, which is the major driving force behind adaptive evolution in plants (Schaffner and Sabeti, 2008; Bose and Bartholomew, 2013). Genetic identification of those beneficial alleles is essential for answering fundamental questions concerning plant adaptive evolution as well as to utilize them in crop improvement.

Genome-wide scan has been proven to be an effective approach for studying adaptive genetic variation (Nosil et al., 2009). Classically, this approach uses different genotyping protocols to assay a large number of DNA marker polymorphisms across the genome to associate them with different traits and environmental factors (Bonin et al., 2006; Eckert et al., 2010a,b; Brachi et al., 2011; Wang et al., 2012; Westengen et al., 2012). Recently, advances in next generation sequencing technologies have resulted in the development of newer methods of high-throughput genotyping such as genotyping by sequencing (GBS). This method brought out clear advantages

to genotype highly diversified and complex genomes in lesser time and at a low cost per sample (Elshire et al., 2011). GBS generates thousands of sequence tags and single nucleotide polymorphisms (SNPs) across the genome. It has been used successfully in a number of plant species like barley, maize (Elshire et al., 2011; Poland et al., 2012; Larsson et al., 2013), sorghum (Morris et al., 2013), soybean (Sonah et al., 2013), and Brachypodium (Dell'Acqua et al., 2014).

Genome-wide scan generally rely on the assumption that the loci involved in adaptation exhibit stronger differentiation among populations and lower diversity within a population when compared with selectively neutral regions of genome (Storz, 2005). Such loci are considered outlier loci and can be detected among populations using molecular marker data by calculating the population differentiation coefficient (F_{ST}) (Excoffier et al., 2009). Therefore, F_{ST} analysis has the ability to determine the signatures of divergent selection evolving under the pressure of ecological factors. This selection is the fundamental process in adaptive differentiation and speciation among the natural populations of plants (Schluter, 2001, 2009; Funk et al., 2006).

Landscape genomics is a relatively new approach that combines landscape factors and genomics to scan for the presence of a signature of selection (Allendorf et al., 2010; Schoville et al., 2012). This approach attempts to detect the loci that underlie observed adaptive genetic variation and hence called adaptive loci. Currently, there is a growing body of literature demonstrating the feasibility of landscape genomics in detecting loci related to adaptation. For instance, Westengen et al. (2012) detected adaptive loci that respond to the precipitation and maximum temperature of a given habitat by analyzing African maize landrace populations using association analysis. Eckert et al. (2010b) found significant correlations between genetics and climatic variables indicating the evidence of natural selection in loblolly pine (*Pinus taeda* L.). Similarly, Poncet et al. (2010) identified ecological relevant genes linked to minimum temperatures in *Arabis alpina*. Recently, De Kort et al. (2014) reported a clear association among outlier loci, temperature and latitude in the tree species *Alnus glutinous* across Europe. These reports clearly advocate the utility of the landscape genomics in detecting and understanding the adaptive biology of plants. Dell'Acqua et al. (2014) studied local adaptation in Brachypodium and found genes related to environmental adaptation in natural populations. However, until now, the utilization of landscape genomics to dissect the fundamental components of adaptation in crops like wheat and barley has not been studied well.

Ethiopia, with its diverse agro-ecological and climatic features, is known for being one of the 12 Vavilovian centers of diversity (Vavilov, 1951; Harlan, 1969). It contains a tremendous range of altitudes spanning from 110 m below sea level in areas of the Kobar Sink to 4620 meter above sea level (m.a.s.l.) at Ras Dashen. In addition, Ethiopian regions experience huge temperature and rainfall differences, which are coupled with highly variable edaphic factors. This diverse topography and environmental heterogeneity may be the major reasons behind the highly diversified plant species across Ethiopia. These diverse climatic conditions and rich biodiversity make Ethiopia a model environment to dissect the genetic basis of ecological adaptations in plants.

Barley (*Hordeum vulgare* L.) is an important cereal for subsistence farmers in Ethiopia. These farmers typically grow barley without any application of inputs such as fertilizers, pesticides, and insecticides (Lakew et al., 1997). They usually sow their own harvested grain as seeds each year. Sowing their own seeds from year to year, these farmers have established farmer varieties (landraces) that are adapted to different ecological environments across Ethiopia. It is not possible to neglect the role of farmer-driven artificial selection to fit these landraces to a particular ecological condition. However, the prevalence and diverse adaptive differentiation of barley landraces across Ethiopia clearly suggests that these genetic resources have successfully undergone natural selection (Zeven, 1998).

The present study was aimed at detecting the signatures of local adaptation in a state of the art barley population using the landscape genomics approach. Here, we report the first insight into the identification of putative adaptive loci by combining molecular data of diverse barley landraces with highly divergent climatic variables. The detection of these signatures of local adaptation in a long-lasting native barley gene pool of the farming communities, will help in understanding the mechanisms of plant adaptation in barley and beyond in major crops like wheat.

Materials and Methods

Plant Material and Genotyping

In the present study, we selected 130 diverse barley landraces originating from 10 major barley-growing regions of Ethiopia (**Figure 1**). These landraces are not only described with altitude and geographic coordinates but also with the vernacular name given by the local community. This germplasm and its detailed information were provided by the Institute of Biodiversity Conservation (IBC) in Ethiopia. We genotyped two samples from each landrace resulting in 260 total samples (Table S2), which were analyzed using the genotyping by sequencing (GBS) approach. In addition, a German spring barley cultivar Barke was included in two replications as an internal control for the GBS analysis and data control. Initially, all samples were planted in a glass house, and after 2 weeks, the leaves were harvested for DNA extraction using the Qiagen DNeasy plant mini kit (Qiagen, Hilden, Germany) to ensure high-quality DNA, which was required for the GBS analysis. After DNA extraction, GBS libraries were prepared and analyzed at the Institute for Genomic Diversity (IGD), Cornell, USA, according to Elshire et al. (2011) using the enzyme *PstI* for digestion and creating a library with 96 unique barcodes. These libraries were sequenced using the Illumina HiSeq2000 platform. GBS analysis pipeline ver. 3.0.139, an extension to the Java program TASSEL (Bradbury et al., 2007), was used to call SNPs from the sequenced GBS library with the following options. Tags were aligned with the barley reference genome of cv. Morex (International Barley Genome Sequencing Consortium, 2012). VCF tools ver. 0.1.8 (Danecek et al., 2011) was used to summarize and filter data as well as to generate input files for PLINK (Purcell et al., 2007), which were used

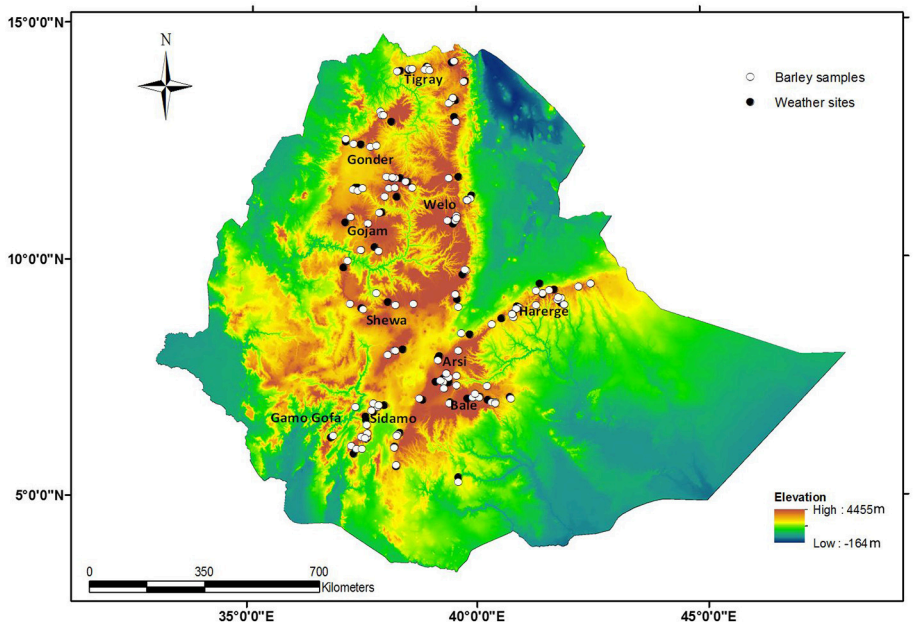


FIGURE 1 | Distribution of barley landraces and weather sites across Ethiopia on an altitudinal map.

for MDS (multidimensional scaling). The output was visualized using basic plotting functions in R ver. 2.15.0 (R Development Core Team, 2008). Before using these SNP markers for analysis, the original SNP data were filtered by applying different criteria. The first criterion was the SNP call rate for which SNP markers showing less than 10% missing values were passed to the next step. Among these, SNPs with a minor allele frequency (MAF) of less than 5% and monomorphic SNPs were excluded from the data. However, two barley samples (1%) were excluded in the final analyses because of missing genotypic data.

Climatic Data

The climate data from 61 weather sites were provided by the Ethiopian Meteorological Agency (**Figure 1**). The weather data were collected over multiple years, for an average of 21 years. The weather sites supplied monthly rainfall (lm^{-2}) and maximum and minimum temperature (°C) data. The three main seasons of Ethiopia, *Kiremt* (June–September), *Bega* (October–January), and *Belg* (February–May) were the basis for the grouping of the annual climatic data (USDA, 2002). *Kiremt* is the main rainy season all over Ethiopia, whereas *Bega* is the dry season, and *Belg* is considered the short rainy season. The altitude data were obtained from the passport data of the barley samples procured from the Institute of Biodiversity Conservation of Ethiopia (IBC). The altitudes of the sampling sites were grouped into four classes according to the traditional agro-ecological classification of Ethiopia. These classes are cold temperate, cool sub-humid highlands (Classes I and II, 1500–2500 m.a.s.l.), cool humid highlands (Class III, 2500–3000 m.a.s.l.) and highlands (Class IV, over 3000 m.a.s.l.) (USDA, 2002). The temperate, cool sub-humid highland was further divided into two classes because it covers a wide range of altitudes (Table S1).

