2017

# DISPERSAL, GENETIC STRUCTURE, NETWORK CONNECTIVITY AND CONSERVATION OF AN AT-RISK, LARGE-LANDSCAPE SPECIES

Todd Bartholomew Cross

DISPERSAL, GENETIC STRUCTURE, NETWORK CONNECTIVITY AND CONSERVATION OF AN AT-RISK, LARGE-LANDSCAPE SPECIES

By
TODD BARTHOLOMEW CROSS
B.S. Biology, Wheaton College, Wheaton, Illinois, 2006

Dissertation
presented in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in Wildlife and Fisheries Biology

University of Montana
Missoula, MT

May 2017

Approved by:

Scott Whittenburg, Dean of The Graduate School
Graduate School

Dave E. Naugle, Co-Chair
Wildlife Biology Program

Michael K. Schwartz, Co-Chair
Wildlife Biology Program

Fred W. Allendorf
Division of Biological Sciences

Jeffrey M. Good
Division of Biological Sciences

Jon Graham
Department of Mathematical Sciences

Cross, Todd B., Ph.D., Spring 2017 Wildlife and Fisheries Biology

Abstract Title: Dispersal, genetic structure, network connectivity and conservation of an at-risk, large-landscape species

Co-Chairperson: David E. Naugle
Co-Chairperson: Michael K. Schwartz

## ABSTRACT

Wide-ranging species face many threats to genetic connectivity. In light of these threats, one major challenge is the efficient use of scarce resources for the conservation of these species. Setting conservation priorities for landscapes and connectivity can be informed using molecular genetics, and can ensure the efficient use of scare resources to maximize returns in biodiversity conservation.

The greater sage-grouse (*Centrocercus urophasianus*; hereafter sage grouse) is a species of conservation concern that spans eleven state boundaries, land managed by multiple agencies, and one international boundary. Across the species' distribution, the threats to the genetic connectivity range from agricultural conversion to energy development, to catastrophic wildfire. In order to prioritize management as threats loom, there is considerable interest in gaining insight into the species' population genetic substructure, dispersal capabilities, and range-wide genetic connectivity. The insights gained and be used to prioritize management efforts to preserve or restore genetic diversity and connectivity. This dissertation is composed of an investigation of population genetic substructure, breeding season dispersal, and the characterization of a range-wide genetic network for conservation prioritization.

Limitations in greater sage-grouse dispersal have resulted in the existence of five subpopulations across the northeastern range of the species, none of which appears to be genetically isolated. The genetic structure discovered appears to have been shaped by the natural landscape and ecological features. However, recent disturbances associated with human alteration of the landscape may have increased subpopulation divergence. Existing state conservation areas align well with genetic subpopulation structure allowing straightforward translation of management planning to the conservation of genetic diversity and connectivity. Simulation-based evaluation of the analytical methods used to detect subpopulation structure provided insight into interpretation of the evolutionary history of subpopulation divergence.

While many individuals remained philopatric to the same breeding sites (leks) year after year, more individuals dispersed to alternate leks. Evidence for sex-biased dispersal did not exist: either in tendency to disperse nor in distances traveled. Dispersal appears costly, as there was a greater occurrence of mortality among farther dispersing individuals. Individuals dispersed within, into and out of designated conservation areas, providing additional evidence that these areas are not isolated. Breeding dispersal likely counteracts the effect of philopatry, fostering gene flow.

Using network theory, I characterized the patterns of range-wide genetic connectivity among spring breeding congregations (leks), finding that connectivity is greatest among neighboring leks. The entire network is connected such that there are no isolated subunits. Hubs of genetic connectivity exist, evidenced by increased measures of both local and global network centrality, indicative of their importance to maintaining gene flow across the entire species'

range. These high-centrality hubs are centrally located within the species' distribution, with concentrations within the Upper Snake River Basin of Idaho and the Green River Basin of Wyoming. Conservation efforts to protect these areas could prove essential to securing range-wide genetic connectivity into the future. Overall, this research provides insight into how to use molecular genetic analyses of substructure, dispersal, and connectivity of a continuously distributed species across a vast landscape to inform management and prioritize conservation actions.

## ACKNOWLEDGEMENTS

The following body of research was made possible by the collaboration of more people than I could list here, many of whom I have never met. Throughout my time working on this project, I have been fueled by the enthusiasm and insights of the people I have interacted with along the way.

I am very grateful to both my advisors: Dave Naugle, a master of communication, humor, and seeing the big picture, and Mike Schwartz, a man who takes interest in both the personal and professional development of those he mentors. Both are true servants at heart, invested in the future of conservation both personally and professionally. I would also like to thank Fred Allendorf, Jeff Good, and Jon Graham for their investment in my research and service on my committee.

I have also benefitted personally and professionally from the relationships formed with my fellow lab members, peers, and adventure compatriots: Gretchen Roffler, Keith Slauson, Jody Tucker, Taylor Wilcox, Katie Zarn, Marisa (Lipsey) Sather, Joe Smith, Rebecca (Smith) Newton, Jason Tack, Kellie Carim, Pat Cross, Hilary Eisen, Blake Lowry, Adam Moreno, Ed O'Donnell, and Robin Steenweg. These relationships formed over the course of my time in Missoula have been essential to my personal and professional development. I am blessed to call you all friends.

My time working at the USFS Rocky Mountain Research Station's National Genomics Center for Wildlife and Fish Conservation has enabled interaction with some of the brightest and most helpful minds. Specifically, Cory Engkjer, Kevin McKelvey, Lucretia Olson, Kristy Pilgrim, Roberta Steele, David Wright, and Mike Young. Furthermore, I have had the pleasure of leading a team of dedicated laboratory technicians: Kara Bates, Nasreen Broomand, Taylor Dowell, W. Scott Hampton, Randi Lesagonicz, Inga Ortloff, Sara Schwarz, and Kate Welch. I am so appreciative of their dedication to and enthusiasm for this project.

This research would not have been possible without the many state, federal, and non-governmental organization (NGO) biologists and technicians who have spent lifetimes working tirelessly in the field and behind desks for the conservation of the Greater Sage-Grouse. Thanks to John Carlson [Bureau of Land Management (BLM)], Jake Chaffin [U.S. Forest Service (USFS)], Ben Deeble (Big Sky Upland Bird Association), Brad Fedy (University of Waterloo), Don Kemner [Idaho Fish and Game (IDFG)], Steve Knick (U.S. Geological Survey [USGS], retired), Rick Northrup [Montana Fish, Wildlife and Parks (MTFWP)], Sara Oyler-McCance (USGS), Aaron Robinson [North Dakota Game and Fish (NDGF)], Jeff Row (University of Waterloo), Travis Runia [South Dakota Game, Fish and Parks (SDGFP)], Dale Tribby (BLM, retired), Susan Werner (IDFG), Catherine Wightman (MTFWP), and David Wood (USGS).

This research was supported by grants from the Montana and Dakotas Bureau of Land Management, the Great Northern Landscape Conservation Cooperative, and the Natural Resources Conservation Service—Sage Grouse Initiative. Additional support was provided by the U.S. Forest Service, Montana Fish, Wildlife and Parks, North Dakota Game and Fish, South Dakota Game, Fish & Parks, as well as the Western Association of Fish & Wildlife Agencies, California Department of Fish & Game, Colorado Division of Wildlife, Idaho Fish & Game, Nevada Division of Wildlife, Oregon Fish & Wildlife, Utah Wildlife Resources, Washington Department of Fish and Wildlife, Wyoming Game & Fish Department.

Finally, I would like to thank my family for their unbounded love and support. To my sister, Katie, in whom I have always had the closest friend to share life with and to explore the

wilds outside of doors. To my parents for teaching me the value of hard work, and for their love and imparted wisdom, which I am certain is one of the greatest blessings I will ever receive in this life. To my mother, for teaching patience, compassion, and the joy of keeping lists. To my father, the embodiment of a true renaissance man, for continually sharing with me the love for all creatures great and small, all things bright and beautiful, and all things wise and wonderful. There's no doubt in my mind that our family time together in and on the mountains, woods, lakes, and streams is what seeded and continues to fuel my love of wild things and wild places.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1

## INTRODUCTION AND OVERVIEW

One major challenge in conservation is the efficient use of scarce resources for the protection of biodiversity. Optimization of resource use is often referred to as conservation triage (Bottrill et al. 2008). There is a large body of science on the approach by which to identify landscapes to target for conservation (Possingham *et al.* 2000). Within the field of landscape prioritization, where to target first is based on biodiversity targets and the cost of conservation action. This approach assumes different costs and different amounts of biodiversity within landscapes, and balances the two to select an optimal conservation target. However, this approach does not consider connectivity among landscapes.

Recently, conservation prioritization methods have been extended to consider prioritization of connectivity among patches, as isolated landscapes are undesirable (Dilkina *et al.* 2016). When prioritizing connectivity, the assumption is often made that all parts of the landscape are producing the same number of successful dispersers and so are of equal value to connectivity. Under this approach, the highest priority is to preserve connectivity among all landscapes. However, not all landscapes are of equal value, nor do all landscapes produce the same number of successful dispersers. At the forefront of conservation planning is the fusion of these two methods: landscape and connectivity prioritization. The overarching question of my research is how can we use molecular genetics to prioritize landscapes and connectivity for wide-ranging species?

At the forefront of North American species of conservation concern is the greater sage-grouse. The sage grouse is an iconic species of the American west and a sentinel species for sagebrush landscapes (Rowland *et al.* 2006, Thacker *et al.* 2012, Holloran & Anderson 2005, Hagen *et al.* 2007). In fact, they are so reliant upon sagebrush that through the winter months, their diet consists of over 97% sagebrush (Connelly *et al.* 2000). The remaining intact sagebrush ecosystems of the west face multiple large-scale threats that vary across the range of the species. These threats include agricultural conversion, wildfire, exurban development, pinyon-juniper expansion, and energy development. Sage grouse are very sensitive to the alteration of their sagebrush habitat such that they are considered the canary in the coalmine for the sagebrush ecosystem. The birds will vacate when disturbances become too great. Sage grouse once ranged across 1.2 million square kilometers extending from the southern part of Canada's prairie provinces of Alberta, Saskatchewan & British Columbia south to Arizona, from the Pacific states east to Nebraska. Now, the species occupies less than 0.7 million square kilometers across eleven Western states and two Canadian provinces. Only 56% of pre-settlement range remains (Schroeder *et al.* 2004). Currently, these iconic birds exist in decreased density and increasingly isolated populations. In an unprecedented collaborative effort, state, federal, and private partners banded together to delineated priority areas for conservation (PACs) composed of the highest quality habitat with the greatest number of birds. These PACs encompass 271,000 square kilometers. However, there is concern that these PACs not become isolated islands of sage grouse, especially as threats increase and resources are limited (Finch *et al.* 2016). Therefore, there is an increasing desire to quantify population structure and connectivity for this wide-ranging and continuously distributed species, and to prioritize crucial landscapes required to maintain connectivity.

My dissertation research asks the question: how can molecular genetic approaches be used to prioritize landscapes and connectivity for wide-ranging species? I address this question by coupling large landscape-scale sampling with population genetic analyses to better understand patterns of genetic connectivity. In the four interlocking chapters of this dissertation, I first, identify population substructure and use the findings to provide insight into management boundary delineation; second, test using simulations whether a common analytical approach used to detect substructure are valid; third, evaluate dispersal using genetic mark recapture; and finally, use network analysis to simultaneously quantify substructure and gene flow. Each of these chapters addresses the overarching goal of enhancing conservation prioritization.

## RESEARCH OBJECTIVES AND FINDINGS

*Hierarchical population structure and insight into management boundary delineation*

The previous chapter showed that sage grouse breeding dispersal occurs for both sexes across broad areas. Cumulatively dispersal patterns shape population structure. Therefore, in chapter 2 I asked the questions:

- *What patterns of genetic substructure exist across the northeastern range of the species?*
- *Is the observed substructure consistent with management boundaries and landscape features?*

I genotyped 1,499 individuals from 297 leks across the northeastern range of the sage grouse (Montana, North Dakota and South Dakota), and found five subpopulations separated by mountain ranges or valleys occupied by river-roadway corridors. Subpopulations also occupied different elevations and were surrounded by different dominant sagebrush subspecies. I detected "isolation by distance," the genetic phenomena of increased genetic distance among individuals as distance increases. I also discovered genetic admixture among subpopulations at the most distal portions of the study area, which suggests that the cumulative effect of individual dispersal translates into long-range connectivity. Existing protected conservation areas aligned well with genetic subpopulations, and could be grouped in accordance with genetic subpopulation structure.

This chapter has been published in *Conservation Genetics* under the title "Hierarchical population structure in sage grouse provides insight into management boundary delineation," and is co-authored by David Naugle, John Carlson, and Michael Schwartz.

*Validation of hierarchical analysis to detect genetic substructure in natural populations*

The previous chapter relied heavily on a commonly used approach to delineate population substructure often referred to a hierarchical substructure analysis. Despite being used in over dozens of publications (e.g., Coulon *et al.* 2008; Balkenhol *et al.* 2014; Vähä *et al.* 2008; Lukoschek *et al.* 2008; Cheng *et al.* 2014; Warnock *et al.* 2010; Cross *et al.* 2016; Viricel & Rosel 2014) this approach has never been validated. In Chapter 3, I examined the validity of analyzing genetic subpopulation structure using hierarchical substructure analysis. This method entails first using Bayesian clustering analysis to detect genetic substructure and to group

samples into putative subpopulations by minimizing both gametic disequilibrium and deviation from Hardy-Weinberg Proportions within groups. Next, each subpopulation is independently analyzed to detect further genetic substructure. This process is repeated until each subpopulation appears panmictic. This method is suggested to reveal additional subpopulations that go undiscovered when only performing the first round of analysis and subdivision. I used this method in Chapter 3 after having seen its common use in the literature. In chapter 3, I simulated populations with known substructure to investigate whether using this method can reveal true subpopulation structure with greater sensitivity and precision. In this chapter, I addressed the following questions:

- *Does hierarchical substructure analysis reveal true subpopulation structure?*
- *If so, does it do so with greater sensitivity and precision than traditional approaches?*

I found that hierarchical substructure analysis does not reveal true subpopulation structure that would go undetected if only one round of analysis were completed. Instead, I found that the $\Delta K$ statistic reveals the most deeply rooted substructure, while the mean LnP($K$) statistic reveals more recently subdivided subpopulations. I also found that with a greater the number of subpopulations involved (with increased complexity), there is lower sensitivity and precision in analysis results. My findings suggest that the hierarchical approach to population substructure analysis is not necessary to discover complex population substructure. With these insights, I still find that, despite the use of hierarchical analysis in Chapter 2, the population substructure discovered using a combination of the $\Delta K$ and mean LnP($K$) statistics is defensible given similar population structure results with both the hierarchical method and the use of maximum mean LnP($K$).

*Long-distance breeding dispersal*

Due to the difficulty of documenting long movements in animal species, long distance-dispersal is poorly understood. Field research has documented sage grouse natal dispersal distances, seasonal migration distances, and breeding behavior, but little is known about breeding dispersal distances. Furthermore, most of these studies have been limited in geographic extent and sample size. Therefore, in Chapter 4 I asked the following questions:

- *Is there evidence for breeding dispersal among genetically recaptured sage grouse?*
- *If so, is there evidence for sex-bias in dispersal predisposition or distances traveled?*

I addressed this question by using a dataset of 3,244 individuals from 763 leks throughout Idaho, Montana, North Dakota, and South Dakota to examine the occurrence of breeding dispersal and philopatry among spring breeding congregations (leks). I recaptured over 2% of individuals, finding 41 instances of breeding dispersal, with seven dispersal events > 50 km, including one of 194 km. I also documented 39 instances of philopatry, with capture and recapture spanning up to 5 years. I found no difference between the sexes in distance travelled or in the tendency to disperse versus remain philopatric. I also documented many movements within and among existing protected conservation areas. However, I did not observe any individuals moving among the five subpopulations discovered in Chapter 2. My genetic

approach to detecting dispersal is unprecedented in the number of captures of long-distance exchanges among animal populations largely due to both sample size and geographic scope.

This chapter has been published in *The Condor: Ornithological Applications* under the title "Genetic recapture identifies long-distance breeding dispersal in Greater Sage-Grouse (*Centrocercus urophasianus*)," and is co-authored by David Naugle, John Carlson, and Michael Schwartz.

*High-connectivity hubs of gene flow across the species' range*

Within this final chapter, I used recently developed network models to provide insight into how gene flow among leks relates to the population substructure evaluation in chapter 2 when informed by the dispersal findings of chapter 4. More specifically, I sought to identify which leks were most important to maintaining gene flow locally and range-wide. In Chapter 5, I addressed the following questions:

- *Of the four common network classes, to which is the range-wide genetic network most similar?*
- *What does network structure reveal about genetic connectivity?*
- *Which leks are keystone (i.e., most important to maintaining network connectivity and persistence when ranked by centrality measures)?*

I used a dataset of 6,723 individuals from 1,417 leks across the range of sage grouse coupled with network theory to model patterns of gene flow in order to gain insight into the importance of leks' contributions to range-wide genetic connectivity. I found a completely connected network without any isolated subnetworks. This network was structured most similarly to the small-world class, which revealed that connectivity was greatest among neighboring leks, and that gene flow traverses the range quite rapidly often through leks that function as hubs of genetic connectivity. I also used measures of network centrality to discover multiple leks that appeared to function as the hubs of genetic connectivity. These leks were located within several major river basins across the West and are centrally located within the species' range. Protecting these critical hubs would likely help to secure range-wide genetic connectivity into the future.

**SYNTHESIS AND SIGNIFICANCE**

One of the greatest challenges in conservation is the efficient use of limited resources for the protection of biodiversity. To meet this challenge, managers must perform conservation triage, where decisions are made to prioritize conservation targets of highest importance (Bottrill et al. 2008). My research fuses two methods at the forefront of conservation prioritization: landscape and connectivity prioritization. By coupling large landscape scale sampling with population genetic analyses and simulations, I advance the approaches for discovering and interpreting patterns of genetic connectivity and the subsequent use of these patterns for the efficient prioritization of landscape and connectivity conservation for wide-ranging species

I implement a multi-pronged approach based in molecular genetics that provides a dynamic, cross-referenced means by which to prioritize conservation actions. I use network theory to identify crucial conservation targets in a way that can be scaled to any management-

relevant extent, while considering each target's importance to maintaining range-wide connectivity. Previous work has measured overall network structure (Bunn *et al.* 2000; Garroway *et al.* 2008), and here I provide new insight into the utility of network analysis by ranking populations by importance to maintenance of genetic connectivity and by identifying keystone populations that contribute disproportionately to connectivity, but which might be undervalued by measures of geographic centrality or abundance. I show that multiple hubs essential to genetic connectivity exist across the range, and that if conserved, these populations will have the greatest conservation benefit. I show that genetic connectivity discovered using network approaches can be validated by studying both genetic substructure and dispersal. I provide novel insight by demonstrating this validation: I first identify biologically meaningful groups based on genetic substructure, then show that genetic substructure is in agreement with both network structure and existing management boundaries, allowing for straightforward translation between management planning and the conservation of genetic diversity and connectivity. Furthermore, I provide insight into the use and interpretation of results from one of the most common Bayesian approaches to detecting genetic population substructure by demonstrating that the common method of hierarchical substructure analysis should be reconsidered given the method's tendency to generate spurious results. Through genetic recapture, I provide novel insight into avian dispersal and movements among protected areas and show that dispersal patterns validate genetic substructure, network structure, and the resultant patterns of connectivity. Overall, my dissertation research provides a better understanding of the means by which population prioritization can be accomplished using molecular genetics.

These approaches and strategies can be applied with any species. Herein, I used greater sage-grouse as a model system. Persistence of sage grouse on the western landscape is a major conservation priority spanning eleven states and one international boundary, and involving many stakeholders including state and federal land and wildlife management agencies (U.S. Fish and Wildlife Service 2010, U.S. Fish and Wildlife Service 2015). In order to manage sage grouse effectively, these stakeholders seek science relevant to their conservation efforts. Given the scale of the conservation efforts, prioritizing conservation actions across the species' range to maximize efficacy and efficiency of limited resources is a primary objective. My research advances the field of conservation prioritization from multiple angles: through an evaluation of population substructure, an evaluation of individual dispersal across vast landscapes, and an evaluation of a range-wide genetic network that simultaneously quantifies substructure and gene flow.

## DISSERTATION FORMAT

Chapters 2 and 4 were formatted and published in accordance with the guidelines of two specific peer-reviewed scientific journals. Chapter 2 was published in *Conservation Genetics* (Cross *et al.* 2016), and Chapter 4 was published in *The Condor: Ornithological Applications* (Cross *et al.* 2017). Chapters 3 and 5 are formatted for *Molecular Ecology Resources* and *Molecular Ecology*, respectively. Each one of these chapters is the product of important contributions from many other scientists. It is for this reason that I refer to this collaborative effort by using the collective "we" throughout the remainder of the dissertation.

# CHAPTER 2

## HIERARCHICAL POPULATION STRUCTURE IN GREATER SAGE-GROUSE PROVIDES INSIGHT INTO MANAGEMENT BOUNDARY DELINEATION


## ABSTRACT

Understanding population structure is important for guiding ongoing conservation and restoration efforts. The greater sage-grouse (*Centrocercus urophasianus*) is a species of concern distributed across 1.2 million $km^2$ of western North America. We genotyped 1499 greater sage-grouse from 297 leks across Montana, North Dakota, and South Dakota using a 15-locus microsatellite panel, then examined spatial autocorrelation, spatial principal components analysis, and hierarchical Bayesian clustering to identify population structure. Our results show isolation by distance, suggesting that the cumulative effect of short-range dispersal translates to long-range connectivity. We found primary and secondary hierarchical genetic substructure. These subpopulations occupy significantly different elevations and are surrounded by divergent vegetative communities with different dominant subspecies of sagebrush, each with its own chemical defense against herbivory. We propose five management groups reflective of genetic subpopulation structure. These genetic groups are largely synonymous with existing priority areas for conservation. On average, 85.8 % of individuals within each conservation priority area assign to a distinct subpopulation. Our results largely support existing management decisions regarding subpopulation boundaries.


## INTRODUCTION

The evolution of species and their ecological communities is influenced by the effect of historic physiogeographic features (Carrol *et al.* 2007; Lomolino *et al.* 2006; Wiens 2007). Across northwestern North America, the advance and recession of glaciers, formation of mountains, and carving of river valleys have all sculpted modern landscapes into ecologically distinctive areas composed of distinct assemblages of soils, vegetation, and wildlife (Carstens *et al.* 2005; Shafer *et al.* 2010; Soltis *et al.* 1997). Many of these landscapes are being rapidly altered by anthropogenic forces (Ricketts 1999), which may affect genetic connectivity within and among wildlife populations (Short Bull *et al.* 2011). Big sagebrush (*Artemisia tridentata*) likely colonized and diversified in North America from Eurasia via Beringia during the late Tertiary or early Quaternary (McArthur and Plummer 1978; Stanton *et al.* 2002).

Big sagebrush is a topographic climax dominant species that provides soil stability and ground cover and functions as critical habitat for at least 350 species of birds, reptiles, and mammals (Allen *et al.* 1984; Green *et al.* 2001; Green and Flinders 1960; Monsen and Shaw 2000; Wambolt 1996). However, the geographic extent of sagebrush and the plant and animal communities it supports has been drastically reduced and fragmented by anthropogenic disturbances including cultivation, energy development, invasive species, wildfire, and exurban development (Braun 1998; Braun *et al.* 2002; Copeland *et al.* 2009; Knick *et al.* 2003; Murphy *et al.* 2013; Naugle *et al.* 2004, 2006, 2011). The combined effect of sagebrush fragmentation and loss poses a major threat to the greater sage-grouse (*Centrocercus urophasianus*) a sentinel species for sagebrush ecosystem integrity (Smits and Fernie 2013).

Greater sage-grouse, henceforth sage grouse, is a species of conservation concern, an icon of sage-steppe ecotypes, an umbrella species for shrub-grassland communities (Rowland *et al.* 2006), and an indicator species for landscape scale connectivity (Aldridge *et al.* 2008). Sage

grouse are sagebrush obligates—relying on sagebrush for every aspect of their life history: food (Thacker *et al.* 2012; Doherty *et al.* 2008; Wallestad and Eng 1975), nesting (Holloran and Anderson 2005), brood rearing (Hagen *et al.* 2007), and spring breeding congregations known as leks (Connelly *et al.* 2000; Wallestad and Schladweiler 1974). Males battle with one another to claim the center of the lek and energetically display to potential mates. Females appear to select their mate based on phenotypic traits (Gibson *et al.* 1991); this mate selection process can occur many times with multiple mates during a single breeding season (Semple *et al.* 2001). Most females nest within 5 km of the lek (Holloran and Anderson 2005). Lek attendance by males is significantly correlated with female lek attendance (Bradbury *et al.* 1989) and despite long seasonal migratory movements (up to 240 km; Smith 2012) and large home ranges (4–195 $km^2$; Connelly *et al.* 2011a, b), fidelity to leks and stability in lek location is well documented (Dalke *et al.* 1963; Dunn and Braun 1985; Emmons and Braun 1984; Patterson 1952; Wallestad and Schladweiler 1974). However, sage grouse may shift or abandon leks because of persistent disturbance or alteration of sagebrush cover (Holloran *et al.* 2010; Walker *et al.* 2007).

Sage grouse once occupied over 1.2 million $km^2$ (Edminster 1954; Schroeder *et al.* 2004). The species now occupies less than 0.67 million $km^2$ across 11 western states and two Canadian provinces (Patterson 1952; Schroeder *et al.* 2004)—56 % of its range compared to pre-western settlement (Schroeder *et al.* 2004). An additional 29 % of the remaining species' range is likely at risk of extirpation (Aldridge *et al.* 2008). Increased geographic isolation and declines of sage grouse populations range-wide coincides with fragmentation and loss of sagebrush (Copeland *et al.* 2009; Schrag *et al.* 2011). Due to loss of habitat and subsequent population declines, the U.S. Fish and Wildlife Service was petitioned to consider listing the species under the Endangered Species Act in 2010. The species was found warranted for listing but precluded by higher priority actions (U.S. Fish and Wildlife Service 2010); however, as a condition of a court approved settlement agreement, a status review was required. In September 2015, the USFWS determined the species is not warranted for listing, due to the species' relative abundance (which increased since 2010), widespread distribution, and reduced extinction threat (U.S. Fish and Wildlife Service 2015).

In light of the USFWS listing decisions, state and federal agencies have drafted comprehensive conservation planning strategies. Using the current large-scale understanding of population structure, expert opinion, and published research, management agencies in all western states have collaborated to draft management plans that identify and protect the areas deemed most important for sage grouse survival and reproduction (U.S. Fish and Wildlife Service 2013). As part of their planning strategy, the Western Association of Fish and Wildlife Agencies (WAFWA), whose membership is composed of twenty-three states, has delineated seven Management Zones (MZs; Stiver *et al.* 2006) by grouping sage grouse populations and subpopulations that occur within common floristic provinces identified by Connelly *et al.* (2004). In addition, Montana Fish, Wildlife & Parks (MTFWP), North Dakota Game and Fish (NDGF) and South Dakota Game, Fish & Parks (SDGFP) have collectively delineated 20 Priority Areas for Conservation (PACs) to protect the highest densities of sage grouse based on male lek attendance.

The goal of these PACs is to protect important lek complexes and to conserve associated habitat (Montana Fish, Wildlife & Parks 2014). On public lands, the BLM and USFS completed their largest planning effort in history to implement regulatory mechanisms that safeguard sage-steppe habitats (U.S. Fish and Wildlife Service 2015). Since 2010, the NRCS-led Sage Grouse Initiative (SGI) has reduced fragmentation of large and intact sage-steppe habitats, increasing the

acquisition of conservation easements by 1809 %, totaling 361,984 acres (NRCS 2015b). Through 2018, SGI has committed another US$211 million to conserve an additional 3.5 million acres range-wide (NRCS 2015a). In Montana and the Dakotas, 870,000 acres of priority habitat will be conserved by acquiring additional conservation easements, prioritizing restoration of intervening croplands and by grazing livestock sustainably across the landscape (NRCS 2015a. In Montana, an executive order and US$10 million for conservation projects back the implementation of conservation actions within and among these PACs to benefit sage grouse and sage grouse habitat via the Greater Sage Grouse Stewardship Act (State of Montana 2014, 2015). The agencies involved have recognized the need for continued incorporation of the best available science in executing their conservation strategies such that their focus is not only directed at conservation management within individual PACs, but also at planning for connectivity among PACs to prevent isolation and divergence of existing populations in the future (Finch *et al.* 2016).

Relatively little is known about sage grouse genetic variation, genetic population structure, and population connectivity at a high-resolution, regional scale relevant to state and regional federal managers. At the broad scale, mitochondrial DNA analysis yielded 80 haplotypes, distributed into two distinct monophyletic clades, both of which showed signals of allopatric fragmentation and gene flow restricted by distance (Oyler-McCance *et al.* 2005), also known as isolation by distance (IBD; Wright 1943). All of the populations across the northeastern range belonged to clade II, which is hypothesized to have expanded northward after the last glacial maximum (Zink 2014). The most comprehensive evaluation of population structure and genetic diversity examined range-wide genetic structure using samples from 46 locations. Nuclear marker analysis showed that geographic distance was the most significant factor shaping ten genetic subpopulations across the species' range (Oyler-McCance *et al.* 2005). Three of the subpopulations occupy the northeastern portion of the species' range (hereafter ''the northeastern range'')—two in Montana, and one in North Dakota and South Dakota (Oyler-McCance *et al.* 2005). In a high-resolution, small-scale population structure analysis of northern Montana and southern Canada two subpopulations were discovered that were located north and south of the Milk River (Bush *et al.* 2011).

In this study, we had three objectives. Our primary objective was to quantify sage grouse genetic population structure and gene flow at a high resolution. Our secondary objective was to compare existing management group boundaries (MZs and PACs) to genetic population structure and characterize gene flow among management groups. Our final objective was to initiate exploration of the relationship between population structure and major landscape features (vegetation and elevation) across the northeastern range of the species. To address our objectives we used a combination of individual and population-based approaches.

## METHODS

### Study area and sampling

We used 3481 spatially referenced sage grouse feather and blood samples representative of the northeastern range of the species in the United States of America (154,800 $km^2$ of sagebrush-dominated ecosystems in Montana, North Dakota, and South Dakota; Fig. 2.1). Feather samples were collected from leks using non-invasive methods (Bush *et al.* 2005, Segelbacher 2002) where they were dropped or plucked by sage grouse during lekking activity, while blood samples were collected from sage grouse on leks as part of radio telemetry research efforts. Samples were collected from 351 leks (median: 9 samples per lek, interquartile range [IQR]: 8) between 2009

and 2012 by field biologists and technicians with the Bureau of Land Management, MTFWP, and the Montana Audubon Society.

**Laboratory analysis**

*DNA Extraction*

Feather DNA was extracted from the quill (calamus) using QIAGEN's DNeasy Blood and Tissue Kit and the user developed protocol for purification of total DNA from nails, hair, or feathers. We modified the protocol by incubating samples for a minimum of 8 h after addition of Proteinase K and by eluting DNA with 100 µl of Buffer AE. Feather samples were extracted in a lab used only for non-invasive DNA extraction in order to avoid potential contamination from samples with higher DNA concentrations. Blood samples were extracted using QIAGEN's DNeasy Blood and Tissue Kit and the protocol for nucleated blood.

*Microsatellite DNA amplification and electrophoresis*

We amplified 15 variable microsatellite loci and one sex-diagnostic locus in eight multiplex polymerase chain reactions (PCR) (Table 2.4 in Supplementary Material). We initially used 16 microsatellite loci; however, we removed TUD3 due to a significant heterozygote excess. Results were nearly identical with and without the inclusion of TUD3. The total PCR volume of 10 µl contained 2.0 µl of DNA template and 8 µl reagent mix. Locus specific reaction mixes, annealing temperatures, and thermal cycler profiles are listed in Tables 2.2 and 2.3 in the Supplementary Material. PCR mixes included 1 µM dye-labeled forward primer (IDT® Custom DNA Oligos), 1 µM reverse primer (Eurofins MWG Operon Custom DNA Oligos), 1 U *Taq* polymerase (Invitrogen™), 1X reaction buffer (Invitrogen™), 200 µM of each dNTP, 2.0 mM $MgCl_2$ (Invitrogen™), and 1.5 mg/ml BSA. Each PCR included two samples of known genotype to allow for calibration of genotypes across gels, for identification of PCR-generated stutter, and for identification of sample contamination.

We loaded PCR product into a 6 % polyacrylamide gel in a Li-Cor Biosciences 4300 DNA Analyzer (Li-Cor Biosciences, Lincoln, Nebraska USA) and electrophoresed for 2 h 30 min at 1500 V, with a current of 40 mA, and a power of 40 W. PCR products were visualized and genotypes were determined using Li-Cor Biosciences' e-Seq software.

*Genotyping*

To ensure correct genotypes from low quality and low quantity feather DNA samples, each sample was PCR amplified at least twice across the 15 variable microsatellite loci to screen for allele dropout, stutter artifacts, and false alleles. To minimize genotyping error, two independent observers scored each genotype. If there was any discrepancy between the first two genotypes or if samples failed to amplify in both replicates, samples were PCR amplified and genotyped an additional two times. If successful in this repeat analysis, genotypes were retained. If the repeat analysis failed, the sample was assigned a missing score for that locus.

To screen samples for quality control, we removed from analysis any individual for which amplification failed at five or more loci. After removal of poor quality samples, genotypes were screened to ensure consistency between allele length and length of the microsatellite repeat motif. We used program DROPOUT v2.3 (McKelvey and Schwartz 2005) and package ALLELEMATCH v2.5 (Galpern *et al.* 2012) in R (R Core Team 2016) to screen for genotyping error and to identify and remove multiple captures of the same individual from the same lek in the same year. We quantified the power of our microsatellite locus panel to discern individuals using probability identity ($P_{ID}$; Evett and Weir 1998): the probability that two individuals drawn at random from the population could have the same genotype across all loci.

**Population genetic descriptive statistics**

We calculated the average number of alleles across 15 loci (*A*), expected heterozygosity ($H_e$), and observed heterozygosity ($H_o$) in the R package GSTUDIO (Dyer 2014). We calculated $F_{IS}$ for each locus, and tested for deviation from Hardy–Weinberg proportions (HWP) and gametic disequilibrium among loci correcting for multiple tests for significance using Bonferroni corrected p-values using program GENEPOP v4.5.1 (Rousset 2008). Finally, we calculated allelic richness (*AR*; El Mousadik and Petit 1996) using program FSTAT v2.9.3.2 (Goudet 1995).

Sampling scheme has been shown to have an effect on Bayesian clustering analyses in the presence of spatial autocorrelation (Schwartz and McKelvey 2009). Therefore, we tested for IBD using Mantel tests for correlation (Mantel 1967) between matrices of genetic distance (i.e., individual-based analysis of molecular variance [AMOVA; Excoffier et al. 1992] distance calculated in the R package GSTUDIO) and Euclidean geographic distance (calculated in the R package VEGAN; Oksanen *et al.* 2015) using the Pearson product-moment correlation coefficient (*r*). We calculated *r* for 80 even-distance classes and tested for significance of spatial autocorrelation using 999 permutations of the distance matrices with a Bonferoni corrected α of 0.05 in the R package VEGAN. We visualized spatial autocorrelation between samples using a Mantel correlogram. Within the correlogram, each individual bin distance was determined using the Sturges equation (Sturges 1926) to optimize the number of data points in each bin.

**Individual-based analyses**

We tested for spatial structure using individual-based spatial principal components analysis (sPCA) calculated in the R package ADEGENET (Jombart and Ahmed 2011). We also conducted a group-based principal component analysis (PCA) conducted in the R package GSTUDIO. The group-based PCA was calculated by computing the mean component scores for all individuals located within each PAC ($n$ = 1116; individuals located outside of PACs were excluded from this analysis). To estimate the number of subpopulations within the study area, we used the Bayesian clustering program STRUCTURE v2.3.4 (Pritchard *et al.* 2000). STRUCTURE groups samples so that gametic disequilibrium and deviation from HWP are both minimized within each cluster (*K*). Therefore, it can be used to evaluate the most probable *K* from the pool of individual genotypes, and to summarize genetic admixture calculated as the percent assignment to each of the clusters (*Q*-values). We used STRUCTURE to analyze all individuals for values of *K* from one to seven. We used settings recommended by Falush *et al.* (2003) for detecting subtle population subdivision using the admixture model, correlated allele frequencies among populations, and with the allele frequency distribution parameter (λ) set to 1. We allowed STRUCTURE to infer the value of the model's Dirichlet parameter (α), for the degree of admixture, from the data. We set the length of burn-in period before the start of data collection to 1,000,000, the number of MCMC repetitions after burn-in to 1,000,000, and used no prior. That is, we did not use user-defined population-of-origin for each individual, nor user-defined sampling location for each individual. We completed ten replicate runs for each value of *K*.

To determine the most probable value of *K*, we used STRUCTURE HARVESTER v0.6.94 (Earl and vonHoldt 2012) to plot the mean and standard deviation of the natural log of the probability of each value of *K* [Ln P(*K*)] and to plot Δ*K*, a second-order statistic formulated by Evanno *et al.* (2005). We selected the most probable *K* by examining the plots generated and selecting the value of *K* at which Δ*K* is greatest and/or at which the Ln P(*K*) plot asymptotes. The most probable number of *K* clusters was determined as one if the Δ*K* plot indicated a *K* of two but the Ln P(*K*) plot clearly exhibited the highest Ln P(*K*) at *K* = 1. Once we determined the most

probable value of *K*, we used program CLUMPP v 1.1.2 (Jakobsson and Rosenberg 2007) to average individual assignment to each of the *K* clusters across the ten STRUCTURE replicates for each value of *K*. In CLUMPP, we used 30,000 repeats of the greedy method with greedy option two and the pairwise matrix similarity statistic, G.

We examined hierarchical subpopulation structure following determination of the primary substructure. First, we split the pool of sample genotypes based on each individual's maximum percent ancestry (*Q*-value). Next, we independently analyzed each of the primary hierarchical clusters for additional substructure using STRUCTURE and the same settings as above. We continued our hierarchical analysis in this way until we inferred that the most likely number of clusters was one ($K = 1$) for each sample pool analyzed (according to our interpretation of the Ln P(*K*) plot), as suggested by Coulon *et al.* (2008). Within each inferred primary and secondary cluster we calculated HWP and gametic disequilibrium among loci, *A*, *AR*, $H_e$, $H_o$, and $F_{IS}$ as described above. Among clusters, we calculated $F_{ST}$ (Weir and Cockerham 1984) using program GENEPOP v4.5.1 (Rousset 2008), and constructed an $F_{ST}$-based dendrogram by first transforming the pairwise $F_{ST}$ matrix to a maximum distance matrix (using the supremum norm) and then via the single linkage method (closely related to a minimal spanning tree) using the R package STATS (R Core Team 2016).

As an independent verification of subpopulation structure, we compared the results of STRUCTURE to those of program TESS v. 2.3 (Francois *et al.* 2006; e.g., Chen *et al.* 2007). TESS is another Bayesian analysis, which uses not only each sample's genotype but also a prior informed by both global trend surfaces and spatial autocorrelation to determine the most probable number of population clusters. In TESS we completed 10 iterations for each value of *K* from two to ten, allowing admixture, using the conditional autoregressive (CAR) Gaussian model, the default program values for spatial interaction parameter (0.6) and degree of trend [Linear (1)], with 12,000 MCMC sweeps and a burn-in of 2,000 sweeps for each run. From the output of TESS, we selected *K* by calculating the change in mean Deviance Information Criterion (ΔDIC: averaged across all ten replicates for each value of *K*) between successive *K* values and selecting the value of *K* for which we calculated the greatest value of ΔDIC (largest decrease). After selecting the most probable value of *K*, we again used CLUMPP to average individual assignment to each of the *K* clusters across the 20 TESS replicates for the most probable value of *K*. Here we used the same CLUMPP parameter options as stated above.

**Comparison of population structure to priority areas for conservation and landscape characteristics**

After we determined the most probable hierarchical population structure via the individual-based analyses described above, we calculated the percent of each genetic subpopulation within each of the current sage grouse PACs recognized by the U.S. Fish and Wildlife Service (2013) and within management zones recognized by WAFWA (Stiver *et al.* 2006). We then compared genetic population structure to landscape characteristics to explore factors that may have influenced population structure, as follows. First, we buffered all sampled leks by 5 km, a management buffer suggested for the protection of nesting habitat (Connelly *et al.* 2000). Next, we calculated the percent area within the buffer covered by the three big sagebrush subspecies (basin big sagebrush: *Artemisia tridentata tridentata*, mountain big sagebrush: *A. t. vaseyana*, and Wyoming big sagebrush: *A. t. wyomingensis*) and associated vegetation, basing our calculations on a broad-scale remote-sensed vegetation map of the western USA (Comer *et al.* 2002). Second, we used a digital elevation map (Gesch *et al.* 2002) to measure elevation—a proxy for abiotic factors such as temperature and precipitation–within each 5 km buffer. We

calculated the range and variability of elevation across all 5 km buffers in each subpopulation. Finally, we compared mean elevation across all 5 km buffers in each subpopulation using the Kruskal–Wallis rank sum test.

## RESULTS

### Genotyping

After removing inferior quality samples (those that did not genotype across at least 2/3 of all loci; $n$ = 718) and all duplicate genotypes (the same individual sampled from the same lek in the same year; $n$ = 1260) we retained 1499 of 3481 samples genotyped (43.1 %). The 1499 genotypes from feathers ($n$ = 1445) and bloods ($n$ = 54) were from 324 leks (mean of 4.6 samples per lek ± 3.69 [SD]): 1393 samples from 297 leks in Montana (mean = 4.69 samples per lek ± 3.75 [SD]), 15 samples from eight leks in North Dakota (mean = 1.88 samples per lek ± 0.99 [SD]), 69 samples from 17 leks in South Dakota (mean = 4.06 samples per lek ± 2.98 [SD]), and 22 samples from three leks across the northern border of Wyoming (mean = 7.33 samples per lek ± = 2.08 [SD]). Samples sizes per lek in North Dakota are low; however, our sampling reflects the lek counts, where there was an average of 4.82 males per lek counted on an average of 14 leks during the time we collected samples (Robinson 2014). Using our 15-locus panel, P ID was $8.24 \times 10^{-18}$, providing substantial power to discern individuals. We were able to determine sex for 1487 (99.2 %) of the final individual genotypes (Table 2.1). Each occurrence of an individual recaptured at more than one lek in the same year ($n$ = 22), or in more than one year at the same ($n$ = 18) or at different leks ($n$ = 7) was retained for all analyses.

### Population genetic descriptive statistics

Before taking into account any population structure, 12 of the 15 loci showed a deficit of heterozygotes (Table 2.2), and 47 out of 120 pairwise locus comparisons showed gametic disequilibrium. Within the sample pool there were an average of 13.60 alleles per locus ranging from seven alleles at TUT3 to 32 alleles at MSP11 (Table 2.4 in Supplementary Material), with an expected heterozygosity of 0.752, and an $F_{IS}$ of 0.045 (Table 2.3a).

### Individual-based analyses

We found substantial genetic spatial structure using the PAC-based PCA. The sampled PACs sorted by geographic location from west to east as PC1 values increase and from north to south as PC2 values increase (Fig. 2.2). Together, the first (35.6 %) and second (17.2 %) principal components (PC) captured 52.8 % of the variance in the data.

The first eigenvector (Fig. 2.5a in Supplementary Material) and second eigenvector (Fig. 2.5b in Supplementary Material) of the sPCA show the major north–south and east–west genetic divergence patterns. We retained both the first and second eigenvectors, as they composed 5.1 and 2.9 % of the variation in the data respectively. The eigenvalues of both the first and second eigenvector were highly spatially auto-correlated. When decomposed into genetic variance and spatial autocorrelation components (measured by Moran's $I$), the first eigenvector was composed of 76.1% genetic variance and 26.6% spatial autocorrelation while the second eigenvector was composed of 60.4% genetic variance and 19.3% spatial autocorrelation (Fig. 2.6 in Supplementary Material).

We used the Bayesian clustering results—both the natural log of the probability of $K$ [Ln P($K$)] and $\Delta K$ statistic plots—to infer that the sample genotypes represented three spatially distinct population clusters (Fig. 2.7 in Supplementary Material): a subpopulation in southwestern Montana (SW), a subpopulation in northern Montana (N), and a subpopulation in

southeastern Montana, North Dakota, and South Dakota (SE). Genetic admixture was present among these subpopulations (Fig. 2.1b).

After assigning each individual to one of the three primary subpopulations by maximum *Q*-value, we proceeded with STRUCTURE analysis of each subpopulation independently. We discovered secondary hierarchical population structure within both the SE and the SW subpopulations (Fig. 2.1c and d), but not within the N subpopulation (Fig. 2.4a and d in Supplementary Material). The SE group divided into secondary eastern (SE-E) and western (SE-W) subpopulations (Figs. 4b, e and 5b all in Supplementary Material) while the SW subpopulation split into secondary northern (SW-N) and southern (SW-S) subpopulations (Figs. 4c, f and 5c all in Supplementary Material).

We discovered no evidence for additional hierarchical substructure. The cumulative primary and secondary subpopulations in Fig. 2.1a depict each individual as a member of a subpopulation if its *Q*-value was at least 70% for that cluster. Despite the disparity in sample sex ratio (Table 2.1), the STRUCTURE results for females and males analyzed independently were consistent with substructure identified with the full dataset (results not shown).

After grouping samples into primary hierarchical subpopulations based on maximum *Q*-value, no loci were out of HWP in more than one subpopulation (Table 2.2). However, five loci were out of HWP in the southeastern subpopulation and one locus was out of HWP in the southwestern subpopulation. Only one locus pair was in significant gametic disequilibrium in more than one subpopulation (BG16/SGCA5 in the N and SE subpopulations; $a = 4.167 \times 10^{-4}$; $p < 0.001$). After grouping samples into the secondary hierarchical subpopulations, no loci were out of HWP in more than one subpopulation (Table 2.2, $\alpha = 3.333 \times 10^{-3}$), and only one locus pair was significantly in gametic disequilibrium in more than one population (MS06.8/MSP11 in both the N and SW-S subpopulations). When grouping samples by PAC, no loci were found significantly out of HWP in more than one PAC (only BG16 in Carter, and SGCTAT1 in McCone-Garfield), and none of the locus pairs was in gametic disequilibrium in more than one PAC.

When both primary and secondary hierarchical substructure were considered, the average number of alleles per locus, per population ranged from 8.73 in the SW-N subpopulation to 11.73 in the N subpopulation (mean = 10.39 ± 4.12 [SD]). Averages for $H_e$ across all subpopulations were 0.747 ± 0.16 [SD] (Table 2.3a). The average number of alleles per locus, per PAC ranged from 5.27 in the North Dakota (ND) PAC to 10.07 in the South Phillips (SP) PAC (mean = 8.31 ± 3.17 [SD]). Within PACs, average $H_e$ was 0.723 ± 0.17 [SD] (Table 2.3b).

We found a significant positive correlation between increasing geographic distance between individuals and increasing genetic divergence (Mantel test: $r = 0.271$, $p = 0.001$; Fig. 2.10 in Supplementary Material). Genotypes were significantly positively correlated for samples up to 242.6 km apart and significantly negatively correlated for samples over 311.9 km apart.

Divergence among geographically proximal subpopulations and PACs, as measured by $F_{ST}$, was low compared to those that were distal (Fig. 2.3a, b; raw values in Table 2.7a, b in Supplementary Material, respectively). The dendrogram revealed divergence relationships among the subpopulations detected using Bayesian clustering analyses. The greatest divergence was found between the two most distal subpopulations, the SE-E and SW-N subpopulations ($F_{ST} = 0.0777$), while the least divergence was found between the two most proximal subpopulations, the SE-E and SE-W subpopulations ($F_{ST} = 0.0174$). Mean $F_{ST}$ among all genetic subpopulations discovered using STRUCTURE was 0.0473 ± 0.0243 [SD]. Mean $F_{ST}$ among all PACs was 0.0380 ± 0.0216 [SD]. The dendrogram of the PACs (Fig. 2.3b) highlights that while the

Beaverhead 1 and Beaverhead 2 PACs are less diverged ($F_{ST}$ = 0.004) than are the Beaverhead 2 and Beaverhead 3 PACs ($F_{ST}$ = 0.005), it is the greater divergence between Beaverhead 1 and Beaverhead 3 ($F_{ST}$ = 0.009) that drives Beaverhead 2 and Beaverhead 3 to cluster more closely.

Finally, we found strong concordance between the resulting individual assignment of TESS and that of the primary substructure analysis in STRUCTURE (Fig. 2.11 in Supplementary Material). In TESS, three clusters ($K$ = 3) showed the largest drop in DIC between successive values of $K$ = 2–3 (ΔDIC = 2748.2).

**Comparison of population structure to priority areas for conservation and landscape characteristics**

*Management zones and priority areas for conservation (PACs)*

The N and SE-E subpopulations are located almost entirely within WAFWA MZ-I (99.4 and 99.7 %, respectively; Fig. 2.12 in Supplementary Material). Similarly, the SW-N subpopulation is entirely contained and the SW-S subpopulation is nearly entirely contained within MZ-IV (96.1 %). The SE-W subpopulation spans MZ-I (74.7 %), II (21.1 %) and IV (4.2 %).

On average, 85.8 % of the individuals within each PAC belong to the same subpopulation (Fig. 2.4). Six PACs encompass a majority of the N subpopulation (mean = 80.6 %), five PACs envelop a majority of the SE-E subpopulation (mean = 88.7 %), two PACs envelop a majority of the SE-W subpopulation (mean = 87.7 %), one PAC encompasses a majority of the SW-N subpopulation, and two PACs envelop a majority of the SW-S subpopulation (mean = 92.0 %).

*Vegetation*

Within 5 km of leks in the N subpopulation, grasses dominated the landscape (53 %) and Wyoming big sagebrush (19 %; Table 2.5; Fig. 2.13 in Supplementary Material) was the dominant sagebrush subspecies. Within 5 km of leks in the SE-E subpopulation, grasses dominated the landscape (61 %) and Wyoming big sagebrush (10 %) was the dominant sagebrush subspecies. The 5 km buffer of the SE-W subpopulation leks contained mixed xeric shrubland (39 %) and grasses (31 %) with Wyoming big sagebrush (15 %) and mountain big sagebrush (2 %) as the dominant structural elements. In the SW-N subpopulation, mountain big sagebrush was dominant within the 5 km buffer of leks (63 %) and was interspersed with basin big sagebrush (24 %). Finally, within 5 km of the leks in the SW-S subpopulation, mountain big sagebrush (77 %) and Wyoming big sagebrush (15 %) were the most prevalent subspecies of sagebrush with a minority cover of grasses (4 %).

