Probabilistic graph models for landscape genetics

Brook Milligan
Conservation Genomics Laboratory
Department of Biology
New Mexico State University
Las Cruces, New Mexico 88003 USA
brook@nmsu.edu

OGRS 2016
October 13, 2016
Landscape genetics: the broad goal and central question

Identify the mechanisms by which landscape and environmental factors influence genetic and genomic variation.

Given data on intraspecific genetic variation across landscapes, what inferences are possible regarding the functional mechanisms and factors causing that variation?
Pinus ponderosa

Third in production volume (4.5 million m$^2$); second in value (WWPA 2001)
Pinus ponderosa geographic range

Hipkins et al. (2013)
Pinus ponderosa mtDNA genotypes

Hipkins et al. (2013)
Extrapolating allele frequencies across the landscape
What tools are available for landscape genetics?

What tools are available for landscape genetics?

4.6 Recommendations for optimal sampling strategies in landscape genetics, 71
4.7 Conclusions and future directions, 73
References, 74

5 BASICS OF SPATIAL DATA ANALYSIS: LINKING LANDSCAPE AND GENETIC DATA FOR LANDSCAPE GENETIC STUDIES, 77
Helene H. Wagner and Marie-Josée Fortin

5.1 Introduction, 77
5.2 How to model landscape effects on genetic variation, 84
5.3 How to model isolation-by-distance, 93
5.4 Future directions, 95
Acknowledgments, 96
References, 96

PART 2: METHODS

6 SIMULATION MODELING IN LANDSCAPE GENETICS, 101
Erin Landguth, Samuel A. Cushman, and Niko Balkenhol

6.1 Introduction, 101
6.2 A brief overview of models and simulations, 101
6.3 General benefits of simulation modeling, 102
6.4 Landscape genetic simulation modeling, 103
6.5 Examples of simulation modeling in landscape genetics, 104
6.6 Designing and choosing landscape genetic simulation models, 108
6.7 The future of landscape genetic simulation modeling, 111
References, 111

7 CLUSTERING AND ASSIGNMENT METHODS IN LANDSCAPE GENETICS, 114
Olivier François and Lisette P. Waits

7.1 Introduction, 114
7.2 Exploratory data analysis and model-based clustering for population structure analysis, 115
7.3 Spatially explicit methods in landscape genetics, 119

7.4 Spatial EDA methods: spatial PCA and spatial factor analysis, 119
7.5 Spatial MBC methods, 120
7.6 Habitat and environmental heterogeneity models, 121
7.7 Discussion, 123
References, 125

8 RESISTANCE SURFACE MODELING IN LANDSCAPE GENETICS, 129
Stephen F. Spear, Samuel A. Cushman, and Brad H. McRae

8.1 Introduction, 129
8.2 Techniques for parameterizing resistance surfaces, 133
8.3 Estimating connectivity from resistance surfaces, 137
8.4 Statistical validation of resistance surfaces, 139
8.5 The future of the resistance surface in landscape genetics, 142
8.6 Conclusions, 144
References, 144

9 GENOMIC APPROACHES IN LANDSCAPE GENETICS, 149
Andrew Storfer, Michael F. Antolin, Stéphanie Manel, Bryan K. Epperson, and Kim T. Scribner

9.1 Introduction, 149
9.2 Current landscape genomics methods, 150
9.3 General challenges in landscape genomics, 157
9.4 Spatial autocorrelation, 157
9.5 Applications of landscape genomics to climate change, 159
References, 160

10 GRAPH THEORY AND NETWORK MODELS IN LANDSCAPE GENETICS, 165
Melanie Murphy, Rodney Dyer, and Samuel A. Cushman

10.1 Introduction, 165
10.2 Background on graph theory, 167
10.3 Landscape genetic applications, 170
10.4 Recommendations for using graph approaches in landscape genetics, 175
10.5 Current research needs, 176
10.6 Conclusion – potential for application of graphs for conservation, 176
References, 177
A comprehensive theory for landscape genetics is currently missing

One aspect related to all of the above conclusions is the current lack of a comprehensive theory that links landscape heterogeneity in space and time to patterns in neutral and adaptive genetic variation, by considering the many different processes that affect gene flow, drift, and selection.

— Balkenhol et al. (2016), page 252
Inspiration for a comprehensive theory: Bayesian analysis

**Landscape genetics:** Given data on intraspecific genetic variation across landscapes, what inferences are possible regarding the functional mechanisms and factors causing that variation?