Inference of Population Structure

Correction of the confounding effect of population structure in association studies plays a major role in reducing false positives (Pritchard et al., 2000a; Yu et al., 2006; Kang et al., 2008). Similarly, detecting adaptive loci without considering the impact of population structure will lead to false positive loci. Therefore, the analysis of hierarchical population structure was computed using the Bayesian-based program STRUCTURE ver. 2.3.5 (Pritchard et al., 2000b). For the analysis, an admixture model with correlated allele frequencies was chosen (Falush et al., 2003). The analysis was performed for a number of subpopulations varying from $K = 2$ to $K = 20$. For each value of K 20 independent runs were performed. For each run a burn-in of 10,000 and 50,000 iterations was specified. Finally, the Evanno et al. (2005) method was applied to determine the number of K. For this function, a web-based program, STRUCTURE HARVESTER ver. 0.9.93 (Earl and Vonholdt, 2012) was employed to infer the level of population structure. Ultimately, CLUMPP (Jakobsson and Rosenberg, 2007) was used to combine and average the individual's assignment across 20 runs for the determined number of K. To identify barley landraces that were admixed, each individual sample was assigned to its respective group based on a membership coefficient. The samples with a membership coefficient of ≥90% were assigned to a single group, whereas those that were smaller than the threshold were considered admixed. The membership coefficients (Q) were calculated using administrative regions instead of considering LOCPRIOR option during structure analysis. Eventually, altitude classes were used as basis of

grouping to test if the detected sub-populations were influenced by altitude. This was determined by, assigning each barley accession to its origin of altitude class and plot the structure graph using the membership coefficient.

Principal Component Analysis

A principal component analysis (PCA) was conducted using SNP markers data to reduce the number of variables into fewer components that explain the maximum variance. These components were then plotted in a two-dimensional plot for ease of viewing the existing genetic pattern. Before computing the PCA, the missed marker data were replaced with the mean values calculated over the markers. Subsequently, the analysis was carried out with the Proc princomp procedure using SAS software ver. 9.3 (SAS, 2011). A parallel analysis (PA) (Franklin et al., 1995) was then carried out to decide the number of principal components to retain for further analysis. PA is a method based on the generation of random eigenvalues to determine the number of components to retain. The eigenvalues are computed from the permutations of the observed data rather than from simulated data. This is an advantage not to keep the assumption of multivariate normality since the null reference set is conditioned on the observed data (Ledesma and Valero-Mora, 2007). In this analysis the covariance matrix was decomposed in which the parallel analysis restricted random matrices to have variable means and standard deviation of the real data (Franklin et al., 1995). Hence, a permutation test of 100 replications was used to run covariance matrices to calculate the eigenvalues. Afterwards, principal components which showed higher observed eigenvalues than their randomly generated associated values were retained for further analysis.

Spatial Genetic Structure

Isolation by distance (IBD) was computed using the "Spatial" option implemented in GenAlEx ver. 6.41 (Peakall and Smouse, 2006). The autocorrelation coefficient (r) obtained was similar to Moran's I (Moran, 1950), which ranges from -1 to 1. The spatial autocorrelation analysis was computed based on the pairwise comparison of the genetic distances derived from the genetic markers and geographic distance (km). Prior to performing the correlation analysis, the coordinates were converted into the Universal Transverse Mercator (UTM) system, and autocorrelation was computed first for all accessions from all regions followed by another analysis excluding accessions collected from Tigray. Accessions collected from Tigray region were excluded because of the geographic distance of the region and the sole grouping of the accessions during structure analysis. The significance of the spatial autocorrelation value was tested by constructing a two-tailed 95% confidence interval around the null hypothesis of no spatial genetic structure, which is $r = 0$. The analysis was performed with an option of an even distance class of 20 km based on a study that reported the distances traveled by Ethiopian farmers to obtain seeds (Bishaw, 2004). Permutations of 9.999 and a bootstrap of 1000 were used to compute the confidence interval around the null hypothesis.

Partitioning of Genomic Variation due to Climate Variables and Geographic Distance

Partial redundancy analysis (RDA), a constrained ordination technique, attempts to explain differences in species composition by combining a regression analysis with a principal component analysis (Borcard et al., 2011). It is based on genetic and environmental matrices (climate and geography). Partial constrained ordinations determine relationship between desired environmental and biological variables by removing the effect of known and uninteresting factors. Whereas unconstrained partial RDA considers the residual variance (Peres-Neto et al., 2006). In the present study, RDA was computed using XLSTAT ver. 2014.05.1 and *vegan* function in R package to disentangle the relative contribution of climate variables and geographic coordinates in driving genetic structure (Legendre and Fortin, 2010). For this, Hellinger transformed SNP allele frequencies were used as the response variable, and climate and coordinates as explanatory variables (Liu et al., 2011; De Kort et al., 2014). Before running the analysis the climate data were standardized using the *Proc* Stand procedure in the SAS software. The geographic variables were also normalized using a square root transformation of the geographic coordinates (Borcard et al., 2011). To examine how much of the genetic variation in barley landraces explained by climate variables, geographic coordinates and the combination of both, the variance components of the RDA were partitioned by running three different models. The first model considered all climate and geographic variables as explanatory variables (Model 1); the second model was a partial model in which the climate variables explained the genetic data conditioned on geographic coordinates (Model 2); and the third model was a partial model in which geographic coordinates explained the genetic data conditioned on climate variables (Model 3). For all models redundancy analysis was followed by significance test using Monte Carlo permutations test with 500 runs. For determination of best model forward selection with permutation of 999 and $\alpha = 0.01$ were computed. This process of model determination was improved by the introduction of adjusted R^2 by Peres-Neto et al. (2006), and the analysis was conducted using *ordistep* function of *vegan* in R package (Oksanen et al., 2010). Subsequently, the variation partitioning was followed when more than one significant explanatory variables were found (Legendre and Legendre, 1998).

Association Analysis of Climatic Variables

At present, a number of statistical tools are available for detection of outlier loci that are possibly affected by selection (Pérez-Figueroa et al., 2010; Narum and Hess, 2011). In the present study, we used two different software for the associations, between the environments and SNPs, and one for detection of the outlier loci. Bayenv2 and latent factor mixed model (LFMM) were used to identify association of climate factors with genetic markers whereas outlier loci were detected using BayeScan software. A detailed description of each statistical method is presented below.

The detection of loci correlated with different climatic variables was carried out using Bayenv2 (Coop et al., 2010) and LFMM (Frichot et al., 2013). Bayenv is a Bayesian method that

estimates the empirical pattern of covariance in allele frequencies between populations from a set of markers and then uses this as a null model for testing individual SNPs. Genome scans for SNPs with allelic correlations with climate variables were performed using Bayenv2 (Coop et al., 2010; Günther and Coop, 2013). This program runs in two steps. First, it creates a covariance matrix of relatedness between populations. Then, in the second step, it runs the correlation between the covariance matrix and the environmental variables generating a Bayes factor (BF) and non-parametric Spearman's rank correlation coefficient [ρ (Rho)]. The null model assumes that allele frequencies in a population are determined by the covariance matrix of relatedness alone against the alternative model, where allele frequencies are determined by a combination of the covariance matrix and an environmental variable, producing a posterior probability (Coop et al., 2010). Before running a null model estimation, the exclusion of outlier loci and loci which are in linkage disequilibrium, is recommended to ensure independence between SNPs on a chromosome (Bayenv2 Manual). Hence, we excluded outlier loci which were detected using BayeScan and LFMM program followed by loci which were in linkage disequilibrium ($r^2 > 0.2$) within each linkage group. The rest (801 neutral SNPs) were used to estimate the covariance matrix with 50,000 iterations. To control the variation across the covariance matrix, the average was calculated for the outputs of 10 matrices. Covariance matrices were compared after three independent runs with different seed numbers to ensure that the matrix was well-estimated. According to the recommendation of Blair et al. (2014) the BF of each SNP was calculated by averaging five independent runs of Bayenv2 at 50,000 Markov chain Monte Carlo (MCMC) for both the covariance matrix and Bayes factor analysis. For detection of outlier loci, Günther and Coop (2013) recommended considering the Spearman correlation coefficient, which measures the correlation between ranks of SNP allele frequencies and environmental factors, in addition to BF. BF is considered to have a slightly higher power, and SNPs, which fall in the top x% of BF and y% (where x < y; Bayenv2 Manual) of absolute values of spearman rank correlation coefficient ρ, are suggested to be robust candidate loci. Thus, we considered loci which were commonly detected in the top 1% of the *BF*-values ($BF > 3$) and top 5% of the absolute correlation values as a significant putative adaptive loci.

The other correlative method used for adaptive loci detection was LFMM, a software package that is a newly developed statistical model (Frichot et al., 2013). According to the study conducted by de Villemereuil et al. (2014), LFMM provided the best compromise between power and error rate across different scenarios. LFMM tests the association between environmental and genetic markers while estimating the hidden effect of population structure. The LFMM implemented fast algorithms using a hierarchical Bayesian mixed model based on a variant of PCA, in which the residual population structure is introduced via unobserved or latent factors. All SNP markers (1370) and the original climate variables were used for association analysis. The principal components of environmental variables are recommended when the summary of the variables is required because of their numbers (personal communication with Dr. Eric Frichot). The first three principal components generated for genetic markers were used as latent factors to estimate the population structure effect. The SNPs, which showed an association with environment, were determined based on the z-score. To estimate the z-scores for the environmental effect, the Gibbs sampler algorithm was run for 50,000 sweeps after a burn-in period of 10,000 sweeps. The threshold for the z-scores was determined after applying the Bonferroni correction for type I error $\alpha = 0.01$. Loci exhibiting z-scores above the absolute value of four and corresponding to $P < 10^{-5}$ were retained as significant loci.