*Elevation*

Mean elevation within 5 km of leks in each subpopulation was significantly different among subpopulations (Kruskal–Wallis rank sum test, H = 642.77, 4 d.f., $p < 0.001$), and there was little to no overlap in elevations occupied (Fig. 2.14 in Supplementary Material). Most notably, the southwestern subpopulations occupy far higher elevations than do the other three subpopulations. Elevation increased from N to SE-E to SE-W to SW-N to SW-S subpopulations (IQR: 781–998, 910–1022, 1007–1258, 1858–2003, 2043–2202 m). The subpopulations spanned an elevation range of 671–2411 meters.

**DISCUSSION**

We detected discontinuities in sage grouse genetic connectivity across the northeastern range of the species. These discontinuities indicate divergence in allele frequencies between three primary subpopulations, and secondary hierarchical subpopulations within two of the primary subpopulations (Fig. 2.4). The greater divergence among the primary subpopulations compared

to divergence among the secondary subpopulations may be indicative of the chronology in which these subpopulations diverged.

We detected the pattern of population structure using both spatially independent (PCA: Fig. 2.2, and STRUCTURE: Fig. 2.1a) and spatially dependent (sPCA: Fig. 2.5 in Supplementary Material, and TESS: Fig. 2.11 in Supplementary Material) analyses. Concordance between the non-spatial and spatial analyses supports the patterns of genetic subpopulation structure detected, as does the concordance between the two Bayesian clustering methods (STRUCTURE and TESS) and the two ordination methods (PCA and sPCA). Using the hierarchical approach to analyze genetic structure, we detected genetic substructure within the SE and SW subpopulations that otherwise would have remained cryptic. If we had not independently analyzed the primary substructure subpopulations, we would not have found the substructure resulting from low divergence among secondary subpopulations. This oversight may occur if divergence among primary subpopulations is far greater than divergence among secondary subpopulations (Latch *et al.* 2006).

The primary population structure we detected (i.e., three subpopulations across the northeastern range: N, SE, SW) supports results of Oyler-McCance *et al.* (2005). However, our sampling intensity allowed us to define spatial subpopulation structure with greater resolution. Furthermore, we detected additional substructure within two of these primary subpopulations: both in the southwestern Montana subpopulation (SW) and in the southeastern Montana/ North Dakota/South Dakota subpopulation (SE).

Within the southeastern portion of our study area, our results align with those of Schulwitz *et al.* (2014) who discovered hierarchical structure within sage grouse populations in northwestern Wyoming and southeastern Montana. They found that their Southeastern Montana, South Powder River Basin, and North Powder River Basin sampling locations constituted a single subpopulation (that which we have called the SE-E subpopulation). Furthermore, their results indicate that there is genetic discontinuity between this subpopulation and subpopulations throughout the rest of Wyoming. The magnitude of genetic divergence among the subpopulations identified in both studies were similar (see Table 2.7b in Supplementary Material and Table 2 in Schulwitz *et al.* 2014).

We did not find the same subpopulation substructure that Bush *et al.* (2011) found within the N subpopulation via their analysis of thirteen microsatellite loci. Our results may differ due to our lack of samples from sage grouse in Alberta and Saskatchewan. Alternatively, there may have been a change in substructure between the years during which Bush *et al.* sampled and the years during which we sampled.

Similar to prior studies, we detected a signal of IBD. Prior research has suggested IBD as one of the primary drivers of genetic divergence for sage grouse (Bush *et al.* 2011; Davis *et al.* 2015; Oyler-McCance *et al.* 2005; Schulwitz *et al.* 2014). We found IBD, which suggests that despite the relatively short dispersal distances documented (7–9 km; Dunn and Braun 1985), the cumulative effect of these dispersals translates into long-range connectivity.

**Physiogeographic correlates with genetic subpopulations**

We found evidence for the effect of pre-European physiogeographic landscape processes on contemporary subpopulation structure. First, elevations occupied by the subpopulations are significantly different (Fig. 2.14 in Supplementary Material). The two southeastern populations—most notably the SE-W subpopulation—occupy elevations more similar to the two southwestern subpopulations, which may facilitate gene flow among these subpopulations if individuals are locally adapted to habitats at this elevation. The smaller divergence among these

subpopulations ($F_{ST}$ = 0.0508) compared to the divergence of the N and southwestern subpopulations ($F_{ST}$ = 0.0590) may be due to gene flow among the southwestern and southeastern clusters through subpopulations located in Idaho and Wyoming which are outside the geographic extent of this study. Future research could examine whether genetic connectivity through subpopulations in Idaho and Wyoming fosters gene flow among the southeastern and southwestern subpopulations.

Second, there are stark differences in the vegetative community assemblage concomitant with subpopulations. Most notable are the differences in dominant subspecies of sagebrush, a plant vital to the species' requirements for food, cover, and nest success (Connelly *et al.* 2011a, b). We found that the two southwestern subpopulations occur on a landscape dominated by mountain big sagebrush, a subspecies that grows on high-elevation mountain slopes (Jaeger *et al.* 2016). These two subpopulations are also surrounded by basin big sagebrush, a subspecies that grows in deep-soil drainage basins (Jaeger *et al.* 2016). Basin big sagebrush does not occur within proximity of leks in any other subpopulation and mountain big sagebrush only composes 2 % of the vegetation proximal to leks in the SE-W subpopulations. The SW-N subpopulation has far more basin big sagebrush proximal to leks (24 %) than does the SW-S subpopulation (3 %) which occupies significantly higher elevation.

The three sagebrush subspecies within our study area differ in terpene composition and quantity. Terpenes are chemical compounds found in plants that function as anti-herbivory agents (Byrd *et al.* 1999; Thacker *et al.* 2012; Jaeger *et al.* 2016). Differences in terpenes among sagebrush subspecies could be an important factor in diet selection (Frye *et al.* 2013). Sage grouse prefer the palatability of mountain big, to Wyoming big, over basin big sagebrush (Rosentreter 2004), and diet selection appears linked to both lower monoterpene content and higher crude protein content within the preferred sub-species (Remington and Braun 1985).

Given the stark differences in both elevation and in prevalent sagebrush subspecies among subpopulations, it is possible that these physiogeographic factors have affected both subpopulation divergence and adaptive divergence. For example, it is possible that adaptations to terpene metabolism have arisen within subpopulations, allowing for improved digestion of the most prevalent sagebrush subspecies. Similar to our findings, Schulwitz *et al.* (2014) found that sage grouse subpopulations across northwest Wyoming occupied distinct ecoregions, composed of distinct assemblages of species and shaped by different environmental processes.

**Major river-highway corridors**

In the Rocky Mountain west, the landscape is composed of basin and range topography. Major highways are routed along large river valleys across this landscape. The combination of direct and indirect effects of these river-highway corridors and associated landscape alteration may have shaped sage grouse genetic population structure. The vast majority (95.9 %) of the N subpopulation is north of the Interstate 90/Interstate 94/Yellowstone and Bighorn River corridor, while 89 % individuals in the southern subpopulations (SE-E, SE-W, SW-N, and SW-S) existed south of the corridor (individually, 69.4 % of the southeastern subpopulation and 100 % of the southwestern subpopulation).

We found the greatest admixture within individuals on leks closest to these major river-highway corridors. That is, there was a significant negative correlation between distance from major highways that co-occur with major rivers and percent admixture ($r(1499) = -0.164$, $p < 0.001$, 95 %CI -0.212 – -0.115). This indicates that gene flow is occurring across these corridors but that these areas appear to be where subpopulations interface. However, it is difficult to

discern whether the effects of transportation or that of the natural landscape features within the corridors influence genetic structure.

Major river-highway corridors may have increased genetic divergence among subpopulations. Divergence among the N and SE-E/SE-W subpopulations (Figs. 2 and 5a in Supplementary Material) may have been influenced by impeded gene flow across the Yellowstone River and Interstate 94 corridor. Much of this area has been converted to cultivation, lek attendance has declined, and lek extirpation has been observed. The effect of highways on genetic connectivity and genetic diversity are detailed across multiple taxa (Holderegger and Di Giulio 2010; Jackson and Fahrig 2011; Proctor *et al.* 2005), especially that of wider highways with high traffic volume such as the Interstate 90/Interstate 94 corridor. Highways surrounded by cultivated land are a documented barrier to sage grouse gene flow (Bush *et al.* 2011). Distance to highways can lead to nest failure (Webb *et al.* 2012) and can decrease peak male sage grouse lek attendance by over 70 % (Blickley *et al.* 2012). Furthermore, road noise has been linked to the reduction of breeding bird densities in multiple avian species (Parris and Schneider 2009; Slabbekoorn and Ripmeester 2008).

**Management and conservation implications**

The five subpopulations we detected span the boundaries of the MZs recognized by WAFWA. The main discrepancy we found between genetic population structure and the boundaries of these MZs is that the SE-W population overlaps the boundaries of the three northeastern MZs: MZ-I, MZ-II and MZ-IV (Fig. 2.12 in Supplementary Material). Depending on the goal of managing populations, it may be useful to use genetic structure in addition to common floristic provinces to set boundaries of MZs.

MTFWP, NDGFD, and SDGFP delineated management areas, called PACs, to protect areas with the greatest abundance of birds (U.S. Fish and Wildlife Service 2013). Our results provide insight into how these units can be functionally managed at multiple scales that reflect genetic connectivity. Existing PACs align very well with genetic population structure, allowing relatively simple assembly into management groups (Fig. 2.1a). These management groups could serve as the conservation units used when setting conservation goals, modifying policy, targeting conservation resources, evaluating connectivity, translocation, and regulation of harvest. For example, translocations may wish to use sage grouse from within a genetically identified subpopulation. We suggest grouping PACs into five management groups based on genetic structure: a Northern management group composed of the Fergus, McCone-Garfield, Musselshell, North Rosebud, North Valley, and South Phillips PACs; a Southeastern-East management group composed of the Carter, Cedar Creek, PRB, North Dakota and South Dakota PACs; a Southeastern-West management group composed of the Carbon 3 and Golden Valley PACs; a Southwestern-North management group composed of solely the Beaverhead 3 PAC; and a Southwestern-South management group composed of the Beaverhead 1 and Beaverhead 2 PACs (Fig. 2.4).

Each PAC encompasses a group of individuals almost entirely from a single subpopulation, with the exception of the North Rosebud PAC (Fig. 2.4). North Rosebud is the most diverse PAC by population membership, being composed mostly of individuals in the N subpopulation and nearly equal parts SE-E and SE-W (Fig. 2.4). When measuring divergence among PACs, values of F ST were smaller than when measuring divergence among genetic subpopulations. This is likely the result of mixed genetic subpopulation membership when pooling individuals by PAC.

In the PAC-based dendrogram (Fig. 2.3b), the Golden Valley PAC, located in the Southeastern-West management group, aligns more closely with PACs in the Northern

management group (Fergus, McCone-Garfield, Musselshell, North Rosebud, North Valley, and South Phillips) than it does with the PACs in the Southeastern-East management group (Carter, Cedar Creek, PRB, North Dakota, and South Dakota). We might expect Golden Valley to align most closely with the PACs within the southeastern primary subpopulation. However, the unexpected dendrogram alignment is likely due to grouping individuals from within each PAC, which for Golden Valley means 30 % of the individuals in the pack assign to the N subpopulation. Despite these few caveats, state-delineated PACs align very well with genetic subpopulation structure.

**Future research directions**

We sampled genetic data from sage grouse subpopulations located across multiple landscapes with different physiogeographic histories. These subpopulations exist at different elevations, surrounded by different assemblages of vegetation, and persist in spite of novel landscape disturbances. Maintaining connectivity and minimizing habitat fragmentation within and among these subpopulations is critical to the persistence of the species. A thorough analysis of which landscape features and disturbances have the greatest effect on genetic connectivity would lead to a better understanding of which areas on the landscape are essential to maintaining connectivity. These areas could be the targets for management action. Therefore, we suggest that future research should focus on discerning whether anthropogenic landscape alteration or natural features have had the greatest influence on genetic structure.

The coincidence of breaks in genetic continuity and underlying physiogeography suggests that it is possible that adaptive differences have arisen within subpopulations. These adaptive differences could in turn restrict contemporary gene flow among subpopulations. Therefore, quantification of adaptive variation across not only the northeastern portion, but also the entire range could lend a much better understanding of the species' phylogeography. A comparison of our analysis of subpopulation structure, an understanding of adaptive variation, and an analysis of fine-scale connectivity among leks within and among subpopulations—using network theory or a similar approach—would provide a powerful set of tools to assist managers as they set management objectives to preserve genetic diversity across multiple scales (Manel *et al.* 2010).

**ACKNOWLEDGMENTS**

This research would not be possible without the many state, federal, and NGO biologists and technicians who have spent countless hours working tirelessly in the field and behind desks for the conservation of the greater sage-grouse. We would specifically like to acknowledge Jake Chaffin (BLM, now USFS), Ben Deeble (Montana Audubon), Brad Fedy (University of Waterloo), Rick Northrup (MTFWP), Sara Oyler-McCance (USGS), Aaron Robinson (NDGFD), Travis Runia (SDGFP), Dale Tribby (BLM), Catherine Wightman (MTFWP), and David Wood (BLM). Our molecular genetic analyses and manuscript preparation would not have been possible without the assistance of Kevin McKelvey, Kristy Pilgrim, Cory Engkjer and the tireless laboratory efforts of Kara Bates, Nasreen Broomand, Taylor Dowell, Scott Hampton, Randi Lesagonicz, Inga Ortloff, Sara Schwarz, Kate Welch, and Katie Zarn, all of the USFS Rocky Mountain Research Station National Genomics Center for Wildlife and Fish Conservation. Use of trade names does not imply endorsement by the U.S. Government. The views in this article are those of the authors and do not necessarily reflect those of their employers.

**FUNDING**

This study was supported by grants from the Montana and Dakotas Bureau of Land Management (07-IA-11221643-343, 10-IA11221635-027, 14-IA-11221635-059), the Great Northern Landscape Conservation Cooperative (12-IA-11221635-132), and the Natural Resources Conservation Service—Sage Grouse Initiative (13-IA11221635-054).

**CONFLICT OF INTEREST**

The authors declare that they have no conflict of interest.

**Table 2.1.** Sample representation by sex across the 4 years of collection (2009–2012)

| Sex | 2009 | 2010 | 2011 | 2012 | Total |
|---|---|---|---|---|---|
| Females | 34 | 86 | 147 | 22 | 289 |
| Males | 533 | 176 | 119 | 370 | 1198 |
| Unknown | 5 | 7 | 0 | 0 | 12 |
| Total | 572 | 269 | 266 | 392 | 1499 |

**Table 2.2.** Per-locus tests for deviation from Hardy–Weinberg proportions. Results for subpopulations for both primary and secondary hierarchical substructure are shown. Primary and secondary subpopulations are abbreviated as follows: N (northern), SE (southeastern), SW (southwestern), SE-E (southeastern-east), SE-W (southeastern-west), SW-N (southwestern-north), SW-S (southwestern-south). Indicative adjusted nominal level (5 %) for multiple comparisons was 0.0033; all loci that significantly deviated from expected proportions are bold and marked with an asterisk

| Locus | All | Primary Subpopulations | | | Secondary Subpopulations | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | N | SE | SW | N | SE-E | SE-W | SW-N | SW-S |
| BG6 | **0.000*** | 0.310 | 0.006 | 0.043 | 0.306 | 0.140 | 0.064 | 0.063 | 0.379 |
| BG16 | **0.000*** | 0.015 | **0.000*** | 0.224 | 0.017 | 0.311 | **0.000*** | 0.729 | 0.357 |
| BG18 | 0.007 | 0.032 | 0.047 | 0.903 | 0.031 | 0.449 | 0.169 | 0.724 | 0.924 |
| MS06.4 | **0.002*** | 0.509 | 0.068 | 0.072 | 0.496 | 0.108 | 0.115 | 0.024 | 0.272 |
| MS06.6 | **0.000*** | 0.045 | 0.138 | **0.003*** | 0.037 | 0.036 | 0.819 | 0.684 | **0.000*** |
| MS06.8 | 0.0247 | 0.511 | 0.218 | 0.045 | 0.515 | 0.297 | 0.594 | 0.034 | 0.785 |
| MSP11 | **0.000*** | 0.386 | 0.209 | 0.308 | 0.364 | 0.309 | 0.463 | 0.162 | 0.489 |
| MSP18 | **0.002*** | 0.024 | **0.000*** | 0.468 | 0.026 | **0.001*** | 0.335 | 0.765 | 0.448 |
| SGCA5 | **0.002*** | 0.138 | **0.003*** | 0.074 | 0.135 | 0.251 | 0.016 | 0.589 | 0.090 |
| SGCA11 | **0.000*** | 0.116 | 0.671 | 0.048 | 0.111 | 0.791 | 0.526 | 0.188 | 0.063 |
| SGCTAT1 | **0.000*** | 0.006 | **0.000*** | 0.122 | 0.005 | 0.068 | **0.001*** | 0.695 | 0.192 |
| TTD6 | **0.000*** | 0.049 | 0.015 | 0.485 | 0.052 | 0.006 | 0.572 | 0.877 | 0.258 |
| TTT3 | **0.000*** | 0.279 | 0.215 | 0.008 | 0.282 | 0.833 | 0.010 | 0.012 | 0.318 |
| TUT3 | 0.014 | 0.263 | 0.207 | 0.418 | 0.260 | 0.537 | 0.169 | 0.524 | 0.782 |
| TUT4 | **0.002*** | 0.133 | **0.003*** | 0.760 | 0.136 | 0.007 | 0.013 | 0.368 | 0.974 |

**Table 2.3.** Measures of genetic diversity across 16 microsatellite loci within each of the subpopulations detected using STRUCTURE for both primary and secondary hierarchical substructure (**a**) and within each of the greater sage-grouse PACs (**b**)

a

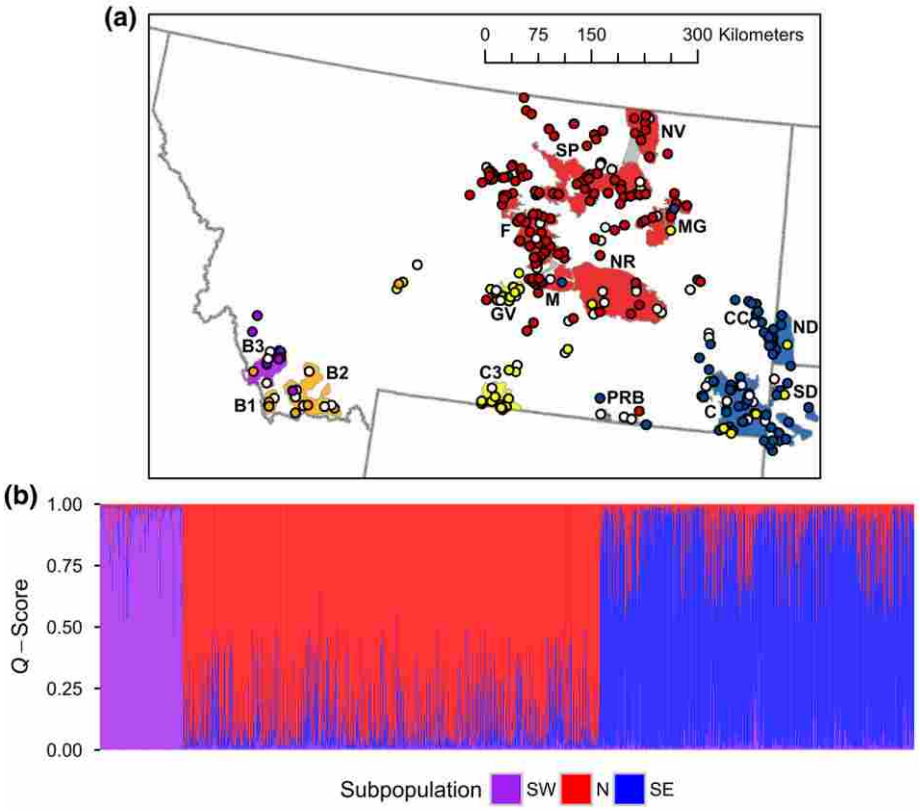
| Hierarchical Substructure | Subpopulation (abbreviation) | *n* | *A* | *AR* | $H_o$ | $H_e$ | $F_{IS}$ |
|---|---|---|---|---|---|---|---|
| All Samples | | 1499 | 13.60 | 13.51 | 0.718 | 0.752 | 0.045 |
| | Northern (N) | 767 | 11.73 | 10.26 | 0.711 | 0.727 | 0.023 |
| Primary | Southeastern (SE) | 579 | 11.93 | 10.63 | 0.713 | 0.737 | 0.033 |
| | Southwestern (SW) | 153 | 10.53 | 10.51 | 0.774 | 0.788 | 0.021 |
| | Northern (N) | 767 | 11.73 | 9.14 | 0.711 | 0.727 | 0.023 |
| | Southeastern-east (SE-E) | 323 | 10.33 | 8.72 | 0.710 | 0.722 | 0.019 |
| Secondary | Southeastern-west (SE-W) | 256 | 10.87 | 9.08 | 0.718 | 0.739 | 0.031 |
| | Southwestern-north (SW-N) | 62 | 8.73 | 8.70 | 0.747 | 0.754 | 0.017 |
| | Southwestern-south (SW-S) | 91 | 10.27 | 9.91 | 0.792 | 0.793 | 0.008 |

b

| PAC | Abbreviation | *n* | *A* | *AR* | $H_o$ | $H_e$ | $F_{IS}$ |
|---|---|---|---|---|---|---|---|
| Beaverhead 1 | B1 | 19 | 7.47 | 6.36 | 0.813 | 0.766 | -0.035 |
| Beaverhead 2 | B2 | 69 | 10.00 | 6.87 | 0.806 | 0.787 | -0.017 |
| Beaverhead 3 | B3 | 53 | 8.93 | 6.51 | 0.746 | 0.765 | 0.035 |
| Carbon 3 | C3 | 37 | 7.80 | 6.06 | 0.751 | 0.735 | -0.008 |
| Carter | C | 112 | 9.80 | 6.07 | 0.716 | 0.727 | 0.020 |
| Cedar Creek | CC | 39 | 7.20 | 5.59 | 0.716 | 0.709 | 0.003 |
| Fergus | F | 208 | 9.67 | 5.74 | 0.706 | 0.707 | 0.007 |
| Golden Valley | GV | 120 | 8.87 | 5.81 | 0.730 | 0.727 | -0.000 |
| McCone-Garfield | MG | 39 | 7.60 | 5.69 | 0.704 | 0.702 | 0.010 |
| Musselshell | M | 62 | 8.53 | 5.70 | 0.715 | 0.713 | 0.007 |
| North Dakota | ND | 11 | 5.27 | 5.17 | 0.665 | 0.660 | 0.040 |
| North Rosebud | NR | 44 | 8.73 | 5.98 | 0.701 | 0.709 | 0.023 |
| North Valley | NV | 49 | 8.20 | 5.76 | 0.722 | 0.713 | -0.001 |
| Powder River Basin | PRB | 25 | 6.86 | 5.65 | 0.677 | 0.707 | 0.060 |
| South Dakota | SD | 59 | 7.93 | 5.54 | 0.690 | 0.691 | 0.010 |
| South Phillips | SP | 196 | 10.07 | 6.00 | 0.732 | 0.733 | 0.003 |

*n* sample size, *A* average number of alleles across 15 loci, *AR* allelic richness, $H_e$ expected heterozygosity, $F_{IS} = 1-(H_o/H_e)$—a measure of departure from HWP within groups/subpopulations (positive values indicate a deficit of heterozygotes, negative values indicate an excess of heterozygotes)

**Figure 2.1.** Primary and Secondary genetic population structure for greater sage-grouse sampled in Montana, North Dakota and South Dakota as determined using STRUCTURE. *Points* show >70 % membership of individuals to each of the primary *K* = 3 clusters [N (northern): red, SE (southeastern): *blue/yellow*, and SW (southwestern): *orange/purple*], and the primary/secondary *K* = 5 clusters [N (northern): *red*, SE-E (southeastern-east): *blue*, SE-W (southeastern-west): *yellow*, SW-N (southwestern-north): *purple*, SW-S (southwestern-south): *orange*, and unassigned: *white*]. Individuals with <70 % assignment to any subpopulation are unassigned. Individuals and PACs (*colored polygons*) are colored by suggested management group membership in accordance with genetic subpopulations. PACs, listed from west to east by centroid: B3 (Beaverhead 3), B1 (Beaverhead 1), B2 (Beaverhead 2), GV (Golden Valley), C3 (Carbon 3), F (Fergus), M (Musselshell), SP (South Phillips), NR (North Rosebud), NV (North Valley), PRB (Powder River Basin 1, 2 and 3), MG (McCone-Garfield), C (Carter), CC (Cedar Creek), ND (North Dakota), SD (South Dakota). Also shown are state lines (*dashed grey lines*). Admixture plots are shown for the primary *K* = 3 clusters **b** and the secondary *K* = 2 clusters discovered within both the southeastern **c** and southwestern **d** primary subpopulations. Admixture plot subpopulation abbreviations are as listed above, and colors correspond to map (**a**)



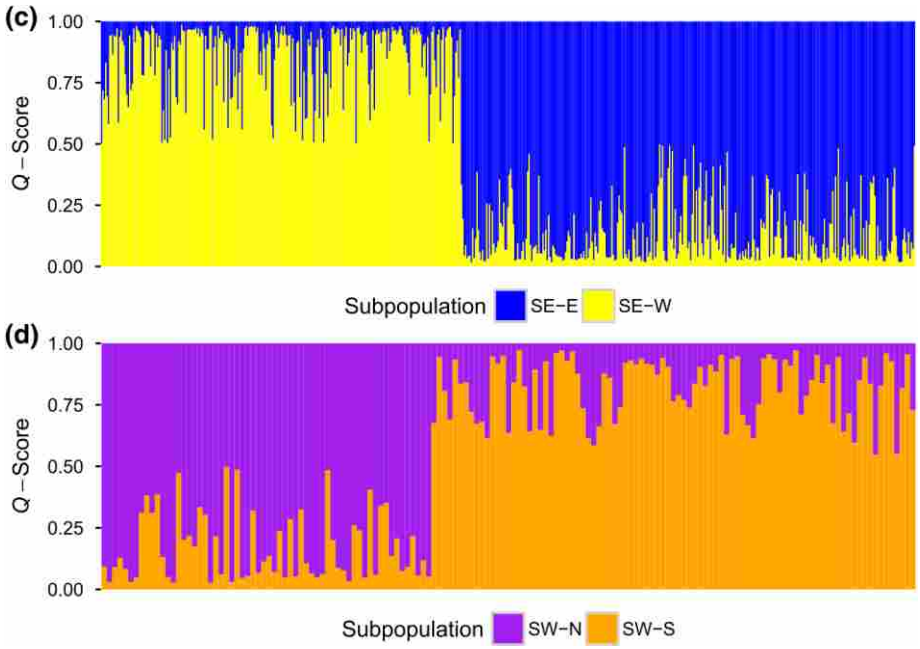

(c)
1.00
0.75
0.50
0.25
0.00
Q – Score
Subpopulation
SE–E
SE–W
(d)
1.00
0.75
0.50
0.25
0.00
Q – Score
Subpopulation
SW–N
SW–S

**Figure 2.2.** Microsatellite-genotype-based PCA plot of the mean PC score for all individuals within each of the 16 PACs. PAC abbreviations listed alphabetically: Beaverhead 1 (B1), Beaverhead 2 (B2), Beaverhead 3 (B3), Carter (C), Carbon 3 (C3), Cedar Creek (CC), Fergus (F), Golden Valley (GV), Musselshell (M), McCone-Garfield (MG), North Dakota (ND), North Rosebud (NR), North Valley (NV), Powder River Basin (1, 2 and 3) (PRB), South Dakota (SD), South Phillips (SP)

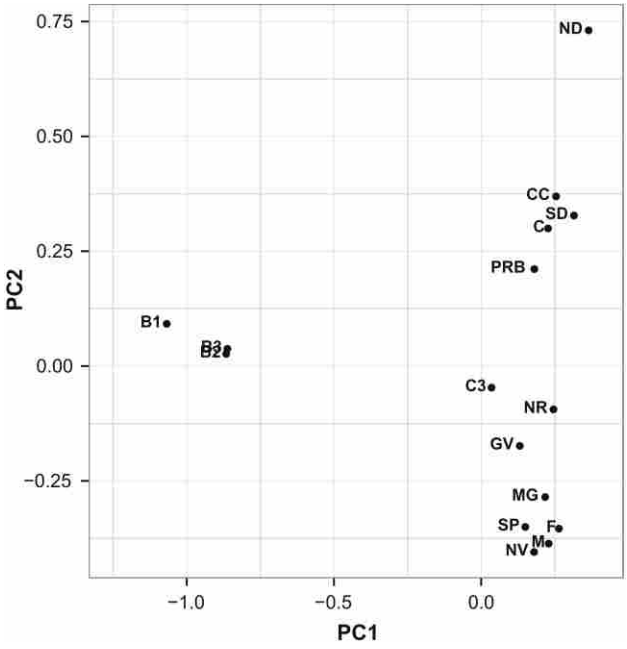
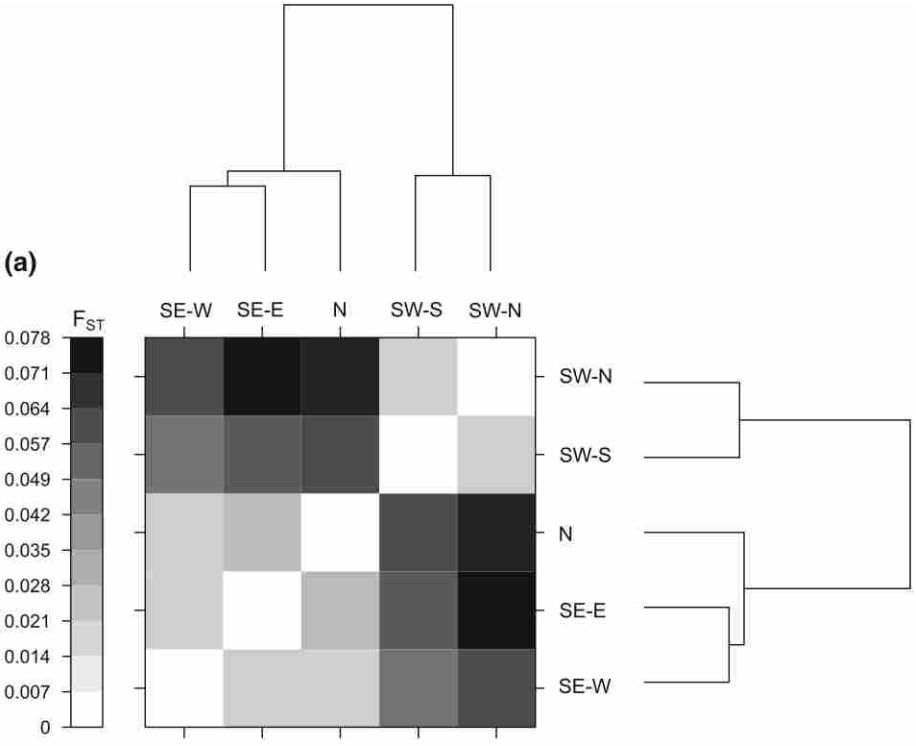


**Figure 2.3.** Genetic divergence among the five greater sage-grouse subpopulations detected using STRUCTURE (subpopulation abbreviations provided in caption to Fig. 2.1) (**a**), and among the 16 PACs sampled (PAC abbreviations provided in caption to Fig. 2.2) (**b**). Divergence is measured in pairwise comparisons using Wright's $F_{ST}$ (Wright 1949). Greater divergence—higher $F_{ST}$ —is shown as darker shades of grey. Subpopulation abbreviations listed alphabetically: N (northern), SE-E (southeastern-east), SE-W (southeastern-west), SW-N (southwestern-north), SW-S (southwestern-south). PAC abbreviations listed alphabetically: B1 (Beaverhead 1), B2 (Beaverhead 2), B3 (Beaverhead 3), C (Carter), C3 (Carbon 3), CC (Cedar Creek), F (Fergus), GV (Golden Valley), M (Musselshell), MG (McCone-Garfield), ND (North Dakota), NR (North Rosebud), NV (North Valley), PRB (Powder River Basin 1, 2 and 3), SD (South Dakota), SP (South Phillips)



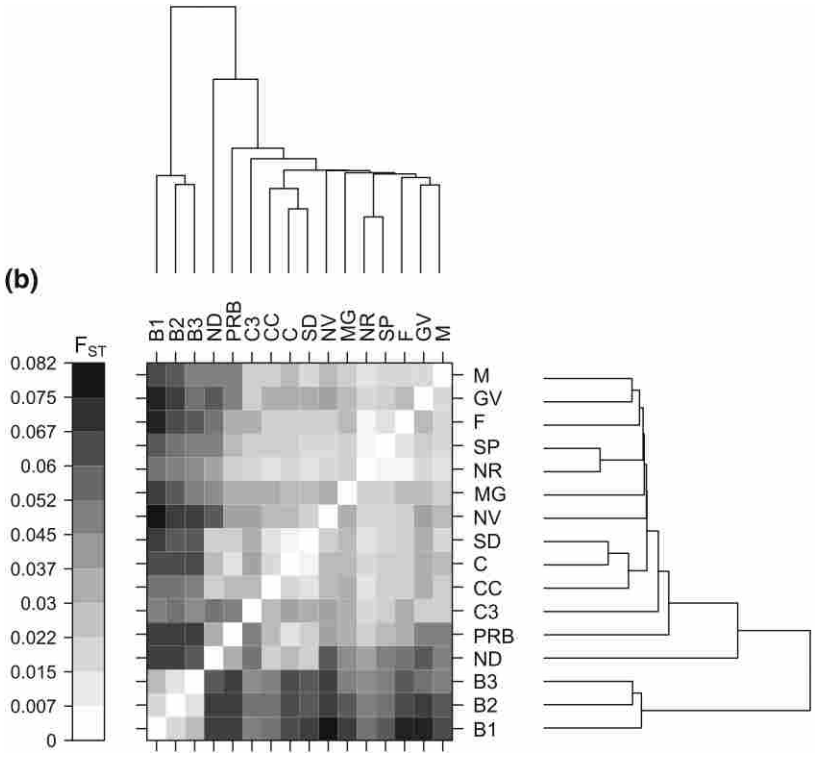
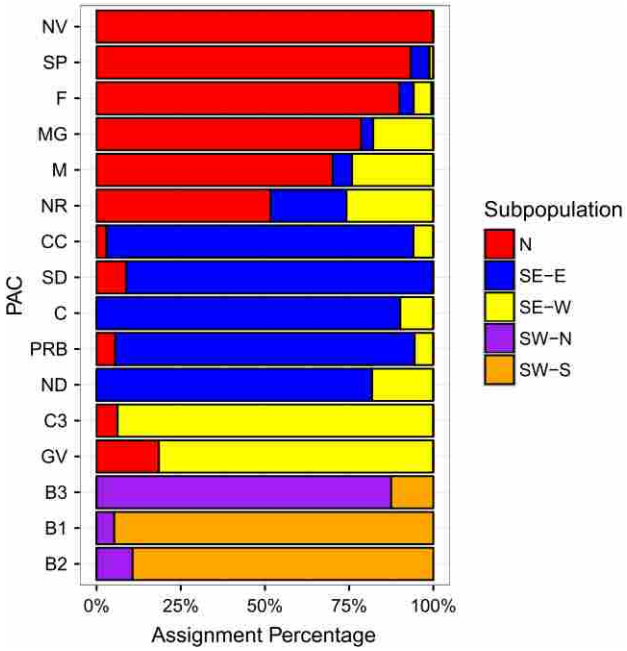

(b)
$F_{ST}$
0.082
0.075
0.067
0.06
0.052
0.045
0.037
0.03
0.022
0.015
0.007
0
B1
B2
B3
ND
PRB
C3
CC
C
SD
NV
MG
NR
SP
F
GV
M

**Figure 2.4.** Subpopulation composition within 16 PACs sorted by majority assignment percentage. PACs listed in order of display: NV (North Valley), SP (South Phillips), F (Fergus), MG (McCone-Garfield), M (Musselshell), NR (North Rosebud), SD (South Dakota), CC (Cedar Creek), C (Carter), ND (North Dakota), PRB (Powder River Basin 1, 2 and 3), C3 (Carbon 3), GV (Golden Valley), B3 (Beaverhead 3), B1 (Beaverhead 1), B2 (Beaverhead 2)



**SUPPLEMENTARY MATERIAL**

**Table 2.4.** Forward and reverse primer sequences. Asterisks indicate primers that required a redesign to increase efficacy with non-invasive samples. Repeat motif, size range, number of alleles, probability identity—$P_{ID}$, probability identity sibling—$P_{IDsib}$, and the source for the primer sequence are shown for the 16 variable microsatellite loci and 1 diagnostic sex locus (1237). After finding a heterozygote excess at TUD3 (italicized) across both primary and secondary hierarchical substructure, we dropped the locus and repeated all analyses

| Locus | F Primer Sequence (5'–3') | R Primer Sequence (5'–3') | Repeat | Size Range | # Alleles | $P_{ID}$ | $P_{IDsib}$ | Source |
|---|---|---|---|---|---|---|---|---|
| 1237 | GAGAAACTGTGCAAAACAG | TAAAGCTGATCTGGAATTTCA* | Bi-Allelic | 224 / 252 | 2 | — | — | 3 |
| BG6 | AAAGAGGCAAGCACTCACAATG | CCCTTGGAATATCCTTTAACAAAAC | $(GATA)_{15}$ | 199–311 | 18 | 0.02 | 0.30 | 5 |
| BG16 | GTCATTAGTGCTGTCTGTCTATCT | TGCTAGGTAGGGTAAAAATGG | $(CTAT)_{15}$ | 125–177 | 11 | 0.07 | 0.37 | 5 |
| BG18 | CCATAACTTAACTTGCACTTTC | CTGATACAAAGATGCCTACAA | $(CTAT)_{17}$ | 135–171 | 10 | 0.65 | 0.82 | 5 |
| MS06.4 | CCTGGAGCAACTTGAGG | GTGACATTICCCCCCAC | $(GATA)_2(GGTA)_6(GATA)$ | 130–178 | 11 | 0.06 | 0.36 | 4 |
| MS06.6 | CAAACAACTGTCTTCCAGTAAGAC | AGAGCCTICATTTCTGGCAG | $(CAT)_{16}$ | 125–185 | 20 | 0.03 | 0.33 | 4 |
| MS06.8 | GCAAAATCAATAGAAGTAGAGAGG | CAGTAGCAGCTTTGTTTGG | $(GATA)_{17}$ | 107–159 | 13 | 0.04 | 0.33 | 4 |
| MSP11 | GGTGAAAAGTGTGGCAACTG* | CATTGTCAGCTTGCAGAC | $(CA)_{22}$ | 206–272 | 32 | 0.02 | 0.31 | 4 |
| MSP18 | CAATGACAGTATTTCCCAGATTA | GAATGGTAATATACTAAGCACAGG | $(CA)_{14}$ | 98–120 | 12 | 0.03 | 0.33 | 4 |
| SGCA5 | CGGACAGGTACATCCTGGAA* | GGGAAAAGATGTCAGAATCTACAAA* | $(CA)_{12}$ | 120–140 | 11 | 0.06 | 0.36 | 7 |
| SGCA11 | GCAGTAAAGAAAATTTGGAAGCA* | TCTTGAACTGATGTTGGATTTG* | $(AC)_{14}(AG)_2(CA)_5$ | 181–203 | 11 | 0.08 | 0.38 | 7 |
| SGCTAT1 | GCGACACTGCTCCCACCT | GAAAGGTTGTAAGAGGTCGT | $(CTAT)_{11}$ | 93–133 | 11 | 0.06 | 0.36 | 7 |
| TTD6 | GGACTGCTTGTGATACTTGCT | CATGCAGATGACTTTCAGCA | $(CA)_{17}$ | 115–151 | 15 | 0.05 | 0.35 | 1 |
| TTT3 | TAGCAAACGAACCAGCCAAC* | GCTCTGAATCTGCCCATCTCT | (CATC)n | 196–228 | 11 | 0.13 | 0.44 | 2 |
| *TUD3* | *TCCAAGGGGAAAATATGTGTG* | *TTCTTCCAGCCCTAGCTTTG* | *$(TG)_{12}$* | *154–222* | *26* | *0.01* | *0.29* | *6* |
| TUT3 | CAGGAGGCCTCAACTAATCACC | CGATGCTGGACAGAAGTGAC | $(TATC)_{11}$ | 140–168 | 7 | 0.16 | 0.46 | 6 |
| TUT4 | GGAGCATCTCCCAGAGTCAG* | TCAGCTGTGAACCAGCAATC* | $(TATC)_8$ | 163–199 | 8 | 0.04 | 0.34 | 6 |

[1]Caizergues *et al.* 2001, [2]Caizergues *et al.* 2003, [3]Kahn *et al.* 1998, [4]Oyler-McCance and St. John 2010, [5]Piertney and Höglund 2001, [6]Segelbacher *et al.* 2000, [7]Taylor *et al.* 2003

**Table 2.5.** Microsatellite locus multiplexes, primer annealing temperatures, and reagent mixes used in polymerase chain reactions (PCR). Values in columns 3 through 12 are measured in µl. Columns F1–F3 indicate the amount of forward primer added to the reaction, as columns R1-R3 indicate the amount of reverse primer added to the reactions. All Reactions use: 1µM IDT® Custom DNA Oligos Forward Primer, 10µM Eurofins MWG Operon Custom DNA Oligos Reverse Primer, Invitrogen™ 5 U/µL AmpliTaq® Gold DNA *Taq* Polymerase, Invitrogen™ GeneAmp 10x PCR Buffer II (100 mM Tris-HCl, 1.5 mL pH 8.3, 500mM KCI), New England Biolabs® Inc. Deoxynucleotide Set (25 µmol 100mM ultrapure dATP, dCTP, dGTP, dTTP)—dNTP; Invitrogen™ 25mM $MgCl_2$, Bovine Serum Albumen (~66kDA, used to stabilize enzymes during digestion of DNA—to prevent adhesion of the enzyme to reaction tubes, to inactivate contaminating nucleases and proteases, to stabilize nucleic acid modifying enzymes, as a blocking agent to minimize background, and to increase PCR yield from low purity templates), and nuclease-free water. TUD3 is in italics, as we dropped this locus and repeated all analyses after finding a heterozygote excess at this locus for subpopulations in both the primary and secondary population substructure

| Primer Multiplex | Annealing Temperature (°C) | F1 | F2 | F3 | R1 | R2 | R3 | *Taq* | 10x Buffer | dNTP | $MgCl_2$ | BSA | $H_2O$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1237 / BG18 / MSP18 | 54 | 0.10 | 0.10 | — | 0.20 | 0.20 | — | 0.26 | 1.00 | 1.00 | 0.80 | 0.10 | 4.24 |
| BG16 / MS06.8 / MSP11 | 52 | 0.10 | 0.07 | 0.10 | 0.20 | 0.20 | 0.20 | 0.26 | 1.00 | 1.00 | 0.80 | 0.10 | 3.97 |
| BG6 / MS06.4 / SGCA5 | 54 | 0.10 | 0.20 | 0.07 | 0.20 | 0.20 | 0.20 | 0.26 | 1.00 | 1.00 | 1.20 | 0.10 | 3.47 |
| MS06.6/ TUT4 | 53 | 0.10 | 0.10 | — | 0.20 | 0.20 | — | 0.26 | 1.00 | 1.00 | 0.80 | 0.10 | 4.24 |
| SGCA11 / SGCTAT1 | 60 | 0.10 | 0.10 | — | 0.20 | 0.20 | — | 0.26* | 1.00 | 1.00 | 0.80 | 0.10 | 4.50 |
| TTD6 / TUT3 | 56 | 0.13 | 0.20 | — | 0.20 | 0.20 | — | 0.26 | 1.00 | 1.00 | 1.20 | 0.10 | 3.71 |
| TTT3 / *TUD3* | 55 | 0.20 | 0.10 | — | 0.20 | 0.20 | — | 0.26 | 1.00 | 1.00 | 0.80 | 0.10 | 4.14 |

**Table 2.6.** Phase, temperature, time, and number of cycles used to amplify microsatellite loci using a PCR thermocycler. Specific loci multiplexes and primer annealing temperatures are listed in Table 2 in supplementary material

| Phase | Temperature (°C) | Time (min) | Cycles |
|---|---|---|---|
| Initial Denaturation | 94 | 11:00 | 1 |
| Denaturation | 94 | 1:00 | |
| Primer Annealing | *see Table 2* | 1:00 | 44 |
| Extension | 72 | 1:00 | |
| Holding | 12 | ∞ | 1 |

**Table 2.7.** Genetic divergence among greater sage-grouse subpopulations detected using STRUCTURE—abbreviated as: northern (N), southeastern (SE), southwestern (SW), southeastern-east (SE-E), southeastern-west (SE-W), southwestern-north (SW-N), southwestern-south (SW-S) (**a**), and among the 16 priority areas for conservation (PACs)—abbreviated as: B1 (Beaverhead 1), B2 (Beaverhead 2), B3 (Beaverhead 3), C (Carter), C3 (Carbon 3), CC (Cedar Creek), F (Fergus), GV (Golden Valley), M (Musselshell), MG (McCone-Garfield), ND (North Dakota), NR (North Rosebud), NV (North Valley), PRB [Powder River Basin (1, 2 and 3)], SD (South Dakota), SP (South Phillips) (**b**). Divergence is measured in pairwise comparisons using Wright's $F_{ST}$ (Wright 1951) calculated across 15 microsatellite loci

| Hierarchical Substructure | Subpopulation Comparison | | $F_{ST}$ |
|---|---|---|---|
| Primary | N | SE | 0.0174 |
| | N | SW | 0.0615 |
| | SE | SW | 0.0547 |
| Secondary | N | SE-E | 0.0227 |
| | N | SE-W | 0.0205 |
| | N | SW-N | 0.0735 |
| | N | SW-S | 0.0618 |
| | SE-E | SE-W | 0.0174 |
| | SE-E | SW-N | 0.0777 |
| | SE-E | SW-S | 0.0603 |
| | SE-W | SW-N | 0.0645 |
| | SE-W | SW-S | 0.0547 |
| | SW-N | SW-S | 0.0196 |

(**a**)

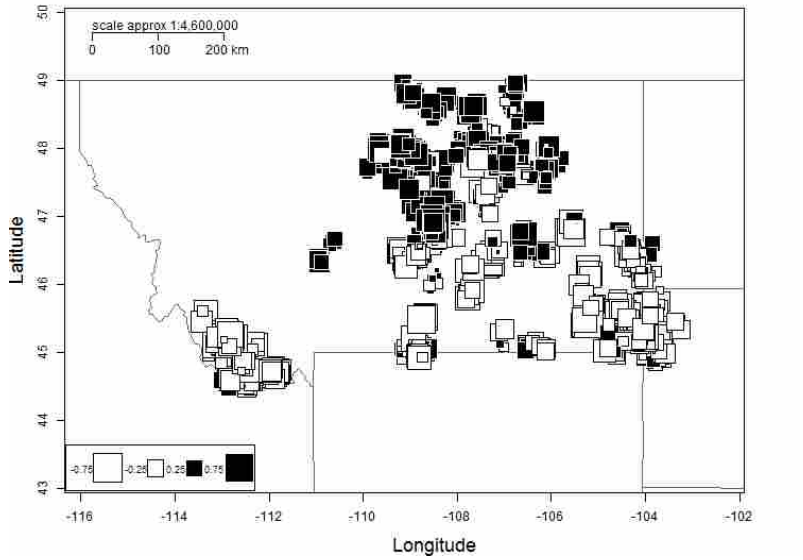
| PAC | B1 | B2 | B3 | C3 | C | CC | F | GV | MG | M | ND | NR | NV | PRB | SD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| B1 | 0.000 | | | | | | | | | | | | | | |
| B2 | 0.018 | 0.000 | | | | | | | | | | | | | |
| B3 | 0.027 | 0.011 | 0.000 | | | | | | | | | | | | |
| C3 | 0.052 | 0.059 | 0.044 | 0.000 | | | | | | | | | | | |
| C | 0.069 | 0.069 | 0.066 | 0.040 | 0.000 | | | | | | | | | | |
| CC | 0.059 | 0.055 | 0.052 | 0.028 | 0.016 | 0.000 | | | | | | | | | |
| F | 0.079 | 0.066 | 0.064 | 0.033 | 0.023 | 0.022 | 0.000 | | | | | | | | |
| GV | 0.077 | 0.074 | 0.058 | 0.022 | 0.030 | 0.032 | 0.025 | 0.000 | | | | | | | |
| MG | 0.072 | 0.063 | 0.051 | 0.034 | 0.026 | 0.034 | 0.025 | 0.025 | 0.000 | | | | | | |
| M | 0.070 | 0.061 | 0.048 | 0.021 | 0.025 | 0.022 | 0.017 | 0.015 | 0.018 | 0.000 | | | | | |
| ND | 0.071 | 0.071 | 0.064 | 0.055 | 0.024 | 0.019 | 0.054 | 0.064 | 0.041 | 0.048 | 0.000 | | | | |
| NR | 0.058 | 0.052 | 0.045 | 0.012 | 0.013 | 0.010 | 0.005 | 0.017 | 0.018 | 0.008 | 0.037 | 0.000 | | | |
| NV | 0.082 | 0.072 | 0.074 | 0.041 | 0.026 | 0.025 | 0.018 | 0.040 | 0.032 | 0.026 | 0.064 | 0.013 | 0.000 | | |
| PRB | 0.074 | 0.074 | 0.076 | 0.048 | 0.011 | 0.028 | 0.031 | 0.049 | 0.035 | 0.047 | 0.035 | 0.023 | 0.035 | 0.000 | |
| SD | 0.074 | 0.061 | 0.059 | 0.035 | 0.005 | 0.008 | 0.021 | 0.032 | 0.026 | 0.017 | 0.022 | 0.010 | 0.018 | 0.022 | 0.000 |
| SP | 0.061 | 0.057 | 0.049 | 0.020 | 0.021 | 0.018 | 0.007 | 0.023 | 0.019 | 0.013 | 0.048 | 0.002 | 0.013 | 0.028 | 0.017 |

(**b**)