**Bayesian analysis:** Given a set of interrelated random variables and data on a subset of them, what is the probability distribution for the remainder?
Assigning a single sampled allele to a population

Question of interest  From which population was this allele sampled?
Assigning a single sampled individual to a population

Question of interest  From which population was this individual sampled?
Assigning alleles from two sampled individuals to a population

Question of interest  From which population were these individuals sampled?
Absence of genetic admixture  Populations are completely isolated so an individual’s entire genome has ancestry in only one population so assignment ($Z$) depends on the individual.
Structure with admixture: Pritchard et al. (2000)

Genetic admixture  Interbreeding between formerly isolated populations leading to individuals with ancestry in more than one population so assignment (Z) depends on the allele.
Structure with correlated populations: Falush et al. (2003)

Correlated populations Sampled populations share a common ancestor $P_A$. 
Geneland: Guillot et al. (2005)

Geneland  Both genetic and spatial data.
Sparse coverage of landscape genetics models

Space of landscape genetics models

- Structure correlated pops.
- Geneland
- Structure with admixture
- Structure no admixture
Adding a new model to landscape genetics models

Space of landscape genetics models
Gene flow
Structure correlated pops.
Geneland
Structure with admixture
Structure no admixture

Space of landscape genetics models
A model of gene flow

Gene flow  Populations are genetic mixtures due to gene flow, so cycles exist.
A model of gene flow: C++

```c++
for (auto l : geographic_region)
{
    // alleles
    for (auto i : individuals(l))
        for (auto a : alleles(l,i))
            allele(l,i,a) =~ multinomial(P(l));

    // sampled populations
    P(l) =~ descendant(P_a(l),F(l));

    // ancestral populations
    P_a(l) = gene_flow(P(l),l,l,m);
    for (auto n : neighbors(l))
        P_a(l) += gene_flow(P(n),l,n,m);
}
```
Complete coverage of landscape genetics models

Space of landscape genetics models
Gene flow
Structure correlated pops.
Geneland
Structure with admixture
Structure no admixture

Space of landscape genetics models
Interested? Excited? Have data?

This is all possible only because of an underlying C++ library. Please collaborate! Please help with development!

- Applications to data
- Additional probability distributions
- Additional variates: e.g., genotypes, spatial layers
- Testing
- Documentation
- ...

Contact:
Brook Milligan
Department of Biology, New Mexico State University
brook@nmsu.edu
A foundation for a comprehensive theory

- Graphical models of arbitrary structure
- Random variables representing genetic variation
- Random variables representing spatial covariates
- Probability distribution influenced by population genetic processes
  - Gene flow
  - Natural selection
  - Finite population size
  - Spatial dependence

*Given a complete toolbox covering the model space, the task for landscape geneticists is to explore alternative models to discover generalities in the genetic–environment interaction that affect patterns of genetic or genomic variation.*
Ideal software requirements for landscape genetics modeling

- Arbitrary graph structures
- Broad range of data types for landscape genetics: genetic, spatial
- Easy accommodation of third-party data types
- Extensible algorithms
- High performance execution environment
- Open-ended and scalable
- High levels of abstraction to improve ease-of-use
Open source graph model software

<table>
<thead>
<tr>
<th>Name</th>
<th>Graph types</th>
<th>Primitive variables</th>
<th>Preprocessing</th>
<th>Implementation language</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenBUGS</td>
<td>DAGs</td>
<td>scalars</td>
<td>interpreted</td>
<td>Component Pascal</td>
</tr>
<tr>
<td>JAGS</td>
<td>DAGs</td>
<td>scalars</td>
<td>interpreted</td>
<td>C++</td>
</tr>
<tr>
<td>Stan</td>
<td>DAGs</td>
<td>scalars</td>
<td>interpreted</td>
<td>C++</td>
</tr>
<tr>
<td>Darwin</td>
<td>FGs</td>
<td>scalars</td>
<td>compiled</td>
<td>C++</td>
</tr>
<tr>
<td>HYDRA</td>
<td>DAGs, MRFs, FGs, HMMs</td>
<td>Java classes</td>
<td>compiled</td>
<td>Java</td>
</tr>
<tr>
<td>Infer.NET</td>
<td>FGs</td>
<td>C# classes</td>
<td>compiled</td>
<td>C#</td>
</tr>
<tr>
<td>JavaBayes</td>
<td>DAGs</td>
<td>scalars</td>
<td>interpreted</td>
<td>Java</td>
</tr>
<tr>
<td>libDAI</td>
<td>FGs</td>
<td>discrete</td>
<td>compiled</td>
<td>C++</td>
</tr>
<tr>
<td>Mocapy++</td>
<td>DAGs, HMMs</td>
<td>C++ classes</td>
<td>compiled</td>
<td>C++</td>
</tr>
<tr>
<td>Nimble</td>
<td>DAGs</td>
<td>scalar</td>
<td>interpreted</td>
<td>C++</td>
</tr>
<tr>
<td>OpenGM</td>
<td>DAGs, MRFs, FGs</td>
<td>discrete</td>
<td>compiled</td>
<td>C++</td>
</tr>
<tr>
<td>PNL</td>
<td>DAGs, MRFs</td>
<td>C++ classes</td>
<td>compiled</td>
<td>C++</td>
</tr>
<tr>
<td>RISO</td>
<td>DAGs</td>
<td>Java classes</td>
<td>compiled</td>
<td>Java</td>
</tr>
<tr>
<td>Vibes</td>
<td>DAGs</td>
<td>scalar</td>
<td>compiled</td>
<td>Java</td>
</tr>
</tbody>
</table>

See https://www.cs.ubc.ca/~murphyk/Software/bnsoft.html
Open source graph model software

Limitations:

- Limited types of graphs
- Limited types of data: genetic and spatial data types are lacking
- Performance often constrained by execution environment
- Difficult to use

See the following for a list of graph analysis software:
https://www.cs.ubc.ca/~murphyk/Software/bnsoft.html