Outlier Loci Detection

BayeScan is the tool that we used to detect outlier loci. It is a Bayesian based method that depends on a highly differentiated locus (Foll and Gaggiotti, 2008). It is the most conservative method with the least type I error compared to other outlier loci detection methods (Narum and Hess, 2011). However, it may detect high false positive loci if demographic history is not included in the analysis (Lotterhos and Whitlock, 2014). BayeScan identifies loci that are characterized by higher or lower levels of population divergence than neutral loci, suggesting a diversifying or purifying selection. It estimates the probability that a given SNP is under selection by calculating the posterior odds (PO). The PO are the ratio of the posterior probabilities of the two models (selection/neutral) for each locus based on the allele frequency. Before running the outlier loci analysis, the barley landraces were assigned to their respective K groups, thus supporting the comparison of the discrete groups in the process of candidate loci detection. To compare the result of outlier analysis, the individuals were assigned twice based on admixture Q coefficients of ≥70 and ≥90. Outlier loci detection was conducted by setting the prior probability of the model with a selection of 1/10, assuming a priori that the neutral model is 10 times more likely than the model including selection. During this run, all of the default values of 10 pilot runs of 5000 iterations with 50,000 additional burn-in steps were retained. We used false discovery rate (FDR = 0.05) as significance level for detection of the outlier loci. The FDR was controlled using the *q*-value which is the FDR analog of the *p*-value (Storey, 2002).

Detection of Candidate Genes

Candidate genes were found using the BLASTn function of DNA sequence analysis where the DNA sequences of SNP markers showing significant association were searched against the barley genome sequence using the NCBI and IPK databases. Genomic contigs showing the best hits were selected based on highly significant and maximum similarity percentages (>95%) and an *E*-value cut-off of 1E-15. The putative candidate genes across the contigs and relative distance of the associated SNP marker and candidate genes were found using BARLEX database and alignment package of Lasergene core suit of DNASTAR program. The gene ontology (GO) terms of the putative candidate genes were assigned using the Uniprot database.

Results

Genotyping by Sequencing and SNP Detection

The genotyping by sequencing (GBS) pipeline resulted in a total of 2,028,787 sequence tags, of which 1,548,708 (76.3%) were aligned with unique positions across the barley chromosomes. The sequence reads aligned with unique positions were subjected to SNP calling across the genotypes, founding 67,508 (unfiltered) Hapmap SNPs. After applying the filtering criteria as described in material and methods, a total of 1370 polymorphic SNPs were retained and utilized in further analyses. These SNP were distributed across all seven barley chromosomes. The highest number of SNP (214) were found on chromosome 7H and the lowest on chromosome 4H (108). The details of these SNPs, their corresponding chromosomes and contigs information are summarized in Figure S1.

Population Structure

The population structure analysis grouped barley landraces into three subpopulations (Figures S2A–C). The membership coefficient assignment (≥90%) indicated that most of the individuals were grouped in the first two groups, whereas the third inferred cluster contained few individuals. The membership coefficient assignment also revealed that most of the landraces from different geographic regions were clustered in group 1 (30 accessions) and group 2 (63 accessions). However, all landraces that were assigned to group 3 (18) originated from Tigray, but one accession from this region was assigned to group two. Bale (89%), Arsi (83%), Sidamo (79%), Harerge (68%), and Welo (63%) were the regions that contained highly admixed individuals. In contrast, less than half of the accessions collected from Shewa, Tigray, Gonder, and Gojam contained less than 10% admixtures within each individual, which was derived from historical ancestors. This percentage value indicated that more than half of the barley individuals from these regions have a membership coefficient that assigned these accessions to a distinct group. After the membership coefficient was assigned to each individual, we also tested whether altitude classes (Class I: below 2000; Class II: 2001–2500; Class III: 2501–3000; Class IV: above 3000 m.a.s.l) were the basis for the detection of the three sub-populations. All but one of the barley accessions in group 3 and 80% of the accessions in group 1 were collected from altitude classes I and II; the rest (20%) were collected from altitude class III (Figure S2D). Unlike other groups, barley landraces in group 2 were collected from altitude class II (13%) and class III (68%), and all accessions collected from altitude class IV (19%).

To visualize the geographic distribution of the population structure, we plotted the pie chart of the membership coefficient on an Ethiopian map (**Figure 2A**). The distribution of the barley landraces based on their area of origin was associated with their groupings. Most of the landraces from the eastern part of Ethiopia (Harerge), Gojam, Sidamo, and Welo were clustered in group 1, whereas the landraces collected from the rest of the regions were assigned to group 2, except Tigray, which was assigned to group 3.

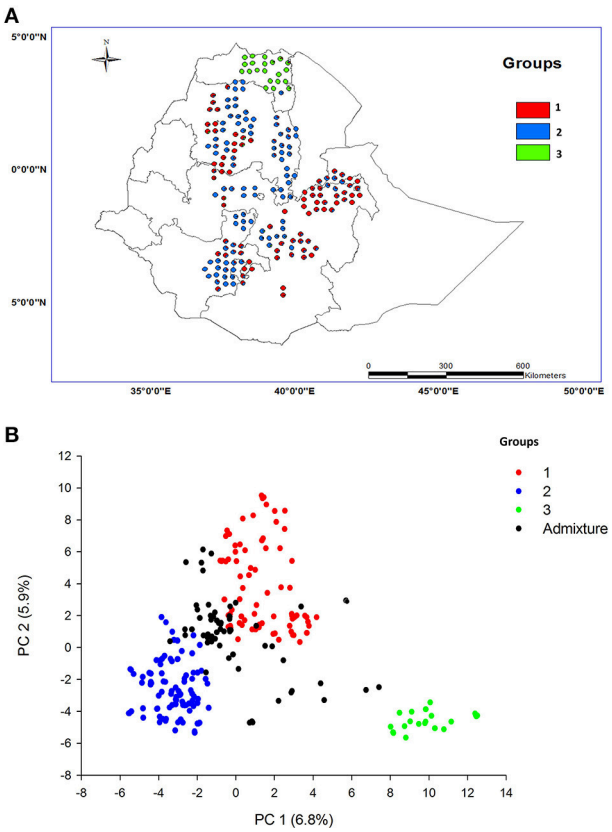


FIGURE 2 | Distribution of barley landraces and assignment of population membership coefficient along Ethiopian map (A). Each barley landrace was assigned to its respective inferred cluster based on the membership coefficient obtained from population structure analysis carried out using the model based software STRUCTURE. The scatter plot of principal components represented by the first and second principal components depicting the groupings of the barley individuals based on the subpopulations **(B)**.

Principal Component Analysis

Principal component analysis (PCA) reduced the variables into fewer components to explain most of the variation. Despite many eigenvalues, which were greater than one, we retained the first three principal components with variance of 15.03, 13.29, and 10.83. The proportions of variance explained by the respective principal components were 6.8, 5.9, and 4.9%. According to parallel analysis, the first three eigenvalues were sufficient for describing the grouping of the population. In order to visualize the pattern of the population grouping the first two principal components were plotted in 2-D. An assignment of individuals to their respective groups based on a ≥90% membership coefficient from population structure analysis resulted in approximately 57% of the individuals being categorized as admixtures (**Figure 2B**). Consequently, we assigned each barley individual to its respective group by considering its membership coefficient from the structure analysis and plotted the individuals based on the principal component values. In general, the first principal component separated groups one and three from group two, whereas the

second principal component separated group one from the rest of the groups.

Spatial Population Structure

A spatial analysis was computed using the entire data and excluding the accessions collected from Tigray. First, the analysis was performed for the accessions from all regions, and it showed a significant spatial autocorrelation (**Figure 3A**). Further, this analysis revealed a significant and positive spatial autocorrelation for the closest accessions and a negative correlation for the accessions collected from a wide distance. The positive and weak correlation between genetic similarity and geographic distance in the first dataset was observed for the genotypes collected in a range of 180 km ($r = 0.017$, $p = 0.001$). The presence of negative correlation for accessions collected in a geographic distance range of 780 km ($r = -0.013$) to 960 km ($r = -0.037$) was observed. However, after the accessions collected from Tigray were removed, the positive correlation was detected at short distances ranging from 20 km ($r = 0.095$, $p = 0.001$) to 40 km ($r = 0.006$, $p = 0.1$) (**Figure 3B**). Although most of the distance classes showed no spatial autocorrelation, the overall result of the spatial analysis revealed the presence of weak spatial population structure at the shortest distances, thus indicating genetic similarity.

Partitioning of Genomic Variation due to Climate Variables and Geographic Distance

A partial redundancy analysis (RDA) was performed to partition the variations accounted by climatic and geographic variables. The RDA analysis for model 1, which used climate and geographic variables as explanatory variables, indicated that the variation due to climate and geographic variables (constrained) explained most of the variation compared with the residual variance (unconstrained) (**Figure 4A**). Partitioning of the total variance indicated that the climatic variables accounted for 40% of the explainable total variance after removing the effect due to geographic variables, whereas geographic variables explained 29% of the total variance after the effect of climatic variables was controlled. The combination of climate and geographic effects explained 61% of the total explainable variation. The variance partitioning indicated that in Model 1 (F1 = 38.4%, F2 = 34.64%, F3 = 26.71%) and Model 2 (F1 = 69.64%, F2 = 22.62%, F3 = 7.74%) the first three eigenvalues contributed 100% to the variation, in contrast to Model 3, where two of the eigenvalues contributed to the total explainable variation (F1 = 81.83%, F2 = 18.17%; Figures S3A–C). The RDA result obtained after excluding Tigray indicated the importance of the region in shaping the genetic diversity pattern of the entire population. Executing the RDA analysis without conditioning on any of the