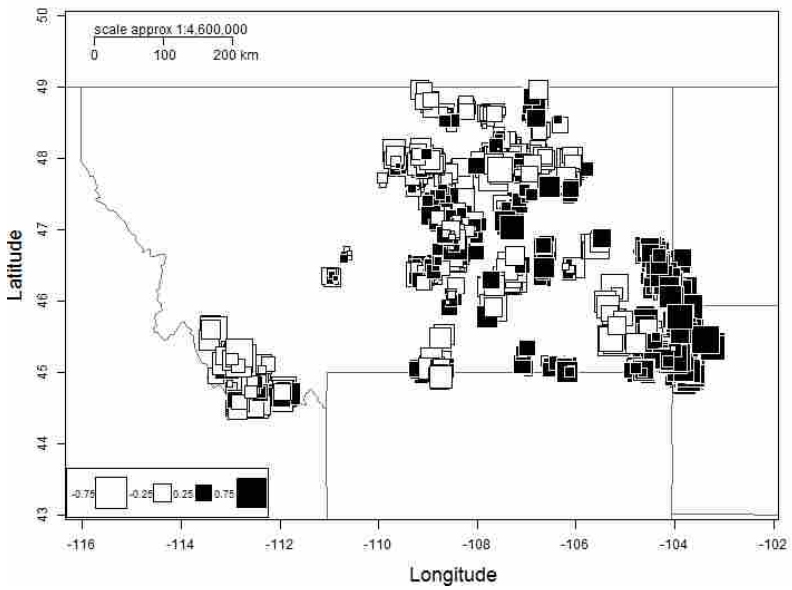
**Table 2.8.** Percent vegetative land cover for 14 classes of land cover within each of the five subpopulations identified. Subpopulations are abbreviated as northern (N), southeastern (SE), southwestern (SW), southeastern-east (SE-E), southeastern-west (SE-W), southwestern-north (SW-N), southwestern-south (SW-S)

| Vegetative Land cover | N | SE-E | SE-W | SW-N | SW-S |
|---|---|---|---|---|---|
| Agriculture | 0.02 | 0.00 | 0.00 | 0.02 | 0.00 |
| Aspen | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 |
| Badlands | 0.07 | 0.14 | 0.12 | 0.00 | 0.00 |
| Basin big sagebrush | 0.00 | 0.00 | 0.00 | 0.24 | 0.03 |
| Grasses | 0.53 | 0.61 | 0.31 | 0.03 | 0.04 |
| Forest | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| Juniper | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 |
| Mesic shrubs | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Mixed xeric shrubland | 0.18 | 0.05 | 0.39 | 0.02 | 0.00 |
| Mountain big sagebrush | 0.00 | 0.00 | 0.02 | 0.63 | 0.77 |
| Riparian | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 |
| Salt desert scrub | 0.00 | 0.03 | 0.01 | 0.00 | 0.00 |
| Wyoming big sagebrush | 0.19 | 0.10 | 0.15 | 0.05 | 0.15 |
| Unknown | 0.02 | 0.01 | 0.00 | 0.00 | 0.00 |

**Figure 2.5.** Spatial PCA of all sampled individuals' PC1 scores (a) and PC2 scores (b), which capture 74.2% and 61.2% of the genetic variation in the data and each of which are spatially autocorrelated (0.22 and 0.20 as measured by Moran's *I*), respectively. Each square represents an individual genotyped. The shading and size of the squares indicate the principal component score for each sample. The shading of the points indicates principal component scores increasing from negative (white) to positive (black). The size of the square increases with greater magnitude PC scores




**(a)**




**(b)**

**Figure 2.6.** Plot of eigenvalues decomposed into their two components: variance and spatial autocorrelation (measured by Moran's *I*). The vertical dashed line indicates the maximum attainable variance by a linear combination of alleles. The range of variation of Moran's *I* is indicated by the horizontal dashed lines. The first two global structures (associated with eigenvalues 1 and 2) ae the largest in terms of variance and of spatial autocorrelation

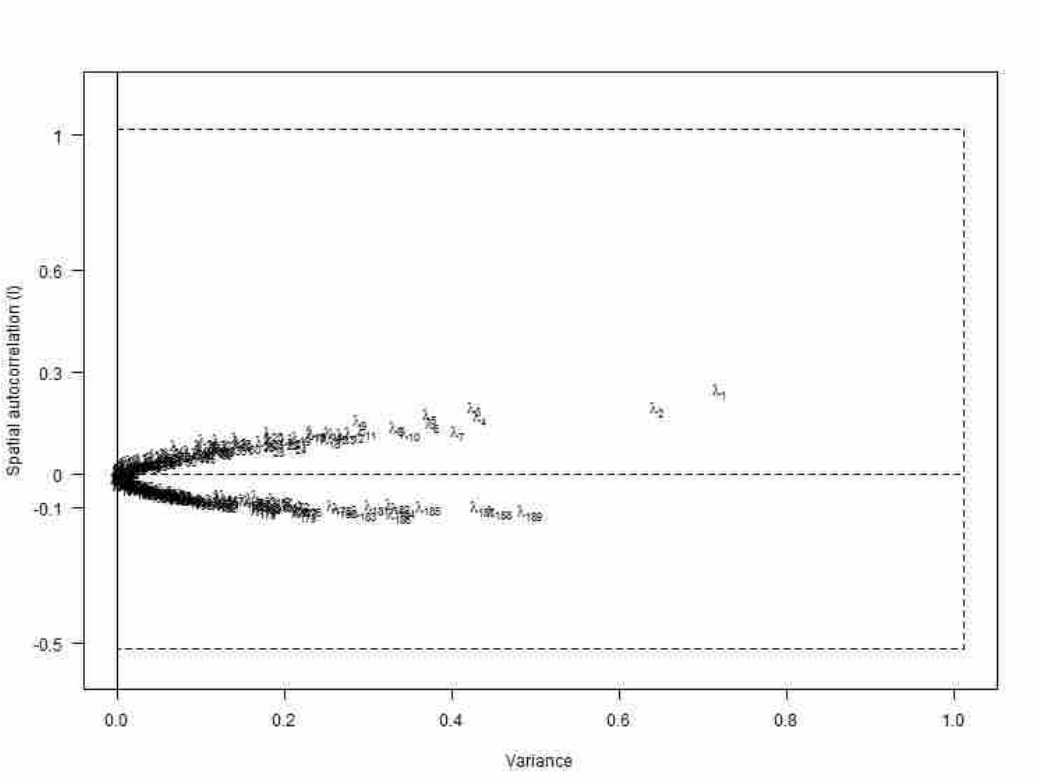
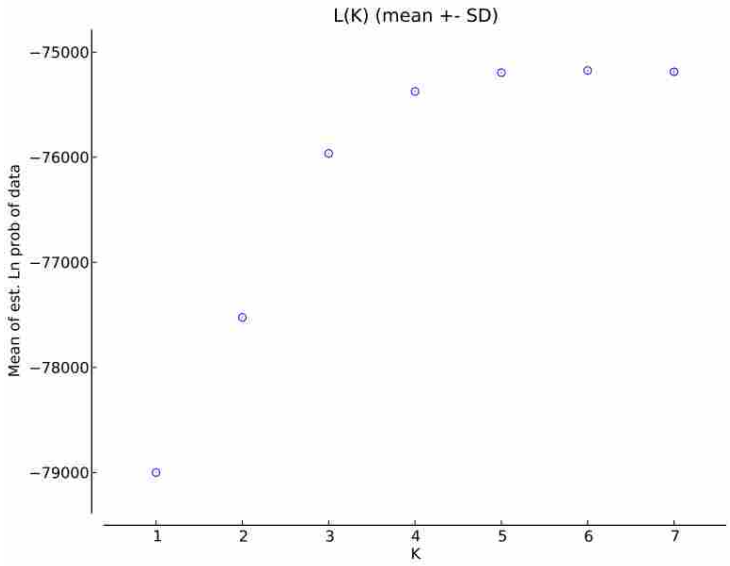
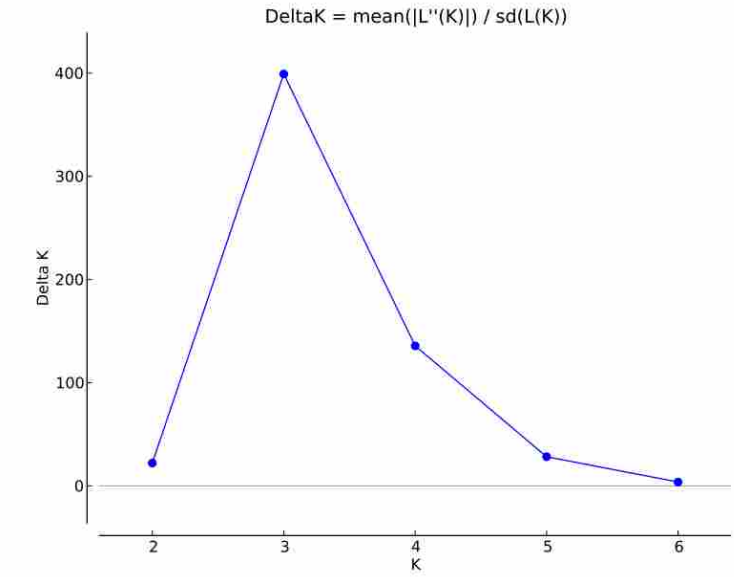


**Figure 2.7.** L(*K*) (**a**) and Δ*K* (**b**) plot for primary hierarchical substructure analysis plotted using STRUCTURE HARVESTER (Earl and vonHoldt 2012). For each value of *K*, 20 independent STRUCTURE runs were performed. In plot (**a**), the mean estimated Ln probability of each value of *K* is plotted with the standard deviation (SD) of the 20 runs shown with the error bars. Plot (**b**) shows the resulting graph from using the Evanno method (Evanno *et al.* 2005) to calculate the change in the likelihood between values of *K* between successive values of *K*. Using both the Ln P(*K*) and Evanno (Δ*K*) methods, *K* = 3 was selected as the most likely number of subpopulations




**(a)**




**(b)**

**Figure 2.8.** Secondary hierarchical STRUCTURE analysis likelihood plots (**a**–**c**) and Δ*K* plots (**d**–**f**) of *K* clusters within each of the primary *K* = 3 groups: northern (**a**, **d**), southeastern (**b**, **e**), and southwestern (**c**, **f**). For each value of *K*, 20 independent STRUCTURE runs were performed. The mean estimated Ln probability of each value of *K* is plotted with the standard deviation (SD) of the 20 runs shown with the error bars. The most likely subpopulation structure was determined to be *K* = 1 within the northern subpopulation, *K* = 2 within the southeastern subpopulation, and *K* = 2 within the southwestern subpopulation. Evanno method (Evanno *et al.* 2005) plots were created using STRUCTURE HARVESTER (Earl and vonHoldt 2012) to calculate the change in the likelihood of each value of *K* between successive values of *K*

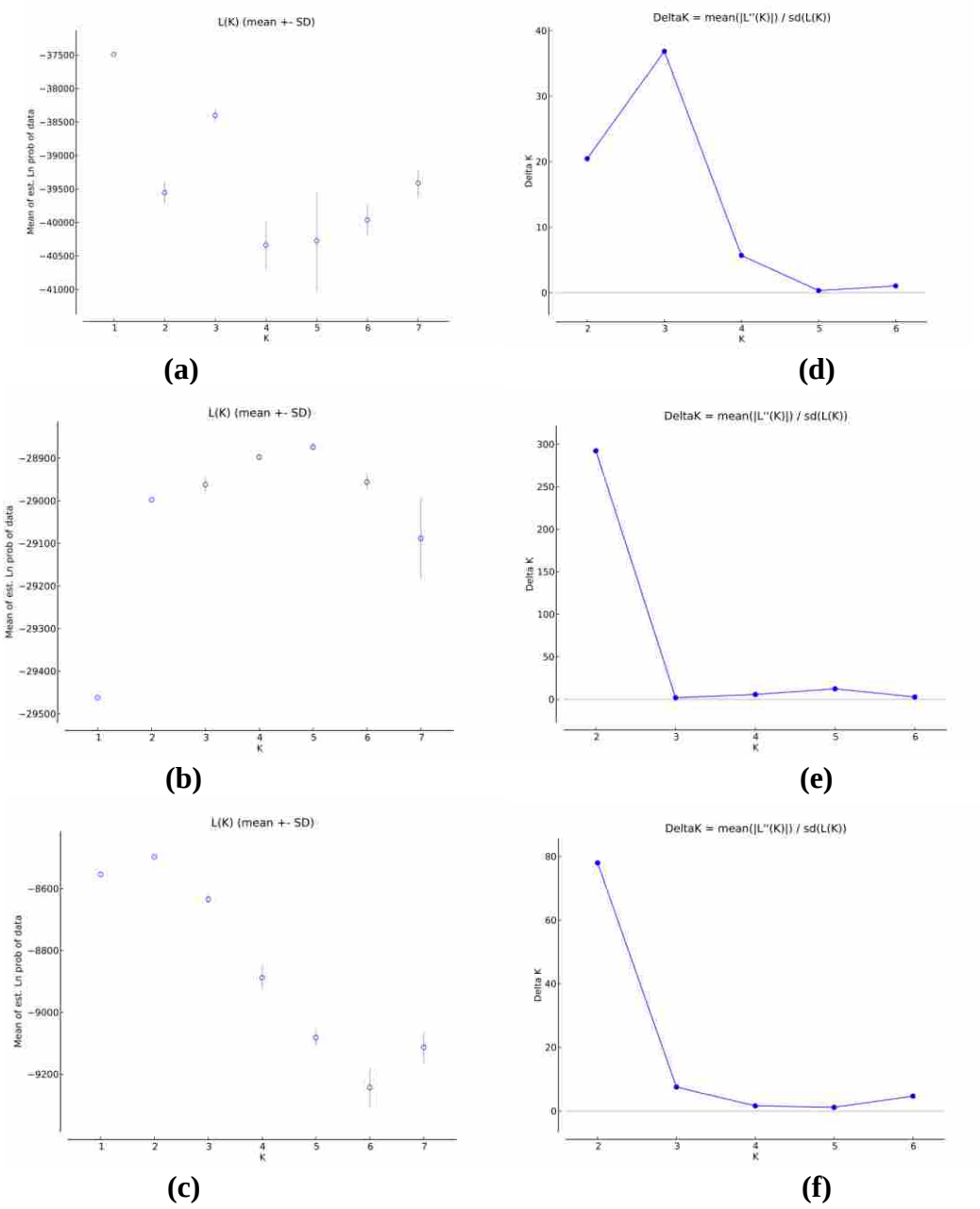
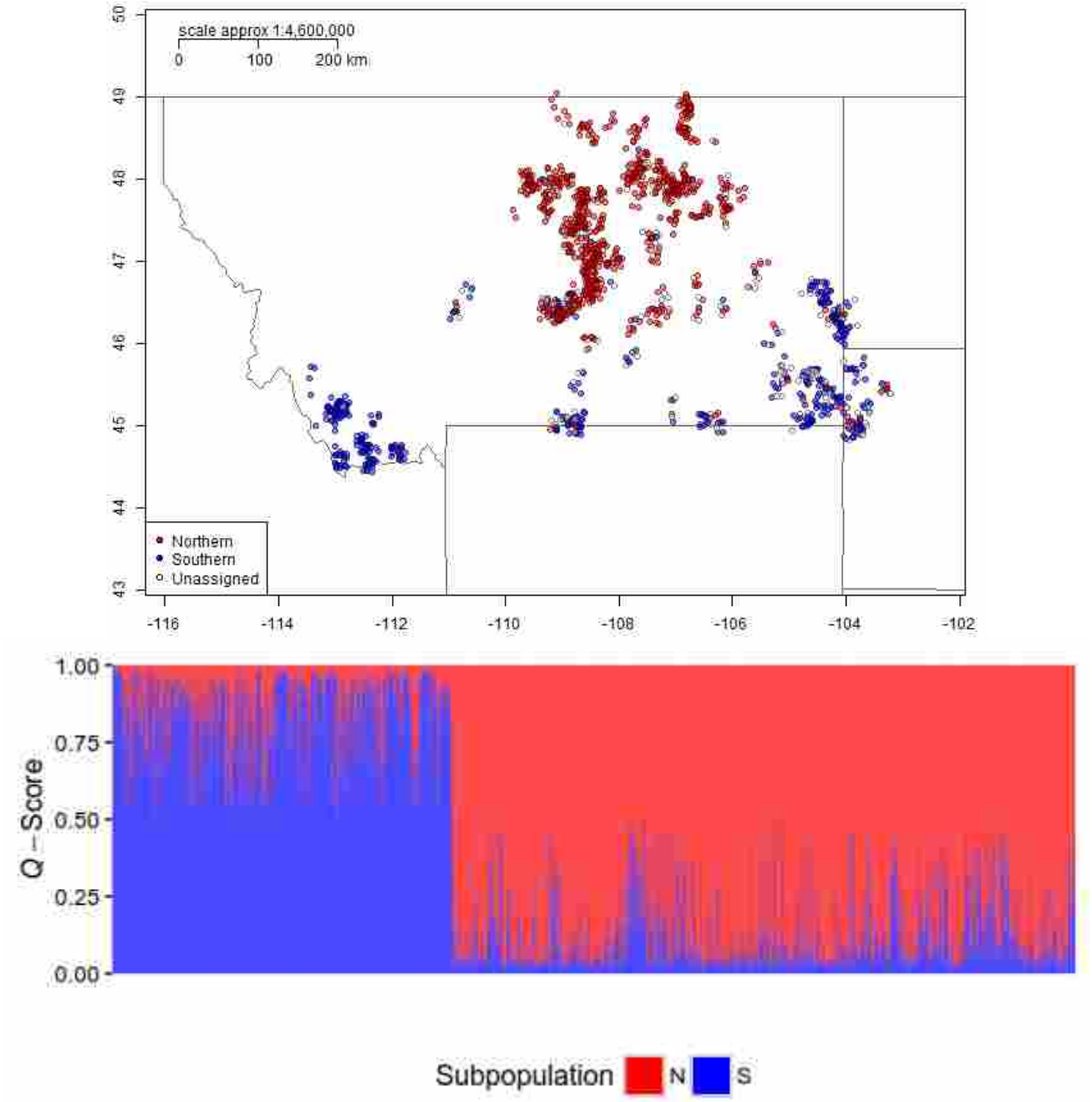
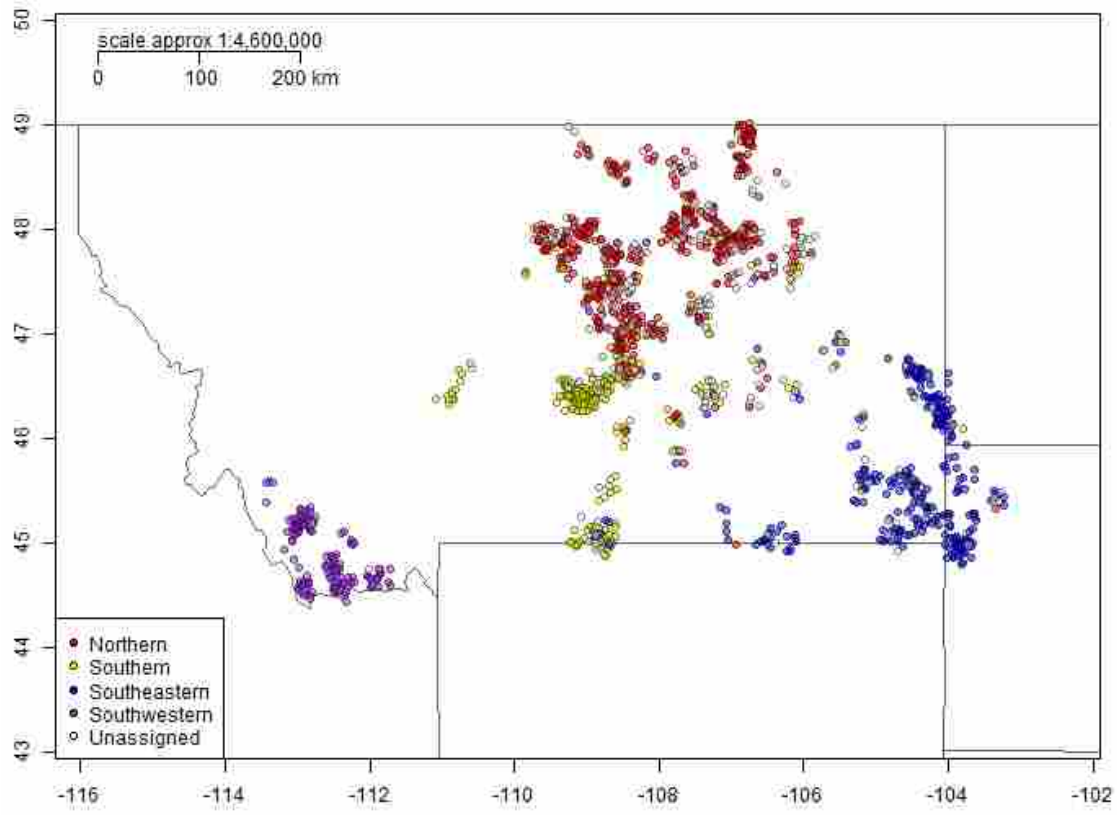
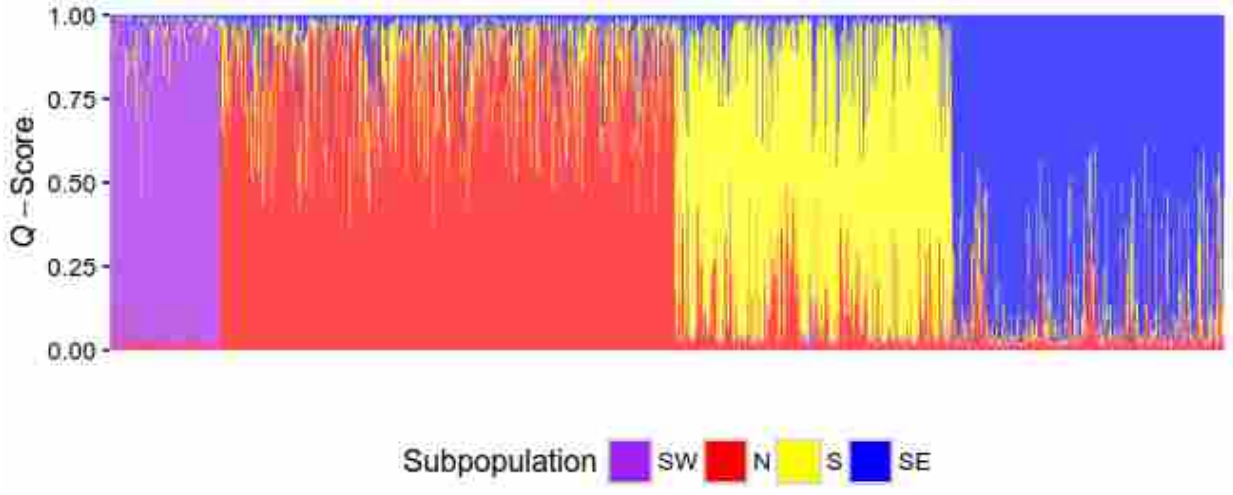


**Figure 2.9.** Genetic population substructure maps and genetic admixture plots of *Q*-values for the primary *K* = 2 (**a**), *K* = 4 (**b**), K = 5 (**c**) and for the secondary *K* = 2 within the southwestern [SW-N (southwestern-north): *purple*, SW-S (southwestern-south): *orange*] (**d**) and southeastern [SE-E (southeastern-east): *blue*, SE-W (southeastern-west): *yellow*] (**e**) hierarchical substructure for greater sage-grouse sampled in Montana, North Dakota and South Dakota as determined using STRUCTURE. Individuals with <70 % membership (measured by *Q*-value) to any subpopulation are unassigned (*open circles*). Also shown are state lines (*dashed grey lines*). Individual genetic admixture plots (of *Q*-values) are shown where subpopulation abbreviations are as listed above, and colors correspond to map

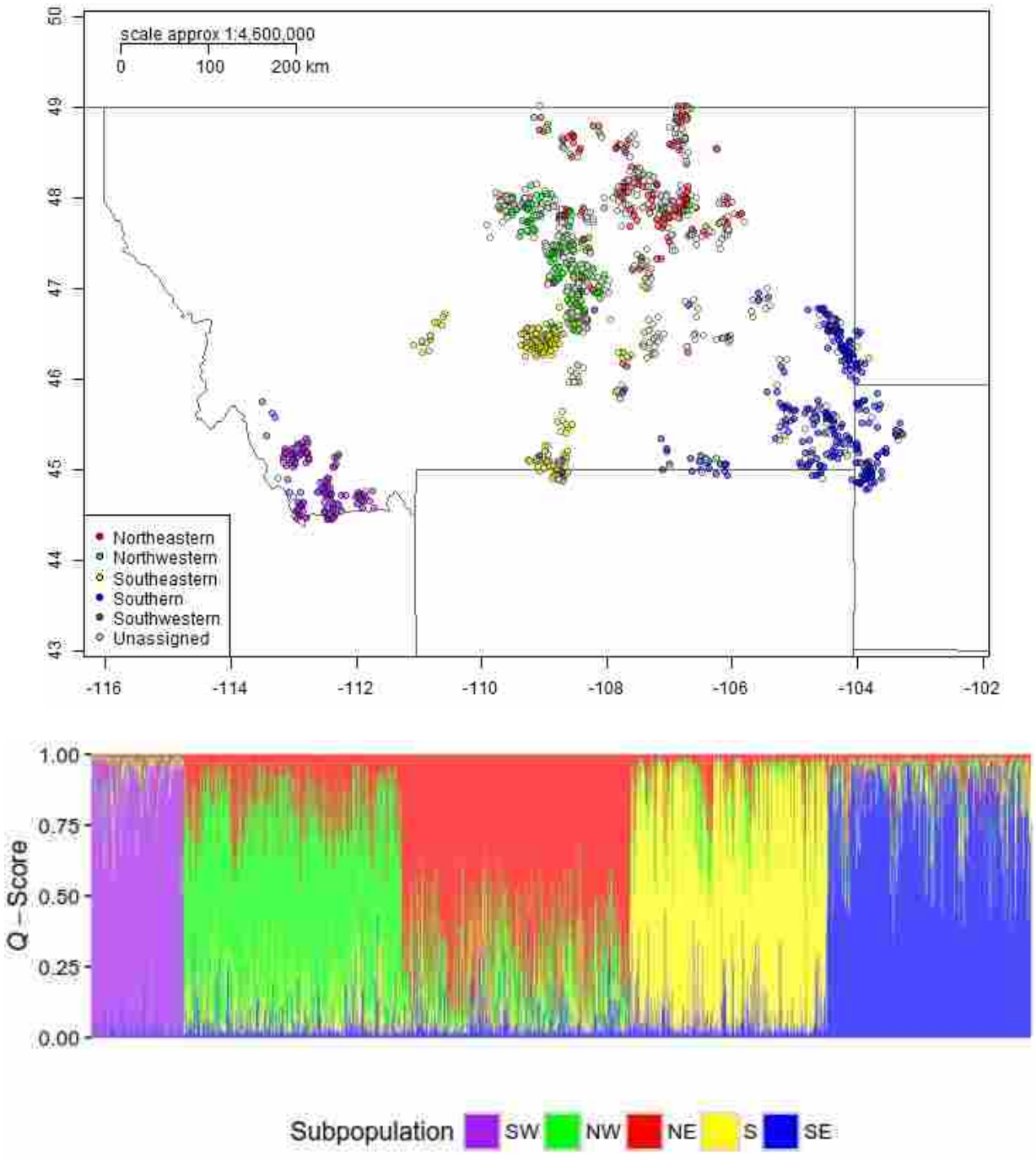



**(a)**

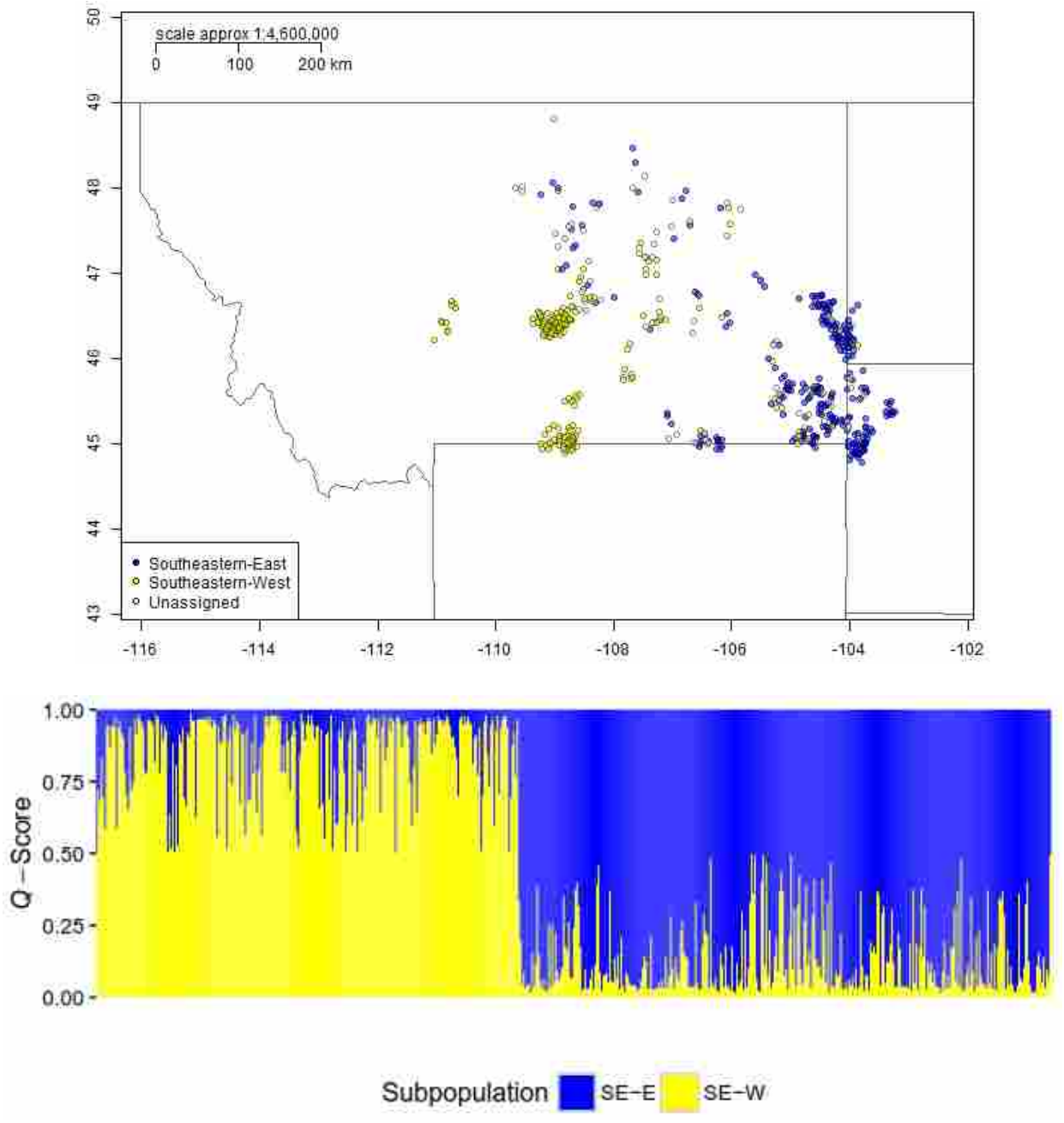

scale approx 1:4,600,000
0
100
200 km
Northern
Southern
Southeastern
Southwestern
Unassigned
50
49
48
47
46
45
44
43
-116
-114
-112
-110
-108
-106
-104
-102


Q – Score
1.00
0.75
0.50
0.25
0.00
Subpopulation
SW
N
S
SE

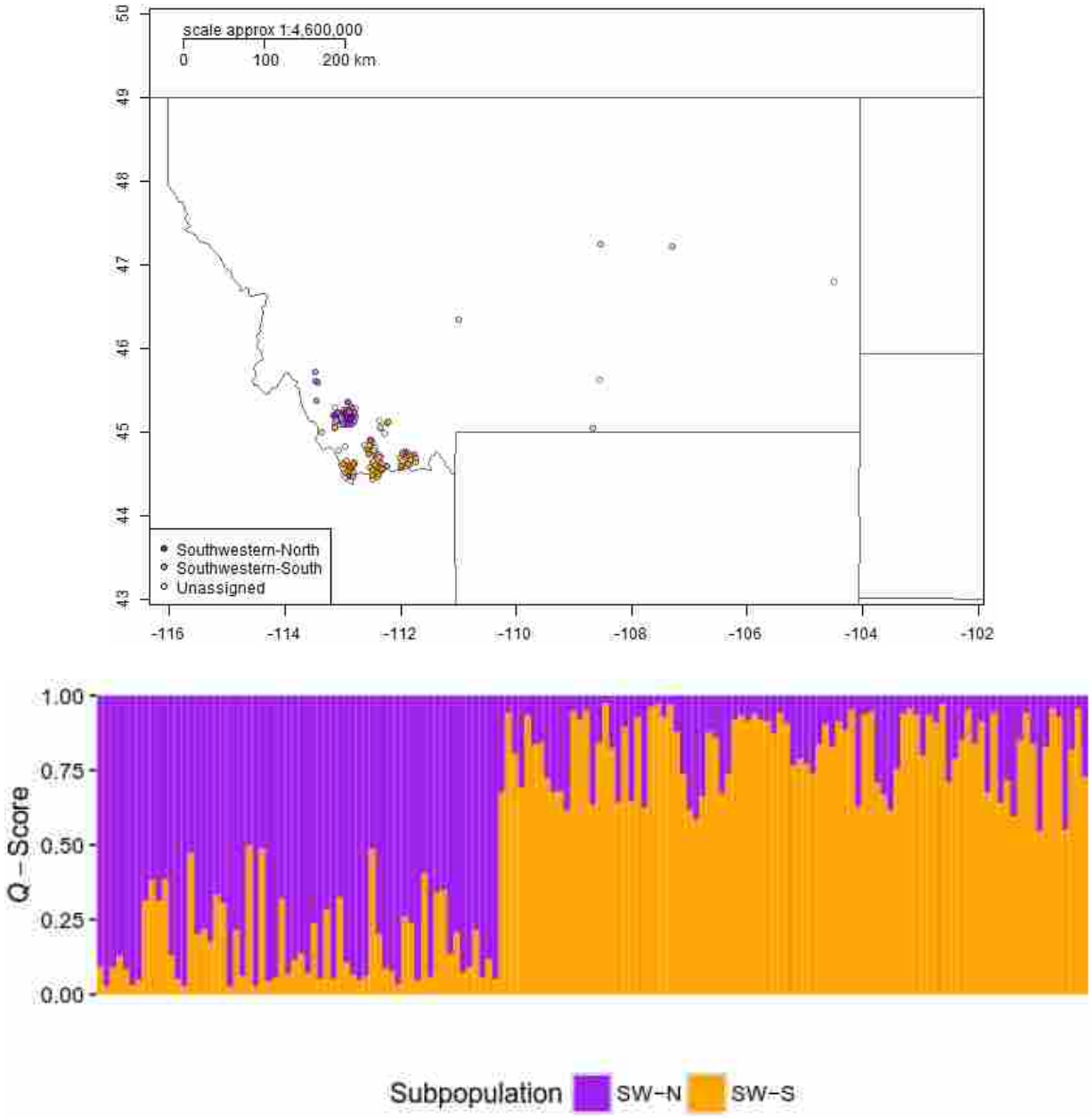
**(b)**


scale approx 1:4,600,000
0
100
200 km
Northeastern
Northwestern
Southeastern
Southern
Southwestern
Unassigned
50
49
48
47
46
45
44
43
-116
-114
-112
-110
-108
-106
-104
-102
Q−Score
1.00
0.75
0.50
0.25
0.00
Subpopulation
SW
NW
NE
S
SE

**(c)**


scale approx 1:4,600,000
0
100
200 km
Southeastern-East
Southeastern-West
Unassigned
50
49
48
47
46
45
44
43
-116
-114
-112
-110
-108
-106
-104
-102
Q−Score
1.00
0.75
0.50
0.25
0.00
Subpopulation
SE−E
SE−W

**(d)**


scale approx 1:4,600,000
0
100
200 km
Southwestern-North
Southwestern-South
Unassigned
50
49
48
47
46
45
44
43
-116
-114
-112
-110
-108
-106
-104
-102
1.00
0.75
0.50
0.25
0.00
Q – Score
Subpopulation
SW–N
SW–S

**(e)**

**Figure 2.10.** Mantel correlogram of spatial autocorrelation among greater sage-grouse microsatellite genotypes across the study area. Each point indicates a 9.9 km distance bin. The dashed line marks the point at which no positive or negative correlation is present in the data. Black squares indicate distance classes at which the spatial autocorrelation, as measured by the Mantel test for significant correlation ($r$), is statistically significant ($p < 0.05$, Bonferroni corrected, 999 permutations); white squares indicate non-significance. Size of square indicates the number of pairwise distances compared for each distance class

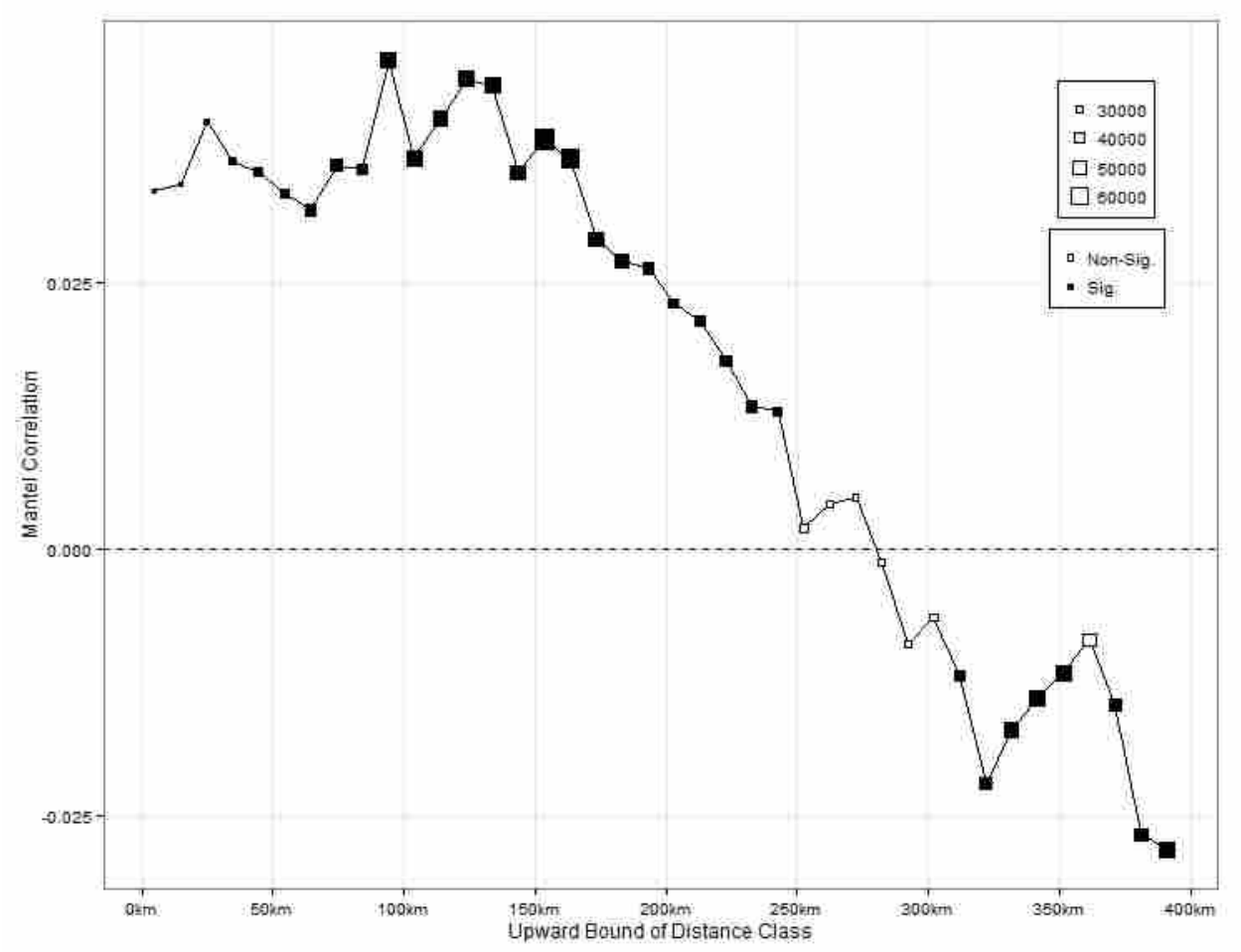


**Figure 2.11.** Genetic structure for greater sage-grouse sampled in Montana, North Dakota, and South Dakota as determined using TESS, which incorporates both the microsatellite genotype and spatial coordinates of each sample. Points show majority membership of individuals to each of the $K = 3$ clusters (determined to be the most likely value of $K$) according to majority TESS admixture value

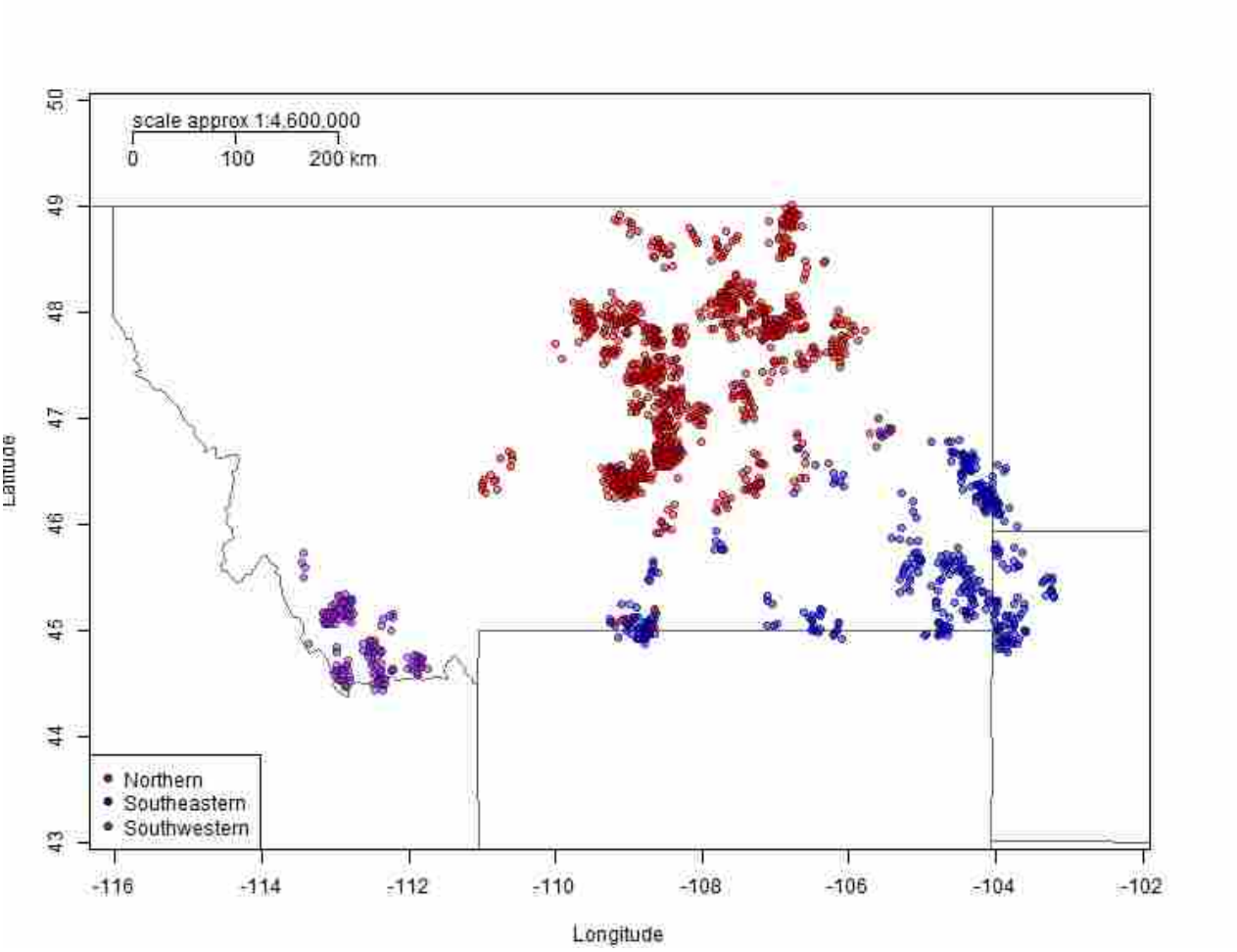


**Figure 2.12.** Percent location of subpopulations within the currently recognized USFWS management zones

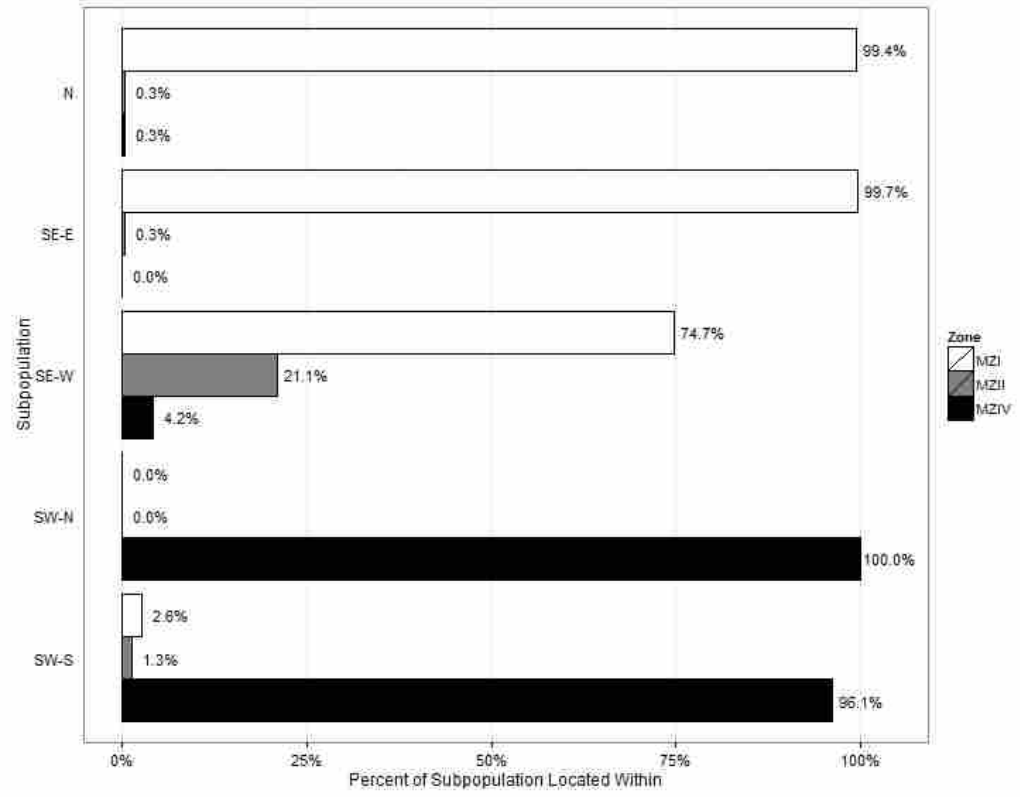


**Figure 2.13.** Percent vegetative land cover within each of the five subpopulations

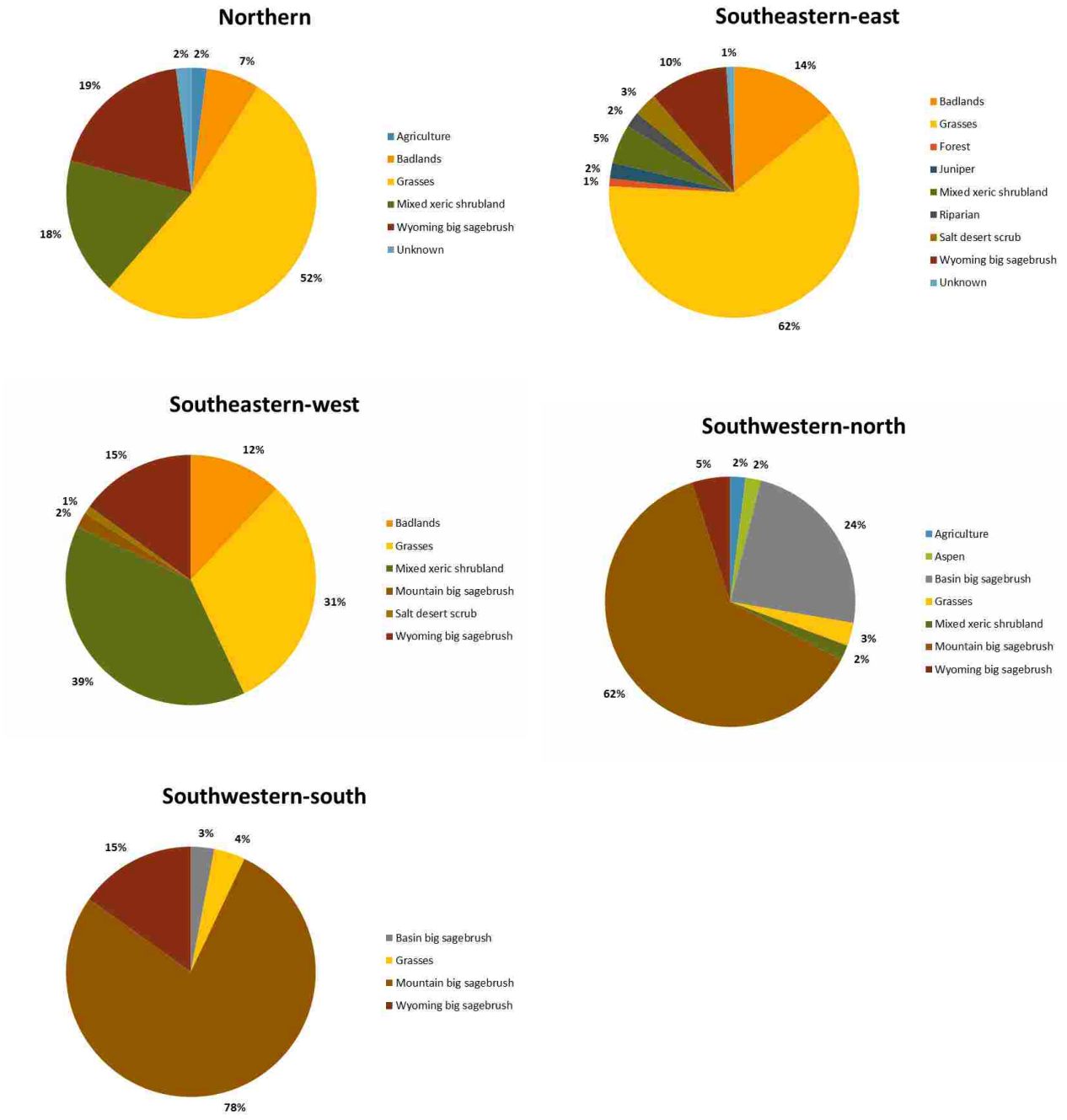
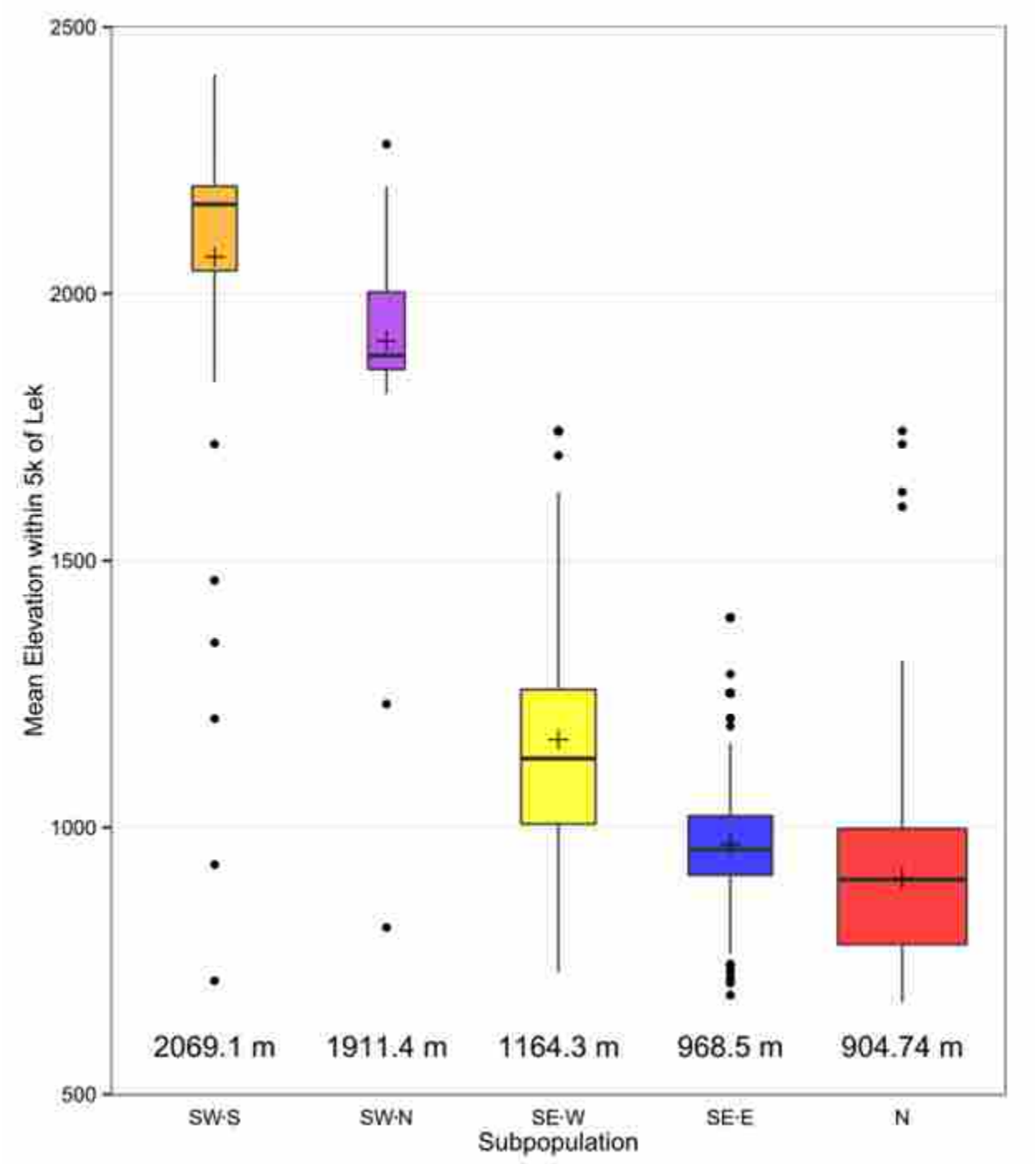


**Figure 2.14.** Elevation ranges occupied by leks within each of the five subpopulations. Mean elevation in meters is displayed above the x-axis for each subpopulation and is depicted with a + on the plot. Box width is proportional to the square root of the number of samples within each group



# CHAPTER 3

## IS HIERARCHICAL GENETIC SUBSTRUCTURE ANALYSIS VALID?


### ABSTRACT

Bayesian clustering has been used extensively by researchers seeking to identify genetic population substructure and is often used to delineate conservation units. It has become commonplace to use Bayesian clustering in a hierarchical fashion. That is, after identifying primary genetic subpopulations in a first round of analysis, each subpopulation is independently analyzed to test for secondary substructure. Whether hierarchical analysis accurately reveals population substructure is unclear; it is unknown whether secondary substructure identified within each primary subpopulation is a product of the population's evolutionary history or whether these signals may arise as an artifact of analysis. Here we simulated multiple scenarios of population divergence and used the known evolutionary histories to test whether hierarchical Bayesian clustering is valid for identifying genetic substructure within populations that would otherwise go undetected during the analysis of the primary substructure alone. Under our simulated conditions, we show that the hierarchical method to analysis does not reveal any additional substructure that cannot be detected using the mean LnP($K$) statistic. We found that while the $\Delta K$ statistic will sometimes reveal the more deeply rooted substructure, the $\Delta K$ statistic can also be misleading, indicating spurious groupings. We also discovered that at lesser divergence among subpopulations or with greater substructure complexity, misassignment of individuals to clusters is common. We conclude that the hierarchical approach to population substructure is misleading, and we provide some guidelines for how to interpret STRUCTURE results from analysis of genetic data from natural populations.