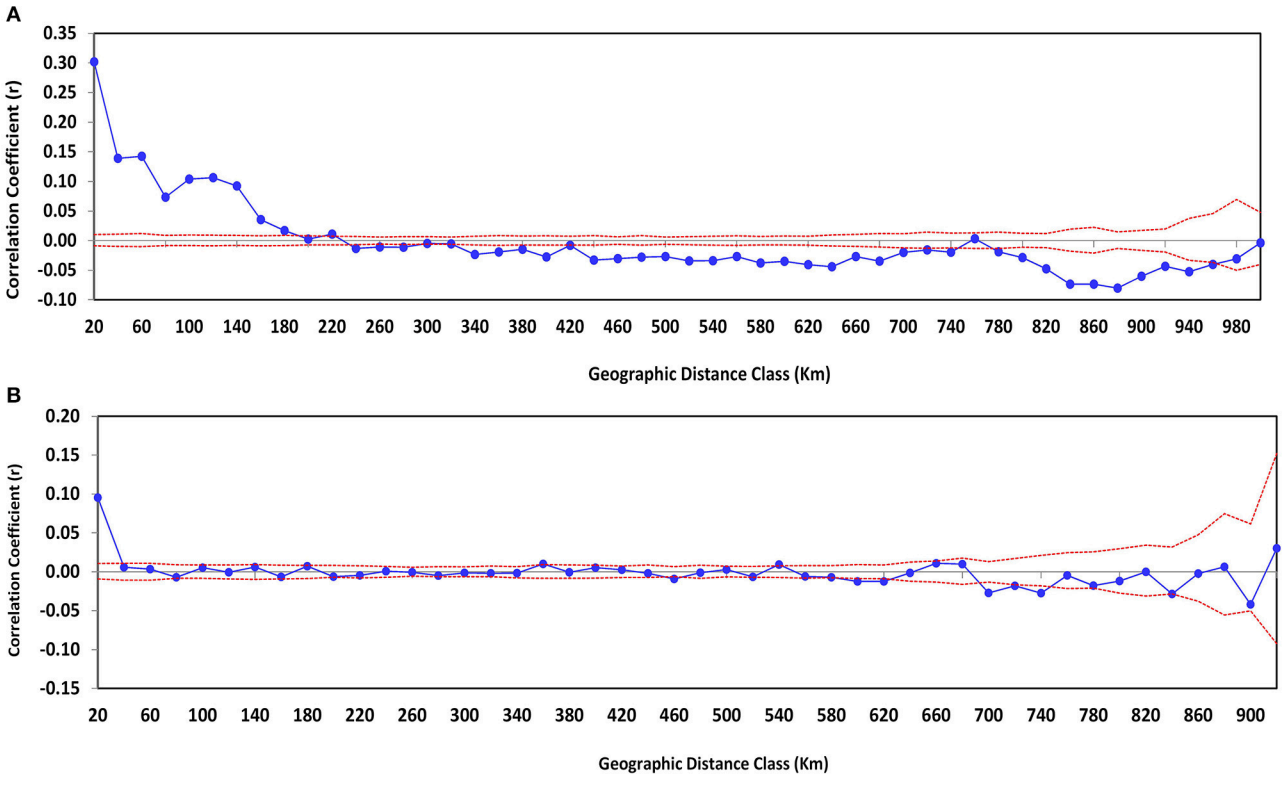


FIGURE 3 | Spatial autocorrelation correlogram plots. The plot depicts results obtained from all geographic regions **(A)**, and after accessions collected from Tigray region were removed and isolation by distance was calculated for accessions from the rest of the regions **(B)**. The analysis considered geographic distances with even distance class of 20 km. Dashed lines encompass the 95% confidence interval of the null hypothesis, and each point represents the autocorrelation coefficient (r).

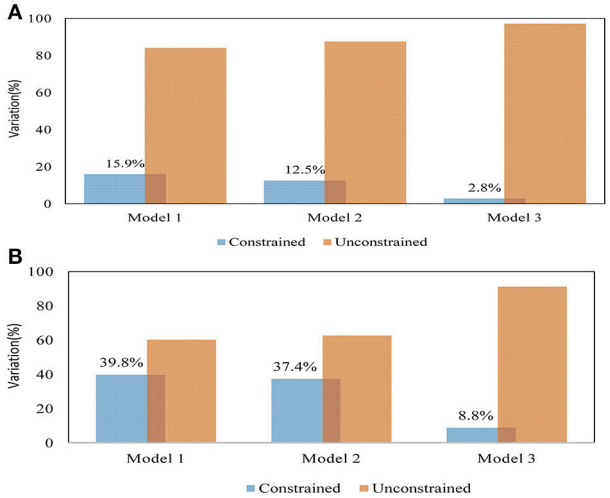


FIGURE 4 | The partial RDA variance partitioning was computed for entire dataset (A) and after accessions from Tigray region were excluded (B). The variance explained due to all variables (Model 1), the variance explained after controlling the effect introduced due to geographic distance (Model 2), and variance explained by geographic coordinates after the variance due to climate variables controlled (Model 3).

variables gave a close cumulative variance both with and without the Tigray region (60.6%; 57.2%) in the dataset (**Figure 4B**). A partial RDA analysis test for the full dataset yielded 40 and 29% for conditioning on geographic and climate variables, respectively. However, excluding Tigray from the dataset gave a value of 14.1 and 4.7% when conditioned on geographic and climate variables, respectively. The relative variances contributed by the presence of Tigray in the entire dataset conditioning on climate and geography were 35.3 and 16.5%, respectively. Furthermore, the eigenvalue results indicated low value and most of the variation was explained by residual variance (Figures S3D–F). We have also computed the partitioning among the climate variables while considering their major proportion in the total variance. It revealed that the variables altitude, Rf_*Kiremt* (rainfall in *Kiremt*) and Rf_annual explained most of the variation across the climatic variables (**Figure 5**, Table S3).

Association Analysis of Climatic Variables

The association analysis of SNP markers and climatic variables was performed using the Bayenv program. This analysis detected a total of 18 loci showing significant association with one or more climatic variables (**Table 1**). Among these, three loci were associated with variable altitude. Similar number of loci were associated with rainfall variables; Rf_*Bega* (1) and Rf_*Kiremt* (2). The highest number of loci were associated with minimum temperature variables; Mintemp_*Bega* (2), Mintemp_*Belg* (3), Mintemp_*Kiremt* (1), and Mintemp_aver (2) followed by maximum temperature variables; Maxtemp_*Bega* (2), Maxtemp_*Kiremt* (1), and Maxtemp_aver (1).

The association of SNP markers and climatic variables was also analyzed using a LFMM analysis. This analysis revealed the detection of 62 loci associated with the 13 selected climatic variables (**Table 2**). The highest number of loci (35) were associated with rainfall variables; Rf_*Bega* (10) and Rf_*Belg* (10), Rf_*Kiremt* (8) and Rf_*annual* (7). The second most number of loci were associated with variable altitude (9). In contrast, Mintemp_*Belg* and Mintemp_*Kiremt* were the only two climate variables that had one significant locus with $z = 5.20$ and $z = 5.57$, respectively. The highest number of common putative adaptive loci (6) were found for Rf_*Bega* and Rf_*Belg* followed by altitude and Rf_*Kiremt* (4). Among the loci commonly detected for altitude and Rf_*Kiremt*, we have selected the SNP locus Hv_SNP27845 with the highest significance level ($z = 6.71$). This locus was further illustrated to examine the allele frequency distribution along the altitude classes (**Figure 6A**) and rainfall as well as allele distribution over the country (**Figure 6A**). It showed that the most prevalent major allele at lowland was gradually decreased with an increase in the altitude and rainfall (**Figures 6B,C**). A complete summary of the LFMM analysis is presented in Table S4.

Outlier Loci Detection

The BayeScan method detected 12 and nine outlier loci (FDR = 0.05, prior 10:1) using a threshold of ≥70 and ≥90% ancestry coefficient of admixture for each barley individual, respectively (Figure S4, for the first approach). Of the nine loci detected using the second approach, six loci were also detected using the first approach. Three of the loci (Hv_SNP23336, Hv_SNP66136, and Hv_SNP27872) that were also detected with 100: one prior were considered for further analysis (**Figure 7**). The detected outlier loci showed a positive alpha value, which indicated directional selection. F_{ST}-values ranged between 0.69 and 0.66 for Hv_SNP53122 and Hv_SNP23336, respectively. Notably, the three detected SNPs were mapped on the same position (70.68 cM) on chromosome 7H.

Altogether, none of the software shared common significant loci among them but one locus (Hv_SNP4131) was commonly detected between LFMM and Bayenv software (Figure S5).

Detection of Candidate Genes

We have made an *in silico* analysis of the associated genomic regions to detect underlying putative candidate genes (**Table 3**). It revealed that all three SNP marker associated to altitude were found in single genomic contig (contig_46879) on chromosome 4H. These SNPs were found in the coding region (+1108 base pairs (bp) from ATG) of the sulfate transporter (*ST3.1*) gene. The SNP markers associated with altitude and Rf_*Kiremt* (rainfall in *Kiremt*) appear to underlie the L-lactate dehydrogenase (LDH) gene. These markers were around at +357 bp from ATG. The SNP locus (Hv_SNP4131) associated with maximum temperature (*Kiremt, Bega* and average) was found in the region of the cation\H+ exchanger (*CAX*) gene, +465 bp from ATG. Additionally, SNP loci on chromosome 2H associated with Maxtemp_*Bega*, were found next to each other in the putative promoter region (−2749 bp from ATG) of the universal stress responsive protein (*USP1*).

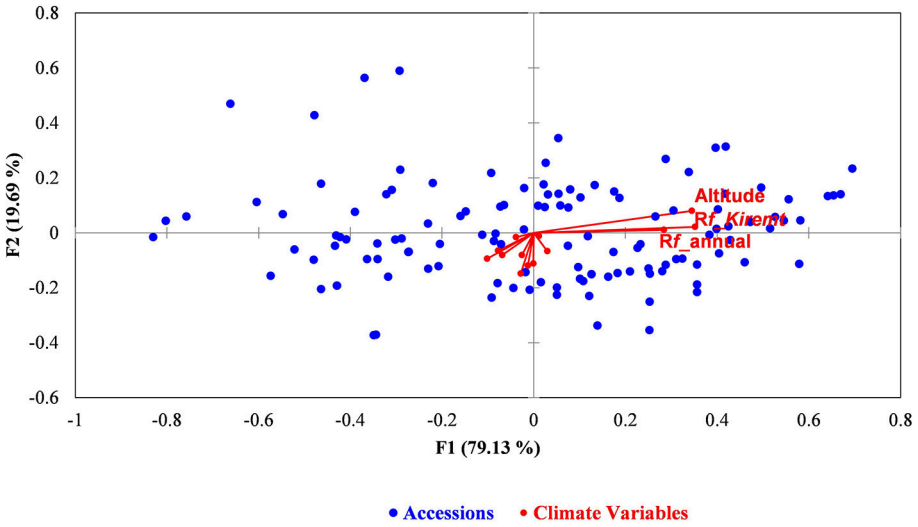


FIGURE 5 | Partial RDA analysis was performed to determine the relative contribution of climate and geographic variables shaping the genetic structure. The biplot depicts the eigenvalues and lengths of eigenvectors for the RDA conditioned on geographic distance.

Discussion

Population Structure

The population structure analysis was computed using the STRUCTURE program and supported by the principal component analysis approach. The detected clusters did not completely reveal a geographically based population structure. Though accessions from 10 geographic regions were analyzed, the population structure analysis detected that three sub-populations contained different regions as one group. Hence, this result suggests the weak impact of geographic boundaries on the genetic structure of the barley population. A weak effect of political regions was reported for the morphological and genetic differences between major barley-growing areas of Ethiopia (Abebe et al., 2010; Abebe and Léon, 2013a). However, the pattern of clustering in the present study, was different compared with previous studies because of the difference in the number of barley genotypes, number of genetic markers and sampling strategy to genotype landraces. In the present study, we also replicated each landrace twice for genotyping to ensure high-quality genotyping data and to determine the genetic purity of the landraces that farmers have selected and established for barley cultivation. Among the inferred groups, the third cluster was aligned with one of the geographic regions. This region was Tigray, which is located in the northern part of Ethiopia and is frequently affected by drought because of a degraded environment and erratic rainfall (Abay et al., 2009). Farmers in this region have selected drought-resistant landraces to grow under water-limited conditions (Meze-Hausken, 2004). In addition, a decrease in rainfall northwards and eastwards from the high rainfall pocket area in the southwest has been reported (USDA, 2002). In the present study, Tigray was one of the regions having low percentage of admixed barley landraces (39%) and over 90% of the accessions from other regions were assigned to group 3. However, more than three quarters of the accessions from Arsi, Bale and Sidamo were considered admixed and were thus not assigned to a single cluster. These regions are known as the cereal belt of Ethiopia, which implies that a considerable amount of cereal production and marketing occurs in these areas. This leads to high genetic diversity in the region and gene flow between farmers' fields, resulting in admixed landraces (Negassa, 1985; Abebe et al., 2013b). The population structure coefficient sorted by altitude classes indicated that the accessions grouped in the first and second sub-populations originated in altitudes less than 2500 m.a.s.l. Except for a few accessions, the third sub-population contained accessions collected from the highlands (above 3000 m.a.s.l) of Ethiopia. In general, geographic regions and altitude classes were associated with different groups; however, the spatial distance was presumably not considered as the basis for the inferred clustering.

Spatial Genetic Structure

Isolation by geographic distance occurs when the gene flow between organisms is restricted because of spatial isolation. The detection of a correlation between the genetic and geographic distance was described as isolation by distance (Wright, 1946). We also detected significant but weak isolation by distance for the dataset consisting of all the barley accessions and in the dataset where the accessions from the Tigray region were removed. The correlogram from the first dataset showed correlation with the geographic distance covering a wide range, whereas after excluding Tigray, a significant correlation was observed over a shorter distance. In this case, the accessions in a 40 km range were considered to be genetically similar and positively associated with geographic distance but the correlation was not different from zero. The population structure analysis grouped most of the accessions from this region in one group, indicating the presence

TABLE 1 | A summary of putative adaptive loci showing association with different climate variables identified using Bayenv analysis.

SNP ID	Chr	cM	BF	Rho (ρ)	Climatic variables												
					A	B	C	D	E	F	G	H	I	J	K	L	M
Hv_SNP785	2H	65.59	3.05	0.52						*							
Hv_SNP4131	2H	123.94	3.26	0.60		*											
Hv_SNP8058	4H	77.48	8.22	0.37								*					
Hv_SNP51899	4H	78.61	3.97	0.44									*				
Hv_SNP51899	4H	78.61	4.21	0.44							*						
Hv_SNP594	4H	79.87	4.06	0.43									*				
Hv_SNP594	4H	79.87	3.00	0.49						*							
Hv_SNP594	4H	79.87	4.27	0.43							*						
Hv_SNP4616	5H	40.07	4.88	0.60													*
Hv_SNP4616	5H	40.07	5.51	0.51										*			
Hv_SNP15799	5H	129.65	5.00	0.45				*									
Hv_SNP56701	5H	169.38	3.68	0.51												*	
Hv_SNP31344	6H	100.42	3.05	0.46							*						
Hv_SNP30323	7H	55.74	4.55	0.47	*												
Hv_SNP23710	7H	109.92	5.44	0.54										*			
Hv_SNP23253	7H	124.58	3.14	0.45	*												
Hv_SNP58	U	U	3.20	0.44	*												
Hv_SNP32903	U	U	4.25	0.54				*									
			Total loci		3	1	–	2	–	2	3	1	2	2	–	1	1

Where: A, Altitude; B, Rf_Bega; C, Rf _Belg; D, Rf_Kiremt; E, Rf_annual; F, Mintemp_Bega; G, Mintemp_Belg; H, Mintemp_Kiremt; I, Mintemp_aver; J, Maxtemp_Bega; K, Maxtemp_Belg; L, Maxtemp_Kiremt; M, Maxtemp_aver; U, Unknown.

**Indicates that the particular SNP showed correlation with that specific climate variable. BF (Bayes factor), Rho (ρ) (Spearman's rank correlation coefficient).*

of less shared ancestors among the accessions. Furthermore, the autocorrelation result revealed that the other regions are spatially isolated from Tigray because of its geographic location. Hence, the location of Tigray influenced the pattern of the spatial genetic structure in the studied population. The low percentage of admixture among the accessions was presumably associated with the low gene flow from the neighbor regions. This is attributed to the location, landscape, social and economic activity of the region. In general, the accessions from Tigray region affected the pattern of isolation by distance when all regions were considered for analysis. But the detected spatial correlation was weak and limited to a short distance to infer the presence of isolation by distance.

Partitioning of Genomic Variation due to Climate Variables and Geographic Distance

The partial RDA was computed to estimate the proportion of variation explained by the environmental variables or by geographic distance alone or as the fraction of the variation shared by both variables. The variance partitioning for partial RDA models indicated that the variation contributed by climate variables were higher than the variation introduced due to geographic variables in both datasets (datasets are explained in material and methods). However, all the models showed significant association between the environmental variables and the genetic variation. The positive association of the climate variables with the genetic markers while controlling the variations due to geographic variables, thus suggests an important influence of climate diversity in shaping genetic variation (Temunovic et al., 2012). Similar findings were reported by Lasky et al. (2012) where they found a significant contribution of climate variables after controlling the spatial structure in *Arabidopsis thaliana*. They propose these variables as the selective gradients related to local adaptation across the species range. Unlike the climate variables the geographic coordinates showed low linear association with the genetic data indicating the influence of the spatial structure on the genetic variation of barley. Previously, Liu (1997) also found that climate factors accounted for 13% of the explained variation, whereas the geographic position was considered less important for algae colony thickness and colonization which are in agreement with the present study. Similar outcome was reported by McGaughran et al. (2014) who suggested the association between geography and genetic distance as an important determinant of genetic structure beyond genetic drift in isolated population. Moreover, comparing the results of both datasets revealed that accessions collected from Tigray region contributed more than a unit variance considering the contribution of the remaining regions to the environmental variation.

Further partitioning of variance explained due to the climate variables revealed altitude, total rainfall and rainfall of the main growing season as the main contributors of the detected genetic variation. Besides, the forward selection process retained altitude twice (in both datasets) as the

TABLE 2 | A summary of putative adaptive loci showing association with different climate variables identified using LFMM analysis.

SNP_ID	Chr	cM	Zscore	-log10 (*p*-value)	Climatic variables												
					A	B	C	D	E	F	G	H	I	J	K	L	M
Hv_SNP57960	1H	7.22	4.26	4.68													*
Hv_SNP57963	1H	7.22	4.43	5.03										*			
Hv_SNP9160	1H	42.71	5.55	7.53		*	*		*								
Hv_SNP28572	1H	48.51	4.86	5.92	*			*									
Hv_SNP28218	1H	49.75	5.6	7.66		*	*										
Hv_SNP6094	1H	70.25	5.13	6.55				*									
Hv_SNP53255	1H	103.82	4.06	4.31			*										
Hv_SNP54198	1H	132.51	5.75	8.06		*											
Hv_SNP3374	2H	18.8	4.2	4.58		*											
Hv_SNP27845	2H	18.91	6.71	10.71	*			*									
Hv_SNP13837	2H	39.66	4.19	4.55										*			
Hv_SNP4499	2H	55.56	5.57	7.59	*			*		*	*	*	*				
Hv_SNP55036	2H	92.21	5.21	6.72					*								
Hv_SNP4131	2H	123.94	4.67	5.52										*		*	*
Hv_SNP25024	2H	138.6	4.32	4.8				*									
Hv_SNP51311	3H	83.59	4.37	4.91			*			*							
Hv_SNP7771	4H	18.48	5.89	8.41		*	*										
Hv_SNP54437	4H	19.9	4.17	4.52					*								
Hv_SNP15569	4H	35.13	5.96	8.6	*												
Hv_SNP19635	4H	60.55	4.48	5.12		*	*										
Hv_SNP25404	4H	91.18	4.18	4.53				*									
Hv_SNP5505	4H	105.49	4.68	5.55		*			*								
Hv_SNP34901	5H	13.77	4.68	5.54						*							
Hv_SNP34783	5H	77.08	6.27	9.44		*	*										
Hv_SNP37305	5H	79.13	4.23	4.64					*								
Hv_SNP30681	5H	80.35	4.06	4.31			*										
Hv_SNP13299	5H	95.9	5.01	6.27	*			*									
Hv_SNP27374	5H	161.08	4.42	5.00					*								
Hv_SNP64267	5H	164.72	5.88	8.39	*												
Hv_SNP8419	5H	164.72	4.57	5.31												*	*
Hv_SNP36036	5H	169.38	5.23	6.76		*	*			*							
Hv_SNP65888	5H	169.38	6.14	9.09			*			*			*				
Hv_SNP28364	6H	15.72	4.04	4.27											*		
Hv_SNP23365	6H	52.2	5.44	7.28		*											
Hv_SNP64219	6H	94.62	5.12	6.51	*												
Hv_SNP8527	7H	12.75	4.33	4.82				*									
Hv_SNP8936	7H	67.37	8.39	16.32	*												
Hv_SNP29190	7H	85.98	4.06	4.31					*								
Hv_SNP8273	7H	109.92	4.28	4.73	*												
				Total loci	9	10	10	8	7	5	1	1	2	3	1	2	3

Where: A, Altitude; B, Rf_Bega; C, Rf _Belg; D, Rf_Kiremt; E, Rf_annual; F, Mintemp_Bega; G, Mintemp_Belg; H, Mintemp_Kiremt; I, Mintemp_aver; J, Maxtemp_Bega; K, Maxtemp_Belg; L, Maxtemp_Kiremt; M, Maxtemp_aver.