### INTRODUCTION

Determination of population genetic substructure is essential to understanding the natural processes that drive genetic divergence and is increasingly relied upon to inform wildlife and natural resource management decisions. Neutral, multi-locus genetic data is commonly used to quantify the substructure present within a population and to delineate management units. One of the most common approaches to evaluate genetic substructure (implemented in program STRUCTURE) uses algorithms to cluster multi-locus genotypic data such that gametic disequilibrium and deviation from Hardy-Weinberg proportions is minimized within clusters (Pritchard *et al.* 2000; Hubisz *et al.* 2009). While this approach can be used to test *a priori* hypotheses about population substructure, it is most commonly used as an exploratory tool to identify population substructure lacking clear *a priori* expectations. Lacking *a priori* expectations, STRUCTURE is used to group a genotypic sample into 1, 2,...,$n$ groups ($K$); where the $K$ with the greatest likelihood is assumed to indicate the most appropriate grouping (Pritchard *et al.* 2000). The most common practice is to use STRUCTURE to analyze a dataset multiple times for each value of $K$, then to calculate and identify the maximum mean log-likelihood of the data given $K$ [equation 12 in Pritchard *et al.* 2000; henceforth: mean LnP($K$)]. This method also allows the user to visualize the "uncertainty" of an estimate of $K$ in the form of the standard deviation around the mean log-likelihood.

However, using mean log-likelihood scores to pick an appropriate $K$ has proven difficult, leading to a variety of alternate *ad-hoc* approaches. Perhaps the most common *ad-hoc* approach is to use the rate of change in log-likelihood between successive $K$ values (henceforth: $\Delta K$),

where *K* is selected as the value with the maximum Δ*K* value (Evanno *et al.* 2005). Evanno *et al.* (2005) found that using the Δ*K* statistic allowed for precise identification of the uppermost or primary substructure in a hierarchical island model (Wright 1931). Once *K* is identified, individuals are assigned to a subpopulation based on the percentage of their genotype attributed to each of the identified subpopulations (*i.e.*, percent admixture, known as *Q*-value).

Following the derivation of the Δ*K* statistic and discovery that it can reveal the primary substructure, it became common practice in peer-reviewed literature to perform hierarchical substructure analysis. This approach entails using STRUCTURE to probe datasets for additional, or secondary, population substructure within each of the primary subpopulations (Coulon *et al.* 2008; Balkenhol *et al.* 2014; Vähä *et al.* 2008; Lukoschek *et al.* 2008; Cheng *et al.* 2014; Warnock *et al.* 2010; Cross *et al.* 2016; Viricel & Rosel 2014). The hierarchical approach, first implemented by Coulon *et al.* (2008), entails first analyzing the complete sample of genotypes to determine the most likely number of primary subpopulations [using Δ*K* and mean LnP(*K*)]. After the samples are subsequently partitioned into subpopulations, each subpopulation is independently analyzed using STRUCTURE, *K* is inferred, the sample divided, and the process repeated until the inferred number of subpopulations detected within each partition is only one as determined by mean LnP(*K*). This method seems to reveal subpopulation substructure that was not evident when initially analyzing the complete sample.

Despite extensive use in published literature, hierarchical substructure analysis has only seen limited cross-examination to determine how best to interpret the additional substructure identified when using the method (see Balkenhol *et al.* 2014). Questions remain, such as whether additional substructure discovered is the result of underlying evolutionary processes undetected in the primary analysis, or the result of inaccurate clustering of individuals to the primary subpopulations.

To test whether hierarchical substructure analysis is required to detect population substructure resulting from evolutionary processes that would be otherwise undetected, we simulated several different population subdivision and divergence scenarios using a model of Wright-Fisher ideal population. We used these evolutionary models to address one primary objectives and two secondary objectives.

In this study, our overarching goal was to answer the question: does the use of hierarchical substructure analysis detect population substructure that would otherwise go undetected when using a single analysis? Nested within this main objective were two objectives. Our first objective was to answer the question: is the sensitivity of the Δ*K* and mean LnP(*K*) statistics to correctly identifying population substructure influenced by the complexity of the evolutionary scenario (i.e., the number of subpopulations involved in analysis and whether the scenario involve a single or multiple branching events)? We define sensitivity as the magnitude of divergence among subpopulations ($F_{ST}$; Weir & Cockerham 1984) required to identify "true *K*." We define true *K* as the number of known subpopulations in the simulation at each generation. Our second objective was to answer the question: is the accuracy with which individuals are assigned to their respective subpopulations influenced by the magnitude of divergence and the complexity of the evolutionary scenario? We hypothesized that accuracy of individual clustering would increase with an increase in divergence, and decrease with an increase in the complexity of the evolutionary scenario.

## SIMULATIONS

To address our two questions, we constructed a simple simulation model composed of sexually reproducing, non-selfing individuals with non-overlapping generations, random mating, and single-step mutation. Within the resulting Wright-Fisher idealized populations, the rate of genetic change (and therefore the degree to which subpopulations diverged) was affected solely by population size (genetic drift), and the rate of mutation. We simulated five different evolutionary scenarios (Fig 3.1). Each simulation began with the same single founder population of 1000 diploid individuals, which was created in EASYPOP (Balloux 2001). These individuals were genotyped across 10 codominant neutral microsatellite loci, with free recombination between loci (i.e., unlinked loci; $r = 0.5$), and a mutation rate of $10^{-3}$ per generation (Ellegren 2000). We used a single-step mutation model, with a maximum of 10 alleles per locus, and allowed our founder population to burn-in for 1000 generations in order to reach mutation-drift equilibrium. For each simulation, we performed three independent replicates. We wrote our simulation models for the program R version 3.2.2 (R Core Team 2016) and scripts are available upon request.

*Non-branching scenario simulations*

We simulated three hierarchical evolutionary scenarios without a branching event to model two, three and four diverging subpopulations (A, B, and C in Fig 3.1). These three simulations started with the same founder population of 1000 individuals which was allowed to randomly mate for one generation, producing a population of 2000 (scenario A), 3000 (scenario B), and 4000 (scenario C) progeny. We randomly divided the first-generation populations into subpopulations of 1000 (scenario A= 2 subpopulations, scenario B = 3 subpopulations, scenario C = 4 subpopulations). In generations following the primary split, we allowed populations to drift without gene flow among subpopulations. We calculated $F_{ST}$ at each generation and allowed each replicate of the model to drift until $F_{ST}$ reached 0.10 for all pairwise subpopulation comparisons. We selected this terminal $F_{ST}$ as it well exceeded the amount of divergence at which STRUCTURE could correctly assign all individuals to their population of origin (discovered in preliminary analyses). We conducted all calculations of $F_{ST}$ using the FSTAT function in R package GENELAND (Guillot & Santos 2009).

*Branching scenario simulations*

We simulated two different branching evolutionary scenarios to model populations that have undergone more than a single subdivision in the past (D and E in Fig 3.1). These evolutionary simulations began with the same founder population of 1000 individuals as the non-branching simulations. This founder population was allowed to randomly mate for one generation, generating a population of 2000 progeny. The first-generation populations were each randomly divided into subpopulations of 1000 (two primary subpopulations). Following this splitting event, we simulated drift as we did for the aforementioned scenarios until $F_{ST}$ among primary subpopulations reached 0.025. We chose this divergence threshold after preliminary non-branching analyses showed accurate detection of true $K$ and accurate clustering of individuals to the true subpopulation at this magnitude of divergence. In the generation following that in which the divergence threshold was met, we simulated a branching event by allowing one 1000 member primary subpopulation to generate 2000 progeny (via random mating with mutation) which we then randomly divided into two independent 1000 individual secondary subpopulations. Meanwhile, the remaining non-branching subpopulations were allowed to drift for that generation. This process resulted in three secondary subpopulations for scenario D, and four secondary subpopulations for scenario E. In subsequent generations, each subpopulation of

1000 individuals was allowed to drift without gene flow among subpopulations until $F_{ST}$ reached 0.10 for all pairwise subpopulation comparisons (Fig 3.1).

## METHODS

### Analysis of population substructure

We sampled 100 individuals from each population, every 10 generations from 1 to 151 generations from each of the 3 replicates within each of the 5 evolutionary scenario simulations. For each sample, we used the Bayesian clustering program STRUCTURE (Pritchard *et al.* 2000) to test for genetic population substructure. We tested for *K* clusters from one to seven. We ran three replicates of STRUCTURE for each value of *K*. We used internal program settings that followed those of Falush *et al.* (2003), and which are recommended for detecting subtle population substructure. Specifically, analyses were run using the default settings under the admixture model with correlated allele frequencies among populations, the allele frequency distribution parameter (λ) set to 1, where STRUCTURE is allowed to infer the value of the model's admixture parameter (α) from the data. We set the length of burn-in period before the start of data collection to 10,000, the number of MCMC repetitions after burn-in to 10,000, and used no prior (no use user-defined population-of-origin, nor user-defined sampling location for each individual).

*Error rate in identification of true K*

We wished to quantify the proportion of times structure identified the true *K*, defined as the number of subpopulations modeled in the simulation at a given generation. For each sample, we identified *K* using the maximum value of the mean LnP(*K*) statistic (Pritchard *et al.* 2000) and the maximum value of the Δ*K* statistic (Evanno *et al.* 2005). To calculate these statistics across three STRUCTURE replicates, we used STRUCTURE HARVESTER (Earl & vonHoldt 2012) and compared the results to the divergence measured among samples.

*Accuracy in individual clustering*

We evaluated if the magnitude of divergence and the complexity of the evolutionary scenario influenced the assignment of individuals to populations. We first identified the most supported value of K by finding the maximum mean LnP(*K*) from *K*=1-7 in STRUCTURE. We subsequently used program CLUMPP v 1.1.2 (Jakobsson and Rosenberg 2007) to average *Q*-values for each individual across the three STRUCTURE replicates for each sample. In CLUMPP, we used 30,000 repeats of the "greedy method" with greedy option two and the pairwise matrix similarity statistic, G. We calculated the cluster to which the maximum proportion of each individual's genome was assigned based on maximum *Q*-value.

We calculated the proportion of individuals correctly assigned for each of the three scenario replicates. Within each of the non-branching scenarios (A–C), true *K* did not change through the generations. For the branching scenarios (D and E), true *K* changed when primary subpopulations underwent the branching event after reaching the divergence threshold of $F_{ST}$ = 0.025 (i.e., splitting to form secondary subpopulations).

*Hierarchical analysis*

For all samples where maximum mean LnP(*K*) indicated a *K* of greater than one, we conducted hierarchical substructure analysis. We divided individuals into a subset of *K* respective groups, assigning each individual to a group based on its maximum *Q*-value. For each subset, we used STRUCTURE to analyze for additional substructure, used STRUCTURE HARVESTER to summarize across STRUCTURE replicates, and used CLUMPP to average *Q*-

values all as detailed above. We continued this process until maximum mean LnP(*K*) indicated a *K* of one.

## RESULTS

### Sensitivity to identification of true *K*

We found a relationship between the magnitude of divergence among subpopulations and the sensitivity of STRUCTURE to correctly identifying true *K* as measured by maximum mean LnP(*K*) (Fig. 3.1). In all three non-branching scenarios (A, B and C), maximum mean LnP(*K*) first correctly detected subpopulation structure when $F_{ST}$ measured from 0.023 to 0.026. In non-branching scenarios, maximum $\Delta K$ often indicated spurious values of *K* when divergence among subpopulations was low (e.g., FST < 0.03). In non-branching scenarios, the maximum $\Delta K$ statistic ranged from 2–6, including all values in between, but never equaling one (as this is impossible). We did not find a relationship between scenario complexity and identification of true *K*.

In all branching scenarios, maximum mean LnP(*K*) revealed both primary and secondary branching structure once $F_{ST}$ within the branches crossed a comparable threshold to that observed in the non-branching scenario A. Across all scenario D replicates, maximum mean LnP(*K*) first indicated three subpopulations at a mean $F_{ST}$ of 0.014 among subpopulations. Within scenario D, we measured $F_{ST}$ within the branch from the point of the secondary split in the simulation (indicated by the horizontal dashed line in Fig. 3.1). Across all scenario E replicates, maximum mean LnP(*K*) first indicated four subpopulations at a mean $F_{ST}$ of 0.017 among subpopulations. Within scenario E, we calculated mean $F_{ST}$ from both within-branch measures (see previous note). For branching scenarios (D and E), $\Delta K$ most often identified *K* corresponding to the primary subpopulation structure (Fig. 3.1D and E). In branching scenarios, when maximum mean LnP(*K*) did not indicate the true *K*, it usually indicated a value of *K* where the resulting clusters were composed of individuals sampled from two of the three (D) or two of the four total populations (E).

### Accuracy of individual clustering

Correct clustering of sampled individuals to their population of origin increased with increasing $F_{ST}$ across all scenarios (Fig. 3.2). While complexity of the scenario did not affect detection of true *K* using either statistic, it did affect the percent correct clustering of individuals with their true population. Greater divergence is required for correct clustering of individuals as complexity increases from A to B to C, and within branches for D and E. For example, mean correct clustering of individuals to their population of origin was 94.0% for scenario A, 88.1% for B, 84.3% for C, 84.2% after secondary split for D, and 86.7% after the branching event for E. The percent correct clustering of individuals was lower at any given value of within-branch (two subpopulations) $F_{ST}$ when compared to percent correct clustering of individuals in scenario A (two subpopulations).

### Hierarchical analysis

Following the first round of analysis of STRUCTURE for all repetitions of all scenarios, use of CLUMPP to average *Q*-values, division of individuals into subsets based on maximum mean LnP(*K*) for all scenarios where maximum mean LnP(*K*) was greater than one, and a second round of STRUCTURE analysis and summary in HARVESTER, we found that maximum mean LnP(*K*) always indicated a *K* of one within subsets. Therefore, there was no evidence for further (hierarchical) population substructure within subsets.

## DISCUSSION

### Hierarchical analysis

We find no evidence for the necessity of hierarchical substructure analysis. We found that the maximum mean value for the LnP(*K*) statistic indicated the true number of clusters in all branching scenarios once the magnitude of divergence among subpopulations within the branches was equivalent to the magnitude of divergence at which substructure was revealed in non-branching scenarios.

We believe that the practice of hierarchical analysis arose from the shift to using the Δ*K* statistic to determine substructure. The maximum value of the Δ*K* statistic was not intended to indicate the value of *K* that would reveal all substructure, but only the secondary substructure, and was originally tested on a model of five-island model populations each composed of five island model subpopulations (Evanno *et al.* 2005). We confirmed that maximum Δ*K* sometimes identifies the deepest-rooted population structure when true *K* is 2 or greater, and propose that if used in concert with maximum mean LnP(*K*), one can gain insight into the evolutionary history of populations. Selecting maximum Δ*K* can reveal deeply rooted substructure. Selecting maximum mean LnP(*K*) most often reveals all substructure.

We also found evidence that the use of hierarchical analysis can be misleading, both because maximum Δ*K* alone can never indicate a *K* of one (Coulon *et al.* 2008), and because this metric spuriously identifies values of *K* when true *K* is equal to one. We found that maximum mean LnP(*K*) more consistently identifies true *K* than does Δ*K*. In non-branching scenarios, maximum Δ*K* also tended to veer away from identifying true *K* more often than did maximum mean LnP(*K*). Only when maximum Δ*K* differs from, and is less than maximum mean LnP(*K*), does it reveal the more deeply rooted structure. When the maximum value of Δ*K* indicated a greater value of *K* than did the value of maximum mean LnP(*K*), the Δ*K* resultant clusters were composed of individuals from multiple different subpopulations.

Given our findings, we believe it is imperative to provide both Δ*K* and LnP(*K*) plots and figures for the both the reviewer's and reader's comparison. For example, in our review of Cross *et al.* (2016), we found that if we selected a *K* of four rather than three (as indicated by maximum Δ*K*) within the primary hierarchical substructure, we would have revealed one of their secondary splits. Divergence among the three primary subpopulations was greater than that within the secondary subpopulations, which agrees with our finding that maximum Δ*K* identifies more deeply rooted substructure and that there is value in comparing Δ*K* and LnP(*K*).

### Sensitivity to identification of true *K* and accuracy of individual clustering

We found that with two, three, and four subpopulations STRUCTURE reliably correctly identified the true number of subpopulations at an $F_{ST}$ of 0.023 and 0.026. Similarly, Latch *et al.* (2006) found that a threshold $F_{ST}$ of 0.030 was required for STRUCTURE to identify the true number of subpopulations with five non-hierarchical subpopulations.

Counter to our expectations, the number of subpopulations involved in a scenario did not seem to affect the sensitivity of LnP(*K*) nor the Δ*K* statistics to detecting true population substructure. However, scenario complexity and divergence of subpopulations did affect the accuracy of individual clustering (i.e., whether sampled individuals clustered with other individuals sampled from the same source population). STRUCTURE had greater accuracy clustering individuals with their sample of origin in scenarios of reduced complexity and at greater divergence. As expected, these results confirmed the findings of Latch *et al.* (2006) that the average number of individuals assigned to the correct subpopulation increased as $F_{ST}$ increased.

**Recommendations: inference of true *K* within natural populations**

We see several potential issues applying hierarchical analysis to natural populations. The first issue is that of poor individual assignment during the initial grouping. This can happen when one selects a poorly supported value of *K*, individual admixture is high, or when the underlying evolution scenario is complex (as we saw here). In these cases, it is possible to have mixed sub-groups from the primary round of STRUCTURE, which might then separate into spurious subpopulations during the secondary analysis. In natural populations, individuals could be incorrectly assigned to a subpopulation at the first determination of population substructure, which could result in the appearance of hierarchical substructure upon independent analysis of the primary subpopulations. However, in these situations, the uncertainty of assignment will likely be show in the standard deviation around estimates of the mean LnP(*K*) and in the lack of clear bins when viewing admixture plots of *Q*-values. Therefore, it is important to review possible values of *K* using both Δ*K* and LnP(*K*), and then to review individual admixture plots of *Q*-values and spatial plots of admixture to examine the biological plausibility of substructure. Our results indicate that if population subdivision among clusters identified using STRUCTURE measured at an $F_{ST}$ of less than 0.01, these resultant clusters may be spurious. However, with more loci, or when natural processes such as migration are involved, these "thresholds" may vary. Still we advise caution.

The second issue is the existence of latent gradients or clines within the subpopulation that are too weak to produce separate groups in the initial clustering. We did not simulate genetic gradients/clines but anticipate that clines will further complicate correct identification of true *K*. Prior research has shown that gradients or clines in natural populations may be the source of spurious hierarchical population structure (e.g., Schwartz and McKelvey 2009).

The third and final issue is that identification of hierarchical structure in other studies may have resulted as an artefact of sampling such that some organisms might group together erroneously (e.g., Tucker *et al.* 2014; for a thorough review, see Schwartz and McKelvey 2009). We sampled the same number of individuals, randomly, from each population and individuals were in a Wright-Fisher ideal population and with no spatial distribution. Therefore, we do not see an instance where sampling could have factored into our analysis. However, in natural populations, where one does not have a census of the population, sampling effects could lead to the appearance of branching substructure.

**Conclusions**

When using STRUCTURE to analyze natural populations, true *K* could be greater than the greatest value of *K* tested in STRUCTURE. This will result in missing the true primary population substructure, as the mean LnP(*K*) plot will not asymptote. Additionally, because the Δ*K* statistic is a second order analysis, it cannot indicate that *K* is equal to one even if that is the biological "truth" (Coulon *et al.* 2008), and we found that Δ*K* statistic may only recover the uppermost substructure (also suggested by Waples and Gaggiotti 2006). Finally, selecting *K* clusters based on a combination of maximum Δ*K*, maximum mean LnP(*K*), and an interpretation of the biological plausibility of the spatial pattern of individuals in the *K* clusters inherently introduces interpretational bias. This bias is based on one's perception of the drivers of population divergence, or based on a visually discernable spatial pattern at the scale and resolution of the study. For these reasons, Hubisz *et al.* (2009) have incorporated the ability to set a Bayesian prior within STRUCTURE by using sample coordinates. However, using the "LOCPRIOR" setting may bias the analysis of the true genetic signal. For example, when the

location prior is used, the $F_1$ of a long distance disperser might show suppressed admixture do to its spatial proximity to the multi-generational resident individuals with which it is grouped.

Using STRUCTURE with an awareness of the potential for the analytical or interpretational errors above, we encourage researchers to use maximum mean LnP(*K*) to determine the most likely contemporary population substructure. We urge researchers to only interpret maximum Δ*K* in light of the fact that it may indicate the most deeply rooted substructure (only when it differs from and is less than the maximum mean LnP(*K*)). In other words, we recommend comparing maximum mean LnP(*K*) to Δ*K*. Δ*K* will identify deep-rooted substructure and maximum mean LnP(*K*) can identify more recent substructure ($F_{ST}$ = 0.025). Hierarchical analysis as first proposed by Coulon et al (2008), and widely used since, can be misleading when detecting population substructure.

**ACKNOWLEDGEMENTS**

We thank Jamie Sanderlin for her assistance in proofing and improving our simulation scripts. Use of trade names does not imply endorsement by the U.S. Government. The views in this article are those of the authors and do not necessarily reflect those of their employers.

**Figure 3.1.** Scenarios of evolutionary history used to evaluate the validity of the hierarchical analysis method in STRUCTURE. Scenarios are depicted as cladograms where generations, indicated by t, increase from $t_0$ (panmictic founder population of 100 individuals) to the most divergent branches of the scenario (1000 individuals per subpopulation). All scenarios start at generation t = 1. The dotted line indicates the generation at which the hierarchical branching event occurs, where $F_{ST}$ among subpopulations in the previous generation had reached 0.025. For scenarios D and E, the primary substructure is above the dotted line, and the secondary substructure is below the dotted line. All scenarios drifted until all pairwise subpopulations reach a divergence of $F_{ST} \geq 0.10$

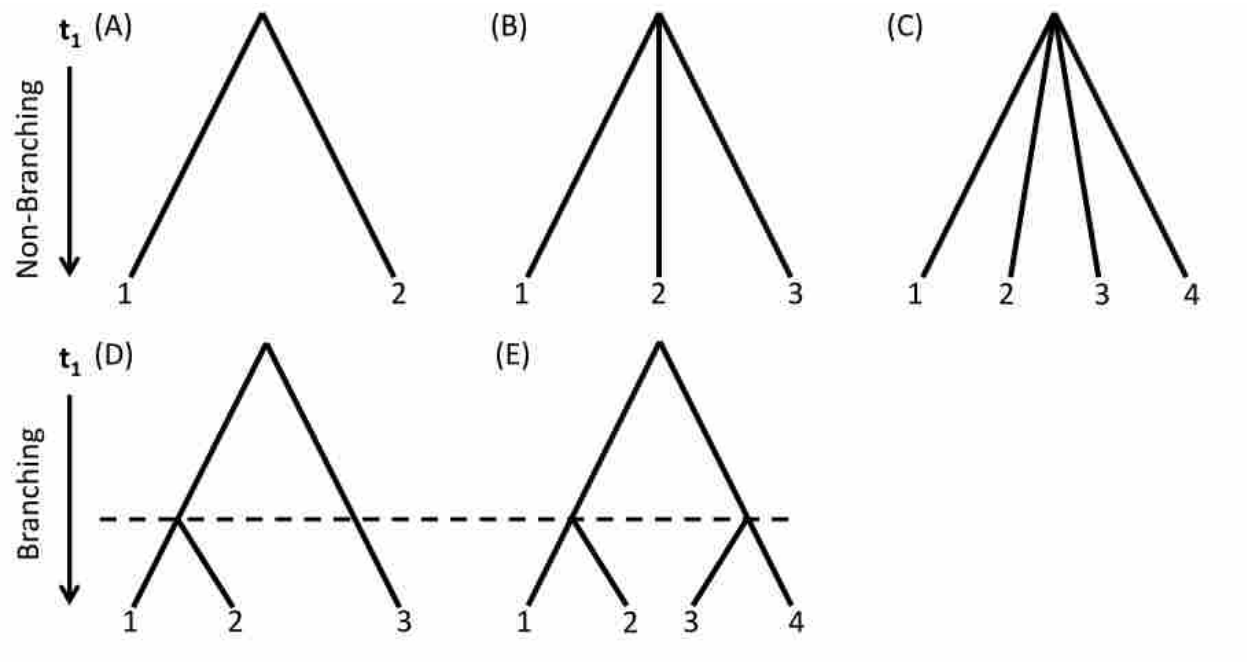
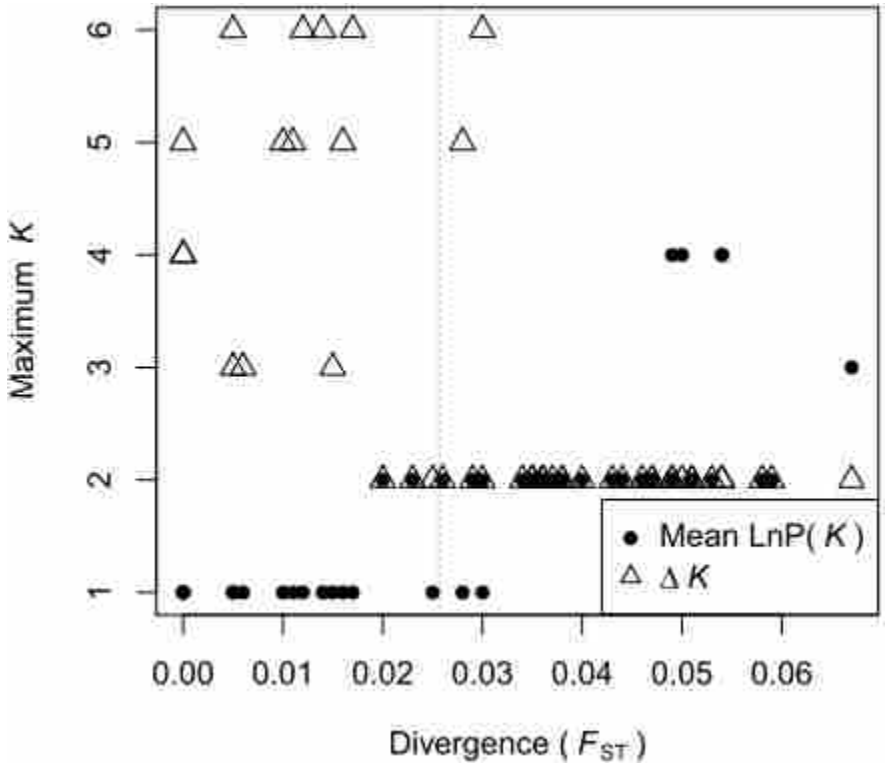
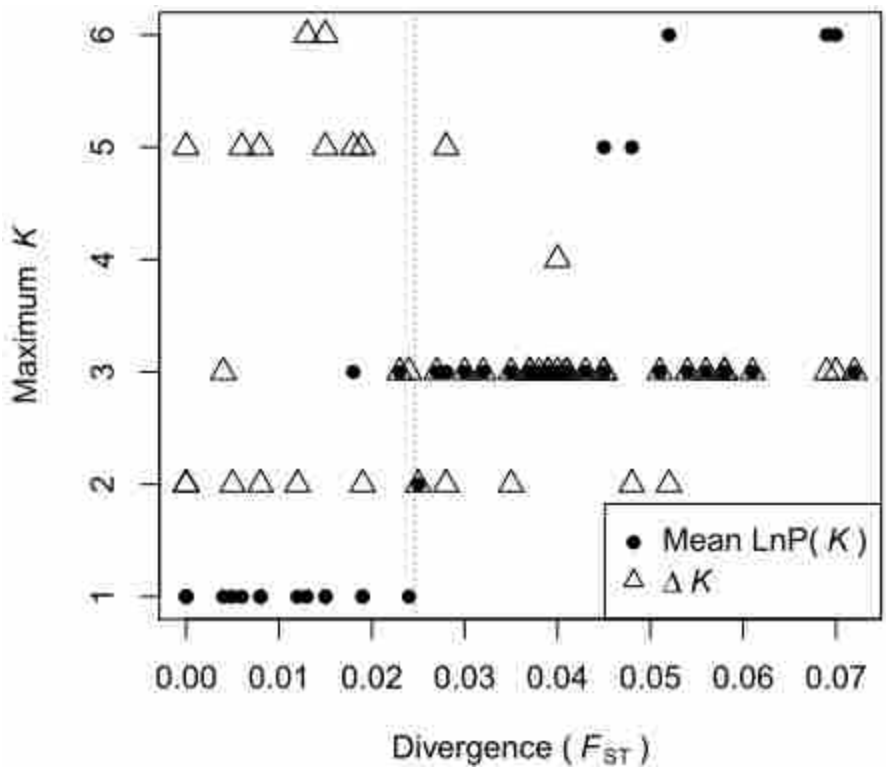


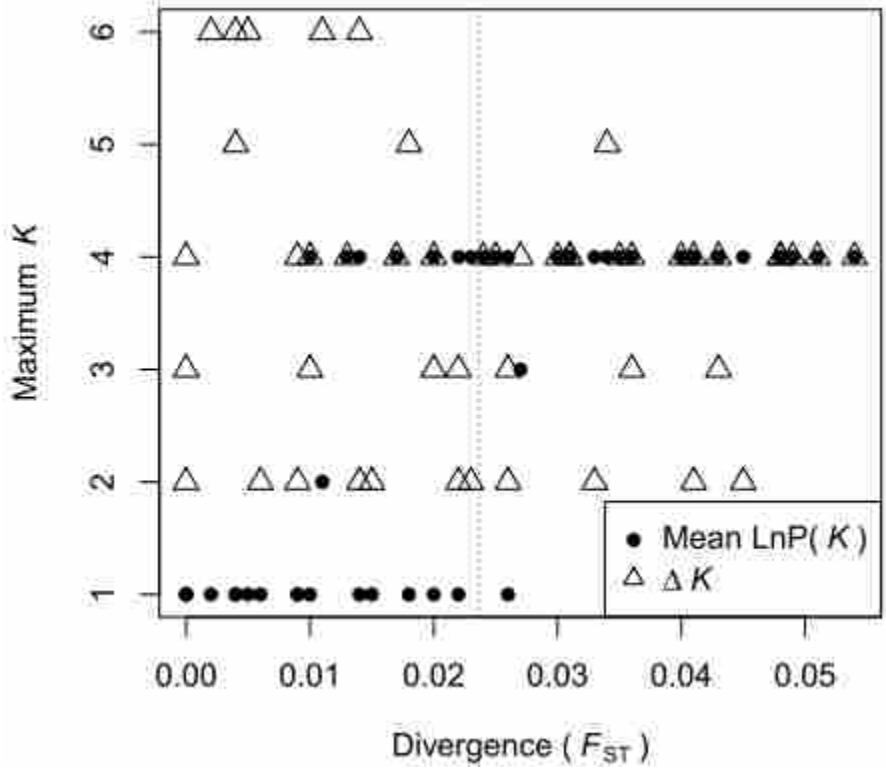
**Figure 3.2.** Effect of divergence ($F_{ST}$) among populations on the maximum value of both the mean LnP($K$) and $\Delta K$ statistics for scenarios (**A**–**E**) in the primary round of hierarchical analysis. Filled black circles indicate maximum mean LnP($K$) and open triangles indicate maximum $\Delta K$ for one sample. For scenarios **D** and **E**, results are shown for analyses after the hierarchical split (before the split, results were similar to scenario **A**). For scenario (**A**) true $K$ = 2, (**B**) true $K$ = 3, (**C**) true $K$ = 4, (**D**) true $K$ = 3, (**E**) true $K$ = 4, where true $K$ is defined as the number of subpopulations modeled in the simulation at a given generation. The dotted line indicates the mean value of $F_{ST}$ at which the maximum mean LnP($K$) statistic correctly identifies true $K$ (the number of simulated subpopulations).



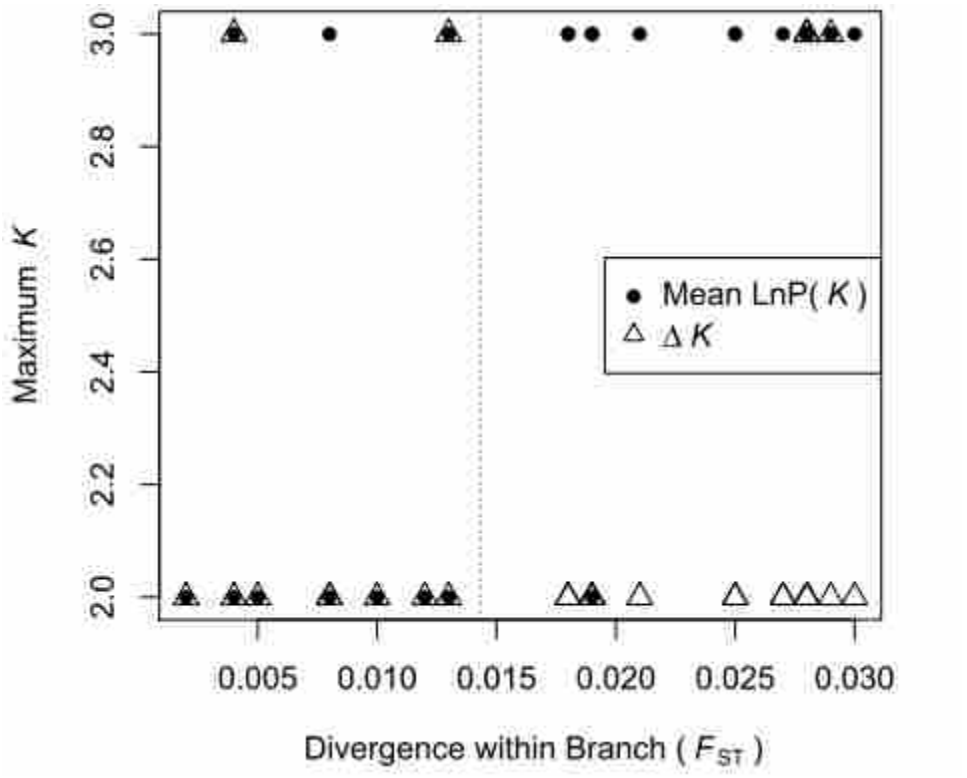

**(A)**


Maximum K
1
2
3
4
5
6
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Mean LnP( K )
Δ K
Divergence ( $F_{ST}$ )

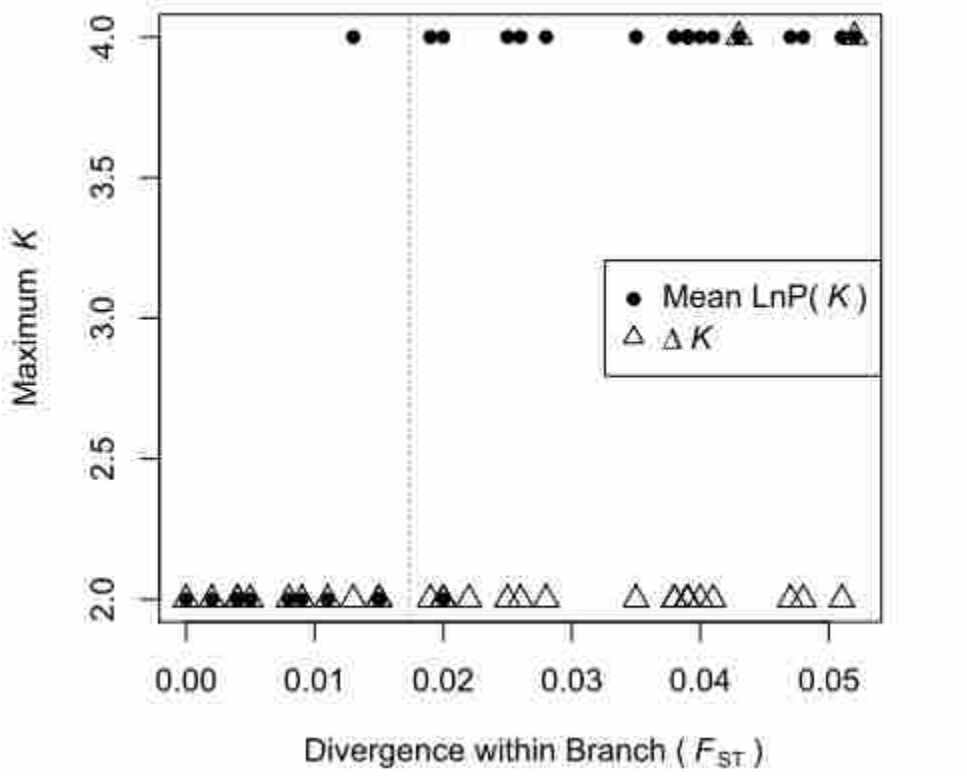
**(B)**


Maximum K
1
2
3
4
5
6
0.00
0.01
0.02
0.03
0.04
0.05
Mean LnP( K )
Δ K
Divergence ( $F_{ST}$ )

**(C)**

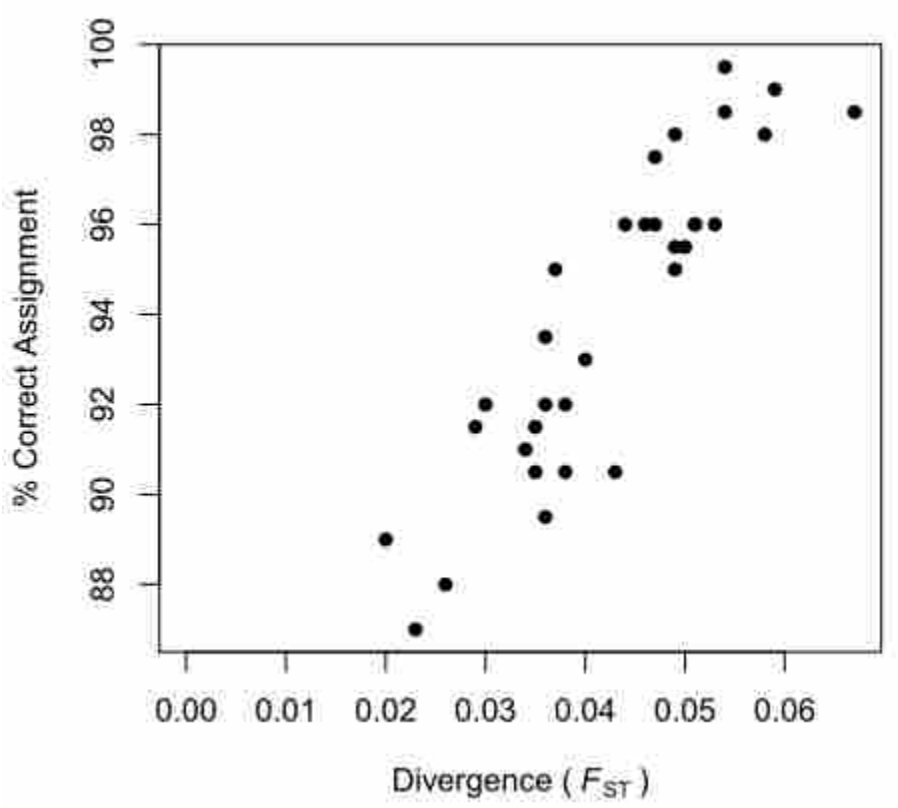

Maximum K
3.0
2.8
2.6
2.4
2.2
2.0
Mean LnP( K )
Δ K
0.005
0.010
0.015
0.020
0.025
0.030
Divergence within Branch ( $F_{ST}$ )

**(D)**

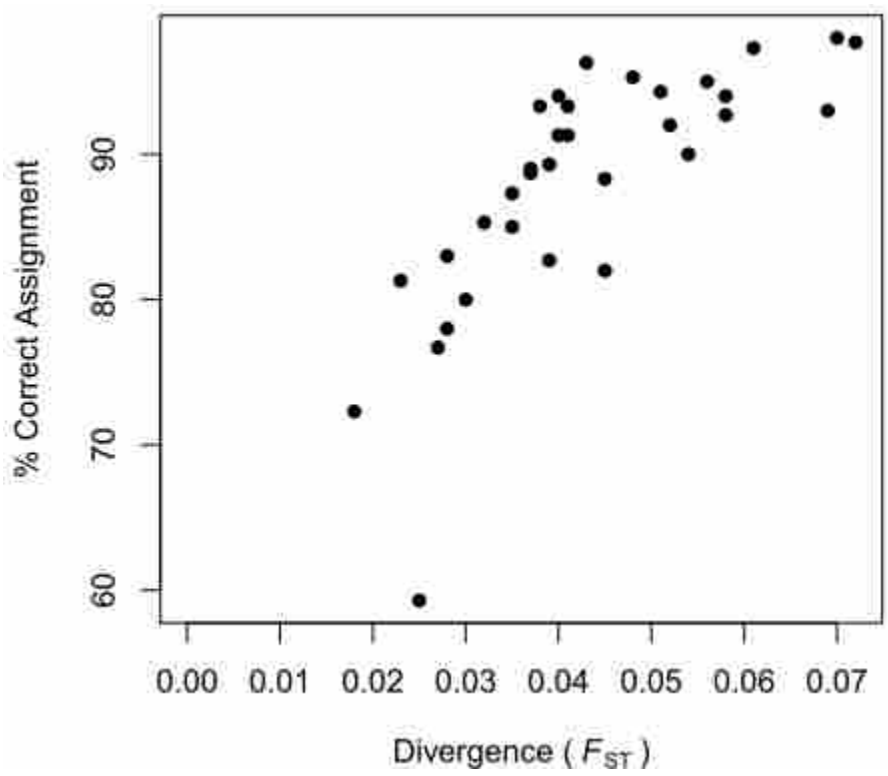

Maximum K
4.0
3.5
3.0
2.5
2.0
Mean LnP( K )
Δ K
0.00
0.01
0.02
0.03
0.04
0.05
Divergence within Branch ( $F_{ST}$ )

**(E)**

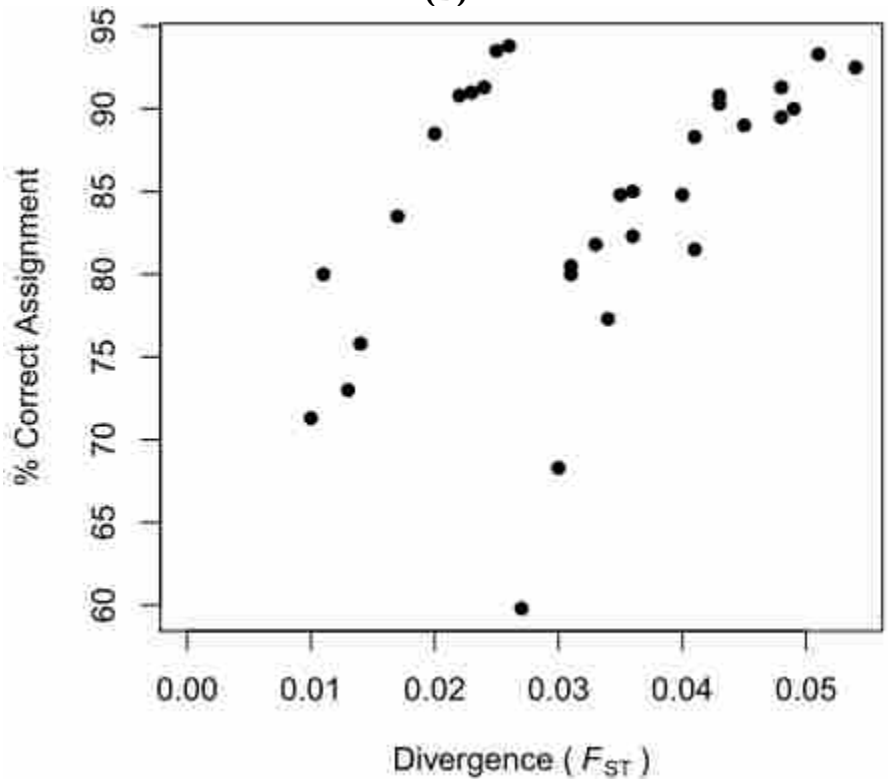
**Figure 3.3.** Effect of divergence ($F_{ST}$) among populations on the percent correct assignment [using maximum mean LnP($K$)] of sampled simulated individuals to their true population of origin by STRUCTURE. For scenarios **D** and **E**, results are shown for analyses after the hierarchical split (before the split, results were similar to scenario **A**). For scenario (**A**) true $K$ = 2, (**B**) true $K$ = 3, (**C**) true $K$ = 3, (**D**) true $K$ = 3, (**E**) true $K$ = 4. Mean $F_{ST}$ is shown when more than two subpopulations are involved (**B**, **C** and **E**) and was calculated among subpopulations within the hierarchical branches for **D** and **E**.



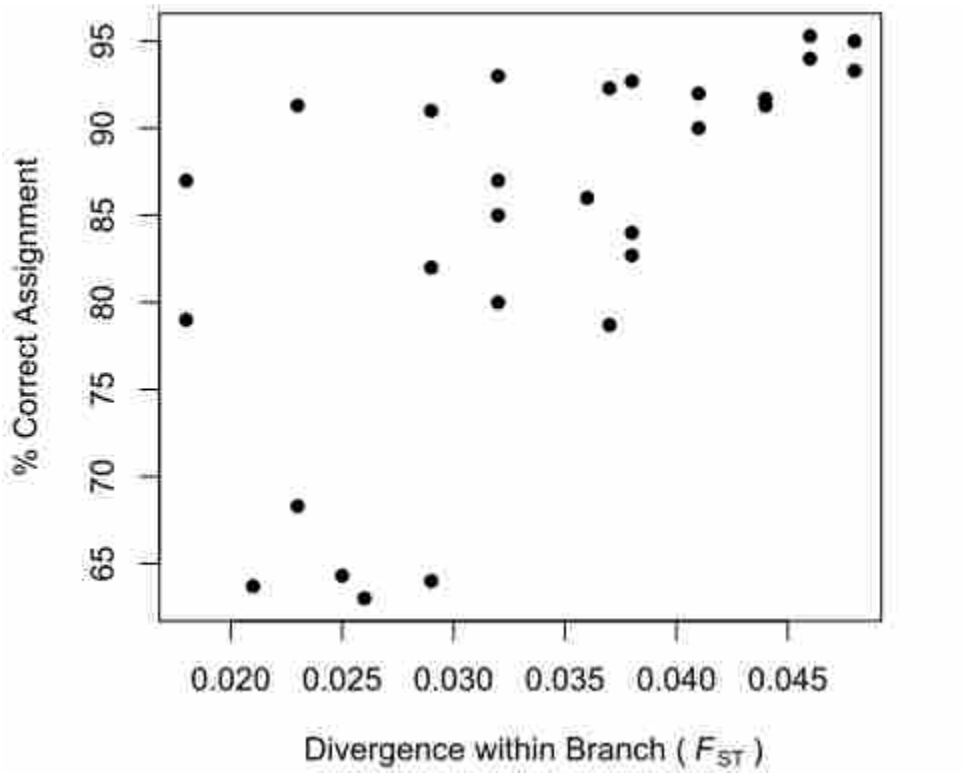

**(A)**


% Correct Assignment
60
70
80
90
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Divergence ( $F_{ST}$ )

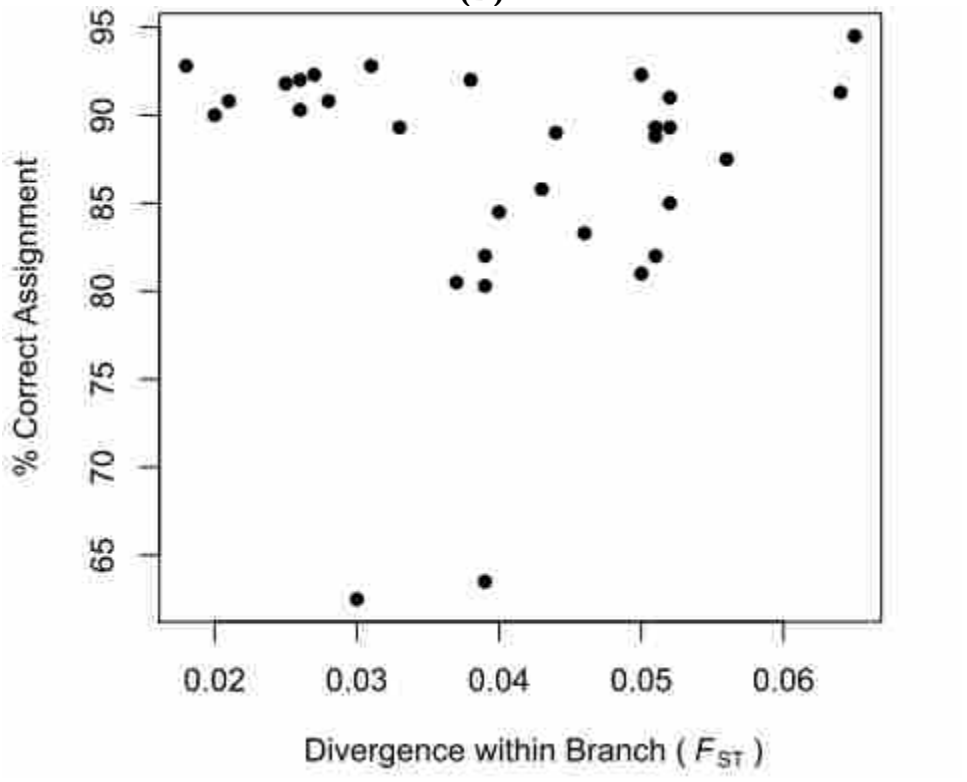
**(B)**


% Correct Assignment
60
65
70
75
80
85
90
95
0.00
0.01
0.02
0.03
0.04
0.05
Divergence ( $F_{ST}$ )

**(C)**


% Correct Assignment
65
70
75
80
85
90
95
0.020
0.025
0.030
0.035
0.040
0.045
Divergence within Branch ( $F_{ST}$ )

**(D)**


% Correct Assignment
65
70
75
80
85
90
95
0.02
0.03
0.04
0.05
0.06
Divergence within Branch ( $F_{ST}$ )

**(E)**

# CHAPTER 4

## GENETIC RECAPTURE IDENTIFIES LONG-DISTANCE BREEDING DISPERSAL IN GREATER SAGE-GROUSE (*CENTROCERCUS UROPHASIANUS*)


## ABSTRACT

Dispersal can strongly influence the demographic and evolutionary trajectory of populations. For many species, little is known about dispersal, despite its importance to conservation. The greater sage-grouse (*Centrocercus urophasianus*) is a species of conservation concern that ranges across 11 western U.S. states and 2 Canadian provinces. To investigate dispersal patterns among spring breeding congregations, we examined a 21-locus microsatellite DNA dataset of 3,244 greater sage-grouse sampled from 763 leks throughout Idaho, Montana, North Dakota, and South Dakota, USA, across 7 yr. We recaptured ~2% of individuals, documenting 41 instances of breeding dispersal (median dispersal distance = 15 km), with seven dispersal events of >50 km, including one of 194 km. We identified 39 recaptures on the same lek up to 5 yr apart, which supports the long-held paradigm of philopatry in lekking species. We found no difference between the sexes in breeding dispersal distances or in the tendency to disperse vs. remain philopatric. We also documented movements within and among state-delineated priority areas of conservation importance, further supporting the need to identify movement corridors among these reserves. Our results can be used to inform the assumptions of count-based population models and the dispersal thresholds used to model population connectivity.


## INTRODUCTION

Dispersal is crucial to maintaining population connectivity. It is the precursor to gene flow, influencing evolutionary processes such as local adaptation and speciation, and demographic processes such as population growth and persistence (Ronce 2007, Ellstrand and Rieseberg 2016). Rates of dispersal are influenced by intrinsic factors such as population density and access to mates, and by extrinsic, or environmental factors such as habitat quality and resource availability (Clobert *et al.* 2009). Offspring often disperse away from parents at a breeding site (natal dispersal) due to kin competition and inbreeding avoidance (Gandon 1999, Platt and Bever 2009). Individuals may also disperse among breeding sites following attempts at reproduction (breeding dispersal) to enhance mating opportunities and increase lifetime reproductive success (Johnson and Gaines 1990).

Long-distance dispersal may be important for population persistence of wide-ranging species (Bohrer *et al.* 2005). This is especially true in naturally fragmented or human-altered landscapes (Bohrer *et al.* 2005), where individuals must disperse through unsuitable habitat with limited available resources along the way. The main limitation to connectivity is the distance between populations that individuals are capable of dispersing; a single successful disperser per generation can transport genes across the landscape, eliminating inbreeding depression and increasing population fitness, survival, and viability (Mills and Allendorf 1996, Schwartz and Mills 2005, Whiteley *et al.* 2015). Therefore, the occurrence of long-distance dispersal can be vital to some species' persistence.

Long-distance dispersal is well documented for plant species (Nathan 2006). However, due to the difficulty of documenting long movements, it is less well known in animals (but see Lowe 2009, Moriarty *et al.* 2009, Hawley *et al.* 2016). The frequency and extent of emigrating

individuals is often underestimated because the quantification of long-distance dispersal is not the primary purpose of many studies (compared with, e.g., fine-scale habitat assessment), because sample sizes are too small to capture rare long-distance dispersal events, and because individuals dispersing long distances may leave the study area and be lost from detection (Koenig *et al.* 1996, Hassal and Thompson 2012). Furthermore, in studies designed expressly to quantify long-distance movements by tracking individuals using global positioning technology or geologgers, cost can be prohibitive (Bridge *et al.* 2013, Earl *et al.* 2016). Noninvasive genetic approaches can help to fill this knowledge gap. Genetic recapture, or the use of molecular genetics to identify individuals captured in different places and times, can be used to estimate dispersal frequency and distance when the focus of such studies is the spatial redistribution of large numbers of marked animals across large areas. However, events between capture and recapture go undetected (Nathan *et al.* 2003). Such approaches allow landscape-scale sampling of great numbers of individuals at relatively low cost per sample, and the collection of data that can additionally be used to plan biodiversity conservation.

The greater sage-grouse (*Centrocercus urophasianus*; hereafter, sage-grouse) is a lekking gallinaceous bird. Every spring, between March and May, individuals congregate on leks across the western United States and southern Canada. Lek locations are highly stable over generations, such that, following natal dispersal, most individuals are thought to exhibit philopatry, returning to the same lek every spring throughout their lifetime (Patterson 1952, Dalke *et al.* 1963, Emmons and Braun 1984, Dunn and Braun 1985). Natal dispersal of females is reported to be greater than that of males (median 8.8 vs. 7.4 km). However this oft-cited dispersal distance is based on a study in which the maximum distance between monitored leks was 13.1 km (Dunn and Braun 1985), and genetic data from 1 northern California population suggest that distances traveled by females may be underestimated (Davis *et al.* 2015).

On leks, a few territorial males may command the vast majority of mating, while nonterritorial adult and second-year males occupy the fringes (Semple *et al.* 2001), leading to high variation in breeding success (Payne 1984). Subdominant males may find mating opportunities by displaying and mating off the lek (Dunn and Braun 1986), or by visiting multiple leks within one breeding season to increase their chances of displacing dominant males or of finding females off the lek (Semple *et al.* 2001). Females are also known to visit multiple leks in a breeding season (Dunn and Braun 1985, Semple *et al.* 2001), occasionally visiting multiple leks within a week (Semple *et al.* 2001). Breeding dispersal and mate selection may occur multiple times with multiple mates during a single breeding season. However, distances traveled during breeding dispersal are unknown.

Field data show that sage-grouse are capable of long-distance movements. Among seasons, migratory sage-grouse may move 20 km (Tack *et al.* 2012), 30 km (Dunn and Braun 1986), or even 80 km (Connelly *et al.* 1988, Leonard *et al.* 2000) depending on habitat availability. Annual, obligate migrations of 122 km and 240 km have been documented by telemetry studies (Tack 2009, Tack *et al.* 2012). Most migratory movements are made in stepping-stone fashion (Tack 2009), but abrupt singular movements are possible when suitable habitat is lacking (Dunn and Braun 1986).

Research has documented sage-grouse natal dispersal distances, seasonal migration distances, and breeding behavior, but little is known about breeding dispersal distances. Furthermore, most studies of lek-site philopatry have been limited in geographic extent and sample size. New dispersal information would come at a critical time for sage-grouse, an imperiled species added to the federal Endangered Species Act (ESA) candidate list in 2010

following several petitions for protection (U.S. Fish and Wildlife Service 2010). A U.S. Fish and Wildlife Service determination in September 2015 found current efforts by state and federal agencies and other partners adequate to obviate the need for listing, but significant conservation challenges remain and the species' status will again be reviewed in 2020 (U.S. Fish and Wildlife Service 2015).

Understanding breeding dispersal is a critical step toward comprehending the relationship between the distribution and abundance of extant populations in fragmented landscapes. State wildlife management agencies across the range of sage-grouse in the western U.S. have collectively delineated Priority Areas for Conservation (PACs) to conserve population strongholds. This conservation strategy is directed at conservation management within individual PACs and planning for connectivity among PACs to prevent isolation and divergence of existing populations in the future (Finch *et al.* 2016). Understanding the distance and frequency of breeding dispersal vs. philopatry is critical to ongoing conservation planning (Crist *et al.* 2017).

As part of a larger study examining genetic substructure and how it relates to PAC delineation (Cross *et al.* 2016), we used molecular genetics tools to analyze thousands of sage-grouse feathers collected from hundreds of leks scattered across Idaho, Montana, North Dakota, and South Dakota, USA, to quantify breeding season dispersal. Our 4 primary objectives were: (1) to evaluate patterns of sage-grouse lek-site philopatry; (2) to quantify distances, frequency, and patterns of breeding season dispersal among leks; (3) to evaluate differences in breeding season dispersal characteristics between the sexes; and (4) to examine the relative cost of breeding season dispersal using known mortalities from a subset of dispersing individuals.

## METHODS

### Study area and sampling

For our analyses we used 7,629 spatially referenced sage-grouse feather ($n$ = 7,399) and blood ($n$ = 230) samples from across the northeastern extent of the species' range in Idaho, Montana, and North and South Dakota. Feather samples were collected noninvasively (Segelbacher 2002, Bush *et al.* 2005) from leks, mostly during the months of March through May. These samples were supplemented by blood and feather samples collected from sage-grouse trapped on leks as part of a radio-telemetry project in central Montana. Samples were collected from 835 leks (median: 9 samples per lek, IQR: 7 samples) of 2,292 known active leks from 2007 to 2013 by field biologists and technicians with the Bureau of Land Management, Montana Fish, Wildlife and Parks, and Montana Audubon (see Cross *et al.* 2016 for detailed methods).

### Laboratory analysis

**DNA extraction.** Feather DNA was extracted from the quill (calamus) using the DNeasy Blood and Tissue Kit (QIAGEN, Valencia, California, USA) and the user-developed protocol for purification of total DNA from nails, hair, or feathers (QIAGEN). We modified the protocol by incubating samples for a minimum of 8 hr after addition of Proteinase K (QIAGEN) and by eluting DNA with 100 µL of Buffer AE (QIAGEN). Feather samples were extracted in a lab used only for noninvasive DNA extraction to avoid potential contamination from samples with greater DNA concentrations. Blood samples were extracted using QIAGEN's DNeasy Blood and Tissue Kit and protocol for nucleated blood.

**Microsatellite DNA amplification and electrophoresis.** We amplified 21 variable microsatellite loci and 1 sex-diagnostic locus in 8 multiplex polymerase chain reactions (PCR). We used procedures detailed in Cross *et al.* (2016), with the addition of the following loci:

SG21, SG28, SG29, SG36, and SG39 (Appendix Table 4.3). Primers and locus-specific reaction mixes, annealing temperatures, and thermal cycler profiles are presented in Appendix Table 4.3, 4.4, and 4.5.

**Genotyping and identification of recaptures**

To ensure correct genotyping from low-quality and low-quantity feather DNA samples, each sample was PCR-amplified across the 22 loci to screen for allele dropout, stutter artifacts, and false alleles (DeWoody *et al.* 2006). To minimize genotyping error, two observers scored each sample. If any locus failed to amplify in either replicate, or if there was a discrepancy between locus genotypes as scored by the 2 observers, PCR-amplification and genotyping was repeated twice more. If a genotype was confirmed by this repeat analysis then it was retained. If a genotype failed again, the sample was assigned a missing score at the failed locus. To screen samples for quality control, we removed from analysis any individual for which amplification failed at >1/3 of the loci (i.e. >7 loci). After removal of poor-quality samples, genotypes were screened to ensure consistency between allele length and length of the microsatellite repeat motif.

To screen for and correct genotyping error, we used DROPOUT 2.3 (McKelvey and Schwartz 2005 as implemented in Schwartz *et al.* 2006) and package ALLELEMATCH 2.5 (Galpern *et al.* 2012) in R 3.3.0 (R Core Team 2016). In ALLELEMATCH, we used the amUnique function to generate a list of all potentially matched sample genotypes, using an alleleMismatch setting of 6 (as calculated using the amUniqueProfile function). The alleleMismatch setting is approximately equivalent to matching samples with up to six pairwise mismatched loci. We used the list of potentially matched samples as a basis for reexamination and repeat analysis to confirm the genotype scores. We reviewed all potentially matching samples, confirming a recapture only when we found no mismatch in genotype across the 22-locus panel. Detection of philopatry required that the same individual attend the same sampled lek in 2 different sampling years, shedding a feather in both years that was then both collected and successfully genotyped in the lab. Detection of a dispersal event required that an individual successfully emigrate from a sampled lek of first capture and successfully immigrate to a sampled lek of recapture, shedding a feather at each lek that was both collected and successfully genotyped in the lab. We calculated great-circle distance between all confirmed pairs of recaptures using the coordinates for each sample collection location with the spDistsN1 function in the SP package in R 3.3.0 (Pebesma and Bivand 2005, Bivand *et al.* 2013).

For all individuals, across all 22 loci, we quantified the power of our microsatellite locus panel to discern individuals by calculating probability identity ($P_{ID}$; Evett and Weir 1998)—the probability that two individuals drawn at random from the population could have the same genotype across all loci—using DROPOUT 2.3. Because we knowingly sampled from multiple populations (Cross *et al.* 2016), we did not test for deviation from Hardy-Weinberg proportions or gametic disequilibrium among loci.

**Sex-biased breeding season dispersal**

Because bias in natal dispersal has been documented (Dunn and Braun 1986), we tested whether the frequency of breeding season dispersal differed between the sexes by performing a Fisher's exact test for count data. To test whether males or females dispersed farther we performed a Mann-Whitney *U*-test, and to evaluate whether breeding season dispersal distances differed between the sexes within or among years we performed a Kruskal-Wallis rank sum test. We performed all tests in R (R Core Team 2016).

**Mortality and breeding season dispersal**

Dispersal can be costly to individuals. Therefore, the cost of dispersal combined with varied reproductive success in different habitats may affect the propensity of individuals to disperse (Leturque and Rousset 2002). In lek breeding systems, reproductive success varies greatly for individuals (Payne 1984). Therefore, individuals may hazard long-distance movements to improve breeding opportunity by visiting distant leks. To test whether breeding season dispersal distances differed between individuals recaptured as known mortalities and all other recaptures, we performed a Mann-Whitney *U*-test in R (R Core Team 2016).

## RESULTS

### Genotyping

After removing samples of inferior quality (those that failed at >7 loci; $n$ = 1,782, ~23%) and recaptures of the same sage-grouse at the same lek on either the same or different day within the same year ($n$ = 2,603), we retained 3,244 of 7,629 samples analyzed (~43%). The 3,244 high-quality genotypes from feathers ($n$ = 3,017) and blood samples ($n$ = 227) were from 763 leks, with a median 3 samples per lek (IQR: 3 samples, range: 1–62 samples per lek). We determined sex for 3,212 (99%) of the final individual genotypes: 600 females (~19%) and 2,612 males (~81%). Using our 22-locus panel (21 autosomal loci and 1 sex-linked locus), $P_{ID}$ was $2.20 \times 10^{-29}$, providing substantial power to discern individuals, given a suggested $P_{ID}$ of 0.001–0.0001 for law enforcement forensic applications in natural populations (Waits *et al.* 2001).

### Identification of recaptures

We recaptured ~2% of captured individuals, with 80 recaptures of 78 individuals from 3,244 total captures of 3,164 individuals. Recaptures matched initial capture genotypes across all 22 loci. Of 78 recaptured individuals, 9 were females (~12%) and 69 were males (~88%); ~2% of the 582 females and ~3% of the 2,472 males were genotyped.

Individuals were either recaptured in different years at the same lek (35 males and 3 females; Table 4.1, Figure 4.1A), in the same year at different leks (26 males and 5 females; Figures 4.1B, 4.1C, 4.2A), or in different years at different leks (10 males and 1 female; Figures 4.1B, 4.1C, 4.2A). Two males were recaptured twice; all other individuals were recaptured just once. One of the males was captured twice in the same year (2012) at different leks (14 km apart) and once one year later (2013) at a different lek 30 km away. We captured the second male three times in the same year (2013) at three different leks that were 21 km, 73 km, and 90 km apart. Five years was the longest time between capture and recapture at the same lek, but we also recaptured three individuals three years apart and another nine individuals two years apart. Median time between capture and recapture was 350 days (IQR: 373 days, range: 0–1,809 days, $n$ = 80).

Thirty-three breeding season dispersal movements were within, among, out of, or into a PAC (Table 4.2). Breeding season dispersal events within sage-grouse PACs was greater than double the number of movements documented outside PACs. Twenty-four movements occurred within PAC boundaries and another 10 occurred outside PAC boundaries. Three movements occurred among PACs, all in Idaho, in which 3 different sage-grouse made 31 km, 35 km, and 70 km movements. Another six individuals dispersed into or out of PACs. Two individuals moved into PACs, both from a distance of 62 km. One other individual was first captured within a PAC and then recaptured outside that PAC, moving 13 km. Three individuals moved into or out of PACs, traveling 14 km, 127 km, and 194 km in a single season.

### Sex-biased breeding season dispersal

The frequency of breeding season dispersal was similar between the sexes (Fisher's exact test for count data: $P = 0.49$, 95% CI = 0.35–12.16, odds ratio = 1.83, 2-tailed). Among dispersing individuals, the distance moved across all years was similar between the sexes (Mann-Whitney *U*-test: W = 133, P = 0.53, 2-tailed), as was distance moved within and among years (Kruskal-Wallis rank sum test: $\chi^2 = 1.54$, df = 3, $P = 0.67$). Dispersing females moved a median 12 km (IQR: 16 km, range: 3–35 km, $n = 6$) and dispersing males moved a median 15 km (IQR: 24 km, range: 0–194 km, n = 38; Figures 4.2A, 4.2B).

**Mortality and breeding season dispersal**

Feathers from five recaptured sage-grouse (all male) were collected from known mortalities found on or near active leks. Four of the carcasses showed evidence of predation. Two of these were recaptured in the same year: one was 127 km from its lek of origin, and the second had been captured previously live at two different leks 73 km and 90 km away. The other two sage-grouse were recaptured a year after initial capture: one near the same lek and the second 13 km from its original capture site. The fifth known mortality had struck a powerline 43 km away from initial capture. Individuals recaptured as mortalities dispersed significantly farther than did all other recaptures (Mann-Whitney *U*-test: $W = 31$, $P = 0.01$, 2-tailed). Individuals recaptured as mortalities moved a median 73 km (IQR: 46 km, range: 13–127 km, $n = 5$), whereas all other recaptured individuals moved a median 14 km (IQR: 18 km, range: 0–194 km, $n = 39$; Figure 4.2A).

## DISCUSSION

Collectively, our findings support the long-held paradigm of lek philopatry in sage-grouse, yet we also identified highly mobile segments of breeding populations that readily dispersed farther than previously known. Long-distance dispersal events are certainly more common than we were able to detect. Individuals showed strong philopatry to leks across the 4-state study region, both within and between years, with evidence of recapture at the same site five years apart. The lek selected during natal dispersal (Dunn and Braun 1985) likely establishes the lek to which most sage-grouse remain philopatric (Schroeder and Robb 2003). However, breeding season dispersal also shapes populations. Our genetic approach to documenting dispersal is the first of its kind in sample size and geographic scope, and is novel for capturing long-distance exchanges in sage-grouse populations, documenting 7 movements of >50 km, 6 of which occurred within a single lekking season. Our estimates of philopatry and dispersal are biased low, given the events required to successfully capture and recapture an individual using genetic sampling. Furthermore, the number of dispersers recaptured is more biased than the number of philopatric individuals recaptured as we sampled only 36% of known active leks ($n = 835$, $N = 2{,}292$), thereby missing any dispersal from or to unsampled leks.

Sage-grouse genetic structure is governed by the process of isolation by distance, whereby the cumulative effects of many short- and fewer long-distance dispersals shape patterns of relatedness (Oyler-McCance *et al.* 2005, Bush *et al.* 2011, Schulwitz *et al.* 2014, Davis *et al.* 2015, Cross *et al.* 2016). Lek-based mating systems should result in inbreeding depression, but dispersal as documented here may alleviate such a deleterious effect (Tallmon *et al.* 2004, Whiteley *et al.* 2015) and may also function to extend the neighborhood of advantageous adaptations (Richardson *et al.* 2014).

Breeding season dispersal may present additional mating opportunities as grouse visit multiple leks, although further study is needed because breeding outcomes for dispersing sage-grouse are unknown. Dominant males may simply be maximizing reproductive opportunity, their

subdominant counterparts may be displaced by dominant males, or males may disperse following unsuccessful mating attempts. Females may disperse following nest failure near their first attended lek, or may disperse to mate with males at other leks, which could result in multiple-paternity broods (Bush *et al.* 2010). Regardless of mechanism, we could not detect a difference in breeding season dispersal behavior between the sexes, likely due to low statistical power. Male-biased sampling is an artifact of the greater amount of time that males spend on leks compared with females, and of energetic male display and fighting on leks compared with relatively quiescent female behavior; male behaviors are more likely to result in a great number of dropped feathers, which can be collected. Widely assumed, but undocumented until now, however, is a potential tradeoff for dispersing to increase mating opportunities: based on a small sample of mortalities, heightened risk appears to accompany long-distance breeding season dispersal.

Greater breeding season dispersal distances within years compared with across years is likely not an underlying biological driver but instead an artifact of higher sampling intensity in later years (Hassall and Thompson 2012). We may have misclassified year of capture or recapture because feathers can persist for multiple years on a lek. However, feathers weather poorly, and we avoided extracting DNA from feathers which appeared aged (feathers that were dirty, physically damaged, or for which the calamus had become opaque from extended UV exposure). In harsh sagebrush environments, high UV radiation and freeze–thaw cycles should rapidly degrade and shear DNA, rendering most old samples incapable of producing a viable genotype (Segelbacher 2002). Regardless, we likely underestimated the frequency and maximum distance of breeding season dispersal events, estimates of which will increase with continued feather collection (Hassall and Thompson 2012). Still, our estimates are valuable for parameterizing individual-based models (Wood *et al.* 2015) and for seeding scenarios that evaluate connectivity under variable dispersal rates and distances (Knick and Hanser 2011, Knick *et al.* 2013, Crist *et al.* 2017). Furthermore, population trends are monitored by counting males annually on hundreds of leks across the 11-state, 2-province sage-grouse range, and our dispersal estimates may be incorporated into trend or density modeling to account for breeding season dispersal among leks and the reality that some individuals are counted more than once (McCaffery *et al.* 2016).

Having documented breeding season dispersal within, between, and outside PACs, we recommend that future connectivity research focus on resistance surface–based modeling (Wade *et al.* 2015) to identify low-cost paths that facilitate continued movement. To date, conservation planning for imperiled sage-grouse has relied on the findings of a few localized dispersal and migratory behavior studies as a surrogate for understanding long-distance dispersal. Localized dispersal studies have revealed short-distance movements—mostly <10 km, with few >20 km—and have drawn lek fidelity into question (Dalke *et al.* 1963, Wallestad and Schladweiler 1974, Dunn and Braun 1985, Hanf *et al.* 1994, Schroeder and Robb 2003). Research into sage-grouse migratory behavior has revealed that some individuals move in stepwise fashion among stepping-stones of intact habitat, and are capable of 250-km round trips (Smith 2012, Tack *et al.* 2012). Now, in addition to the understanding that behavioral research has provided, it is possible to parameterize connectivity models using dispersal distances and genetic data gained from this large-scale, high-sampling-resolution research. Resultant findings will allow managers to quantify connectivity and prioritize leks for conservation according to those that contribute the most to gene flow (Jacoby and Freeman 2016).

## ACKNOWLEDGMENTS

We thank Joe Smith, Jason Tack, Taylor Wilcox, and Katie Zarn for their comments on this manuscript. This research would not have been possible without the many state, federal, and non-governmental organization (NGO) biologists and technicians who have spent lifetimes working tirelessly in the field and behind desks for the conservation of the Greater sage-grouse. We specifically acknowledge Jake Chaffin (U.S. Forest Service [USFS]), Ben Deeble (Montana Audubon), retired, Brad Fedy (University of Waterloo), Don Kemner (Idaho Fish and Game [IDFG]), Steve Knick (U.S. Geological Survey [USGS], retired), Rick Northrup (Montana Fish, Wildlife and Parks [MTFWP]), Sara Oyler-McCance (USGS), Aaron Robinson (North Dakota Game and Fish [NDGF]), Travis Runia (South Dakota Game, Fish and Parks [SDGFP]), Dale Tribby (Bureau of Land Management [BLM], retired), Susan Werner (IDFG), Catherine Wightman (MTFWP), and David Wood (USGS). Our molecular genetics analyses and manuscript preparation would not have been possible without Cory Engkjer, Kevin McKelvey, and Kristy Pilgrim, with laboratory work by Kara Bates, Nasreen Broomand, Taylor Dowell, Scott Hampton, Randi Lesagonicz, Inga Ortloff, Sara Schwarz, Kate Welch, and the USFS Rocky Mountain Research Station National Genomics Center for Wildlife and Fish Conservation.

## FUNDING STATEMENT

This study was supported by grants from the Montana and Dakotas Bureau of Land Management (07-IA-11221643-343, 10-IA-11221635-027, 14-IA-11221635-059), the Great Northern Landscape Conservation Cooperative (12-IA-11221635-132), and the Natural Resources Conservation Service—Sage Grouse Initiative (13-IA-11221635-054). Use of trade names does not imply endorsement by the U.S. Government. The views in this article are those of the authors and do not necessarily reflect those of their employers. The authors declare no conflict of interest.

## ETHICS STATEMENT

Handling of sage-grouse and collection of feathers was conducted by our partners, under appropriate required permitting. For our laboratory research and analysis, no permitting was required.

## AUTHOR CONTRIBUTIONS

T.B.C. conceived the idea, collected the data, analyzed the data, and conducted the research. T.B.C. and M.K.S. developed and designed methods. J.C.C. oversaw sample collection and contributed substantial materials, resources, and funding. T.B.C., D.E.N., and M.K.S. wrote the paper or substantially edited the paper.

**Table 4.1.** Summary of genetic capture of Greater sage-grouse in Idaho, Montana, North Dakota, and South Dakota, USA, including year of genetic sample collection (Year), total number of individuals genotyped each year (*N*), total number of captures ($n_c$) and recaptures ($n_r$) each year, other collection years in which captured individuals were recaptured at the same lek as their lek of initial capture and how many were recaptured in each year (Recaptured same lek (*n*)), and collection years in which captured individuals were recaptured at a different lek from their lek of initial capture and how many were recaptured in each year (Recaptured different lek (*n*)).

| Year | *N* | $n_c$ | $n_r$ | Recaptured same lek (*n*) | Recaptured different lek (*n*) |
|---|---|---|---|---|---|
| 2007 | 85 | 2 | 0 | (0) | (0) |
| 2008 | 25 | 1 | 0 | (0) | (0) |
| 2009 | 590 | 14 | 3 | 2007 (1), 2008 (1) | 2009 (1) |
| 2010 | 276 | 9 | 9 | 2009 (5) | 2009 (3), 2010 (1) |
| 2011 | 269 | 9 | 9 | 2009 (3), 2010 (2) | 2009 (1), 2010 (1), 2011 (2) |
| 2012 | 1,045 | 35 | 31 | 2007 (1), 2010 (5), 2011 (6) | 2012 (19) |
| 2013 | 954 | 8 | 28 | 2009 (1), 2011 (1), 2012 (12) | 2012 (5), 2013 (9) |
| Total | 3,244 | 78 | 80 | 38 | 42 |

**Table 4.2.** The number of Greater sage-grouse breeding season dispersal movements among, entering (incoming) or leaving (outgoing), outside, or within priority areas for conservation (PACs) in Idaho, Montana, North Dakota, and South Dakota, USA, 2007–2013. Also shown are summary statistics for distances in each direction of movement.

| Direction | Number of movements | Distance (km) | | | |
|---|---|---|---|---|---|
| | | Minimum | Median | Mean ± SD | Maximum |
| Among | 3 | 31.30 | 34.73 | 45.28 ± 21.31 | 69.80 |
| Incoming or outgoing | 6 | 13.37 | 62.40 | 67.64 ± 59.04 | 194.39 |
| Outside | 10 | 0.27 | 17.40 | 26.60 ± 30.34 | 89.77 |
| Within | 24 | 0.26 | 9.05 | 16.62 ± 22.65 | 109.61 |

**Figure 4.1.** Greater sage-grouse recapture locations based on feather genotypes at (**A**) the same lek in different years (philopatry), and (**B**) in the same or different years at different leks in Idaho, USA, and (**C**) Montana, USA, 2007–2013. Arrows show breeding season dispersal between capture (tail) and recapture (head) locations. The dotted black line represents the North American continental divide, solid black lines represent state boundaries, solid light gray lines represent major rivers, and dashed dark gray lines represent major highways.

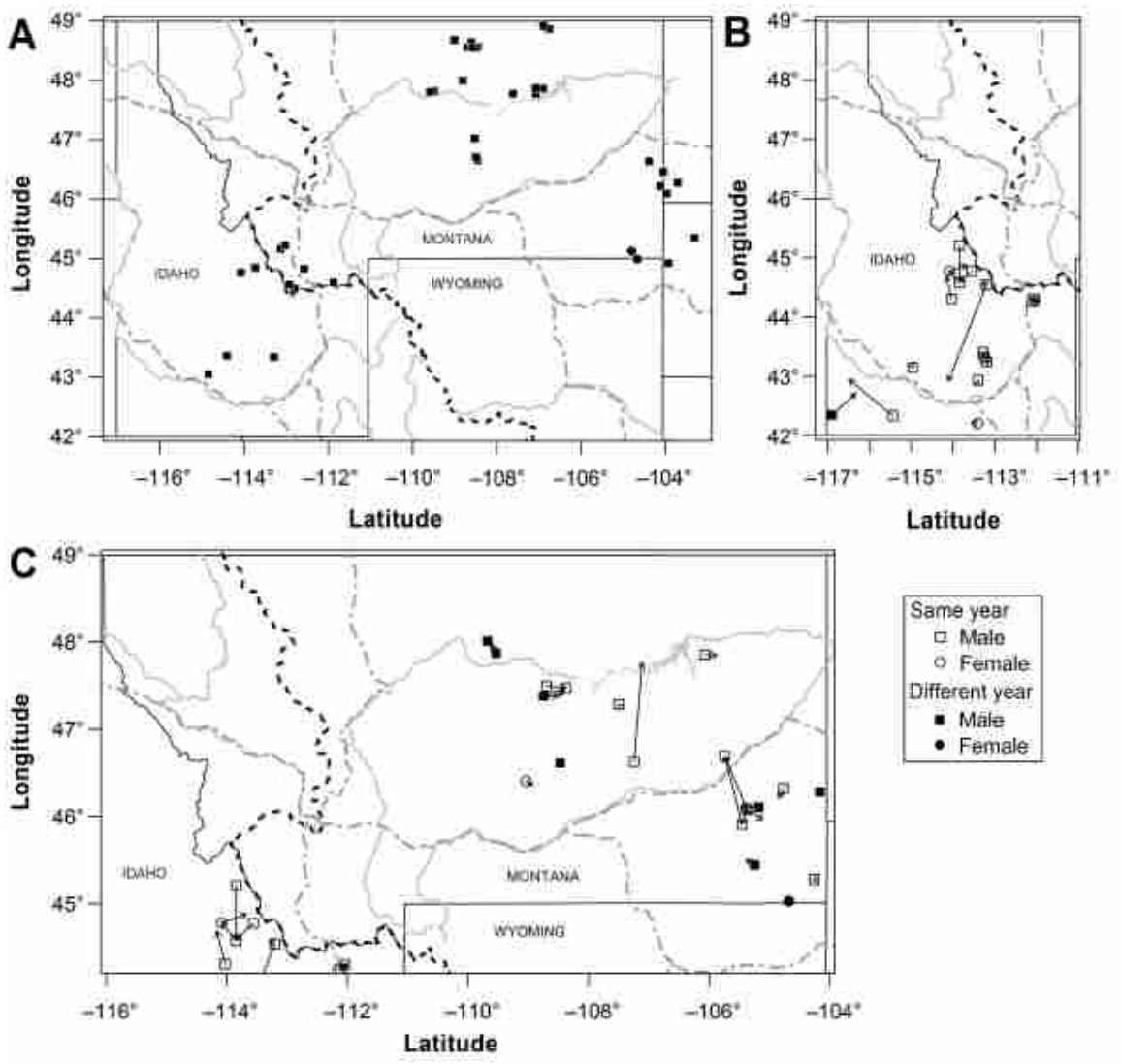
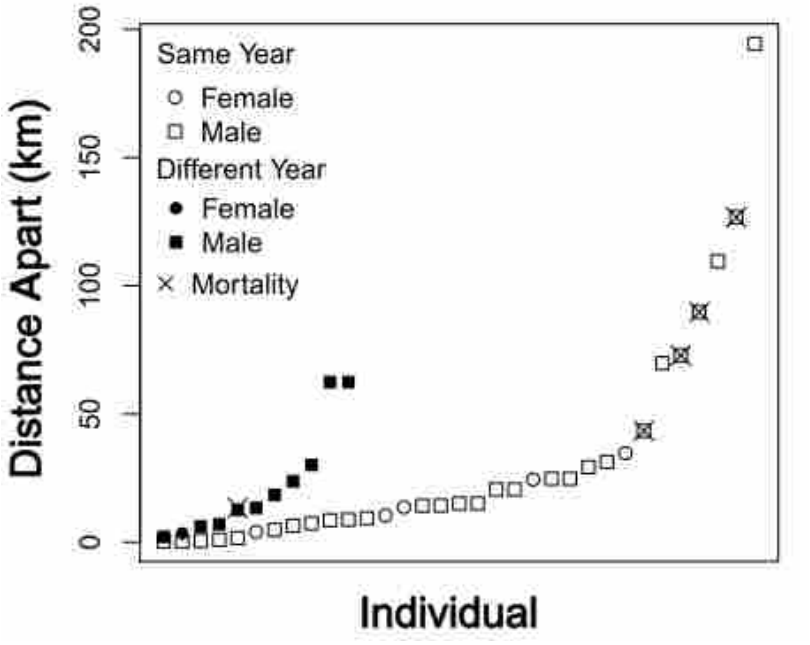
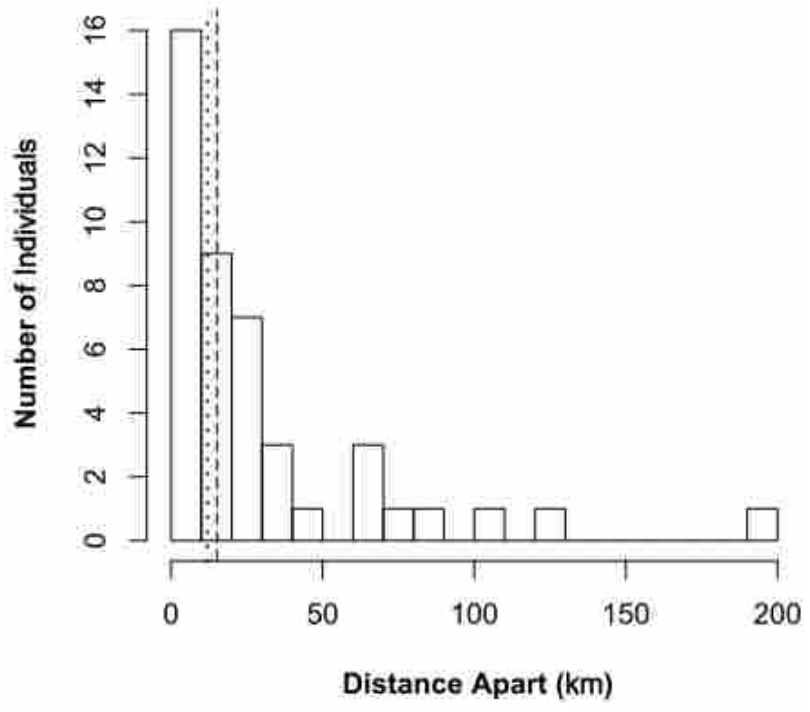


**Figure 4.2.** (**A**) Individual distances between capture and recapture locations and (**B**) distribution of distances travelled by Greater sage-grouse in Idaho, Montana, North Dakota, and South Dakota, USA, 2007–2013. In (**A**), points represent individuals plotted in order of increasing dispersal distance. In (**B**), the dotted line indicates the median dispersal distance for females (12.02 km), and the dashed line indicates the median dispersal distance for males (15.08 km). Philopatry is not plotted.




(**A**)




**(B)**

## APPENDIX

**Appendix Table 4.3.** Forward and reverse primer sequences used to genotype Greater sage-grouse in Idaho, Montana, North Dakota, and South Dakota, USA, 2007–2013. Repeat motif, size range, number of alleles (# alleles), probability identity ($P_{ID}$), probability identity sibling ($P_{IDsib}$), and sources for the primer sequence are shown for the 21 variable microsatellite loci and 1 diagnostic sex locus (1237) used in this study. Primers that required redesign to increase efficacy with noninvasive samples are indicated with an asterisk.

| Locus | Forward primer | Reverse primer | Repeat motif | Size range | # alleles | $P_{ID}$ | $P_{IDsib}$ | Source |
|---|---|---|---|---|---|---|---|---|
| 1237 | GAGAAACTGTGCAAAACAG | TAAAGCTGATCTGGAATTTCA* | Bi-allelic | 224 & 252 | 2 | 0.702 | 0.8401 | a |
| BG6 | AAAGAGGCAAGCACTCACAATG | CCCTTGGAATATCCTTTAACAAAAC | $(GATA)_{15}$ | 199–311 | 18 | 0.0172 | 0.3034 | b |
| BG16 | GTCATTAGTGCTGTCTGTCTATCT | TGCTAGGTAGGGTAAAAATGG | $(CTAT)_{15}$ | 125–185 | 12 | 0.0698 | 0.3712 | b |
| BG18 | CCATAACTTAACTTGCACTTTC | CTGATACAAAGATGCCTACAA | $(CTAT)_{17}$ | 135–171 | 10 | 0.6538 | 0.8166 | b |
| MS06.4 | CCTGGAGCAACTTGAGG | GTGACATTICCCCCCAC | $(GATA)_2(GGTA)_6(GATA)_5$ | 118–178 | 14 | 0.0602 | 0.3604 | c |
| MS06.6 | CAAACAACTGTCTTCCAGTAAGAC | AGAGCCTICATTTCTGGCAG | $(CAT)_{16}$ | 122–185 | 22 | 0.0275 | 0.3208 | c |
| MS06.8 | GCAAAATCAATAGAAGTAGAGAGG | CAGTAGCAGCTTTGTTTGG | $(GATA)_{17}$ | 103–159 | 15 | 0.0336 | 0.3276 | c |
| MSP11 | GGTGAAAAGTGTGGCAACTG* | CATTGTCAGCTTGCAGAC | $(CA)_{22}$ | 206–280 | 35 | 0.0207 | 0.3091 | c |
| MSP18 | CAATGACAGTATTTCCCAGATTA | GAATGGTAATATACTAAGCACAGG | $(CA)_{14}$ | 96–120 | 13 | 0.0333 | 0.3294 | c |
| SG21 | AGGCAAAACAGTCACACATGC | ATCACAAGCAGAGTGCAGGC | $(TC)_{54}$ | 180–138 | 67 | 0.0128 | 0.2956 | d |
| SG28 | ACAGGGGAAGGACAGACTGG | ACCTCTGCTTTTCCATTGCC | $(AC)_{50}$ | 113–169 | 29 | 0.0114 | 0.2926 | d |
| SG29 | AAGGGGCTTAGGGTTTTAATGG | AGTTAACTAAGTTGGGCAGGGG | $(AC)_{50}$ | 110–154 | 21 | 0.0207 | 0.3086 | d |
| SG36 | TTCCAGACATTTTGGGAGCC | CACATGTCCATCCAACCACC | $(ATGG)_{52}$ | 103–251 | 14 | 0.0794 | 0.3816 | d |
| SG39 | GAAAGTCTGAATGCTGGAGAACC | AAGCGTACTGTTTGCTCCCC | $(ATC)_{45}$ | 148–184 | 12 | 0.0441 | 0.3414 | d |
| SGCA5 | CGGACAGGTACATCCTGGAA* | GGGAAAAGATGTCAGAATCTACAAA* | $(CA)_{12}$ | 120–144 | 13 | 0.0604 | 0.3612 | e |
| SGCA11 | GCAGTAAAGAAAATTTGGAAGCA* | TCTTGAACTGATGTTGGATTTG* | $(AC)_{14}(AG)_2(CA)_5$ | 181–203 | 11 | 0.0759 | 0.3806 | e |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| SGCTAT1 | GCGACACTGCTCCCACCT | GAAAGGTTGTAAGAGGTCGT | $(CTAT)_{11}$ | 93–133 | 11 | 0.0607 | 0.3587 | e |
| TTD6 | GGACTGCTTGTGATACTTGCT | CATGCAGATGACTTTCAGCA | $(CA)_{17}$ | 107–153 | 20 | 0.0443 | 0.341 | f |
| TTT3 | TAGCAAACGAACCAGCCAAC | GCTCTGAATCTGCCCATCTCT | $(CATC)_n$ | 192–228 | 10 | 0.1243 | 0.429 | g |
| TUD3 | TCCAAGGGGAAAATATGTGTG | TTCTTCCAGCCCTAGCTTTG | $(TG)_{12}$ | 154–222 | 29 | 0.0108 | 0.2909 | h |
| TUT3 | CAGGAGGCCTCAACTAATCACC | CGATGCTGGACAGAAGTGAC | $(TATC)_{11}$ | 140–172 | 9 | 0.1431 | 0.4462 | h |
| TUT4 | GGAGCATCTCCCAGAGTCAG* | TCAGCTGTGAACCAGCAATC* | $(TATC)_8$ | 163-203 | 11 | 0.0412 | 0.3369 | h |

[a]Kahn *et al.* (1998)
[b]Piertney and Höglund (2001)
[c]Oyler-McCance and St. John (2010)
[d]S. Oyler-McCance personal communication
[e]Taylor *et al.* (2003)
[f]Caizergues *et al.* (2001)
[g]Caizergues *et al.* (2003)
[h]Segelbacher *et al.* (2000)

**Appendix Table 4.4.** Microsatellite locus multiplexes, primer annealing temperatures, and reagent mixes used in polymerase chain reactions (PCR) to genotype Greater sage-grouse samples from Idaho, Montana, North Dakota, and South Dakota, USA, 2007–2013. Columns 3–14 (F1, F2, F3, R1, R2, R3, *Taq*, 10x buffer, dNTP, $MgCl_2$ , BSA, and $H_2O$) are measured in µL. F1–F3 indicate the amount of forward primer added to the reactions, and R1–R3 indicate the amount of reverse primer added to the reactions. All reactions used 1 µM IDT Custom DNA Oligos Forward Primer (Integrated DNA Technologies, Coralville, Iowa, USA), 10 µM Eurofins MWG Operon Custom DNA Oligos Reverse Primer (Eurofins Scientific, Lancaster, Pennsylvania, USA), Invitrogen 5 *U* $\mu L^{-1}$ AmpliTaq Gold DNA *Taq* Polymerase (Thermo Fisher Scientific, Waltham, Massachusetts, USA), Invitrogen GeneAmp 10x PCR Buffer II (100 mM Tris-HCl, 1.5 mL pH 8.3, 500 mM KCI; Thermo Fisher Scientific), New England Biolabs Deoxynucleotide Set (25 µmol 100 mM ultrapure dATP, dCTP, dGTP, dTTP)—dNTP (New England Biolabs, Ipswich, Massachusetts, USA), Invitrogen 25 mM $MgCl_2$ (Thermo Fisher Scientific), bovine serum albumen (~66 kDA, used to stabilize enzymes during digestion of DNA—to prevent adhesion of the enzyme to reaction tubes, to inactivate contaminating nucleases and proteases, to stabilize nucleic acid modifying enzymes, as a blocking agent to minimize background, and to increase PCR yield from low purity templates; Thermo Fisher Scientific), and nuclease-free water (Thermo Fisher Scientific).

| Primer multiplex | Annealing temperature (°C) | F1 | F2 | F3 | R1 | R2 | R3 | *Taq* | 10x Buffer | dNTP | $MgCl_2$ | BSA | $H_2O$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1237 / BG18 / MSP18 | 54 | 0.10 | 0.10 | — | 0.20 | 0.20 | — | 0.26 | 1.00 | 1.00 | 0.80 | 0.10 | 4.24 |
| BG16 / MS06.8 / MSP11 | 52 | 0.10 | 0.07 | 0.10 | 0.20 | 0.20 | 0.20 | 0.26 | 1.00 | 1.00 | 0.80 | 0.10 | 3.97 |
| BG6 / MS06.4 / SGCA5 | 54 | 0.10 | 0.20 | 0.07 | 0.20 | 0.20 | 0.20 | 0.26 | 1.00 | 1.00 | 1.20 | 0.10 | 3.47 |
| MS06.6 / SG21 / SG28 | 60 | 0.10 | 0.10 | 0.10 | 0.20 | 0.20 | 0.20 | 0.26 | 1.00 | 1.00 | 0.80 | 0.10 | 3.94 |
| SG29 / SG36 / SG39 | 60 | 0.10 | 0.07 | 0.20 | 0.20 | 0.20 | 0.20 | 0.26 | 1.00 | 1.00 | 0.80 | 0.10 | 3.87 |
| SGCA11 / SGCTAT1 / TUT4 | 60 | 0.10 | 0.10 | 0.13 | 0.20 | 0.20 | 0.20 | 0.26 | 1.00 | 1.00 | 0.80 | 0.10 | 4.17 |
| TTD6 / TUT3 | 56 | 0.13 | 0.20 | — | 0.20 | 0.20 | — | 0.26 | 1.00 | 1.00 | 1.20 | 0.10 | 3.71 |
| TTT3 / TUD3 | 55 | 0.20 | 0.10 | — | 0.20 | 0.20 | — | 0.26 | 1.00 | 1.00 | 0.80 | 0.10 | 4.14 |

**Appendix Table 4.5.** Polymerase chain reaction (PCR) thermocycler phase, temperature, time, and number of cycles used to amplify microsatellite loci for Greater sage-grouse samples from Idaho, Montana, North Dakota, and South Dakota, USA, 2007–2013. Specific loci multiplexes and primer annealing temperatures are given in Appendix Table 4.

| Phase | Temperature (°C) | Time (min) | Cycles |
|---|---|---|---|
| Initial denaturation | 94 | 11 | 1 |
| Denaturation | 94 | 1 | 44 |
| Primer annealing | See Appendix Table 4 | 1 | 44 |
| Extension | 72 | 1 | 44 |
| Holding | 12 | ∞ | 1 |

# CHAPTER 5

# THE GENETIC NETWORK OF GREATER SAGE-GROUSE: RANGE-WIDE IDENTIFICATION OF KEYSTONE HUBS OF CONNECTIVITY


## ABSTRACT

Genetic networks can be used to characterize and rank nodes that are most important to the overall connectivity of a system allowing scarce resources to be guided towards nodes integral to connectivity. The greater sage grouse (*Centrocercus urophasianus*) is a species of conservation concern that breeds in individual leks that must be connected for population persistence. We genotyped 5,950 individuals, from 1,200 greater sage grouse leks, at 15 microsatellite loci across the entire species' geographic range to quantify connectivity and identify the leks most essential to maintaining connectivity. We found a single robust network composed of 459 interconnected nodes. The network had small-world characteristics where genetic connectivity was influenced by geographic distance among nodes. We identify nodes and connections among nodes that are of high conservation importance. Nodes ranking in the top 1% of betweenness and closeness centrality, both measures of network-wide connectivity, were located toward the center of the species' range. We also discovered that nodes near the northern periphery of the range rank in the top 1% of the clustering coefficient, indicative of groups within which gene flow is relatively unencumbered. Centrality measure distributions indicated the presence of keystone nodes, defined as nodes that have a greater influence on the network relative to the size of the lek. The loss of these hubs could result in the disintegration of the network into smaller, isolated sub-networks. We recommend that, in addition to protecting keystone nodes occurring on public land, conservation and management agencies work to secure the protection of those on private land via conservation agreements. Maintaining existing network connections should ensure a resilient and viable population over time.