**Indicates that the particular SNP showed correlation with that specific climate variable. The underlined loci showed association with two climate variables whereas the underlined and bold loci are associated with three or more climate variables.*

first significant explanatory variable. This result indicated the importance of altitude in affecting the existing genetic variation in barley population. The importance of altitude in shaping and determining the climate variables and thus the genetic diversity in barley has been reported by Abebe et al. (2010) and Demissie and Bjornstad (1996). Similarly, Pyhäjärvi et al. (2013) controlled population structures using partial mantel analysis and found a significant effect of altitude in teosinte, the wild ancestor of maize. Besides, rainfall, which mostly depends on altitude, is one of the determinant factor in

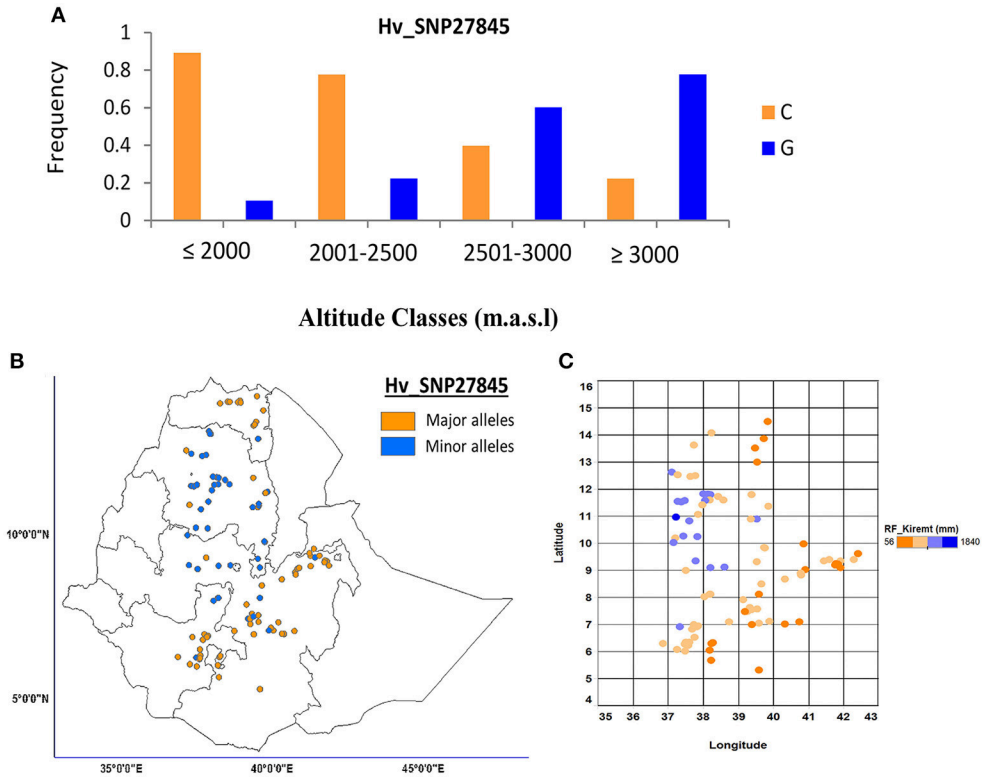


FIGURE 6 | Allele frequency of putative adaptive loci correlated with altitude classes and Rainfall in *Kiremt* as detected by LFMM. The major and minor alleles of putative adaptive loci Hv_SNP27845 showed frequency pattern along the altitude classes **(A)**. The allele distribution depicting on the Ethiopian map **(B)** and the rainfall pattern along the coordinates of Ethiopia displayed in scatter plot **(C)**.

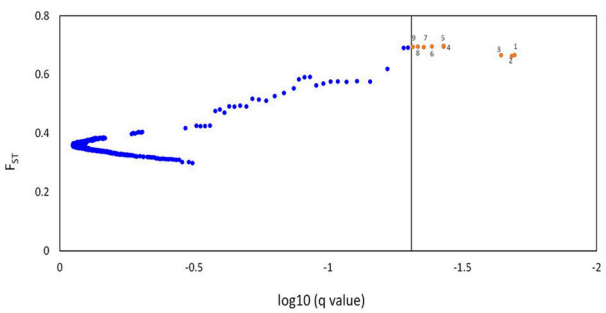


FIGURE 7 | A Bayesian based BayScan program were employed to scan for the presence of putative outlier loci affected by selection. This plot presents F_{ST} against log 10(*q*-value), which is the FDR analog of the *p*-value. The line represents the threshold FDR = 0.05 and the red dots indicate the outlier loci which are potentially affected by directional selection.

the genetic variation. Zhao et al. (2013) proposed annual rainfall as a major factor behind the genetic divergence and adaptation of Chinese wild rice (*Oryza rufipogon*). Hence, the variance partitioning of the significant climate variables emphasized the importance of altitude in shaping the ecological diversity and evolutionary aspect of different plants.

Climatic Adaptations

Natural selection plays a major role in shaping the available genetic variation of a population and thereby determines local adaptation (Kawecki and Ebert, 2004). It also changes the allele frequency when individuals with the same fitness trait survive and increase in number. In this study, we observed a similar situation in allele distribution of the detected putative adaptive loci in response to different climate variables. The association of climate variables with SNP markers using Bayenv and LFMM returned several significant loci in relation to all climate variables, indicating that the variables were the important climate factors that affect selection pressure. Most of the loci detected using LFMM software were associated with rainfall variables followed by altitude, indicating the importance of these variables in determining local adaptation. In Bayenv analysis most loci were correlated with temperature variables followed by altitude and rainfall. Partial RDA analysis also indicated that altitude, Rf_*Kiremt* and Rf_annual were the most important climate variables; most of the variation originated from these variables. De Kort et al. (2014) reported strong associations between outlier loci and temperature using LFMM in the tree species *Alnus glutinosa*. LFMM detected a locus (Hv_SNP27845) showing correlation with altitude and rainfall variables which explained most of the variation in partial RDA. The pattern of decreasing frequency of the major alleles as a function of increasing altitude

TABLE 3 | A summary of detected loci and identification of putative candidate genes.

Climate variables	Chr[a]	Associated SNP[b]	Genomic Contig[c]	Contig length[d]	Identity (%)[e]	Putative genes[f]	SNP to gene[g]	Accession Nr.[h]	Go term[i]		
									Biological function	Molecular function	Cellular components
Altitude	4H	Hv_SNP11857 Hv_SNP11859 Hv_SNP11860	contig_46879	9106	96	Sulfhate transporter (*ST3.1*)	+1108*	AK358393	Sulfhate transport	Secondary active sulfate transmembrane transporter activity	Membrane
Altitude Rf_*Kiremt*	2H	Hv_SNP27843 Hv_SNP27845	contig_136338	8019	98	L-lactate dehydrogenase (*LDH*)	+357*	AK375972	Response to stress	L-lactate dehydrogenase activity	Cytoplasm
Maxtemp_*Bega* Maxtemp_Kiremt Maxtemp_aver	2H	Hv_SNP4131	contig_38530	6129	99	Cation\H$^+$ exchanger (*CAX*)	+465*	AK373169	Cation transport	Cation transmembrane transporter activity	Membrane
Maxtemp_*Bega*	2H	Hv_SNP13837 Hv_SNP13839	contig_49840	10470	99	Universal stress protein (*USP1*)	−2749	AY641412	Response to stress	Nucleotide binding	Cytosol

[a]*Chr. Chromosome*
[b]*Associated SNPs loci*
[c]*Genomic contig containing associated SNPs and putative candidate genes*
[d]*Length of genomic contig in base pairs*
[e]*% similarity of SNP and genomic contig*
[f]*Putative candidate genes*
[g]*Position of SNP to putative genes in base pairs, + and − reveal the number of base pairs downstream and upstream of gene start site (ATG), associated SNP within the candidate genes are indicated as **
[h]*Accession number NCBI*
[i]*Gene ontology terms.*

presumes the presence of directional selection, which leads to local adaptation. The minor alleles were observed in highland areas with high rainfall, indicating the importance of altitude in determining other climate factors. The prevalence of the major allele among the genotypes that were collected below 2500 m.a.s.l. in low rainfall areas was presumably due to local selection (**Figure 6A**). In this case, because of directional selection, the advantageous alleles increased in frequency relative to others and eventually became fixed (Bose and Bartholomew, 2013). Altitude affects phenology, the distribution and type of disease and the prevalence of frost in different crops of Ethiopia. In the highlands, barley matures quite late, and it takes as long as seven to 8 months to mature (Tanto and Demissie, 2001), whereas in the lowlands and in the *Belg* season, barley matures early, within 3–4 months (Mulatu and Grando, 2011). The special adaptation of barley to highlands makes the crop the most valuable cereal for the survival of the farmers living in the highlands as it is the only crop cultivated across those regions (Lakew et al., 1997). The highlands of Ethiopia are described as sunny during the day and cold at night with occurrences of frost, particularly during the *Bega* season (USDA, 2002). In general, altitude plays a major role in the determination of morphological novelties of different crops in Ethiopia (Engels, 1994; Abebe et al., 2010), and it affects the ecological variables and, thus, local adaptation. *Kiremt* (main rainy season) rains occur during June-September, accounting for 50–80% of the annual rainfall over the Ethiopian regions. The most severe droughts are usually related to a failure of the *Kiremt* rainfall to meet Ethiopia's agricultural water demands (Korecha and Barnston, 2007). In general, directional selection occurs when natural selection favors a single phenotype, and the allele frequency thus shifts in one direction. The loci that were identified as adaptive loci presumably underlie the phenotypic variation that affect fitness in different environments (Nunes et al., 2011).