## INTRODUCTION

Genetic network models can be used to gain an understanding of population structure and to quantify gene flow among populations in natural systems (Bunn *et al.* 2000; Dyer 2007; Dyer and Nason 2004; Connelly *et al.* 2011). Networks are constructed of components called nodes and edges, where nodes may represent populations and edges represent gene flow among populations (Sallaberry *et al.* 2013). Each node can be weighted by the genetic diversity within the local population and each edge by the genetic covariance among local populations (Bunn *et al.* 2000). Network models provide a means by which to rank the importance of each network component.

Emerging properties of genetic networks can inform us of the sensitivity of the entire network to disturbances (Dunne *et al.* 2002). There are four common network structures: (1) single-scale ("regular"), (2) broad-scale ("random"), (3) small-world, and (4) scale-free (Amaral *et al.* 2000; Bray 2003). Regular networks are highly structured such that proximal nodes tend to be linked to each other while distant nodes tend not to be linked. This structure is comparable to the isolation by distance pattern commonly discovered in the population genetics literature (Wright 1943). In a regular genetic network, gene flow occurs easily between neighboring nodes and nodes separated by a greater number of edges will be more isolated from one another. Nodes separated by a greater number of edges should be more isolated from one another. Random networks are unstructured such that spatial proximity of nodes is irrelevant to whether nodes are

connected or not and to irrelevant to the strength of connections. This structure is most similar to the theoretical island model first proposed by Wright (1931), and is analogous to the population genetic concept of panmixia. In a random genetic network, gene flow can occur unencumbered across the entire network because the number of steps between any two nodes is relatively small such that close and distant nodes have equal chances of being linked. Physically close and distant nodes have equal chances of being linked. Small-world networks are composed of few highly connected nodes and a greater number of more-isolated nodes, much like the hub-and-spoke model characteristic of the familiar commercial airline model. In the airline industry, hubs are strategically selected to maximize efficiency of air traffic, while spokes are selected based on limited need for services. Most nodes can be reached from every other node by a small number of steps, often routed through central hubs, which foster connectivity among the spokes. Redundancy is an important characteristic of small-world networks. In small-world genetic networks, gene flow is greatest among nearest neighbor nodes, but gene flow can occur between any two nodes by a small number of steps through hubs which are nodes with concentrated gene flow that serve to connect other distal nodes (also known as, spokes). An extreme form of small-world networks are scale-free networks. In scale-free networks, there is less redundancy in inter-node connections and greater centrality for the hubs. Within any network's structure, individual node importance to overall network connectivity can be measured by centrality indices. Therefore, these function-valued measures easily translate to rankings.

In transportation networks, hubs and spokes are known. For wildlife populations, where populations serve as hubs and where spokes are unknown. Qualifying genetic network structure and identifying nodes that act as hubs can be very informative to conservation and management of wildlife species (Garroway *et al.* 2008). Knowledge of which nodes are connected and which nodes rank highly in network centrality can facilitate management prioritization. Network models of wildlife populations have shown identifying hubs and spokes is not intuitive (Bunn *et al.* 2000; Garroway *et al.* 2008; Koen *et al.* 2015; Garroway *et al.* 2011). Populations at the periphery of a species' range can act as critical hubs, connecting populations across the network. Geographically central populations do not necessarily function as hubs (Bunn *et al.* 2000).

The lek mating system of sage grouse caters to network analyses because leks are relatively fixed spatial locations where males display in the spring to attract females and breed with them. Sage grouse disperse short distances from natal to neighboring leks (Dunn and Braun 1985). Following natal dispersal, they are largely philopatric to leks (Cross *et al.* 2017; Dalke *et al.* 1963; Dunn and Braun 1985; Emmons and Braun 1984; Wallestad and Schladweiler 1974). Occasionally, sage grouse disperse large distances during the breeding season (Cross *et al.* 2017). Therefore, we would expect that the resulting network structure would be composed of clustered, hub-like nodes characteristic of a small-world network (Garroway *et al.* 2008).

Prior research has modeled range-wide greater sage-grouse connectivity. Knick and Hanser (2011) weighted nodes using lek attendance and limited edge connections using hypothesized dispersal thresholds. Crist *et al.* (2017) used network approaches to generate several models of hypothesized connectivity among greater sage-grouse priority areas for conservation (PACs), which are areas that protect larger leks (i.e., those with more males visible during breeding) and surrounding area. They characterized the centrality of each PAC and concluded that several sub-networks exist across the species' range. Imposed dispersal thresholds may have affected the resultant network structure. For example, Knick and Hanser (2011) used an exponential decay function to determine the probability of connectivity of leks. Imposing dispersal thresholds likely oversimplified the contribution that each PAC made to network

connectivity by assuming dispersal limitations are equal among all nodes regardless of the internal population dynamics within nodes and environmental conditions within and among nodes. These prior studies generate valuable testable hypotheses about range-wide connectivity, but require empirical validation.

In this study, we had two primary objectives. First, we sought to determine the network structure of connectivity among leks, weighting edges by genetic divergence based on genetic covariance among leks. Second, we sought to identify which leks were important to maintaining overall population connectivity and persistence using measures of network centrality. Within this second objective, we also sought to identify keystone nodes, that is, nodes that act as more important to maintaining gene flow than their size or location within the species range alone would indicate.

## METHODS

### Study area and sampling

We used 16,420 spatially referenced greater sage-grouse feather and blood samples collected from 2,139 leks (mean of 7.68 samples per lek) across the entire contiguous range of the species in the United States of America and Canada from 2005–2014. Feather samples were collected from leks using non-invasive methods (Bush *et al.* 2005; Segelbacher 2002) after having been dropped or plucked by sage grouse during breeding activity, while blood samples were collected from sage grouse on leks as part of radio telemetry field research. Samples were collected by field biologists and technicians with the Bureau of Land Management, California Department of Fish & Game, Colorado Division of Wildlife, Idaho Fish & Game, Montana Fish, Wildlife & Parks, Montana Audubon, Nevada Division of Wildlife, North Dakota Game and Fish, Natural Resources Conservation Service: Sage-Grouse Initiative, Oregon Fish & Wildlife, South Dakota Game, Fish & Parks, Utah Wildlife Resources, U.S. Forest Service (USFS), Washington Department of Fish and Wildlife, and Wyoming Game & Fish Department. The only location throughout the entire distribution of the species that we did not use was Washington State because samples from this location were collected off-lek and during a different time period (from 1992 – 1999) than the rest of the samples.

### DNA extraction

Genetic analysis was conducted at two molecular biological laboratories: the National Genomics Center for Wildlife and Fish Conservation at the USFS Rocky Mountain Research Station (hereafter, NGC) and the Molecular Ecology Lab at the USGS Fort Collins Science Center (hereafter, FORT). Protocols were established at the inception of the study to ensure consistency among lab genotyping and are described below. Feather DNA was extracted from the quill (calamus) using QIAGEN's DNeasy Blood and Tissue Kit and the user developed protocol for purification of total DNA from nails, hair, or feathers. The protocol was modified by incubating samples for a minimum of 8 h after addition of Proteinase K and by eluting DNA with 100 µl of Buffer AE. Feather samples were extracted in a lab used only for non-invasive DNA extraction in order to avoid contamination from samples with higher DNA concentrations. Blood samples were extracted using QIAGEN's DNeasy Blood and Tissue Kit and the protocol for nucleated blood. At FORT, portions of the extraction process were automated using a QiaCube (QIAGEN).

### Microsatellite DNA amplification and genotyping

We amplified 15 variable microsatellite loci and one sex-diagnostic locus [CHD gene, using the primers 1237L and 1272H (Kahn *et al.* 1998)]. Specifics on PCR conditions used at NGC and FORT can be found in Cross *et al.* (2016; 2017) and Row *et al.* (2015).

To ensure correct genotyping from low-quality and low-quantity feather DNA samples, each sample was PCR-amplified twice across the 15 variable microsatellite loci to screen for allele dropout, stutter artifacts, and false alleles (DeWoody *et al.* 2006). To minimize genotyping error, two observers scored each sample. If any locus failed to amplify in either replicate, or if there was a discrepancy between locus genotypes as scored by the 2 observers, PCR-amplification and genotyping was repeated twice more. If a genotype was confirmed by this repeat analysis then it was retained. If a genotype failed again, the sample was assigned a missing score at the failed locus. To ensure consistency among laboratories, both labs genotyped the same 70 individuals. Each lab's genotypes for these individuals were compared, and necessary shifts to synchronize allele calls were made for all samples.

To screen samples for quality control, we removed from analysis any individual for which amplification failed at 1/3 of the loci (i.e. 5 loci). After removal of poor-quality samples, genotypes were screened to ensure consistency between allele length and length of the microsatellite repeat motif using MICROCHECKER v2.2.3 (Van Oosterhout *et al.* 2004). To identify and remove multiple captures of the same individual, and to screen for and correct genotyping error, we used DROPOUT 2.3 (McKelvey and Schwartz 2005 as implemented in Schwartz *et al.* 2006), MICROCHECKER v2.2.3 (Van Oosterhout *et al.* 2004), and package ALLELEMATCH 2.5 (Galpern *et al.* 2012) in R 3.3.0 (R Core Team 2016).

**Network construction**

A minimum node size of four or more individuals is required to calculate within-node genetic variation (Dyer 2014). Therefore, before constructing the network, we performed a hierarchical clustering analysis. First, we built a distance-based tree of the lek locations of the 6,723 individuals using geographic distance among leks (using the HCLUST function in base R), and clustered all leks within 15 km of one another (cut distance implemented using the CUTREE function in base R). We selected 15 km as the cut distance, as this is the best estimate of median breeding dispersal distance among leks for sage grouse (Cross *et al.* 2017). Second, we removed any clusters composed of fewer than four individuals. Following clustering, we constructed a weighted population network among the resulting clusters, which we henceforth refer to as nodes, following the methods of Dyer and Nason (2004), in the packages GSTUDIO (Dyer 2014) and POPGRAPH (Dyer & Nason 2004) in program R (R Core Team 2016).

We estimated genetic covariance, where microsatellite genetic covariance represents the weight of each network edge connecting nodes. These distances were calculated and stored as a genetic distance matrix and represented a saturated population network containing all possible pairwise edge connections. Because a saturated population network is biologically unlikely, we used POPGRAPH to identify and remove, or "prune," network edges that did not significantly contribute to the total among-population genetic covariance structure (Dyer & Nason 2004). This method was based on conditional independence calculated as edge deviance. Wherein each edge in the network was analogous to a predictor variable and the pruning of an edge was contingent upon its removal not causing a significant decrease in the fit of the model to the genetic data (edges among node pairs with a partial correlation coefficient of 0; Dyer & Nason 2004). Two nodes shared an edge if and only if there was significant genetic covariance between the two after removing the covariation each population had with all remaining populations in the data set (Dyer 2007). We pruned the network using the same methods and default pruning parameters

(alpha = 0.05 and tolerance = 1 x $10^{-4}$) as detailed by Garroway *et al.* (2008). Alpha is the significance at which edges were tested in the topology of the Markov network. Tolerance is the lower bound at which a principal component axis of the data was retained in the multivariate data analysis (raw genotypes are converted to multivariate data, and then a principal component rotation is conducted to reduce the number of dimensions). Only columns of the principal component rotation for which the standard deviation is greater than the tolerance setting were retained. A lower tolerance might result in homogeneity within groups in the discriminant rotation, whereas a higher tolerance might result in the loss of relevant information of value to analysis. Following pruning, the resultant minimal incidence matrix contained the smallest set of edges that sufficiently capture the among-node genetic covariance structure (Dyer & Nason 2004). When depicted, the resultant minimum edge set is also known as the minimum spanning tree (MST).

**Network structure determination**

In order to determine the network structure of the greater sage-grouse genetic network, we compared the degree distribution, clustering coefficient, and characteristic path length (CPL) to that of 1000 Erdos-Renyi model random networks with the same number of nodes, edges, and edge weight distribution as the range-wide sage grouse genetic network. The CPL is defined as the average shortest path length between all pairs of nodes in the network. It provides an understanding of how long it takes alleles to traverse the network. We generated the random networks using package IGRAPH (Csardi & Nepusz 2006) in program R, and tested for significant differences between the degree distribution, clustering coefficient, and CPL of the true sage grouse network and the random networks using permutation tests, following the methods of Garroway et al (2008). We used the results of these comparisons to determine whether network structure was purely a function of the number of nodes and edges or whether network structure was a result of non-random processes. For example, a CPL that is similar to the random networks and a clustering coefficient that is high relative to the random networks would indicate that the network has small-world or scale-free characteristics (Watts & Strogatz 1998). Whereas a degree distribution that does not follow the power law (scale-free), is not binomial (random) or fixed (regular), but is fat-tailed would narrow the likely network structure to that of a small-world network. Further evidence for a small-world network is a short diameter which is the longest of all the shortest paths among nodes (i.e., the shortest distance between the two most distant nodes), and which is determined by the order in which the populations are connected (Dyer 2007)

We measured pairwise conditional genetic distance (cGD) among all nodes. cGD is the length of the shortest path connecting each pair of populations conditioned on network structure (Dyer *et al.* 2010) or the relative strength of the genetic covariance between nodes along the connecting edges (Koen *et al.* 2013). When compared to geographic distance among nodes, cGD can provide insight into network structure.

To quantify network connectivity, we calculated each of the centrality measures detailed in Table 5.1, and used these metrics to quantify connectivity and relative isolation of each node. For example, betweenness centrality quantifies the measure of importance of a node or edge in terms of the bottleneck to gene flow it creates, eigenvector centrality quantifies how connected a node is, and strength quantifies how strong the connection is between a node and all its neighboring nodes (Garroway *et al.* 2008). Eigenvector centrality measures both how well a node is connected and how well a node's immediate connections are connected—in essence, measuring both direct and indirect connectivity. These properties make eigenvector a better

measure than betweenness if one is interested in quantifying the strength of connections as it factors in not only how well connected a given node is, but also how well connected the nodes are to which a given node is connected. When quantifying connectivity, we used the centrality measure of node strength rather than node degree, as Koen *et al.* (2015) found that strength more adequately depicts migration and gene flow than does degree. To examine relationships between measures of network centrality, we tested for pairwise correlation between all measures. We also tested for correlation between each measure of network centrality and sample size at each node. All measures of network centrality were calculated using the IGRAPH package in R, and all correlations were calculated in base R.

We screened all nodes within the top 1% of each measure of network centrality (Table 5.1), and within the top 50% of all node centrality measures combined in order to identify those nodes that were most important to local (regional) and global (range-wide) connectivity. These nodes represent the top hubs of genetic exchange that maintain connectivity at all scales.

**Keystone nodes**

We hypothesized that the mostly highly attended and most geographically central nodes would rank highest for centrality (i.e., abundance would be positively correlated with node centrality). We used the number of individuals genotyped per node as a surrogate for overall lek attendance to test for correlation with network centrality. The surrogate was supported by a positive correlation between the number of individual genotypes per lek and the overall high lek count from 2009 to 2014 in Montana (Wilcoxon rank sum test with continuity correction: $W = 657390$, $p < 2.2 \times 10^{-16}$).

We also tested for correlation between geographic centrality of a node and network centrality. We defined geographic centrality as the great-circle distance from the center of the network's geographic center. We calculated geographic centrality as the distance of each node from the centroid of a minimum convex polygon (MCP) enveloping all nodes. We calculated the MCP using the GCONVEXHULL function, calculated the centroid of the MCP using the GCENTROID function in the RGEOS package (Bivand 2017), and calculated the distance of every node from the centroid using the RDIST.EARTH function in the FIELDS package (Nychka *et al.* 2015) all in R.

Finally, we sought to identify nodes with greater importance to genetic connectivity than the magnitude of lek attendance within the node or node location within the species range alone might indicate. That is, we sought to identify nodes that were low in attendance or peripheral to the range, but that still ranked high in centrality. To identify these keystone nodes we plotted both the number of individuals genotyped per node and the geographic centrality of each node against each measure of network centrality.

## RESULTS

**Genotyping and network construction**

Our genotyping efforts resulted in 6,242 individual genotypes from 1,200 leks (median = 3 genotypes per lek, IQR = 4 genotypes per lek, range = 1–62 individuals) following duplicate removal and quality control. Hierarchical clustering and removal of nodes with fewer than four individuals ($n = 153$) yielded 5,950 samples in 459 nodes (median = 10 individuals per node, IQR = 9, range = 4–90 individuals). We constructed a network composed of 459 nodes connected by 14,481 edges. The longest most direct path among nodes in the network (the diameter) spanned just five nodes from the Platte River Basin of Wyoming through the Sevier Lake Basin of Utah and into the Central Nevada Desert Basin. The MST was largely structured

by geography (Fig. 5.1). Montana nodes compose a large group within the MST and are joined to the rest of the network through nodes in Wyoming and Idaho. Nodes in Colorado also group together and are linked by nodes in Wyoming and Utah. Throughout most of the network, connections among many hubs and spokes crossed state lines (Fig. 5.1 and Fig. 5.2). There was strong evidence for a positive correlation between cGD and geographic distance (Wilcoxon rank sum test with continuity correction: $W = 1.1 \times 10^{10}$, $p < 2.2 \times 10^{-16}$; Fig. 5.6 in Supplementary Material), but no discernable geographic pattern for nodes with greater cGD than expected due to geographic distance (the cloud with cGD > 40 in Fig. 5.6 in Supplementary Material).

**Network structure determination**

The sage grouse genetic network showed small-world characteristics. The network deviated from random network structure in both mean clustering coefficient and mean CPL. Both measures were significantly greater for the sage grouse genetic network than for the 1000 random networks with the same number of nodes, edges, and edge weight distribution as the sage grouse network. Furthermore, none of the random networks had a greater mean clustering coefficient or CPL than the greater sage-grouse network (CC: $p < 0.001$, CPL: $p < 0.001$). Both the mean clustering coefficient ($0.19 \pm 0.023$ [SD]) and the mean CPL (1.88) were shorter than has been documented in other species (e.g. 0.254 and 2.26 (Garroway *et al.* 2008). The fat-tailed distribution of node degree for the sage grouse genetic network (Fig. 5.8 in Supplementary Material) confirmed small-world network structure by ruling out scale-free structure, for which the degree distribution follows a power-law. Finally, the effect of node removal on average CPL was not small, as is expected for networks with regular structure where a large decrease in CPL is expected with the removal of each node.

**Node properties**

Nodes within the Upper Snake River Basin in Idaho and the Green River Basin in Wyoming rank highly across multiple measures of centrality indicating the importance of these regions to gene flow across the network. Collectively, these two basins contained the nodes with the maximum measures of centrality (Fig. 5.3, Table 5.2). Many other regions contained high-ranking nodes for one or more measures. Notably, the Bighorn River Basin in Wyoming and the Lower Missouri River Basin in Montana contain nodes that measure high for centrality indices indicative of their importance to local connectivity. The Powder/Tongue River Basin in Wyoming and the Southwest River Basin in Idaho contain nodes that measure high for centrality indices indicative of their importance to global connectivity (Fig. 5.3).

Only five nodes measured in the top $50^{th}$ percentile of all measures of network centrality, indicating importance to gene flow both locally and network-wide and the rarity of this combination of local and network-wide importance. The range of each centrality measure for these nodes was large (betweenness = 766–1314, closeness = $1.49 \times 10^{4}$–$1.55 \times 10^{4}$, clustering coefficient = 0.179–0.202, eigenvector: ≥ 0.556, and strength = 622.1–688.1). Three of the nodes identified were located within the Upper Snake River Basin of Idaho, and two were located in Wyoming: one in the Green River Basin and one in the Powder/Tongue River Basin (Fig. 5.3 and Table 5.2). The locations of these nodes were geographically central to the sampling extent and geographic range of the species.

Nodes with high betweenness centrality act as bridges between different parts of the network, so their loss can have network-wide impacts on gene flow (Garroway *et al.* 2008). We identified several nodes ranking highly in betweenness. Three of the top 1% of nodes ranked by betweenness centrality were located in Wyoming—two in the Powder/Tongue River Basin, and one in the Green River Basin. Of the remaining two nodes, one was located in the Upper Snake

River Basin of Idaho and one in the Bear River Basin of Utah (Fig. 5.3 and Table 5.2). Of these, the node with the greatest measure of betweenness (1521) was located just south of Grand Teton National Park in the Green River Basin. This same node also measured high for strength (691.9). Most nodes in the network were important to network-wide connectivity (right-skewed distribution: median = 92.0, IQR = 258.5, Fig. 5.7A in Supplementary Material; Table 5.3). However, seventeen nodes had a betweenness of zero, indicating relatively low importance to fostering gene flow across the network. These same nodes measured an average strength of 429.2 (± 218.3 [SD]), indicating strong connections to other nodes despite low importance to network-wide gene flow. Most of these nodes with the lowest measures of betweenness were located across the northeastern periphery of the species' range.

To identify nodes with the greatest genetic covariance with all other nodes in the network, we ranked nodes by closeness centrality. Closeness is a measure of the average shortest path between a node and all other nodes in the network. Hence, a smaller measure of closeness indicates shorter paths on average, and therefore, greater connectivity. The top-ranked closeness nodes were mostly located in the northeastern portion of the range and were geographically proximal. Two of the nodes in the top 1% of closeness centrality were located in the Lower Missouri River Basin of Montana, one in the Wind/Bighorn River Basin and one in the Powder/Tongue River Basin in Wyoming, and one in the Central Nevada Desert (Fig. 5.3 and Table 5.2). The node with the lowest closeness ($4.88 \times 10^{-5}$) was located in the far eastern Central Nevada Desert, and had low betweenness (0). The node with the greatest closeness in the network ($1.59 \times 10^{-4}$) was located just south of Grand Teton National Park in the Green River Basin of Wyoming, and had a relatively high betweenness (1521) and strength (691.9). There were a small number of very closely co-varying leks in the network (left skewed distribution: median = $1.37 \times 10^{-4}$, IQR: $1.76 \times 10^{-5}$; Fig. 5.7B in Supplementary Material and Table 5.3).

To identify the most highly networked nodes, we examined node rankings by eigenvector centrality. Eigenvector centrality increases for nodes that are highly connected to other highly connected nodes. All but one of the top one percent of nodes ranked by eigenvector centrality was all located in the Great Basin, indicating increased genetic connectivity. These nodes are located in the Oregon Closed Lakes Basin, one in the Malheur River Basin, one in the South Lahontan River Basin, one in the Central Nevada Desert Basin (Fig. 5.3 and Table 5.2). The exception was a single node outside the Great Basin in the Southwest Basin of Idaho. The node with the greatest eigenvector centrality (1.00) was located in the Southwest River Basin, and had very high strength (1104.8), but very low betweenness centrality (4). The node with the lowest eigenvector centrality (0.064) was located in the Sevier Lake Basin of Utah, and was a terminal-node on the MST (these terminal-nodes on the MST have only one edge, and are referred to as leaves). This same leaf also measured very low for strength (88.1) and betweenness (82), low centrality both locally and network-wide. Eigenvector centrality was normally distributed (mean = 0.54 ± 0.17 [SD]; Fig. 5.7E in Supplementary Material; Table 5.3).

To identify nodes that anchor tightly knit groups connected by a high number of edges, we examined node rankings by clustering coefficient. Increased clustering coefficient is indicative of small-world characteristics. The nodes in the top 1% of clustering coefficient were found across the species' range but mostly toward its periphery (Fig. 5.3 and Table 5.2). The southernmost node in the top 1% was in the Sevier Lake Basin of Utah. Other nodes in the top 1% of centrality were located in the Southwest River Basin of Idaho, the Bighorn River Basin of Wyoming, and the Lower Missouri River Basin of Montana ($n = 2$). The node with the greatest clustering coefficient (0.35) was found in the Southwest River Basin. This node also had low

betweenness (0), strength (191.4), and eigenvector centrality (0.12). The node with the lowest clustering coefficient (0.13) was found in the found in the Bighorn River Basin of north-central Wyoming, and had low betweenness (8), strength (356.3), and eigenvector centrality (0.27). Network-wide, there was a low chance that any two nodes connected to a given node were also connected to one another (right-skewed distribution: mean = 0.19 ± 0.023 [SD], median = 0.18; Fig. 5.7C in Supplementary Material; Table 5.3).

To determine which nodes co-varied closely with many other nodes, we measured the strength of each node. The top 1% of nodes ranked by strength were located in both the central and southern portion of the species' range, indicating increased genetic covariance among a greater number of nodes (Fig. 5.3 and Table 5.2). Of the nodes with the greatest strength, one node was in the South Lahontan River Basin of California, one in the Central Nevada Desert Basin, one in the Southwest River Basin of Idaho, one in the Weber River Basin of Utah, and one in the Green River Basin of Wyoming. All five of the top 1% of nodes by strength were leaves on the MST. The nodes with the greatest and least strength (1105.0 and 88.1), located in the Southwest River Basin of Idaho and the Sevier Lake Basin of Utah, were also similarly ranked for eigenvector centrality (two measures for which there was very strong evidence for a positive correlation ($\rho = 0.98$, $p < 2.2 \times 10^{-16}$). The distribution of node strength was normally distributed (median = 627.1, IQR: 77.3; Fig. 5.7F in Supplementary Material; Table 5.3).

We found evidence for positive correlation between betweenness and closeness centrality ($r_s = 0.9137852$, $p < 2.2 \times 10^{-16}$) and eigenvector centrality and node strength ($r_s = 0.9829317$, $p < 2.2 \times 10^{-16}$).

**Edge properties**

Edge weight is a measure of the magnitude of covariance between connected nodes and can be used to identify nodes most closely linked by gene flow. The strongest genetic covariance existed in the Central Nevada Desert basin. From here the top 0.1% of edges ($n = 15$) all radiate from a single node and connect to nodes across Nevada ($n = 1$), Idaho ($n = 1$), Montana ($n = 6$), Wyoming ($n = 3$), Colorado ($n = 1$) and Utah ($n = 3$). The single edge with the greatest genetic covariance connected a node in the Central Nevada Desert Basin and a node in the Northeast Wyoming River Basin in Wyoming (edge weight = 35.61). This edge linked nodes with the lowest closeness centrality. Both of these nodes also had low eigenvector centrality (0.31 and 0.25) and very low betweenness (0). The edge of least weight connected two nodes within the Dirty Devil River Basin of Utah (in the south central UT group of nodes in Fig. 5.2). These nodes were of moderate importance to network-wide connectivity (betweenness: 119 and 144), but had low connectivity to other nearby nodes (eigenvector centrality: 0.10 and 0.13). Overall, gene flow among nodes has led to increased network connectivity, with the occurrence of some highly connected nodes evidenced by a skewed right distribution of edge weight (9.82 ± 2.23 [SD], median = 9.68, IQR = 2.58); Fig. 5.7G in Supplementary Material; and Table 5.3).

**Keystone nodes**

It was common that the nodes with the highest centrality rankings were also those with lower numbers of individuals per node (Fig. 5.4). However, we found strong evidence for a positive correlation between the number of individuals sampled per node (a proxy for individual abundance) and betweenness, closeness, and clustering coefficient (betweenness: $r_s = 0.66$, $p < 2.2 \times 10^{-16}$, Fig. 5.4A; closeness: $r_s = 0.54$, $p < 2.2 \times 10^{-16}$, Fig. 5.4B; clustering coefficient: $r_s = 0.34$, $p = 4.37 \times 10^{-14}$, Fig. 5.4C). We found a negative correlation between the number of individuals sampled per node and both eigenvector centrality and strength (eigenvector: $r_s = -0.57$, $p < 2.2 \times 10^{-16}$, Fig. 5.4D; strength: $r_s = -0.61$, $p < 2.2 \times 10^{-16}$, Fig. 5.4E).

Across all centrality measures, we discovered 35 nodes that ranked high for network centrality despite having fewer individuals than nodes of similar ranking (Fig. 5.5). These 35 nodes were located across the entire species' range. Twenty-one of these keystone nodes ranked highly for more than one measure of centrality, with high rankings coupled for eigenvector centrality and strength and for betweenness and clustering coefficient. In all cases, these nodes were keystone for betweenness and closeness or for eigenvector and strength (measures for which we found evidence of correlation).

We also found some evidence for correlation between geographic centrality and three of the measures of network centrality: closeness ($r_s$ = -0.013, $p$ = 7.06 x $10^{-3}$), eigenvector centrality ($r_s$ = -0.22, $p$ = 2.4 x $10^{-6}$), and strength ($r_s$ = -0.16, $p$ = 6.4 x $10^{-4}$). However, there were no discernable patterns in the distance of nodes from the geographic center of the network and measure of network centrality.

There was no evidence for correlation between geographic centrality and number of individuals per node ($r_s$ = -0.04, $p$ = 0.35).

## DISCUSSION

### Emergent network properties

We discovered that the greater sage-grouse range-wide genetic network most resembles the structure of a robust small-world network. Many hubs of connectivity within the network are located across the species' range, with most leaves located along the periphery of the range. This hub-and-spoke topology is evident in the MST Fig. 5.2, with important hubs of genetic connectivity occurring in every state across the contiguous range, except North Dakota. Loss of one of these highly connected hubs within several major basins could sever overall network connectivity.

We documented strong connectivity across the entire network. This means that some of the nodes may be able to recover should they be extirpated but the habitat remain intact or be restored, due to the network's traversability (i.e., the apparent low resistance to gene flow). The MST revealed patterns of gene flow and can serve as a powerful guide in making management decisions related to the relative importance of individual nodes to overall landscape connectivity (Urban & Keitt 2001), as it is possible to model which nodes or which portions of the range will most likely be affected by the loss of any given node.

Our results suggest that distance plays an important role in structuring gene flow (isolation by distance, Wright 1943). The vast majority of edges in the MST connect geographically proximal nodes (Fig. 5.1). Similarly, there is a correlation between cGD and geographic distance (Fig. 5.6 in Supplementary Material). These results support prior findings of isolation by distance within smaller portions of the species' range (Cross *et al.* 2016; Bush *et al.* 2011; Davis *et al.* 2015; Oyler-McCance *et al.* 2005; Schulwitz *et al.* 2014). Elevated eigenvector centrality, closeness centrality, and strength were identified toward the geographic center of the range (Fig. 5.3) showing this region is important to both local and network-wide genetic exchange.

Cross *et al.* (2016) found that southwestern Montana is disconnected from the rest of the northeastern range, and is connected to the south into Idaho. The pattern of connectivity shown by our MST supports this finding because all edges emanating from southwestern Montana connect to nodes in Idaho (Fig. 5.2). Cross *et al.* (2016) also found that the population in Northern Montana was highly intra-connected, but was diverged from the subpopulation in Southeastern Montana and the Dakotas, and from the south-central Montana subpopulation (the

SE-W subpopulation in Cross *et al.* 2016). We confirm these findings here, showing nodes with very high clustering coefficient (indicative of highly interconnected network subunits) within the same regions (Fig. 5.2). We expect that the other top ranked nodes for clustering coefficient in the Southwest River Basin of Idaho, and the Sever Lake Basin in Utah might also be embedded at the core of subpopulations. Schulwitz *et al.* (2014) found that the subpopulations in southern and southeastern Montana and the Dakotas were both highly connected to leks in northern Wyoming. This pattern of connectivity is evident in the hub and spoke topology of the MST, where we a hub for Wyoming/south-central Montana is located in the Yellowstone River Basin of south-central Montana, and a hub for Wyoming/south-eastern Montana subpopulations is located within the Powder River Basin of Wyoming (Fig. 5.2). Davis *et al.* (2015) found that the small northern California population known to have experienced population declines had retained genetic diversity. We confirm this understanding by finding that the leks in this area show elevated local connectivity (covariance) within the area. We also found that gene flow into the northern California leks comes from leks to the north in Oregon (Fig. 5.2). Oyler-McCance *et al.* (2014) discovered a northern and a southern subpopulation within the bi-state population in southern California. We found the same break evidenced by a lack of edges connecting these two units in the MST (Fig. 5.2). This lack of inter-connectivity among nodes in the northern and southern groups is especially surprising, given that both groups exhibit greater covariance with far more geographically distant nodes (Fig. 5.2).

**Hubs of genetic exchange**

We identified 35 keystone nodes across the range of sage grouse that stand out with increased importance to genetic connectivity despite being composed of fewer individuals. These keystone nodes do not follow the trend of increased centrality with increased numbers of individuals per node (i.e., a proxy for population size for any given node), and include the highest-ranking nodes for each measure of centrality, regardless of the number of individuals. We believe that these keystone nodes and other nodes ranking high in centrality are top candidates for targeted conservation efforts, as their protection will likely help secure the hubs of range-wide genetic connectivity. The keystone nodes tend to be closer to the center of the range. However, the weak support for a correlation between measures of centrality and distance from the geographic center of the range indicates that while the nodes with the greatest closeness, eigenvector centrality, and strength tend to be located closer to the center of the range, a great deal of variation in this pattern exists. Therefore, neither geographic centrality nor number of individuals sampled alone should be trusted proxies for targeting conservation.

In addition to keystone nodes, we also documented nodes with high importance to large-scale, network-wide genetic connectivity (i.e., nodes with high betweenness; Fig. 5.2), and nodes within the top 50% of all centrality measures important to both network-wide and local connectivity. These top-ranked hubs are located in Idaho, Utah, and Wyoming, toward the geographic center of the species' range. The locations of these hubs are in areas that should foster range-wide genetic connectivity due to their location in the topography of the western landscape. The Upper Snake Basin of Idaho forms a thumb terminating in Southwestern Montana to the northeast, and opening up over the Columbia Plateau into the Great Basin to the south. The Green River Basin of Wyoming sits just west of a low section of the North American Continental Divide connecting the Upper Snake Basin and Great Basin to the rest of Wyoming and farther up into the northeastern portion of the species range. The Green River Basin also sits just north of the Yampa/White River Basin in Colorado and the Bear River Basin in Utah, both connected by low valleys to the south. Similarly, the Powder/Tongue River Basin of Wyoming

connects to the north with both of the Dakotas and eastern Montana, while the Bighorn River Basin, which ranks lower for other connectivity measures, connects to the southeastern-west subpopulation in the Yellowstone River Basin of Montana (Cross *et al.* 2016). Finally, the Southwest River Basin in Idaho connects to the Malheur and Oregon Closed Lakes Basins to the west, and to the South Lahontan River Basin by way of the Central Nevada Desert to the southwest. We suspect that the topology of the genetic network is largely shaped by the topology of the landscape, a hypothesis previously posited for sage grouse (Schulwitz *et al.* 2014; Cross *et al.* 2016), and which has been found to influence genetic structure in other species (e.g., Roffler *et al.* 2014).

**Isolated populations**

We found that some of the nodes with the lowest measures of betweenness were located across the northeastern portion of the species' range (Fig. 5.3). This pattern may indicate that the northeastern portion of the range does not play as large a role in range-wide genetic connectivity and may be more isolated, or at least, genetic exchange with the rest of the range is not as great here as it is toward the central portion of the range. While these northeastern nodes may not foster gene flow across the network, they may still be essential to generating genetic diversity. Population counts are high in this region, and there are high betweenness hubs (top 5%) located in every part of the range that function to connect spokes back into the rest of the network. The top 10% of nodes by eigenvector centrality ($n$ = 49) are mostly in the central and southern portion of the range, with only two occurring in Montana, and none in the Dakotas. Confirming these patterns of connectivity with an evaluation of high-resolution, range-wide population structure and genetic diversity would be prudent to comprehending genetic connectivity.

**Limitations of the study**

Our analysis revealed connectivity and centrality measures determined by the species biology alone. In the analysis by Crist *et al.* (2017)—who determined the network of connectivity among PACs—patch size, shape, and boundary length all had an effect on the pattern of connectivity and centrality. Our analysis is based on genetic covariance, or the genetic network resulting from cumulative dispersal and breeding, a quantitative metric. We have confidence in the cut distance we used to cluster leks into nodes, as it is empirically based on the dispersal distances that the species is capable of (Cross *et al.* 2017). Our clustering approach increased genetic variance within nodes, but also increased covariance among nodes. Choice of cut distance depends on the desired scale of analysis for conservation and management application. While we could have performed this analysis using individual leks, that would have resulted in fewer individuals per node, and we would have to cut many leks altogether due to having fewer than four individuals. By clustering leks into nodes, we are able to make statements about the connectivity of landscapes that exceed the size of an individual lek, and that are representative of leks unsampled within the same landscape.

We found evidence for correlation among measures of network centrality and between node size and centrality measures. Furthermore, the correlation between eigenvector centrality and node strength was also not surprising, as nodes score higher for eigenvector centrality when connected to other highly connected nodes. We also found evidence for a correlation between node size (the numbers of samples within each node) and many of the centrality measures; these correlations switched in sign between measures and were never greater than $r_s$ = 0.66 (range: -0.61 to 0.66 or |0.34–0.66|). There was strong evidence for a negative correlation between strength and the number of individuals per node ($r_s$ = -0.31, $p < 2.2 \times 10^{-16}$), perhaps because when more individuals are clustered at a node, the genetic covariance within the node is greater

than that between that node and others in the network (Dyer & Nason 2004). Some of this correlation could be the result of greater propagule pressure and thus connectivity originating from larger populations, reflected in the number of samples genotyped per node. In this case, larger populations acting as hubs should be expected, as these highly populated hubs should house greater genetic diversity and could be the sources of dispersers. However, as discussed above, the highest-ranking nodes for each measure of centrality were never those with the largest sample size.

**Future directions**

We believe that the greatest utility of our network analysis will be its use in prioritizing and targeting conservation efforts to the nodes most important to maintaining network connectivity. This network approach allows for the ranking of nodes by multiple measures of centrality, indicative of different scales and different patterns of connectivity. Future work may want to examine networks at multiple scales. For example, constructing a genetic network where PACs serve as nodes may help prioritize conservation based on existing management boundaries at a landscape scale. It is worth noting that if PACs are treated as nodes, larger PACs may score higher for measures of centrality due to the proportion of the genetic covariance (and therefore centrality) contained within. Our hope is that the empirically based greater sage-grouse genetic network we constructed will prove a useful tool to conservation planners.

**Table 5.1.** Network parameters used to quantify connectivity, the unit for which each is calculated, and the definition of the parameter, and relation of the parameter as pertains to the greater sage-grouse population network. All but characteristic path length and weight are measures of centrality

| **Network Parameter** | **Network Unit** | **Definition** [source] |
|---|---|---|
| Characteristic Path Length | Network | The mean of all pairwise network distances connecting nodes[2] |
| Closeness Centrality | Node | The average shortest path between node and all other nodes (connected network) |
| Clustering Coefficient | Node | The probability that two nodes connected to a given node are also connected (ranges from 0–1)[2] |
| Degree Centrality | Node | The number of edges connected to a node[2] |
| Eigenvector Centrality | Node | The direct and indirect connectivity: per node and immediate neighbors[1] |
| Betweenness Centrality | Node | The number of shortest paths that a particular node lies on[2] |
| Strength | Node | The sum of all edge weights[2] |
| Weight | Edge | The magnitude of covariance between connected nodes |

[1]Garroway *et al.* 2008, [2]Newman 2006

**Table 5.2.** Major basin location of the maximum node (dark green), top 1% of nodes (light green), and minimum node (red) for each network centrality measure and of nodes in the top 50% (yellow) of all network centrality measures in the greater sage-grouse genetic network. A summary of what the centrality pattern indicates about the particular basin's importance to the overall network is provided (Summary). In cases where both the top 1% and maximum are shown for the same basin, these rankings are both for a single node. Where multiple ranked nodes are located in the same basin, these nodes are identified by split cells within the same column

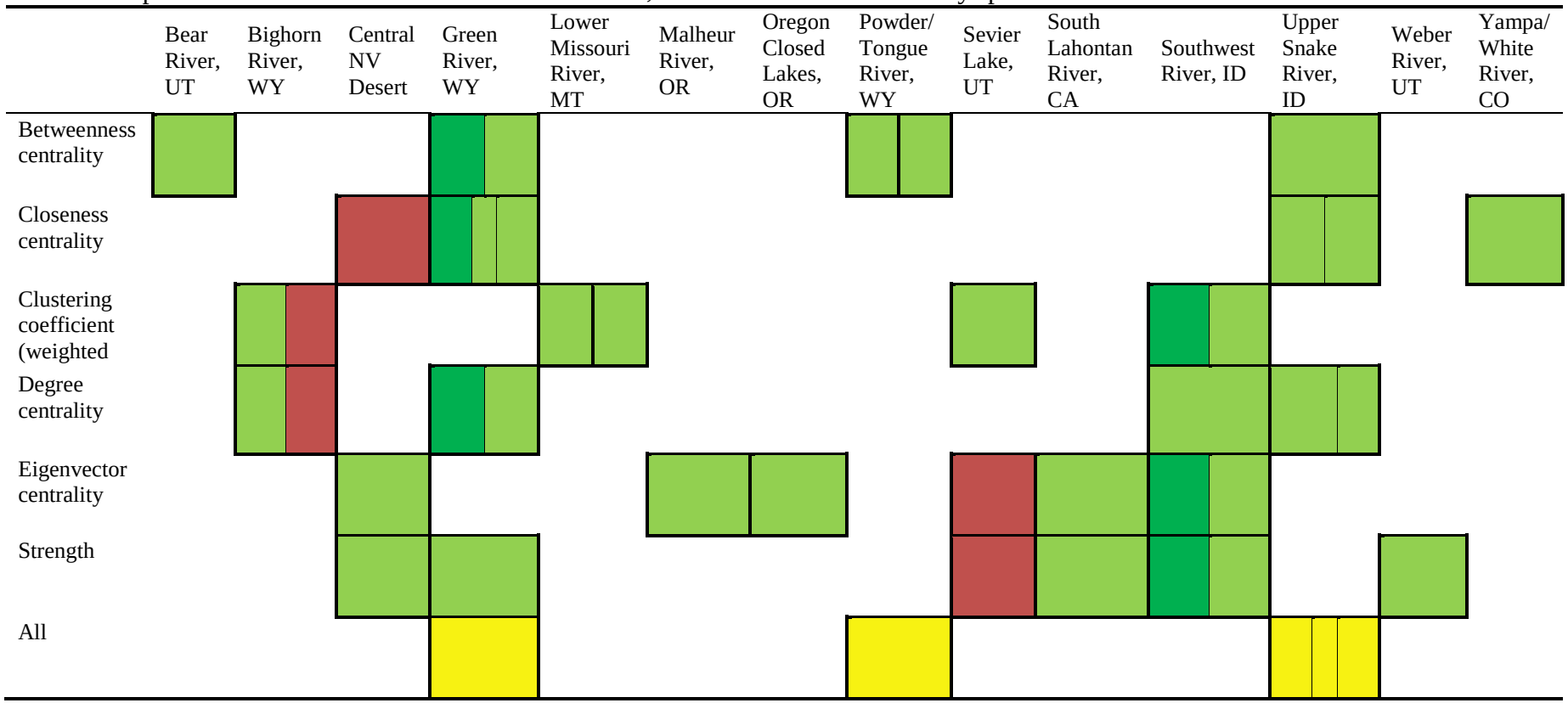


**Table 5.3.** Network parameters of centrality and connectivity for the range-wide greater sage-grouse genetic network, the network unit upon which each was measured, and a four-number summary of each

| Network Parameter | Network Unit | Min | Median | Mean ± SD | Max |
|---|---|---|---|---|---|
| Betweenness centrality | Node | 0.00 | 92.00 | 204.20 ± 251.11 | 1521.00 |
| Closeness centrality | Node | $4.88 \times 10^{-5}$ | $1.37 \times 10^{-4}$ | $1.34 \times 10^{-4} \pm 1.51 \times 10^{-5}$ | $1.59 \times 10^{-4}$ |
| Clustering coefficient | Node | 0.13 | 0.18 | 0.19 ± 0.023 | 0.36 |
| Degree centrality | Node | 9.00 | 66.00 | 63.1 ± 17.04 | 100.00 |
| Eigenvector centrality | Node | 0.06 | 0.56 | 0.54 ± 0.17 | 1.00 |
| Strength | Node | 88.10 | 627.10 | 619.4 ± 181.51 | 1105.00 |
| Weight | Edge | 3.02 | 9.681 | 9.82 ± 2.23 | 35.61 |

**Figure 5.1.** Fruchterman-Reingold plot (layout with minimal edge overlap) of the range-wide greater sage-grouse genetic network minimum spanning tree (MST). The network is pruned such that only the most highly weighted edges are shown between all nodes (i.e., the strongest genetic connections). Node ($n$ = 459) color indicates geographic location by state

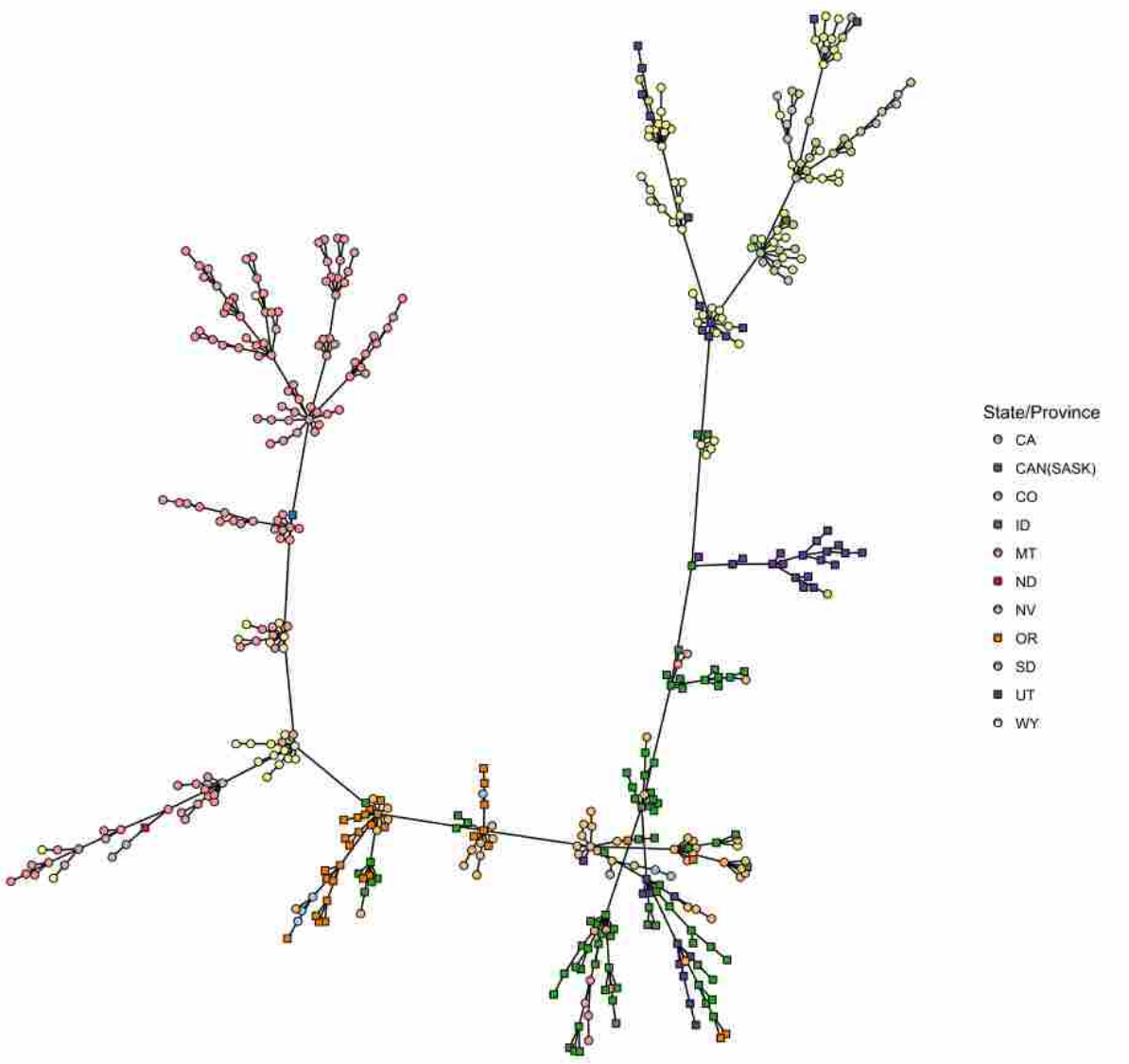


**Figure 5.2.** Map of the range-wide greater sage-grouse genetic network nodes (*n* = 459) connected by edges retained within the minimum spanning tree (i.e., the network is pruned such that only the most heavily weighted edges—those with the strongest genetic connections—are left connecting nodes. Node color indicates geographic location by state

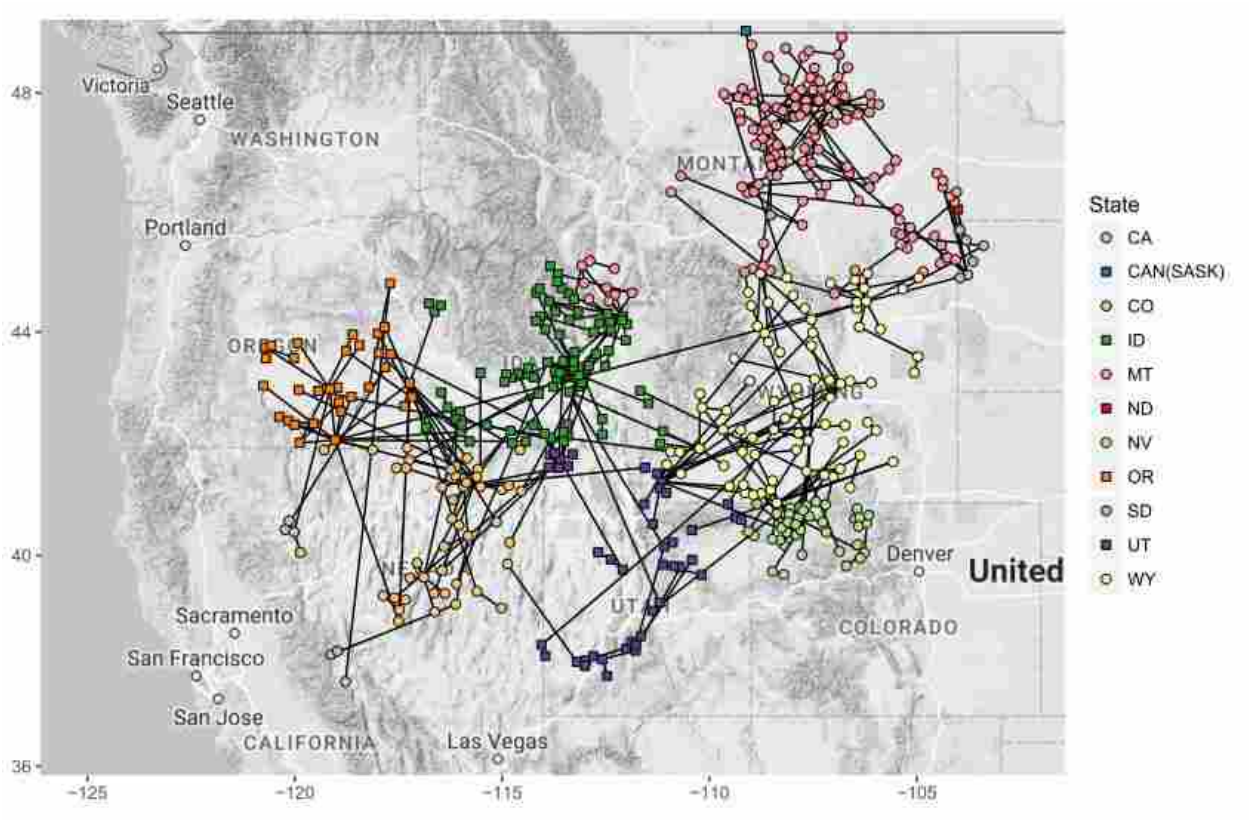


**Figure 5.3.** The top 1% ranking nodes ($n$ = 25) in each of the six centrality measures. Nodes in the top 1% of more than one measure are offset to the right, such that touching points represent the same node

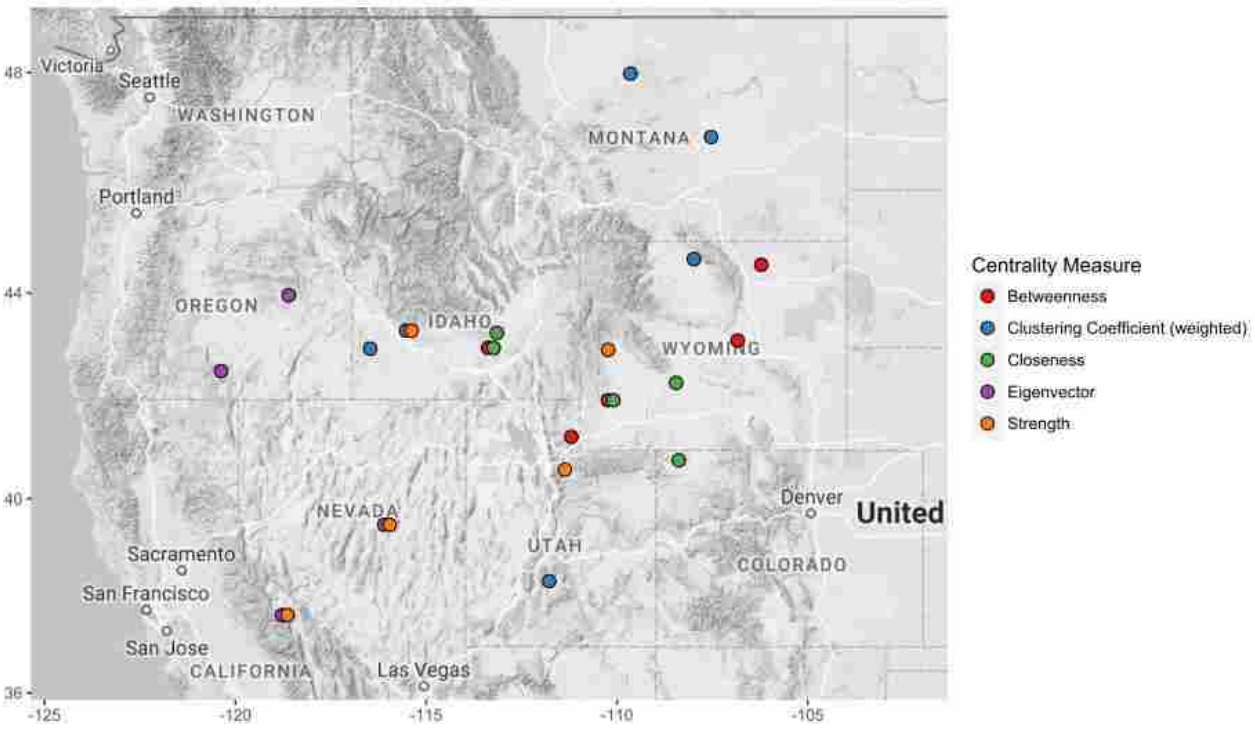
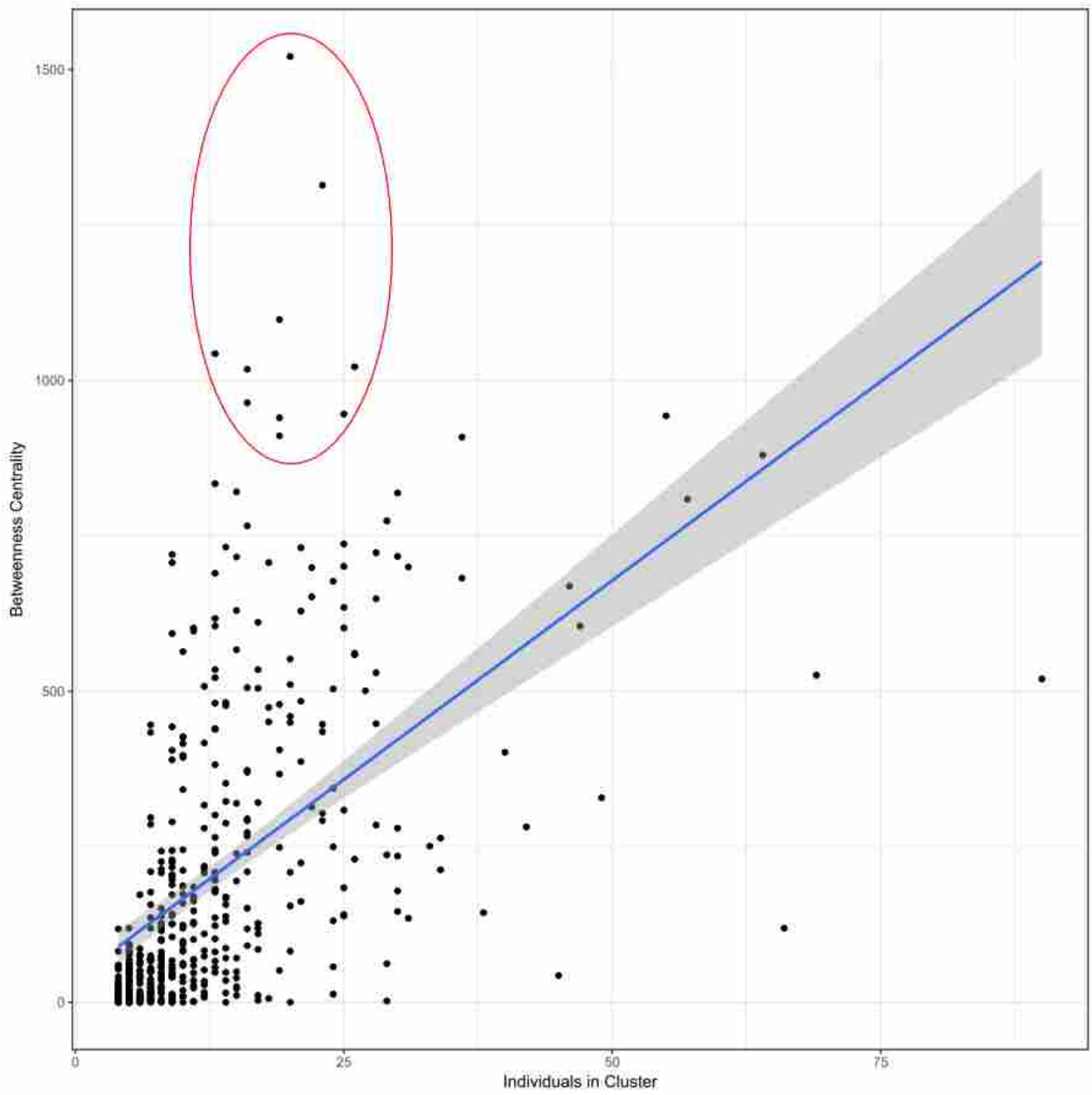
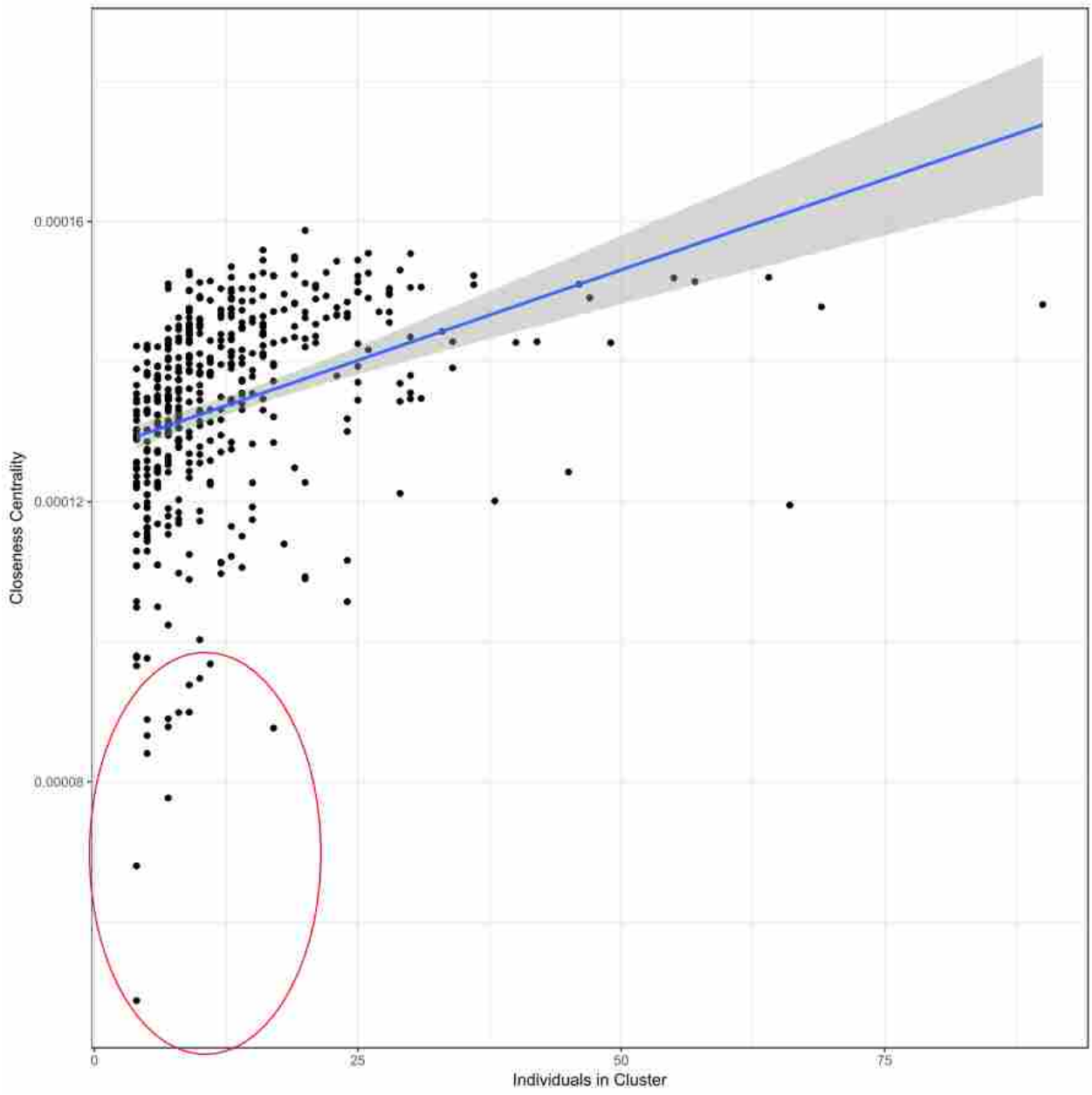


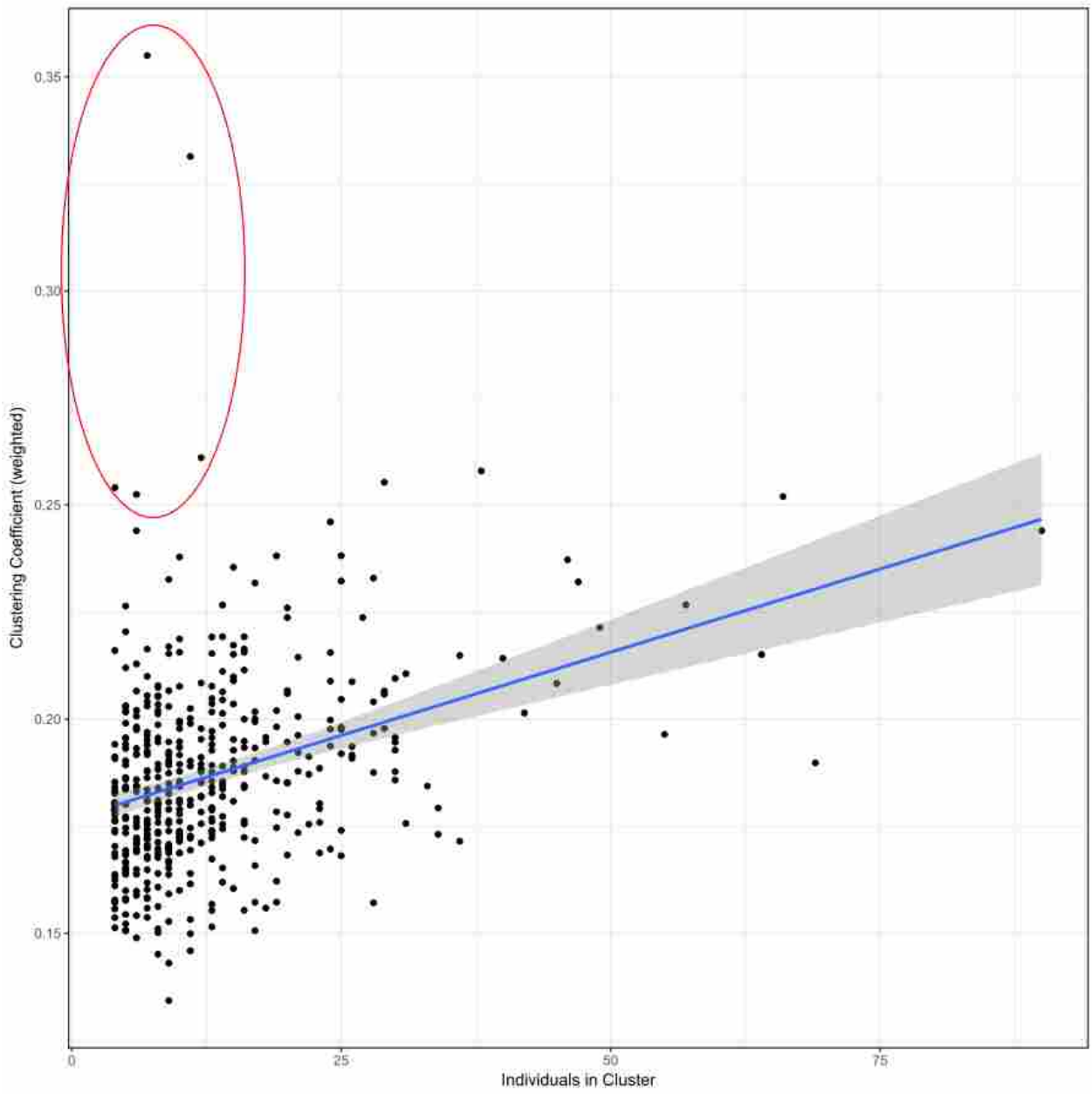
**Figure 5.4.** Relationship between the number of individuals in each node ($n$ = 459) and that node's measure centrality. Red circles show keystone nodes. The fitted linear model and confidence interval are shown (blue line with shaded CI). There was strong evidence for correlation between the two measures [betweenness: $\rho = 0.66$, $S = 5.43 \times 10^{6}$, $p < 2.2 \times 10^{-16}$ **(A)**; closeness: $\rho = 0.54$, $S = 7.35 \times 106$, $p < 2.2 \times 10^{-16}$ **(B)**; weighted clustering coefficient: $\rho = 0.34$, $S = 1.06 \times 10^{7}$, $p = 4.37 \times 10^{-14}$ **(C)**; eigenvector centrality: $\rho = -0.57$, $S = 2.53 \times 10^{7}$, $p < 2.2 \times 10^{-16}$ **(D)**; strength: $\rho = -0.61$, $S = 2.60 \times 10^{7}$, $p < 2.2 \times 10^{-16}$ **(E)**]



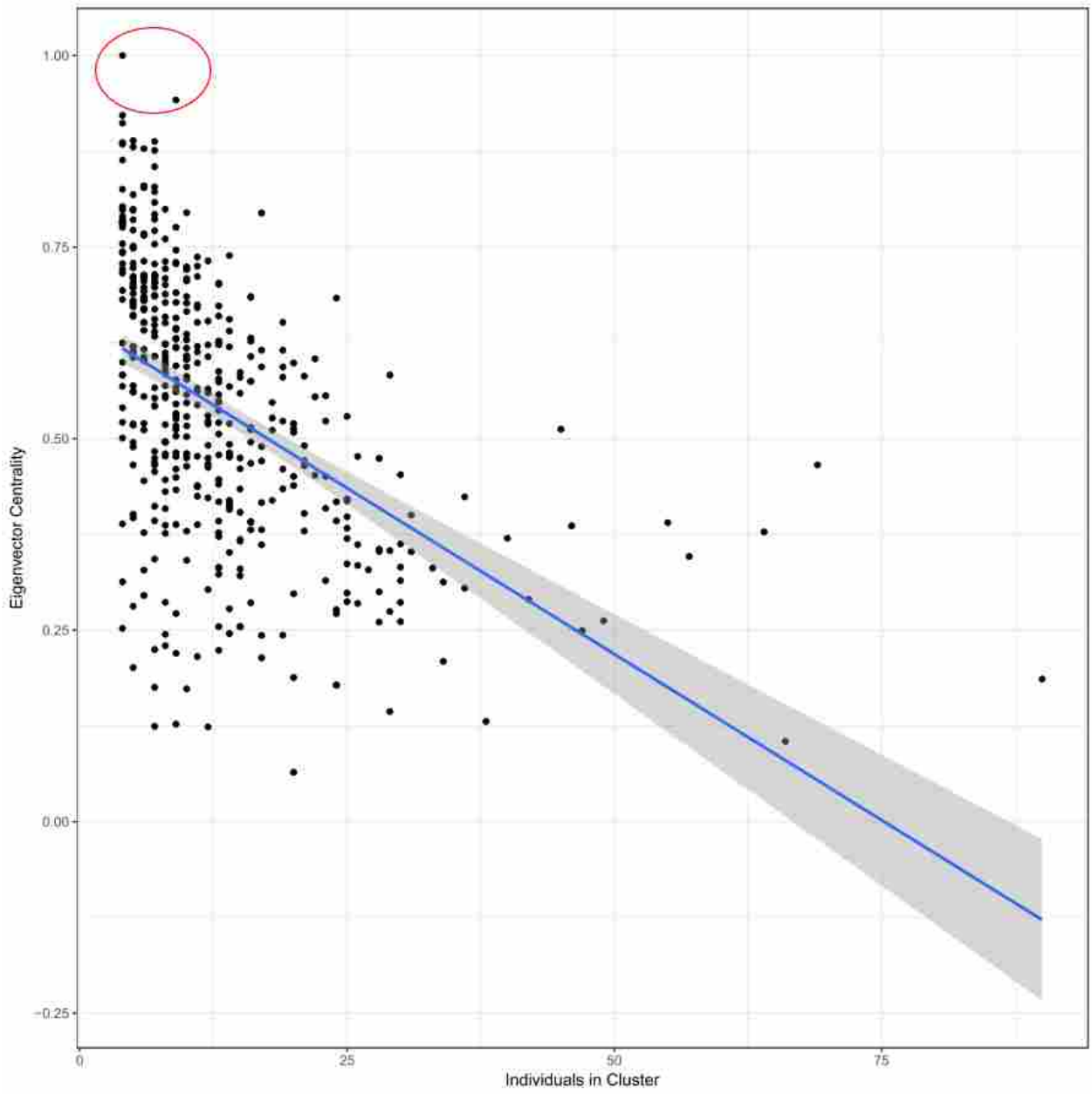

**(A)**


0.00016
0.00012
0.00008
Closeness Centrality
0
25
50
75
Individuals in Cluster

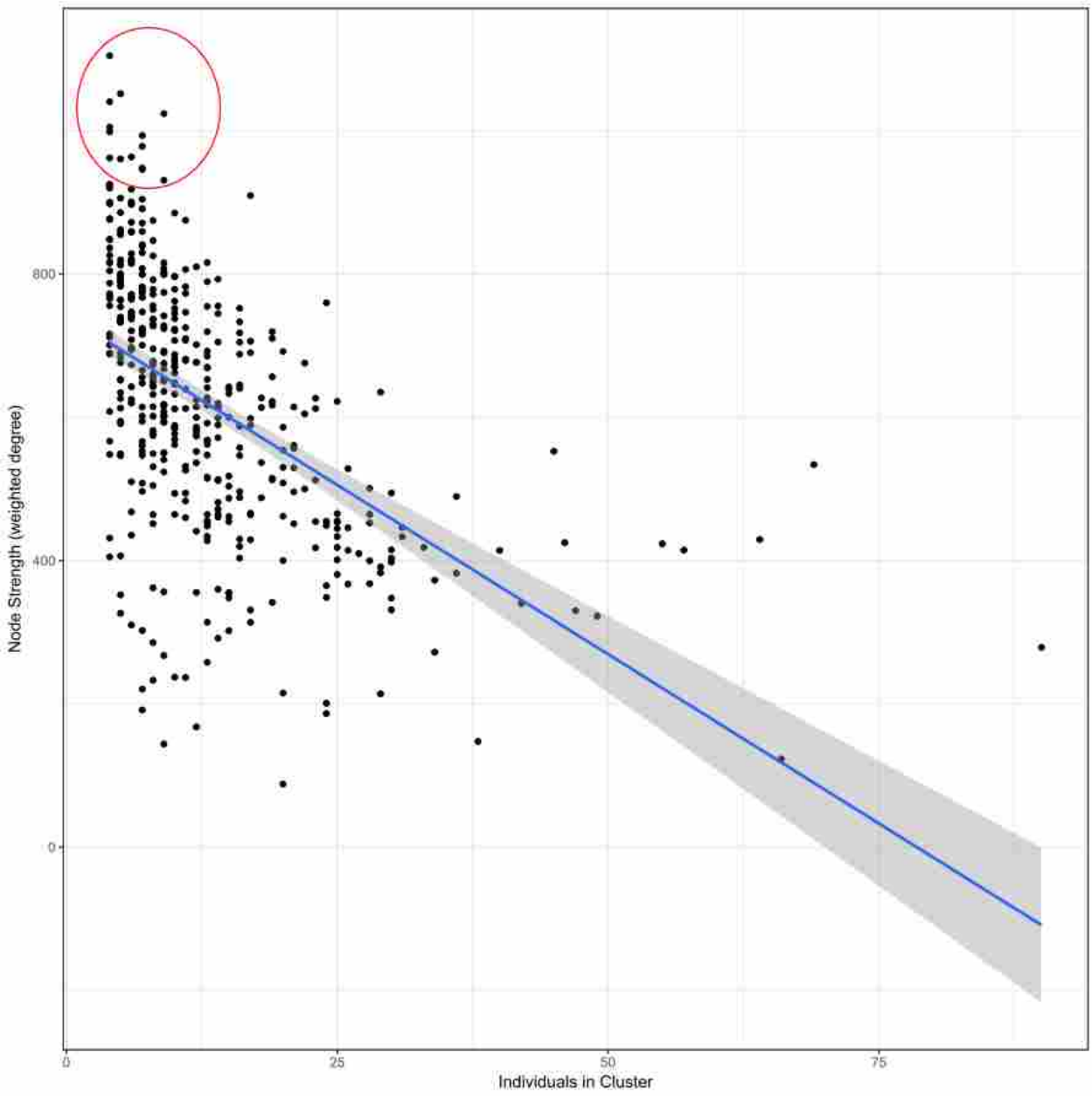
**(B)**


Clustering Coefficient (weighted)
0.35
0.30
0.25
0.20
0.15
0
25
50
75
Individuals in Cluster

**(C)**

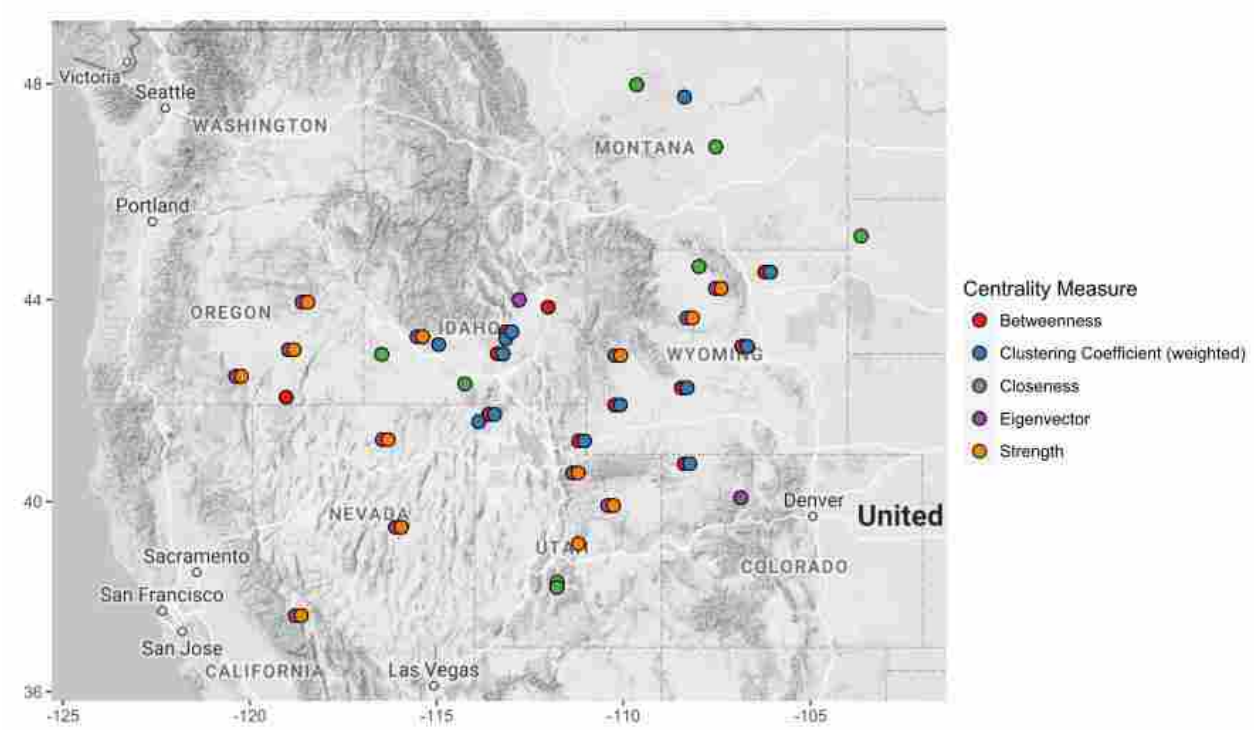

1.00
0.75
0.50
0.25
0.00
−0.25
Eigenvector Centrality
0
25
50
75
Individuals in Cluster

**(D)**


Node Strength (weighted degree)
800
400
0
0
25
50
75
Individuals in Cluster

**(E)**

**Figure 5.5.** Keystone nodes (*n* = 37): nodes with greater importance to genetic connectivity than the magnitude of lek attendance within the node or node location within the species range alone might indicate. These nodes were low in attendance relative to their centrality rankings. Points representing keystone nodes of more than one measure are offset to the right, such that these offset touching points represent the same node



**SUPPLEMENTARY MATERIAL**

**Figure 5.6.** Comparison plot of conditional genetic distance (the length of the shortest path connecting pairs of populations conditioned on network structure) and physical distance, measured as great circle geographic distance. There was very strong evidence for a positive correlation between conditional genetic distance and geographic distance ($\rho = 0.33$, $S = 1.29 \times 10^{14}$, $p < 2.2 \times 10^{-16}$)

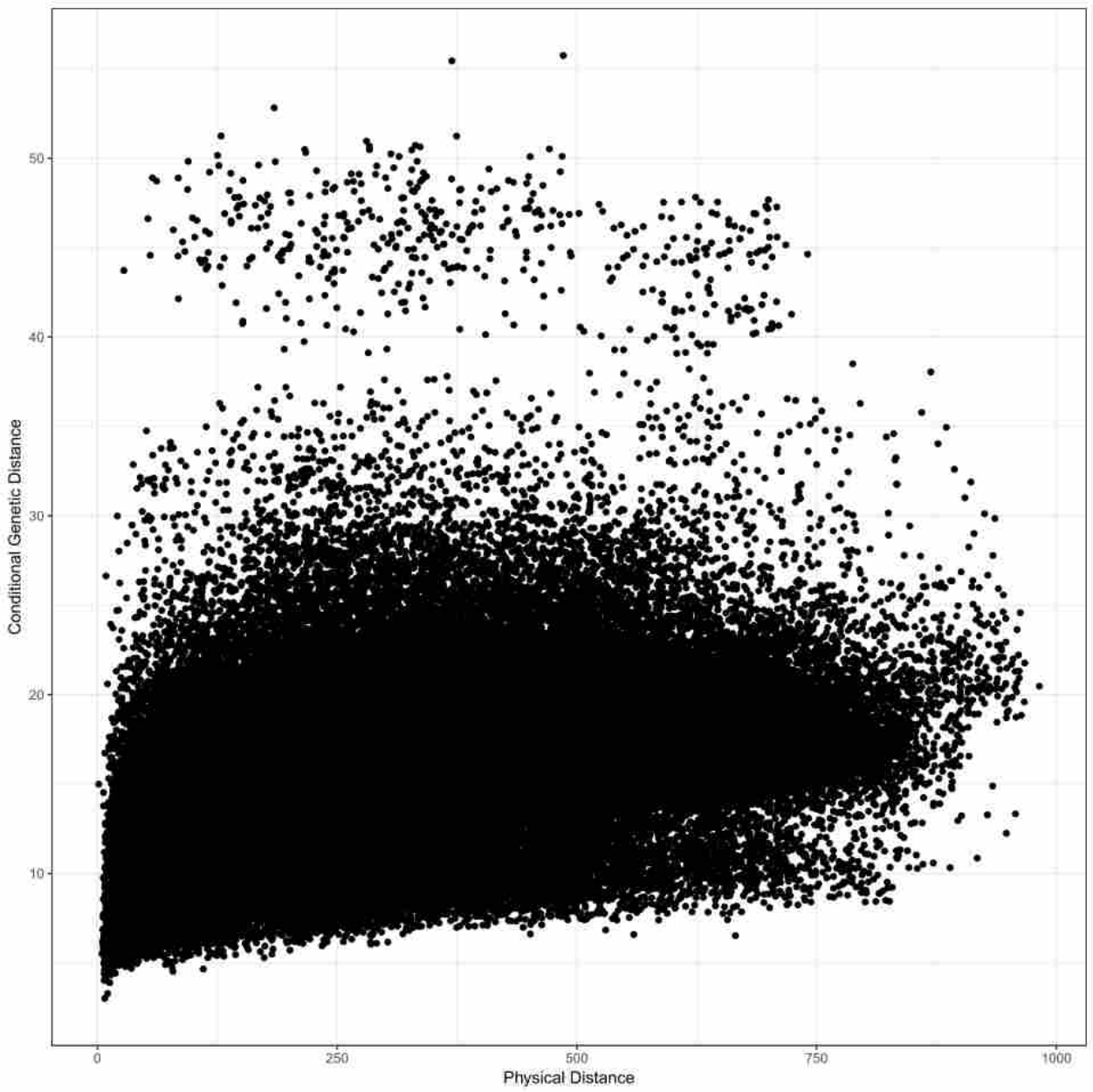
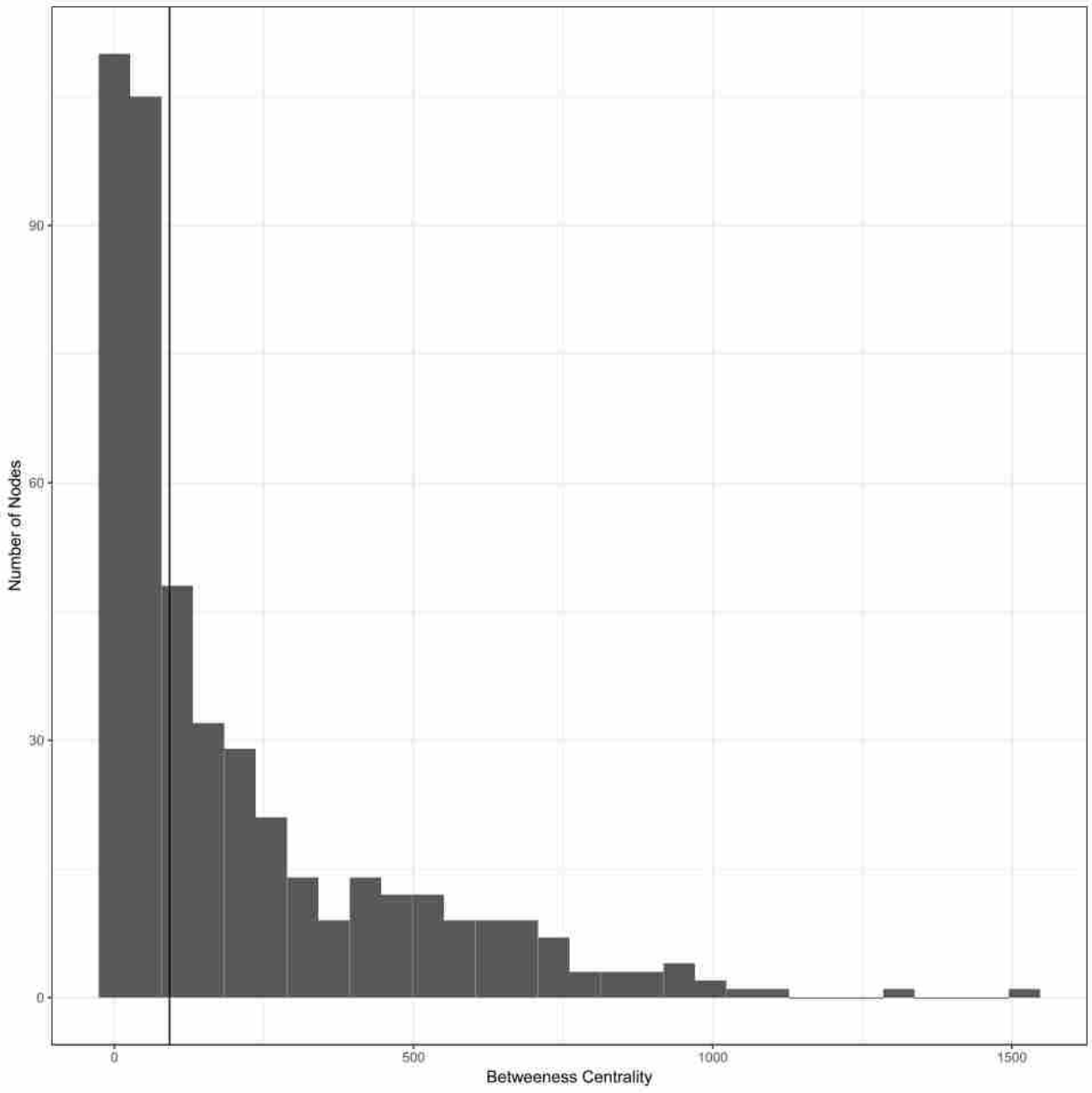
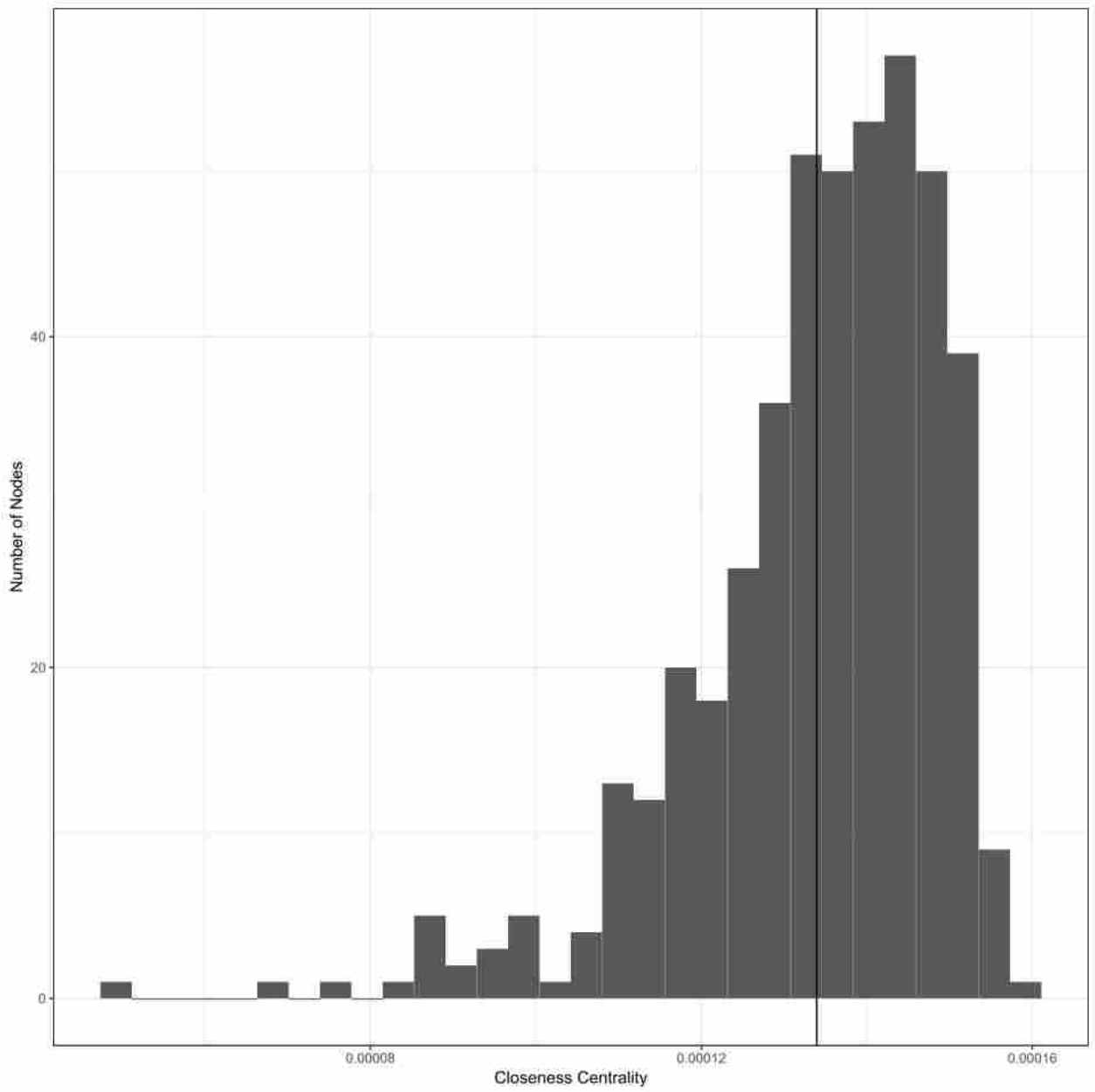


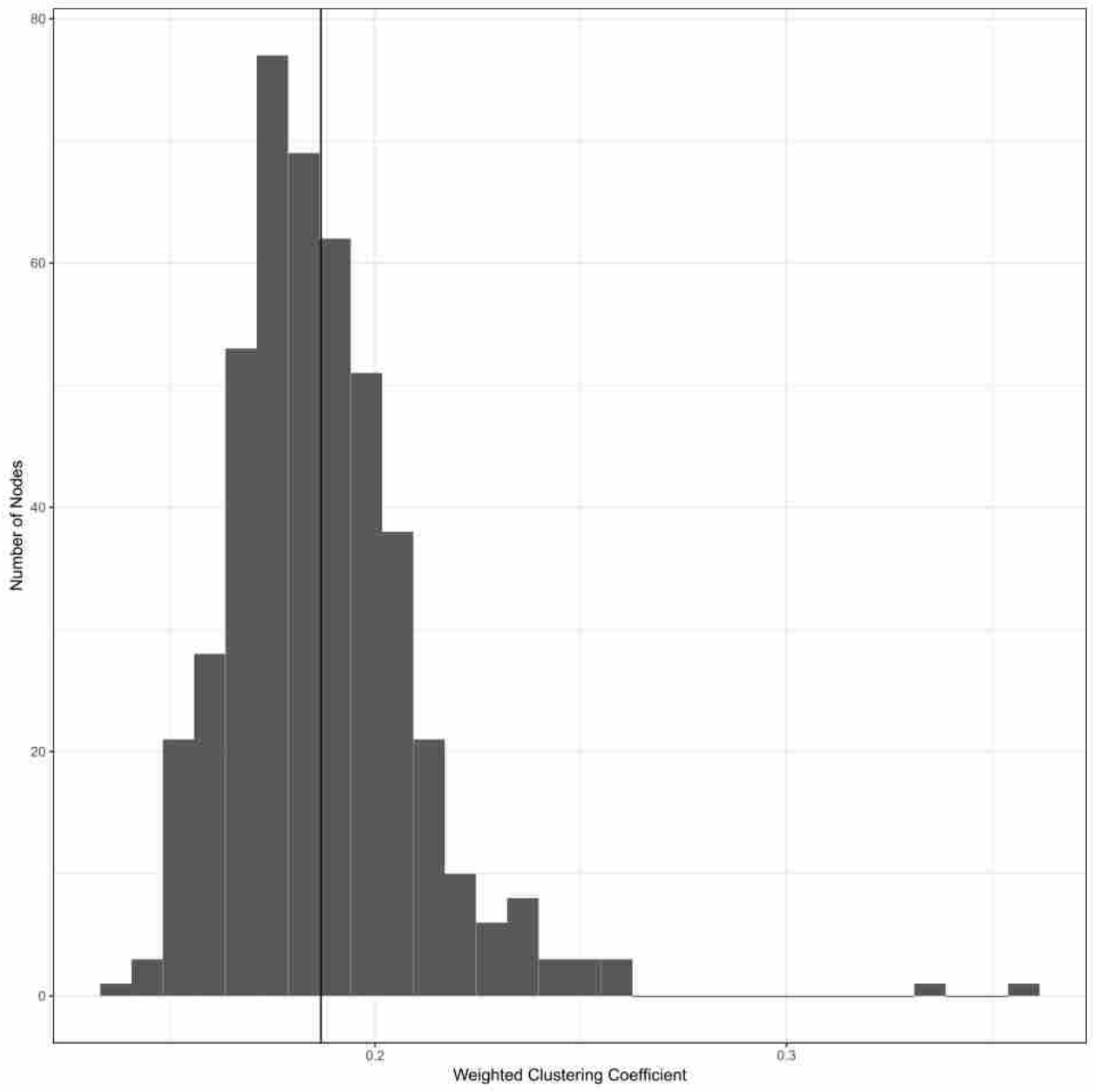
**Figure 5.7.** Centrality measure distributions for all nodes ($n$ = 459) in the greater sage-grouse genetic network. The solid vertical black line shows the reported measure of central tendency for each centrality measure. **(A)** betweenness: the number of shortest paths that a node lies on (median = 92, IQR = 258.5 shortest paths per node); **(B)** closeness: a weighted measure of distance from a given node to all other nodes ($1.34 \times 10^{-4} \pm 1.51 \times 10^{-5}$ [SD]); **(C)** clustering coefficient: the probability that two nodes connected to a given node are also connected (0.19 ± 0.023 [SD]); **(D)** degree distribution: the number of edges emanating from each node (63.1 ± 17.04 [SD] edges per node); **(E)** eigenvector centrality distribution: the direct and indirect connectivity for each node and its immediate neighbors (0.51 ± 0.18 [SD]); **(F)** node strength: the sum of all edge weights (619.4 ± 181.5 [SD]); **(G)** edge weight for all edges ($n$ = 14,481) in the greater sage grouse genetic network (9.82 ± 2.23 [SD])