Detection of candidate genes

Although genes and phenotypes are in a causal relationship, dissecting the genetic components of a phenotype is not simple. Through the advent of genome-wide DNA markers and sequenced genomes, it has become feasible to uncover this relationship precisely and dissect the hidden genetic regulations in the expression of a phenotype at the gene level. In the present study, we utilize genome-wide SNP markers to dissect those footprints associated with barley adaptation to landscape and climatic variables. The associated SNP markers loci were then searched in the database for the putative genes. To this end, we are proposing four putative candidate genes due to their tight linkage with the associated SNP markers as well as due to likelihood of their functional linkage with a given climatic variable. For instance, the significant loci associated with altitude and rainfall variables underlie putative sulfate and L-lactate dehydrogenase genes. A number of studies suggested the role of sulfate genes in nutrient transport for plant growth as well as for environmental adaptation like drought and salinity stress (Hawkesford and Buchner, 2001; Gallardo et al., 2014). In Arabidopsis, lactate dehydrogenase genes are involved in adaptation to hypoxic stress (reduced oxygen because of water logging or higher altitudes) by switching plants from aerobic respiration to anaerobic fermentation (Dolferus et al., 2008). These results, seems to be in line with the present study where we found a putative association of L-lactate dehydrogenase gene with altitude and rainfall. Similarly, cation\H+ exchanger (*CAX*) and universal stress protein (*HvUSP1*) appeared as candidates for adaptation to higher temperature. Plants trigger the expression of a specialized protein called the heat shock or stress protein against climatic conditions such as higher temperature (Vierling, 1991; Parsell and Lindquist, 1993; Gupta et al., 2010). These proteins are then involved in the maintenance of cell membrane stability, capturing the reactive oxygen species (ROS), synthesis of antioxidants, accumulation and osmoregulation of osmoticum (Wahid et al., 2007). We believe that these data reveal a primary insight into the identification of primary evolutionary candidate genes mediating adaptation to important landscape and climatic variables across Ethiopia. However, further experiments are needed to confirm the precise role of these candidate genes in the process of local adaptation in barley.

Taken together, the present study has successfully analyzed the association between genetic markers and environmental factors to determine their effect on the explainable genetic variation. We identified climate and geographic variables as important explanatory aspects of genetic variation followed by altitude and rainfall as underlying cause of climatic variation. Hence, the detected correlation between environmental variables and genetic markers can help to understand the phenomenon of natural selection, yet, conducting the common garden experiment to verify the result will provide the strong evidence for the underlying phenotypic traits. In general, this study has successfully demonstrated how landscape genomics contribute to uncover the genetic components (genes) and evolutionary processes affecting adaptation. In conclusion, we assume that the detected candidate loci were associated with local adaptation that showed selective responses to important climatic variables.

Acknowledgments

We thank Ethiopian Meteorological Agency (EMA) and Institute of Biodiversity Centre of Ethiopia (IBC) for providing us the weather data and the barley landraces with the passport data, respectively. This study was supported by Alexander von Humboldt foundation, George Forster postdoctoral fellowship program. A special thanks to Mr. Arifuzzaman and Mrs. Woitol for reading and correcting the manuscript. Sequence data: The genotyping by sequencing (GBS) data have been submitted to Sequence Read Archive (SRA) of NCBI under accession number SRP063283.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2015.00813

References

Abay, F., Bjørnstad, A., and Smale, M. (2009). Measuring on farm diversity and determinants of barley diversity in Tigray: Northern Ethiopia. *Momona Ethiop. J. Sci.* 1, 44–66. doi: 10.4314/mejs.v1i2.46048

Abebe, T. D., Bauer, A. M., and Léon, J. (2010). Morphological diversity of Ethiopian barleys (*Hordeum vulgare* L.) in relation to geographic regions and altitudes. *Hereditas* 147, 154–164. doi: 10.1111/j.1601-5223.2010.02173.x

Abebe, T. D., and Léon, J. (2013a). Spatial and temporal genetic analyses of Ethiopian barley (*Hordeum vulgare* L.) landraces reveal the absence of a distinct population structure. *Genet. Resour. Crop Evol.* 60, 1547–1558. doi: 10.1007/s10722-012-9941-4

Abebe, T. D., Mathew, B., and Leon, J. (2013b). Barrier analysis detected genetic discontinuity among Ethiopian barley (*Hordeum vulgare* L.) landraces due to landscape and human mobility on gene flow. *Genet. Resour. Crop Evol.* 60, 297–309. doi: 10.1007/s10722-012-9834-6

Allendorf, F. W., Hohenlohe, P. A., and Luikart, G. (2010). Genomics and the future of conservation genetics. *Nat. Rev. Genet.* 11, 697–709. doi: 10.1038/nrg2844

Andrews, C. A. (2010). Natural selection, genetic drift, and gene flow do not act in isolation in natural populations. *Nat. Educ. Knowl.* 3, 5.

Bishaw, Z. (2004). *Wheat and Barley Seed System in Ethiopia and Syria*. Ph.D. Dissertation, Wageningen University, Netherlands.

Blair, L. M., Granka, J. M., and Feldman, M. W. (2014). On the stability of the Bayenv method in assessing human SNP-environment associations. *Hum. Genomics* 8:1. doi: 10.1186/1479-7364-8-1

Bonin, A., Taberlet, P., Miaud, C., and Pompanon, F. (2006). Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Mol. Biol. Evol.* 23, 773–783. doi: 10.1093/molbev/msj087

Borcard, D., Gillet, F., and Legendre, P. (2011). *Numerical Ecology with R*, New York, NY: Springer.

Bose, R., and Bartholomew, A. J. (2013). Macroevolution in deep time. *Springer-Briefs Evol. Biol.* 3, 1–59. doi: 10.1007/978-1-4614-6476-1

Brachi, B., Morris, G. P., and Borevitz, J. O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biol.* 12:232. doi: 10.1186/gb-2011-12-10-232

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J. K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics* 185, 1411–1423. doi: 10.1534/genetics.110.114819

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCF tools. *Bioinformatics* 2, 2156–2158. doi: 10.1093/bioinformatics/btr330

De Kort, H., Vandepitte, K., Bruun, H. H., Closset-Kopp, D., Honnay, O., and Mergeay, J. (2014). Landscape genomics and a common garden trial reveal adaptive differentiation to temperature across Europe in the tree species *Alnus glutinosa*. *Mol. Ecol.* 23, 4709–4721. doi: 10.1111/mec.12813

de Villemereuil, P., Frichot, E., Bazin, E., Francois, O., and Gaggiotti, O. E. (2014). Genome scan methods against more complex models: when and how much should we trust them? *Mol. Ecol.* 23, 2006–2019. doi: 10.1111/mec.12705

Dell'Acqua, M., Zuccolo, A., Tuna, M., Gianfranceschi, L., and Enrico, P.é, M. (2014). Targeting environmental adaptation in the monocot model *Brachypodium distachyon*: a multi-faceted approach. *BMC Genomics* 15:801. doi: 10.1186/1471-2164-15-801

Demissie, A., and Bjornstad, A. (1996). Phenotypic diversity of Ethiopian barleys in relation to geographical regions, altitudinal range, and agro-ecological zones: as an aid to germplasm collection and conservation strategy. *Hereditas* 124, 17–29. doi: 10.1111/j.1601-5223.1996.00017.x

Dolferus, R., Wolansky, M., Carroll, R., Miyashita, Y., Ismond, K., and Good, A. G. (2008). Functional analysis of lactate dehydrogenase during hypoxic stress in Arabidopsis. *Funct. Plant Biol.* 35, 131–140. doi: 10.1071/FP07228

Earl, D. A., and Vonholdt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Eckert, A. J., Bower, A. D., González-Martínez, S. C., Wegrzyn, J. L., Coop, G., and Neale, D. B. (2010a). Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). *Mol. Ecol.* 19, 3789–3805. doi: 10.1111/j.1365-294X.2010.04698.x

Eckert, A. J., van Heerwaarden, J., Wegrzyn, J. L., Nelson, C. D., Ross-Ibarra, J., González-Martínez, S. C., et al. (2010b). Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). *Genetics* 185, 969–982. doi: 10.1534/genetics.110.115543

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379. doi: 10.1371/journal.pone.0019379

Engels, J. (1994). Genetic diversity in Ethiopia barley in relation to altitude. *Genet. Resour. Crop Evol.* 42, 761–768.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Excoffier, L., Hofer, T., and Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity* 103, 285–298. doi: 10.1038/hdy.2009.74

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multi locus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587. doi: 10.1111/j.1471-8286.2007.01758.x

Foll, M., and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective. *Genetics* 180, 977–993. doi: 10.1534/genetics.108.092221

Franklin, S., Gibson, D., Robertson, P., Pohlmann, J., and Fralish, J. (1995). Parallel Analysis: a method for determining significant components. *J. Veg. Sci.* 6, 99–106. doi: 10.2307/3236261

Frichot, E., Schoville, S. D., Bouchard, G., and François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol. Biol. Evol.* 30, 1687–1699. doi: 10.1093/molbev/mst063

Funk, D. J., Nosil, P., and Etges, W. J. (2006). Ecological divergence exhibits consistently positive associations with reproductive isolation across disparate taxa. *Proc. Natl. Acad. Sci. U.S.A.* 103, 3209–3213. doi: 10.1073/pnas.0508653103

Gallardo, K., Courty, P. E., Le Signor, C., Wipf, D., and Vernoud, V. (2014). Sulfate transporters in the plant's response to drought and salinity: regulation and possible functions. *Front. Plant Sci.* 5:580. doi: 10.3389/fpls.2014.00580

Günther, T., and Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics* 195, 205–220. doi: 10.1534/genetics.113.152462

Gupta, S. C., Sharma, A., Mishra, M., Mishra, R. K., and Chowdhuri, D. K. (2010). Heat shock proteins in toxicology: how close and how far? *Life Sci.* 86, 377–384. doi: 10.1016/j.lfs.2009.12.015

Harlan, J. R. (1969). Ethiopia: A centre of diversity. *Econ. Bot.* 23, 309–314. doi: 10.1007/BF02860676

Hawkesford, M. J., and Buchner, P. (2001). *Plant Ecophysiology: Molecular Analysis of Plant Adaptation to the Environment.* Netherlands: Kluwer Academic Publishers.