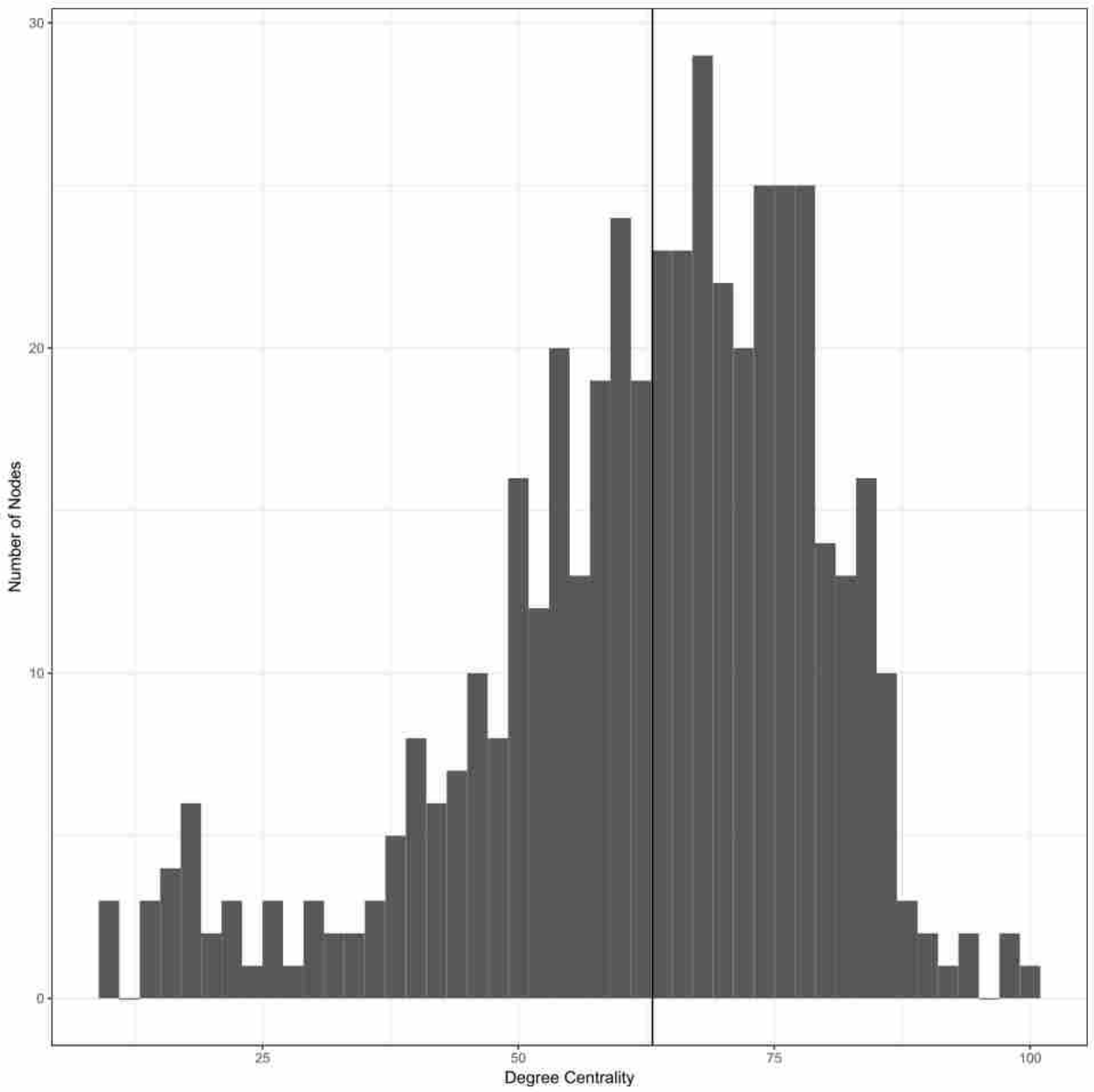

**(A)**


Number of Nodes
40
20
0
0.00008
0.00012
0.00016
Closeness Centrality

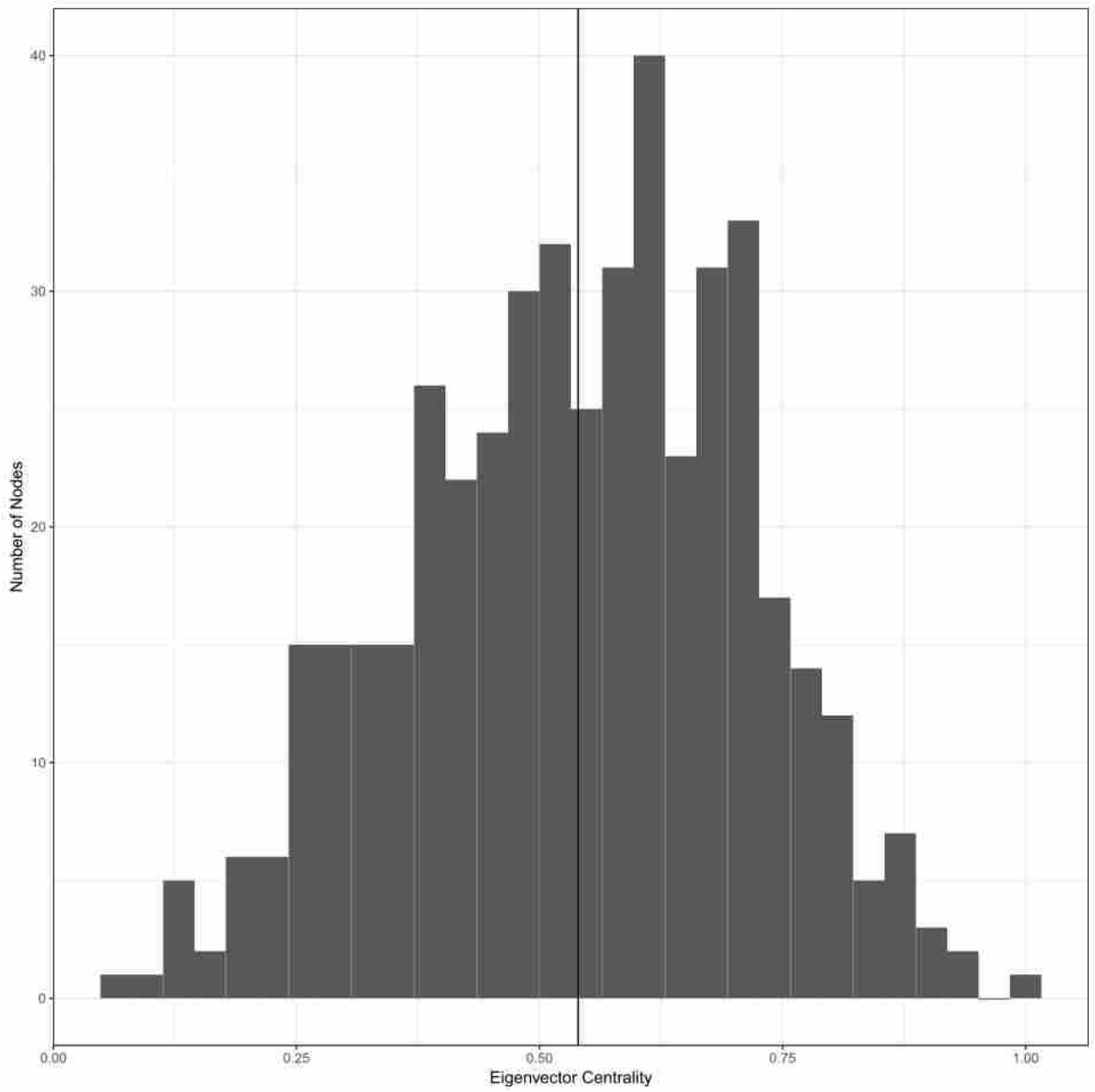
**(B)**


80
60
40
20
0
Number of Nodes
0.2
0.3
Weighted Clustering Coefficient

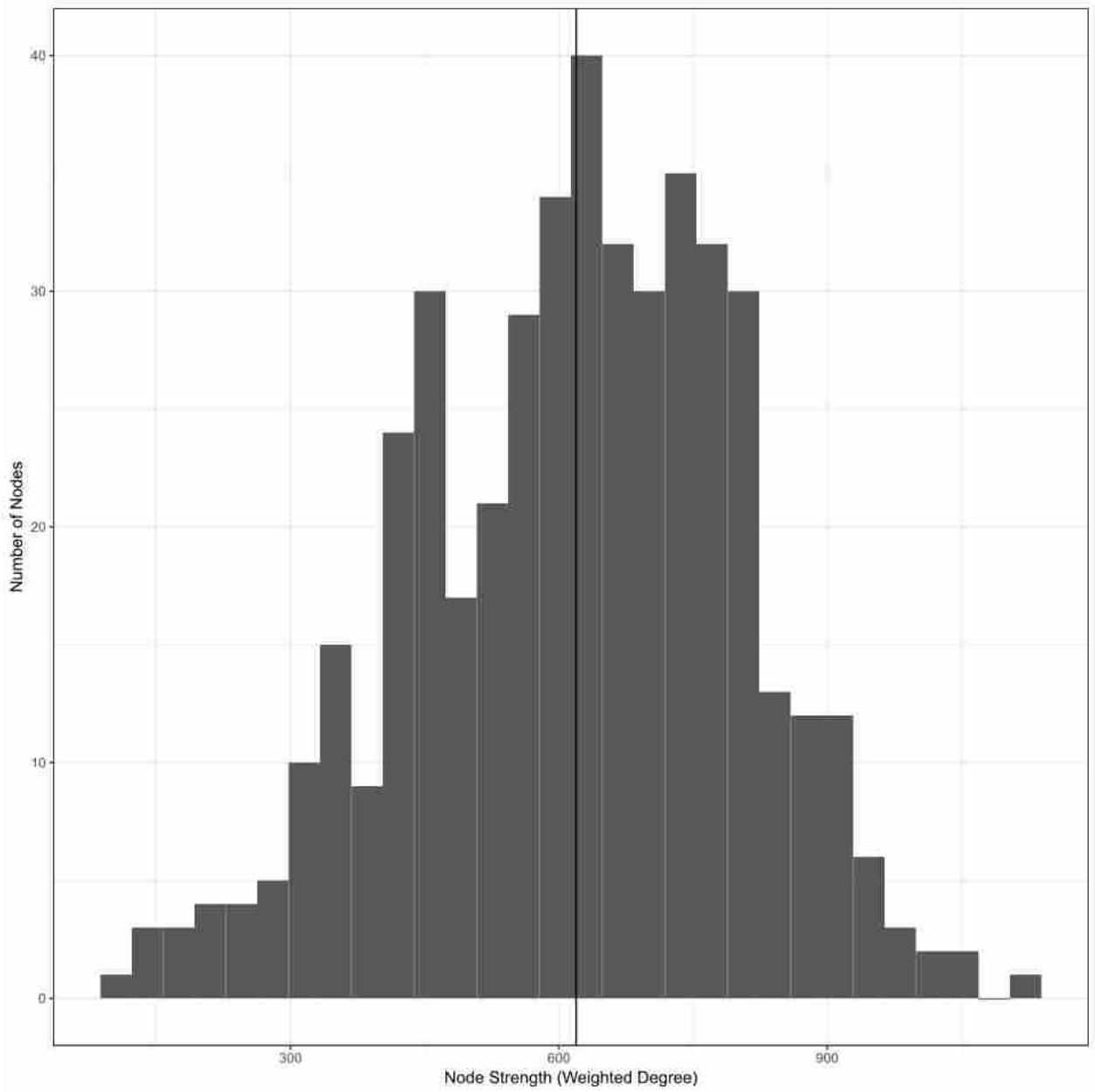
**(C)**



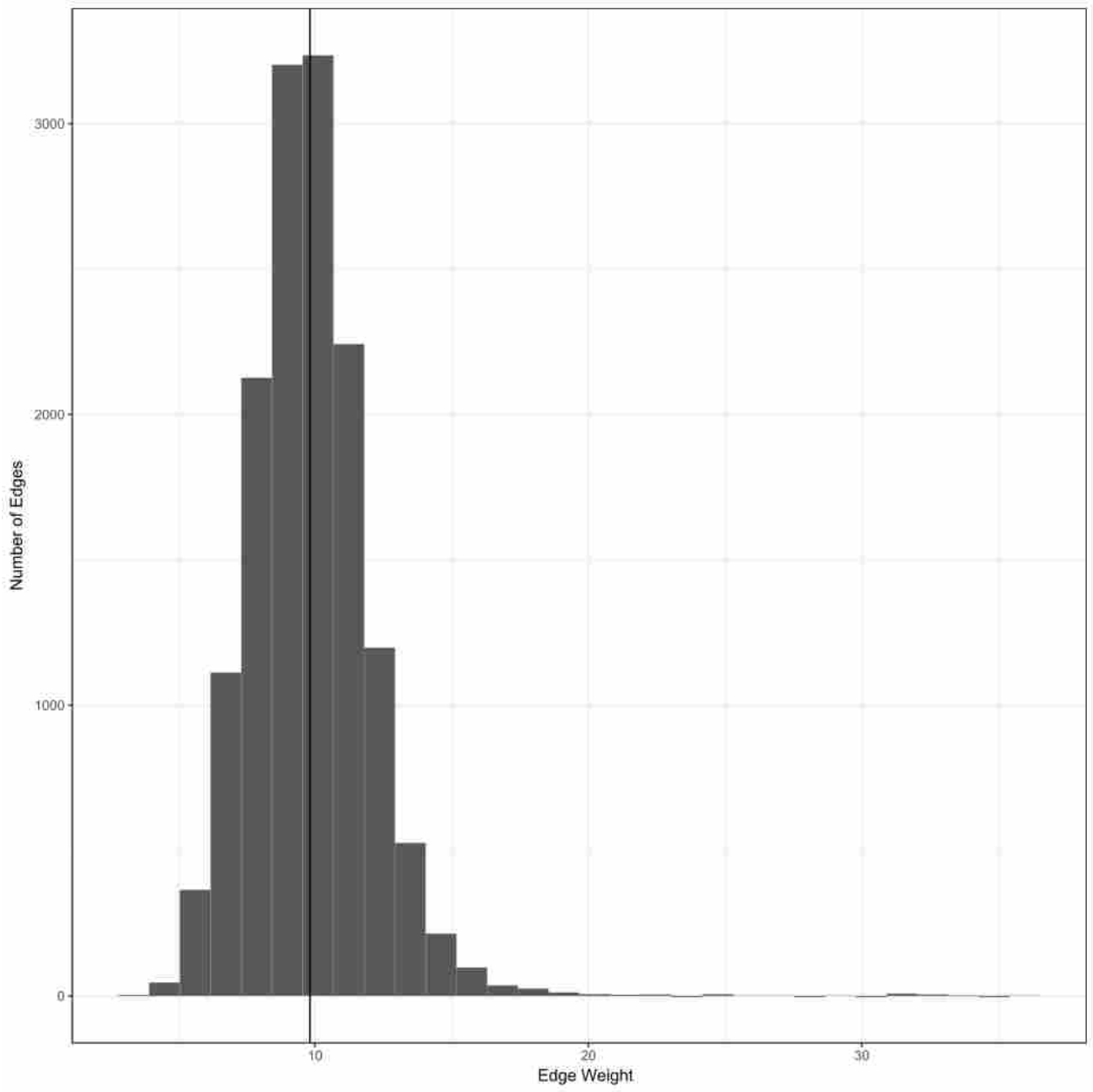

**(D)**


Number of Nodes
40
30
20
10
0
0.00
0.25
0.50
0.75
1.00
Eigenvector Centrality

**(E)**

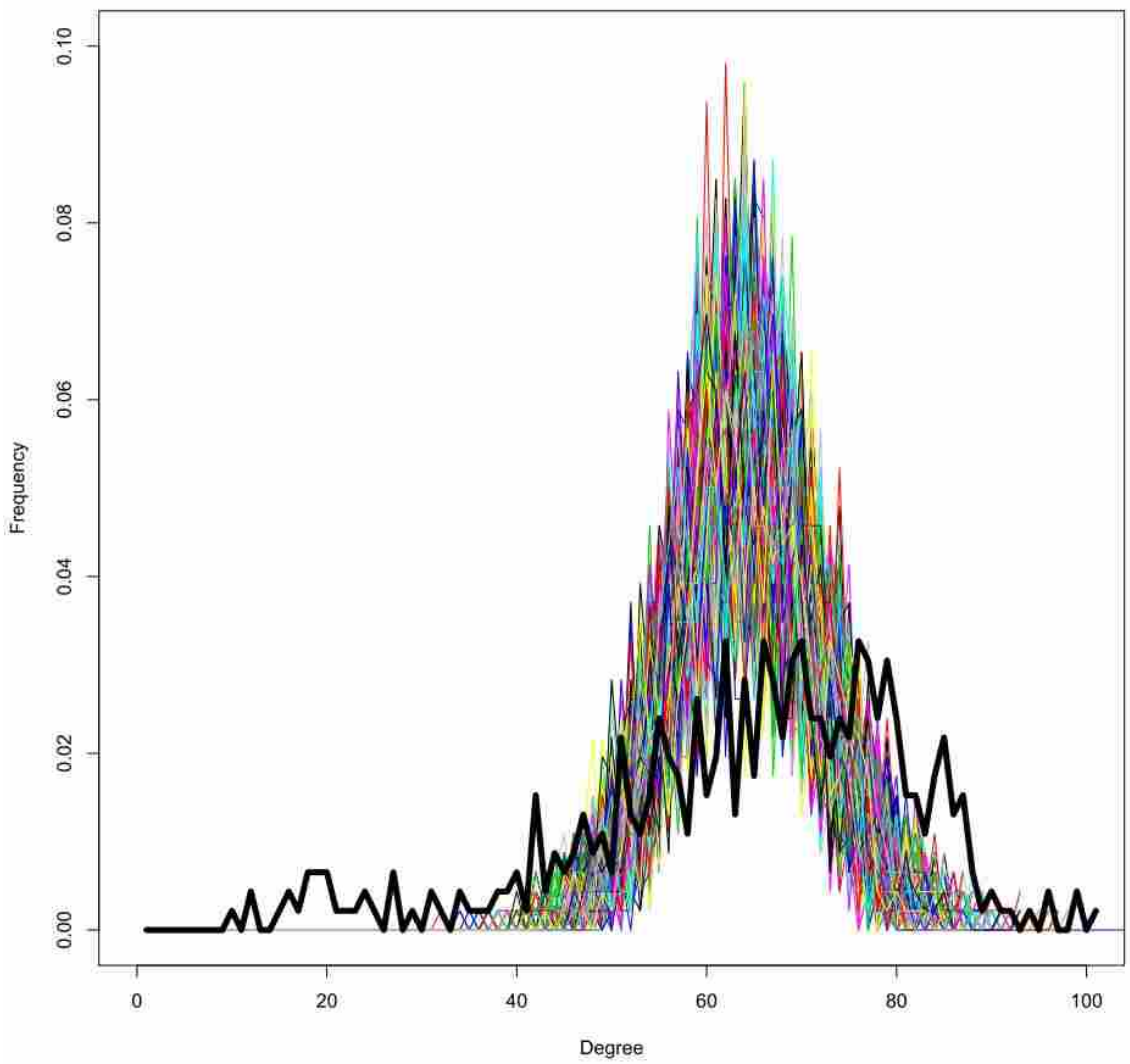

40
30
20
10
0
Number of Nodes
300
600
900
Node Strength (Weighted Degree)

**(F)**




**(G)**

**Figure 5.8.** Degree distribution of 1000 random graphs generated using the same number of nodes and edges as the sage grouse network. The degree distribution of the sage grouse network is shown in black



**LITERATURE CITED**


Aldridge CL, Nielsen SE, Beyer HL, Boyce MS, Connelly JW, Knick ST, Schroeder MA (2008) Range-wide patterns of greater sage-grouse persistence. *Diversity and Distributions,* **14**, 983-994.

Allen AW, Cook JG, Armbruster MJ (1984) Habitat suitability index models: Pronghorn. U.S. Fish and Wildlife Service FWS/OBS-82/10.65, 1–22.

Amaral LAN, Scala A, Barthélémy M, Stanley HE (2000) Classes of small-world networks. *Proceedings of the National Academy of Sciences,* **97**, 11149–11152.

Balkenhol N, Holbrook JD, Onorato D, Zager P, White C, Waits LP (2014) A multi-method approach for analyzing hierarchical genetic structures: a case study with cougars *Puma concolor*. *Ecography,* **37**, 001–012.

Balloux F (2001) EASYPOP (version 1.7): a computer program for population genetics simulations. *Journal of heredity,* **92**, 301–302.

Bivand RS, Pebesma EJ, Gomez-Rubio V (2013) Applied spatial data analysis with R, Second edition. Springer, NY. http://www.asdar-book.org/.

Bivand RS, Rundel, C (2017) RGEOS: Interface to Geometry Engine - Open Source (GEOS). R package version 0.3-22. https://CRAN.R-project.org/package=rgeos.

Blickley JL, Word KR, Krakauer AH, Phillips JL, Sells SN, Taff CC, Wingfield JC, Patricelli GL (2012) Experimental chronic noise is related to elevated fecal corticosteroid metabolites in lekking male greater sage-grouse (*Centrocercus urophasianus*). *PLoS One*, **7**, e50462.

Bohrer G, Nathan R, Volis S (2005) Effects of long-distance dispersal for metapopulation survival and genetic structure at ecological time and spatial scales. *Journal of Ecology*, **93**, 1029–1040.

Bottrill MC, Joseph LN, Carwardine J, Bode M, Cook C, Game ET, Grantham H, Kark S, Linke S, McDonald-Madden E, Pressey RL (2008) Is conservation triage just smart decision making? *Trends in Ecology & Evolution*, **23**, 649-654.

Bradbury J, Vehrencamp S, Gibson R (1989) Dispersion of displaying male sage grouse. *Behavioral Ecology and Sociobiology*, **24**, 1–14.

Braun CE (1998) Sage grouse declines in western North America: what are the problems? *Proceedings of the Western Association of Fish and Wildlife Agencies,* **78**, 139–156.

Braun CE, Oedekoven OO, Aldridge CL (2002) Oil and gas development in western North America: effects on sagebrush steppe avifauna with particular emphasis on sage grouse. *Transactions of the North American Wildlife and Natural Resources Conference*, **67**, 337–349.

Bray D (2003) Molecular networks: the top-down view. *Science,* **301**, 1864–1865.

Bridge ES, Kelly JF, Contina A, Gabrielson RM, MacCurdy RB, Winkler DW (2013) Advances in tracking small migratory birds: A technical review of light-level geolocation. *Journal of Field Ornithology* **84**, 121–137.

Bunn AG, Urban DL, Keitt TH (2000) Landscape connectivity: a conservation application of graph theory. *Journal of Environmental Management,* **59**, 265–278.

Bush KL, Aldridge CL, Carpenter JE, Paszkowski CA, Boyce MS, and Coltman DW (2010) Birds of a feather do not always lek together: Genetic diversity and kinship structure of Greater Sage-Grouse (*Centrocercus urophasianus*) in Alberta. *The Auk,* **127**, 343–353.

Bush KL, Dyte CK, Moynahan BJ, Aldridge CL, Sauls HS, Battazzo AM, Walker BL, Doherty KE, Tack J, Carlson J (2011) Population structure and genetic diversity of greater sage-

grouse (*Centrocercus urophasianus*) in fragmented landscapes at the northern edge of their range. *Conservation Genetics*, **12**, 527–542.
Bush KL, Vinsky MD, Aldridge CL, Paszkowski CA (2005) A comparison of sample types varying in invasiveness for use in DNA sex determination in an endangered population of greater sage-grouse (*Centrocercus uropihasianus*). *Conservation Genetics*, **6**, 867–870.
Byrd DW, McArthur ED, Wang H, Graham JH, Freeman DC (1999) Narrow hybrid zone between two subspecies of big sagebrush, *Artemisia tridentata* (*Asteraceae*). VIII. Spatial and temporal pattern of terpenes. *Biochemical Systematics and Ecology*, **27**, 11–25.
Caizergues A, DuBois S, Loiseau A, Mondor G, Rasplus JY (2001) Isolation and characterization of microsatellite loci in black grouse (*Tetrao tetrix*). *Molecular Ecology Notes*, **1**, 36–38.
Caizergues A, Rätti O, Helle P, Rotelli L, Ellison L, Rasplus JY (2003) Population genetic structure of male black grouse (*Tetrao tetrix L*.) in fragmented vs. continuous landscapes. *Molecular Ecology*, **12**, 2297–2305.
Carrol SP, Hendry AP, Reznick DN, Fox CW (2007) Evolution on ecological time-scales. *Functional Ecology*, **21**, 387–393.
Carstens B, Brunsfeld S, Demboski J, Good J, Sullivan J (2005) Investigating the evolutionary history of the Pacific Northwest mesic forest ecosystem: hypothesis testing within a comparative phylogeographic framework. *Evolution*, **59**, 1639–1652.
Chen C, Durand E, Forbes F, François O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes*, **7**, 747–756.
Cheng E, Hodges KE, Melo-Ferreira J, Alves PC, Mills LS (2014) Conservation implications of the evolutionary history and genetic diversity hotspots of the snowshoe hare. *Molecular Ecology*, **23**, 2929–2942.
Clobert J, Le Galliard JF, Cote J, Meylan S, Massot M (2009) Informed dispersal, heterogeneity in animal dispersal syndromes and the dynamics of spatially structured populations. *Ecology Letters*, **12**, 197–209.
Comer P, Kagan J, Heiner M, Tobalske C (2002) Current distribution of sagebrush and associated vegetation in the Western United States (excluding NM and AZ). Interagency Sagebrush Working Group. http://SAGEMAP.wr.usgs.gov
Connelly J, Rinkes E, Braun C (2011b) Characteristics of greater sage-grouse habitats: a landscape species at micro and macro scales. In: Knick ST, Connelly JW (eds) Greater Sage-Grouse: ecology and conservation of a landscape species and its habitats. Studies in Avian Biology, 38th edn. University of California Press, Berkeley, pp 69–83.
Connelly JW, Hagen CA, Schroeder MA (2011) Characteristics and dynamics of greater sage-grouse populations. In: *Greater Sage-Grouse: Ecology and Conservation of a Landscape Species and Its Habitats* (eds. Knick ST, Connelly JW). *Studies in Avian Biology*, 38th edn. University of California Press, Berkeley, CA, pp 53–67.
Connelly JW, Knick ST, Schroeder MA, Stiver SJ (2004) Conservation assessment of greater sage-grouse and sagebrush habitats. *Western Association of Fish and Wildlife Agencies*. Unpublished report. Cheyenne, Wyoming.
Connelly JW, Schroeder MA, Sands AR, Braun CE (2000) Guidelines to manage sage grouse populations and their habitats. *Wildlife Society Bulletin*, **28**, 967–985.
Connelly JW, Browers HW, Gates RJ (1988) Seasonal movements of Sage Grouse in southeastern Idaho. *The Journal of Wildlife Management*, **52**, 116–122.

Copeland HE, Doherty KE, Naugle DE, Pocewicz A, Kiesecker JM (2009) Mapping oil and gas development potential in the US Intermountain West and estimating impacts to species. *PLoS One*, **4**, e7400.

Coulon A, Fitzpatrick J, Bowman R, Stith B, Makarewich C, Stenzler L, Lovette I (2008) Congruent population structure inferred from dispersal behaviour and intensive genetic surveys of the threatened Florida scrub-jay (*Aphelocoma cærulescens*). *Molecular Ecology*, **17**, 1685–1701.

Crist MR, Knick ST, Hanser SE (2017) Range-wide connectivity of priority areas for Greater Sage-Grouse: Implications for long-term conservation from graph theory. *The Condor: Ornithological Applications*, **119**, 44–57.

Cross TB, Naugle DE, Carlson JC, Schwartz MK (2016) Hierarchical population structure in greater sage-grouse provides insight into management boundary delineation. *Conservation Genetics*, **17**, 1417–1433.

Cross TB, Naugle DE, Carlson JC, Schwartz MK (2017) Genetic recapture identifies long-distance breeding dispersal in Greater Sage-Grouse (*Centrocercus urophasianus*). *The Condor: Ornithological Applications*, **119**, 155–166.

Csardi G, Nepusz T (2006) The IGRAPH software package for complex network research. *InterJournal, Complex Systems*, **1695**, 38.

Dalke PD, Pyrah DB, Stanton DC, Crawford JE, Schlatterer EF (1963) Ecology, productivity, and management of sage grouse in Idaho. *The Journal of Wildlife Management*, **27**, 811–841.

Davis DM, Reese KP, Gardner SC, Bird KL (2015) Genetic structure of Greater Sage-Grouse (*Centrocercus urophasianus*) in a declining, peripheral population. *The Condor: Ornithological Applications*, **117**, 530–544.

DeWoody J, Nason JD, Hipkins VD (2006) Mitigating scoring errors in microsatellite data from wild populations. *Molecular Ecology Notes*, **6**, 951–957.

Dilkina B, Houtman R, Gomes CP, Montgomery CA, McKelvey KS, Kendall K, Graves TA, Bernstein R, Schwartz MK (2016) Trade-offs and efficiencies in optimal budget-constrained multispecies corridor networks. *Conservation Biology*, 1-11.

Doherty KE, Naugle DE, Walker BL, Graham JM (2008) Greater sage-grouse winter habitat selection and energy development. *Journal of Wildlife Management*, **72**, 187–195.

Dunn PO, Braun CE (1985) Natal dispersal and lek fidelity of sage grouse. *The Auk*, **102**, 621–627.

Dunn PO, Braun CE (1986) Late summer–spring movements of juvenile Sage Grouse. *The Wilson Bulletin*, **98**, 83–92.

Dunne JA, Williams RJ, Martinez ND (2002) Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecology Letters*, **5**, 558–567.

Dyer RJ (2007) The evolution of genetic topologies. *Theoretical Population Biology*, **71**, 71–79.

Dyer RJ (2014) GSTUDIO: Analyses and functions related to the spatial analysis of genetic marker data. R package version 3.2.1.

Dyer RJ, Nason JD (2004) Population graphs: the graph theoretic shape of genetic structure. *Molecular Ecology*, **13**, 1713–1727.

Dyer RJ, Nason JD, Garrick RC (2010) Landscape modelling of gene flow: improved power using conditional genetic distance derived from the topology of population networks. *Molecular Ecology*, **19**, 3746–3759.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 1–3.
Earl JE, Fuhlendorf SD, Haukos D, Tanner AM, Elmore D, Carleton SA (2016) Characteristics of Lesser Prairie-Chicken (*Tympanuchus pallidicinctus*) long-distance movements across their distribution. *Ecosphere*, **7**, e01441. doi:10.1002/ecs2.1441
Edminster FC (1954) American game birds of field and forest: their habits, ecology, and management. Scribner, New York
El Mousadik A, Petit R (1996) High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco. *Theoretical and Applied Genetics*, **92**, 832–839.
Ellegren H (2000) Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics*, **16**, 551–558.
Ellstrand NC, Rieseberg LH (2016) When gene flow really matters: gene flow in applied evolutionary biology. *Evolutionary Applications*, **9**, 833–836.
Emmons SR, Braun CE (1984) Lek attendance of male sage grouse. *The Journal of Wildlife Management*, **48**, 1023–1028.
Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.
Evett IW, Weir BS (1998) Interpreting DNA evidence: statistical genetics for forensic scientists. Sinauer Associates, Sunderland, MA, USA.
Excovvier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479-491.
Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.
Finch DM, Boyce DA, Jr., Chambers JC, Colt CJ, Dumroese K, Kitchen SG, McCarthy C, Meyer SE, Richardson BA, Rowland MM, Rumble MA, Schwartz MK, Tomosy MS, Wisdom MJ (2016) Conservation and restoration of sagebrush ecosystems and sage-grouse: An assessment of USDA Forest Service Science. *General Technical Report*, **RMRS-GTR-348**, 1–54.
Francois O, Ancelet S, Guillot G (2006) Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics* **174**:805–816
Frye GG, Connelly JW, Musil DD, Forbey JS (2013) Phytochemistry predicts habitat selection by an avian herbivore at multiple scales. *Ecology* **94**, 308–314
Galpern P, Manseau M, Hettinga P, Smith K, Wilson P (2012) ALLELEMATCH: an R package for identifying unique multi-locus genotypes where genotyping error and missing data may be present. *Molecular Ecology Resources*, **12**, 771–778.
Gandon, S. (1999). Kin competition, the cost of inbreeding and the evolution of dispersal. Journal of Theoretical Biology 200:345–364.
Garroway CJ, Bowman J, Carr D, Wilson PJ (2008) Applications of graph theory to landscape genetics. *Evolutionary Applications*, **1**, 620–630.
Garroway CJ, Bowman J, Wilson PJ (2011) Using a genetic network to parameterize a landscape resistance surface for fishers, Martes pennanti. *Molecular Ecology*, **20**, 3978-3988.
Gesch D, Oimoen M, Greenlee S, Nelson C, Steuck M, Tyler D (2002) The National Elevation Dataset. *Photogrammic Engineering & Remote Sensing*, **68**, 5–11.

Gibson RM, Bradbury JW, Vehrencamp SL (1991) Mate choice in lekking sage grouse revisited: the roles of vocal display, female site fidelity, and copying. *Behavioural Ecology*, **2**, 165–180.
Goudet J (1995) FSTAT (version 1.2): a computer program to calculate F-statistics. *American Journal of Heredity*, **86**, 485–486.
Green GA, Livezey KB, Morgan RL (2001) Habitat selection by northern sagebrush lizards (*Sceloporus graciosus graciosus*) in the Columbia Basin, Oregon. N West Nat 82:111–115
Green JS, Flinders JT (1960) Habitat and Dietary Pygmy Rabbit. *Society of Range Management*, **33**, 136–142.
Guillot G, Santos F (2009) A computer program to simulate multilocus genotype data with spatially auto-correlated allele frequencies. *Molecular Ecology Resources*, **9**, 1112–1120.
Hagen CA, Connelly JW, Schroeder MA (2007) A meta-analysis of greater sage-grouse *Centrocercus urophasianus* nesting and brood-rearing habitats. *Wildlife Biology*, **13**, 42–50.
Hanf JM, Schmidt PA, Groshens EB (1994) Sage Grouse in the high desert of central Oregon: Results of a study, 1988–1993. PT-95/002-4120.7, U.S. Department of Interior, Bureau of Land Management, Prineville, OR, USA.
Hassall, C., and D. J. Thompson (2012). Study design and mark-recapture estimates of dispersal: A case study with the endangered damselfly *Coenagrion mercuriale*. *Journal of Insect Conservation*, **16**, 111–120.
Hawley JE, Rego PW, Wydeven AP, Schwartz MK, Viner TC, Kays R, Pilgrim KL, Jenks JA (2016) Long-distance dispersal of a subadult male cougar from South Dakota to Connecticut documented with DNA evidence. *Journal of Mammalogy*, **97**, 1435–1440.
Holderegger R, Di Giulio M (2010) The genetic effects of roads: a review of empirical evidence. *Basic Applied Ecology*, **11**, 522–531.
Holloran MJ, Anderson SH (2005) Spatial distribution of greater sage-grouse nests in relatively contiguous sagebrush habitats. *The Condor*, **107**, 742–752.
Holloran MJ, Kaiser RC, Hubert WA (2010) Yearling greater sage-grouse response to energy development in Wyoming. *Journal of Wildlife Management*, **74**, 65–72.
Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322–1332.
Jackson ND, Fahrig L (2011) Relative effects of road mortality and decreased connectivity on population genetic diversity. *Biological Conservation*, **144**, 3143–3148.
Jacoby DMP, Freeman R (2016) Emerging network-based tools in movement ecology. *Trends in Ecology & Evolution*, **31**, 301–314.
Jaeger DM, Runyon JB, Richardson BA (2016) Signals of speciation: volatile organic compounds resolve closely related sagebrush taxa, suggesting their importance in evolution. *New Phytologist*, **211**, 1393-1401.
Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–1806.
Johnson ML, Gaines MS (1990) Evolution of dispersal: Theoretical models and empirical tests using birds and mammals. *Annual Review of Ecology and Systematics*, **21**, 449–480.
Jombart T, Ahmed I (2011) ADEGENET 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, **27**, 3070–3071.

Kahn N, St. John J, Quinn TW (1998) Chromosome-specific intron size differences in the avian CHD gene provide an efficient method for sex identification in birds. *The Auk: Ornithological Applications*, **115**, 1074–1078.
Kahn NW, St. John J, Quinn TW (1998) Chromosome-specific intron size differences in the avian CHD gene provide an efficient method for sex identification in birds. *The Auk*, **115**, 1074–1078.
Knick ST, Dobkin DS, Rotenberry JT, Schroeder MA, Vander Haegen WM, Van Riper C III (2003) Teetering on the edge or too late? Conservation and research issues for avifauna of sagebrush habitats. *The Condor*, **105**, 611–634.
Knick ST, Hanser SE (2011) Connecting pattern and process in greater sage-grouse populations and sagebrush landscapes. In: *Greater sage-grouse: ecology and conservation of a landscape species and its habitats. Studies in Avian Biology* (eds. Knick ST, Connelly JW). *Studies in Avian Biology*, **38**, 383–405. University of California Press, Berkeley, CA.
Knick ST, Hanser SE, Preston KL (2013) Modeling ecological minimum requirements for distribution of Greater Sage-Grouse leks: Implications for population connectivity across their western range, U.S.A. *Ecology and Evolution*, **3**, 1539–1551.
Koen EL, Bowman J, Garroway CJ, Mills SC, Wilson PJ (2012) Landscape resistance and American marten gene flow. *Landscape Ecology*, **27**, 29–43.
Koen EL, Bowman J, Garroway CJ, Wilson PJ (2013) The sensitivity of genetic connectivity measures to unsampled and under-sampled sites. *PloS one*, **8**, e56204.
Koen EL, Bowman J, Wilson PJ (2015) Node-based measures of connectivity in genetic networks. *Molecular ecology resources*, **16**, 69–79.
Koenig WD, Van Vuren D, Hooge PN (1996) Detectability, philopatry, and the distribution of dispersal distances in vertebrates. *Trends in Ecology & Evolution*, **11**, 514–517.
Latch EK, Dharmarajan G, Glaubitz JC, Rhodes OE Jr (2006) Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics*, **7**, 295–302.
Leonard KM, Reese KP, Connelly JW (2000) Distribution, movements and habitats of sage grouse Centrocercus urophasianus on the Upper Snake River plain of Idaho: Changes from the 1950s to the 1990s. *Wildlife Biology*, **6**, 265–270.
Leturque H, Rousset F (2002) Dispersal, kin competition, and the ideal free distribution in a spatially heterogeneous population. *Theoretical Population Biology*, **62**, 169–180.
Lomolino MV, Riddle BR, Brown JH, Whittaker RJ (2006) Biogeography. Sinauer Associates, Sunderland.
Lowe WH (2009) What drives long-distance dispersal? A test of theoretical predictions. *Ecology*, **90**:1456–1462.
Lukoschek V, Waycott M, Keogh JS (2008) Relative information content of polymorphic microsatellites and mitochondrial DNA for inferring dispersal and population genetic structure in the olive sea snake, *Aipysurus laevis*. *Molecular Ecology*, **17**, 3062–3077.
Manel S, Joost S, Epperson BK, Holderegger R, Storfer A, Rosenberg MS, Scribner KT, Bonin A, Fortin MJ (2010) Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Molecular Ecology*, **19**, 3760–3772.
Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.
McArthur ED, Plummer AP (1978) Biogeography and management of native western shrubs: a case study, section *Tridentatae* of *Artemisia*. *Great Basin Naturalist Memoirs*, **2**, 229–243.

McCaffery R, Nowak JJ, Lukacs PM (2016) Improved analysis of lek count data using *N*-mixture models. *The Journal of Wildlife Management*, **80**, 1011–1021.
McKelvey KS, Schwartz MK (2005) DROPOUT: A program to identify problem loci and samples for noninvasive genetic samples in a capture-mark-recapture framework. *Molecular Ecology Notes*, **5**, 716–718.
Mills LS, Allendorf FW (1996) The one-migrant-per-generation rule in conservation and management. *Conservation Biology*, **10**, 1509–1518.
Monsen SB, Shaw NL (2000) Big sagebrush (*Artemesia tridentata*) communities—ecology, importance and restoration potential. In: Billings Land Reclamation Symposium, 2000: Striving for Restoration, Fostering Technology, and Policy for Reestablishing Ecological Function: March 20–24, 2000, Sheraton Billings Hotel, Billings, Montana. Bozeman: Montana State University, Publication No. 00–01.
Montana Fish, Wildlife & Parks (2014) Metadata for Greater sage-grouse core areas. http://fwp.mt.gov/gisData/metadata/sgcore.htm
Moriarty KM, Zielinski WJ, Gonzales AG, Dawson TE, Boatner KM, Wilson CA, Schlexer FV, Pilgrim KL, Copeland JP, Schwartz MK (2009) Wolverine confirmation in California after nearly a century: Native or long-distance immigrant? *Northwest Science*, **83**, 154–162.
Murphy T, Naugle DE, Eardley R, Maestas JD, Griffiths T, Pellant M, Stiver SJ (2013) Trial by fire. *Rangelands*, **35**, 2–10.
Nathan R, Perry G, Cronin JT, Strand AE, Cain ML (2003). Methods for estimating long-distance dispersal. *Oikos*, **103**, 261–273.
Nathan, R (2006) Long-distance dispersal of plants. *Science*, **313**, 786–788.
Natural Resources Conservation Service (2015a) Outcomes in conservation: sage grouse initiative. U.S. Department of Agriculture, February 2015.
Natural Resources Conservation Service (2015b) Sage grouse initiative 2.0: Investment Strategy, FY 2015–2018. U.S. Department of Agriculture, August 2015.
Naugle DE, Aldridge CL, Walker BL, Cornish TE, Moynahan BJ, Holloran MJ, Brown K, Johnson GD, Schmidtmann ET, Mayer RT (2004) West Nile virus: pending crisis for greater sage-grouse. *Ecology Letters*, **7**, 704–713.
Naugle DE, Doherty KE, Walker BL, Copel HE, Holloran MJ, Tack JD (2011) Sage grouse and cumulative impacts of energy development. In: Naugle DE (ed) Energy development and wildlife conservation in western North America. Island, Washington, pp 55–70.
Naugle DE, Walker BL, Doherty KE (2006) Sage grouse population response to coal-bed natural gas development in the Powder River Basin: interim progress report on region-wide lek-count analyses. University of Montana, Missoula
Nychka D, Furrer R, Paige J , Sain S (2015). FIELDS: Tools for spatial data. doi: 10.5065/D6W957CT (URL: http://doi.org/10.5065/D6W957CT), R package version 8.10, <URL: www.image.ucar.edu/fields>.
Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MH, Wagner H (2015) VEGAN: Community Ecology Package. https://cran.r-project.org/web/packages/vegan/index.html
Oyler-McCance SJ, Casazza ML, Fike JA, Coates PS (2014) Hierarchical spatial genetic structure in a distinct population segment of greater sage-grouse. *Conservation Genetics*, **15**, 1299–1311.

Oyler-McCance SJ, St. John J (2010) Characterization of small microsatellite loci for use in non-invasive sampling studies of Gunnison sage-grouse (*Centrocercus minimus*). *Conservation Genetics Resources*, **2**, 17–20.
Oyler-McCance SJ, Taylor SE, Quinn TW (2005) A multilocus population genetic survey of the greater sage-grouse across their range. *Molecular Ecology*, **14**, 1293–1310.
Parris KM, Schneider A (2009) Impacts of traffic noise and traffic volume on birds of roadside habitats. *Ecology and Society*, **14**, 29–50.
Patterson RL (1952) The Sage Grouse in Wyoming. Sage Books, Denver, CO, USA, for Wyoming Game and Fish Commission, Cheyenne, WY, USA.
Payne RB (1984) Sexual selection, lek and arena behavior, and sexual size dimorphism in birds. *Ornithological Monographs*, **33**.
Pebesma EJ, Bivand RS (2005) Classes and methods for spatial data in R. R News 5 (2), https://cran.r-project.org/doc/Rnews/.
Piertney SB, Höglund J (2001) Polymorphic microsatellite DNA markers in Black Grouse (*Tetrao tetrix*). *Molecular Ecology Notes*, **1**, 303–304.
Platt TG, Bever JD (2009) Kin competition and the evolution of cooperation. *Trends in Ecology & Evolution*, **24**, 370–377.
Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.
Proctor MF, McLellan BN, Strobeck C, Barclay RM (2005) Genetic analysis reveals demographic fragmentation of grizzly bears yielding vulnerably small populations. *Proceedings of the Royal Society of London B*, **272**, 2409–2416.
R Core Team (2016) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/
Remington TE, Braun CE (1985) Sage grouse food selection in winter, North Park, Colorado. *Journal of Wildlife Management*, **49**, 1055–1061.
Richardson JL, Urban MC, Bolnick DI, and Skelly DK (2014) Microgeographic adaptation and the spatial scale of evolution. *Trends in Ecology & Evolution*, **29**, 165–176.
Ricketts TH (1999) Terrestrial ecoregions of North America: a conservation assessment. Island Press.
Robinson AC (2014) Management plan and conservation strategies for greater sage-grouse in North Dakota. North Dakota Game and Fish Department, Bismark.
Roffler GH, Talbot SL, Luikart GH, Sage GK, Pilgrim KL, Adams LG, Schwartz MK (2014) Lack of sex-biased dispersal promotes fine-scale genetic structure in alpine ungulates. *Conservation Genetics*, **15**, 837–851.
Ronce O (2007) How does it feel to be like a rolling stone? Ten questions about dispersal evolution. *Annual Review of Ecology, Evolution, and Systematics*, **38**, 231–253.
Rosentreter R (2004) Sagebrush identification, ecology, and palatability relative to sage grouse. In: Shaw NL, Monsen SB, Pellant M (eds) Sage grouse habitat restoration symposium proceedings; 4–7 June 2001; Boise, ID. Proceedings RMRS-P-000. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Ft. Collins.
Rousset F (2008) Genepop'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.
Row JR, Oyler-McCance SJ, Fike JA, O'Donnell MS, Doherty KE, Aldridge CL, Bowen ZH, Fedy BC (2015) Landscape characteristics influencing the genetic structure of greater sage-

grouse within the stronghold of their range: a holistic modeling approach. *Ecology and evolution*, **5**, 1955–1969.
Rowland MM, Wisdom MJ, Suring LH, Meinke CW (2006) Greater sage-grouse as an umbrella species for sagebrush-associated vertebrates. *Biological Conservation*, **129**, 323–335.
Sallaberry A, Zaidi F, Melançon G (2013) Model for generating artificial social networks having community structures with small-world and scale-free properties. *Social Network Analysis and Mining*, **3**, 1–13.
Schrag A, Konrad S, Miller S, Walker B, Forrest S (2011) Climate-change impacts on sagebrush habitat and West Nile virus transmission risk and conservation implications for greater sage-grouse. *GeoJournal*, **76**, 561–575.
Schroeder MA, Aldridge CL, Apa AD, Bohne JR, Braun CE, Bunnell SD, Connelly JW, Deibert PA, Gardner SC, Hilliard MA (2004) Distribution of sage grouse in North America. *Condor*, **106**, 363–376.
Schroeder MA, Robb LA (2003) Fidelity of Greater Sage-Grouse *Centrocercus urophasianus* to breeding areas in a fragmented landscape. *Wildlife Biology*, **9**, 291–299.
Schulwitz, S, Bedrosian B, Johnson JA (2014) Low neutral genetic diversity in isolated Greater Sage-Grouse (*Centrocercus urophasianus*) populations in northwest Wyoming. *The Condor: Ornithological Applications*, **116**, 560–573.
Schwartz MK, Cushman SA, McKelvey KS, Hayden J, Engkjer C (2006) Detecting genotyping errors and describing American black bear movement in northern Idaho. *Ursus*, **17**, 138–148.
Schwartz MK, McKelvey KS (2009) Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. *Conservation Genetics*, **10**, 441–452.
Schwartz MK, Mills LS (2005) Gene flow after inbreeding leads to higher survival in deer mice. *Biological Conservation*, **123**, 413–420.
Segelbacher G (2002) Noninvasive genetic analysis in birds: testing reliability of feather samples. *Molecular Ecology Notes*, **2**, 367–369.
Segelbacher G, Paxton RJ, Steinbrück G, Trontelj P, Storch I (2000) Characterization of microsatellites in capercaillie *Tetrao urogallus* (AVES). *Molecular Ecology*, **9**, 1934–1935.
Semple K, Wayne RK, Gibson RM (2001) Microsatellite analysis of female mating behaviour in lek-breeding sage grouse. *Molecular Ecology*, **10**, 2043–2048.
Shafer A, Cullingham CI, Cote SD, Coltman DW (2010) Of glaciers and refugia: a decade of study sheds new light on the phylogeography of northwestern North America. *Molecular Ecology*, **19**, 4589–4621.
Short Bull AR, Cushman SA, Mace R, Chilton T, Kendall KC, Landguth EL, Schwartz MK, McKelvey K, Allendorf FW, Luikart G (2011) Why replication is important in landscape genetics: American black bear in the Rocky Mountains. *Molecular Ecology*, **20**, 1092–1107.
Slabbekoorn H, Ripmeester EA (2008) Birdsong and anthropognic noise: implications and applications for conservation. *Molecular Ecology*, **17**, 72–83.
Smith RE (2012) Conserving Montana's sagebrush highway: Long distance migration in sage-grouse. M.S. thesis, University of Montana, Missoula, MT, USA.
Smits JEG, Fernie KJ (2013) Avian wildlife as sentinels of ecosystem health. *Comparative Immunology Microbiology and Infectious Diseases*, **36**, 333–342.
Soltis DE, Gitzendanner MA, Strenge DD, Soltis PS (1997) Chloroplast DNA intraspecific phylogeography of plants from the Pacific Northwest of North America. *Plant Systematics and Evolution*, **206**, 353–373.

Stanton DJ, McArthur ED, Freeman DC, Golenberg EM (2002) No genetic substructuring in *Artemisia* subgenus *Tridentatae* despite strong ecotypic subspecies selection. *Journal of the American Statistical Association*, **30**, 579–593.
State of Montana (2014) Executive order creating the Montana sage grouse oversight team and the Montana sage grouse habitat conservation program. Executive Order No. 10-2014, pp1–29.
State of Montana (2015) Executive order amending and providing for implementation of the Montana sage grouse conservation strategy. Executive Order No. 12–2015, pp 1–38.
Stiver SJ, Apa AD, Bohne JR, Bunnell SD, Deibert PA, Gardner SC, Hilliard MA, McCarthy CW, Schroeder MA (2006) Greater Sage-grouse comprehensive conservation strategy. Western Association of Fish and Wildlife, Cheyenne.
Sturges HA (1926) The choice of a class interval. *Journal of the American Statistical Association*, **21**, 65–66.
Tack JD (2009) Sage-grouse and the human footprint: Implications for conservation of small and declining populations. M.S. thesis, University of Montana, Missoula, MT, USA.
Tack JD, Naugle DE, Carlson JC, Fargey PJ (2012) Greater Sage-Grouse *Centrocercus urophasianus* migration links the USA and Canada: A biological basis for international prairie conservation. Oryx, **46**, 64–68.
Tallmon DA, Luikart G, Waples RS (2004) The alluring simplicity and complex reality of genetic rescue. *Trends in Ecology & Evolution*, **19**, 489–496.
Taylor SE, Oyler-McCance SJ, Quinn TW (2003) Isolation and characterization of microsatellite loci in Greater sage-grouse (*Centrocercus urophasianus*). *Molecular Ecology Notes*, **3**, 262–264.
Thacker ET, Gardner DR, Messmer TA, Guttery MR, Dahlgren DK (2012) Using gas chromatography to determine winter diets of greater sage-grouse in Utah. *The Journal of Wildlife Management*, **76**, 588–592.
Tucker JM, Schwartz MK, Truex RL, Wisely SM, Allendorf FW (2014) Sampling affects the detection of genetic subdivision and conservation implications for fisher in the Sierra Nevada. *Conservation genetics*, **15**, 123–136.
U.S. Fish and Wildlife Service (2010) 12-month findings for petitions to list the greater sage-grouse (*Centrocercus urophasianus*) as threatened or endangered. *Federal Register*, **75**, 13910–14014.
U.S. Fish and Wildlife Service (2010) Endangered and threatened wildlife and plants; 12-month findings for petitions to list the Greater Sage-Grouse (Centrocercus urophasianus) as Threatened or Endangered; proposed rule. *Federal Register*, **75**, 13910–14014.
U.S. Fish and Wildlife Service (2013) Greater sage-grouse (*Centrocercus urophasianus*) conservation objectives: final report. U.S. Fish and Wildlife Service, Denver, February 2013.
U.S. Fish and Wildlife Service (2015) 12-Month finding on a petition to list greater sage-grouse (*Centrocercus urophasianus*) as an endangered or threatened species. *Federal Register* , **80**, 59858–59942.
U.S. Fish and Wildlife Service (2015). Endangered and threatened wildlife and plants; 12-month finding on a petition to list Greater Sage-Grouse (*Centrocercus urophasianus*) as an Endangered or Threatened species; proposed rule. *Federal Register*, **80**, 59858–59942.
Urban DL, Keitt T (2001) Landscape connectivity: a graph-theoretic perspective. *Ecology*, **82**, 1205–1218.

Vähä J, Erkinaro J, Niemelä E, Primmer CR (2008) Temporally stable genetic structure and low migration in an Atlantic salmon population complex: implications for conservation and management. *Evolutionary Applications*, **1**, 137–154.
Van Oosterhout C, Hutchinson WF, Wills DP, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes*, **4**, 535–538.
Viricel A, Rosel PE (2014) Hierarchical population structure and habitat differences in a highly mobile marine species: the Atlantic spotted dolphin. *Molecular ecology*, **23**, 5018–5035.
Wade AA, McKelvey KS, Schwartz MK (Editors) (2015) Resistance-surface-based wildlife conservation connectivity modeling: Summary of efforts in the United States and guide for practitioners. USDA Forest Service General Technical Report **RMRS-GTR-333**.
Waits LP, Luikart G, Taberlet P (2001) Estimating the probability of identity among genotypes in natural populations: Cautions and guidelines. *Molecular Ecology*, **10**, 249–256.
Walker BL, Naugle DE, Doherty KE (2007) Greater sage-grouse population response to energy development and habitat loss. *The Journal of Wildlife Management*, **71**, 2644–2654.
Wallestad R, Eng RL (1975) Foods of adult sage grouse in Central Montana. *The Journal of Wildlife Management*, **39**, 628–630
Wallestad R, Schladweiler P (1974) Breeding season movements and habitat selection of male sage grouse. *The Journal of Wildlife Management*, **38**, 634–637.
Wambolt CL (1996) Mule deer and elk foraging preference for 4 sagebrush taxa. *Journal of Range Management*, **49**, 499–503
Waples RS, Gaggiotti OE (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, **15**, 1419–1439.
Warnock WG, Rasmussen JB, Taylor EB (2010) Genetic clustering methods reveal bull trout (*Salvelinus confluentus*) fine-scale population structure as a spatially nested hierarchy. *Conservation Genetics*, **11**, 1421–1433.
Watts D, Strogatz SH (1998) Collective dynamics of "small-world" networks. *Nature*, **393**, 440–442.
Webb SL, Olson CV, Dzialak MR, Harju SM, Winstead JB, Lockman D (2012) Landscape features and weather influence nest survival of a ground-nesting bird of conservation concern, the greater sage-grouse, in human-altered environments. *Ecological Processes*, **1**, 1–15.
Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. E*volution*, **38**, 1358–1370.
Whiteley AR, Fitzpatrick SW, Funk WC, Tallmon DA (2015) Genetic rescue to the rescue. *Trends in Ecology & Evolution*, **30**, 42–49.
Wiens JA (2007) Foundation papers in landscape ecology. Columbia University Press, New York, p 582.
Wood KA, Stillman RA, and Goss-Custard JD (2015). Co-creation of individual-based models by practitioners and modellers to inform environmental decision-making. *Journal of Applied Ecology*, **52**, 810–815.
Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**: 97–159.
Wright S (1943) Isolation by distance. *Genetics*, **28**, 114.
Wright S (1949) The genetical structure of populations. *Annals of Human Genetics*, **15**, 323–354.

Zink RM (2014) Comparison of patterns of genetic variation and demographic history in the greater sage-grouse (*Centrocercus urophasianus*): relevance for conservation. *Open Ornithology Journal*, **7**, 19–29.