International Barley Genome Sequencing Consortium. (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature* 491, 711–716. doi: 10.1038/nature11543

Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23, 1801–1806. doi: 10.1093/bioinformatics/btm233

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–1723. doi: 10.1534/genetics.107.080101

Kawecki, T. J., and Ebert, D. (2004). Conceptual issues in local adaptation. *Ecol. Lett.* 7, 1225–1241. doi: 10.1111/j.1461-0248.2004.00684.x

Korecha, D., and Barnston, A. G. (2007). Predictability of June-September rainfall in Ethiopia. *Mon. Wea. Rev.* 135, 628–650. doi: 10.1175/MWR3304.1

Lakew, B., Semeane, Y., Alemayehu, F., Gebre, H., Grando, S., van Leur, J. A. G., et al. (1997). Exploiting the diversity of barley landraces in Ethiopia. *Genet. Resour. Crop Evol.* 44, 109–116. doi: 10.1023/A:1008644901982

Larsson, S. J., Lipka, A. E., and Buckler, E. S. (2013). Lessons from *Dwarf8* on the strengths and weaknesses of structured association mapping. *PLoS Genet.* 9:e1003246. doi: 10.1371/journal.pgen.1003246

Lasky, J. R., Des Marais, D. L., McKay, J. K., Richards, J. H., Juenger, T. E., and Keitt, T. H. (2012). Characterizing genomic variation of Arabidopsis thaliana: the roles of geography and climate. *Mol. Ecol.* 21, 5512–5529. doi: 10.1111/j.1365-294X.2012.05709.x

Ledesma, R. D., and Valero-Mora, P. (2007). Determining the number of Factors to Retain in EFA: an easy-to-use computer program for carrying out Parallel Analysis. *Pract. Assess. Res. Eval.* 12, 1–11.

Legendre, P., and Fortin, M. J. (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Mol. Ecol. Resour.* 10, 831–844. doi: 10.1111/j.1755-0998.2010.02866.x

Legendre, P., and Legendre, L. (1998). *Numerical Ecology*. 2nd English Edn. Amsterdam: Elsevier.

Liu, J., Gao, L. M., Li, D. Z., Zhang, D. Q., and Möller, M. (2011). Cross-species amplification and development of new microsatellite loci for *Taxus wallichiana* (Taxaceae). *Am. J. Bot.* 98, e70–e73. doi: 10.3732/ajb.1000445

Liu, Q. H. (1997). Variation partitioning by partial redundancy analysis (RDA). *Environmetrics* 8, 75–85.

Lotterhos, K. E., and Whitlock, M. C. (2014). Evaluation of demographic history and neutral parameterization on the performance of F-ST outlier tests. *Mol. Ecol.* 23, 2178–2192. doi: 10.1111/mec.12725

McGaughran, A., Morgan, K., and Sommer, R. J. (2014). Environmental variables explain genetic structure in a beetle-associated nematode. *PLoS ONE* 9:87317. doi: 10.1371/journal.pone.0087317

Meze-Hausken, E. (2004). Contrasting climate variability and meteorological drought with perceived drought and climate change in northern Ethiopia. *Climate Res.* 27, 19–31. doi: 10.3354/cr027019

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika* 37, 17–23. doi: 10.1093/biomet/37.1-2.17

Morris, G. P., Ramu, P., Deshpande, S. P., Hash, C. T., Shah, T., Upadhyaya, H. D., et al. (2013). Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl. Acad. Sci. U.S.A.* 110, 453–458. doi: 10.1073/pnas.1215985110

Mulatu, B., and Grando, S. (2011). "Barley research and development in Ethiopia: an overview," in *Barley Research and Development in Ethiopia*, *Proceedings of the 2nd National Barley Research Development Review Workshop*, eds B. Mulatu and S. Grando (Holetta Ethiopia; Aleppo: HARC; ICARDA), 1–16.

Narum, S. R., and Hess, J. E. (2011). Comparison of F_{ST} outlier tests for SNP loci under selection. *Mol. Ecol. Resour.* 11, 184–194. doi: 10.1111/j.1755-0998.2011.02987.x

Negassa, M. (1985). Patterns of phenotypic diversity in an Ethiopian barley collection, and the arussi-bale highland as a center of origin of barley. *Hereditas* 102, 139–150.

Nosil, P., Funk, D. J., and Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Mol. Ecol.* 18, 375–402. doi: 10.1111/j.1365-294X.2008.03946.x

Nunes, V. L., Beaumont, M. A., Butlin, R. K., and Paulo, O. S. (2011). Multiple approaches to detect outliers in a genome scan for selection in ocellated lizards (*Lacerta lepida*) along an environmental gradient. *Mol. Ecol.* 20, 193–205. doi: 10.1111/j.1365-294X.2010.04936.x

Oksanen, J., Blanchet, G., Kindt, R., Legendre, P., O'Hara, R., Simpson, G., et al. (2010). *Vegan: Community Ecology Package*. R package version 1. 17–12.

Parsell, D. A., and Lindquist, S. (1993). The function of heat-shock proteins in stress tolerance: degradation and reactivation of damaged proteins. *Annu. Rev. Genet.* 27, 437–496.

Peakall, R., and Smouse, P. E. (2006). GENALEX 6: genetic analysis in Excel. Population genetics software for teaching and research. *Mol. Ecol. Notes* 6, 288–295. doi: 10.1111/j.1471-8286.2005.01155.x

Peres-Neto, P. R., Legendre, P., Dray, S., and Borcard, D. (2006). Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology* 87, 2614–2625. doi: 10.1890/0012-9658(2006)87[2614:VPOSDM]2.0.CO;2

Pérez-Figueroa, A., García-Pereira, M. J., Saura, M., Rolán-Alvarez, E., and Caballero, A. (2010). Comparing three different methods to detect selective loci using dominant markers. *J. Evol. Biol.* 23, 2267–2276. doi: 10.1111/j.1420-9101.2010.02093.x

Poland, J. A., Brown, P. J., Sorrells, M. E., and Jannink, J. L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE* 7:e32253. doi: 10.1371/journal.pone.0032253

Poncet, B. N., Herrmann, D., Gugerli, F., Taberlet, P., Holderegger, R., Gielly, L., et al. (2010). Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. *Mol. Ecol.* 19, 2896–2907. doi: 10.1111/j.1365-294X.2010.04696.x

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000b). Inference of population structure using multi locus genotype data. *Genetics* 155, 945–959. doi: 10.1086/302959

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000a). Association mapping in structured populations. *Am. J. Hum. Genet.* 67, 170–181. doi: 10.1111/j.1471-8286.2007.01758.x

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Pyhäjärvi, T., Hufford, M. B., Mezmouk, S., and Ross-Ibarra, J. (2013). Complex patterns of local adaptation in teosinte. *Genome Biol. Evol.* 5, 1594–1609. doi: 10.1093/gbe/evt109

R Development Core Team. (2008). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org

SAS. (2011). *SAS/STAT TM Users guide*. Cary, NC: SAS Institute Inc.

Schaffner, S., and Sabeti, P. (2008). Evolutionary adaptation in the human lineage. *Nat. Educ.* 1, 14.

Schluter, D. (2001). Ecology and the origin of species. *Trends Ecol. Evol.* 16, 372–380. doi: 10.1016/S0169-5347(01)02198-X

Schluter, D. (2009). Evidence for ecological speciation and its alternative. *Science* 323, 737–741. doi: 10.1126/science.1160006

Schoville, S. D., Bonin, A., François, O., Lobreaux, S., Melodelima, C., and Manel, S. (2012). Adaptive genetic variation on the landscape: methods and Cases. *Annu. Rev. Ecol. Evol. Syst.* 43, 23–43. doi: 10.1146/annurev-ecolsys-110411-160248

Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., et al. (2013). An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS ONE* 8:e54603. doi: 10.1371/journal.pone.0054603

Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Stat. Soc. B.* 64, 479–498. doi: 10.1111/1467-9868.00346

Storz, J. F. (2005). Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol. Ecol.* 14, 671–688. doi: 10.1111/j.1365-294X.2004.02437.x

Tanto, T., and Demissie, A. (2001). "A comparative genetic diversity study for four major crops managed under Ethiopian conditions. Managing biodiversity in agricultural system," in *Proceedings of the International symposium*, (Montreal, QC).

Temunovic, M., Franjic, J., Satovic, Z., Grgurev, M., Frascaria-Lacoste, N., and Fernández-Manjarrés, J. F. (2012). Environmental heterogeneity explains the genetic structure of Continental and Mediterranean populations of *Fraxinus angustifolia* Vahl. *PLoS ONE* 7:e42764. doi: 10.1371/journal.pone.0042764

USDA (2002). Accessed on April 24, 2013: Available online at: http://www.fas.usda.gov/pecad2/highlights/2002/10/ethiopia/baseline/eth_annual_rainfall.htm

Vavilov, N. I. (1951). The origin, variation, immunity and breeding of cultivated plants. *Chron. Bot.* 13, 1–366.

Vierling, E. (1991). The role of heat shock proteins in plants. *Annu. Rev. Plant Biol.* 42, 579–620.

Wahid, A., Gelani, S., Ashraf, M., and Foolad, M. R. (2007). Heat tolerance in plants: an overview. *Environ. Exp. Bot.* 61, 199–223. doi: 10.1016/j.envexpbot.2007.05.011

Wang, T., Chen, G., Zan, Q., Wang, C., and Su, Y. J. (2012). AFLP genome scan to detect genetic structure and candidate loci under selection for local

adaptation of the invasive weed *Mikania micrantha*. *PLoS ONE* 7:e41310. doi: 10.1371/journal.pone.0041310 e41310

Westengen, O. T., Berg, P. R., Kent, M. P., and Brysting, A. K. (2012). Spatial structure and climatic adaptation in African maize revealed by surveying SNP diversity in relation to global breeding and landrace panels. *PLoS ONE* 7:e47832. doi: 10.1371/journal.pone.0047832

Wright, S. (1946). Isolation by distance under diverse systems of mating. *Genetics* 31, 39–59.

Yu, J., Pressoir, G., Briggs, W. H., Vroh, B. I., and Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Zeven, A. C. (1998). Landraces: a review of definitions and classifications. *Euphytica* 104, 127–139.

Zhao, Y., Vrieling, K., Liao, H., Xiao, M., Zhu, Y., Rong, J., et al. (2013). Are habitat fragmentation, local adaptation and isolation-by-distance driving population divergence in wild rice *Oryza rufipogon*? *Mol. Ecol.* 22, 5531–5547. doi: 10.1023/A:1018683119237

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership