Population Structure

Guillot, Gilles; Orlando, Ludovic

Publication date: 2013

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
INTRODUCTION
When considering a population of organisms with sexual reproduction, assuming that mating takes place at random can be convenient from a modelling and computational point of view but it is often not realistic as this may be impeded by various pre- or post-zygotic barriers. The term population structure (or population subdivision) usually refers to the patterns in neutral genetic variation that result from the past or present departure from panmixia of a population. Its study is of interest to evolutionary biologists because it can provide insight into the causes and consequences of restricted gene flow in natural populations. The general populations genetics background is presented by Hartl and Clark 1988 and Nielsen and Slatkin 2013, connection to ecology is presented by Hanski and Gaggiotti 2004 while more detailed investigation of the role of geography is investigated by Epperson 2003 and Rousset 2004. Population structure plays also a central role in biomedical applications because it acts as a confounding factor in the study of the statistical association between a genetic variant and a phenotype. This (and many other aspects) are presented by Balding et al. 2007. Question specific to conservation biology are treated by Allendorf et a. 2012.

The topic of population structure relates to several other questions covered by the present series of commented bibliographies, in particular Landscape Ecology, Conservation Genetics, Population Genetics, Geographic variation, Phylogeography, Interpretation of Phylogenetic Trees, Metapopulations & Spatial Population Processes, Hybrid Zones, and lastly Dispersal.

General overviews
A reference text book on basic population genetics, including on population subdivision. Suitable for any undergraduate students in evolutionary biology.

Reviews methods and applications on spatially structured populations with a focus on simulation-based inference methods.
Summarizes theories used to describe geographically structure population. Written in a fairly mathematical style. Suitable for graduate students in evolutionary biology or statistics.

An overview of various aspects of population structure from an ecological perspective.

An account of the state of the art on the many aspects of populations genetics, with statistical development. Includes a chapter on Inferences from Spatial Population Genetics by F. Rousset and a chapter on Analysis of Population Subdivision by L. Excoffier. The concise style imposed by the book format and the statistical content makes it best suited to graduate students and researchers.

A mine of information in genetics and ecology in a clearly written style with a nice balance between data and methods. A lot of material about population structure, its causes, consequences and how it can be dealt with from a conservation perspective.

Nielsen, R and Slatkin, M., 2013 An introduction to population genetics, Sinauer.
A modern and clearly written introduction to population genetics with concrete examples. Provides introduction to key concepts in an intuitive yet rigorous way without mathematical details.

**Defining population structure**
The term population structure is defined negatively as a departure from random mating and can therefore correspond to different processes and patterns. Departure from panmixia can be due to the existence of demes (such as flocks, colonies or herds) with restricted within- and between-demes gene flow (Manel et al. 2005, Wapples and Gaggiotti 2005). In a limiting case, a deme consists of a single individual, and reproduction occurs preferentially between geographically close individuals, a situation leading to a continuous population structure (Guillot et al. 2009). Modelling aspects are discussed by Hey and Machado 2003, and Lawson 2013.

A review that outlines the main models of population structure and focuses on methods
for detecting subpopulations and their recent common history.

The two papers above discuss the various concepts associated to demic population structure, give an overview of models and algorithms for statistical inference and assess the accuracy they can achieve.

A review with a focus on statistical inference of geographical structure.

The most recent review paper. Includes a discussion of methods for the detection of structure for large SNP data.

Journals
The main journals in the fields are Genetics, PLOS Genetics, Molecular Biology and Evolution, Molecular Ecology and Molecular Ecology Resources. Other journals relevant to the topic of population structure include Applied Ecology, The American Naturalist, Bioinformatics, Conservation Genetics, Evolution, Genome Research, Heredity, Methods in Ecology and Evolution, Nature Genetics, Nature Reviews Genetics, Systematic Biology and Trends in Ecology and Evolution.

Genetics
Edited and published by the Genetics Society of America. Has published the seminal papers by the fathers of modern population genetics; has long been considered the leading journal in the field.

PLOS Genetics
A peer-reviewed, open-access, electronic-only journal that publishes contributions in all areas of genetics. Has played a key role in the publication of the most innovative method papers related to population structure lately.

Molecular Biology and Evolution
A monthly peer-reviewed scientific journal edited by the Society for Molecular Biology and Evolution and published by Oxford University Press. It publishes work in the intersection
of molecular biology and evolutionary biology.

**Molecular Ecology and Molecular Ecology Resources**

*Molecular Ecology* is a twice-monthly scientific journal covering investigations that use molecular genetic techniques to address questions in ecology, evolution, behavior, and conservation. Its sister journal *Molecular Ecology Resources* publishes articles on technical methods, computer programs, and genomic resource development.

**Historical perspective**

The seminal contributions are due to Wright in 1922, 1931, and 1943.


Wright S. (1931) Evolution in Mendelian populations. *Genetics, 16*, 97–159. **Introduces and analyses a set of assumptions now known as the Wright-Fisher model which became a cornerstone of population genetics and on which more complex models can be built.**

Wright, S. Isolation by distance. *Genetics 28*, 114–138 (1943). **Introduces the first spatially explicit model of geographically structured populations.**

**STATISTICAL INFERENCE ABOUT POPULATION STRUCTURE**

The tasks involved in the study of population structure can be split into (i) quantifying and testing the existence of a structure among various pre-defined sub-populations, (ii) detecting the presence of 'pure' or admixed sub-populations, (iii) inferring the evolutionary history of pre-defined populations (in form of a series of splitting and merging events, with possible variation of population sizes and amount of gene flow) and lastly, (iv) jointly estimating the presence of sub-populations and their history.

There has been a permanent tension in statistical genetics between the interest of developing increasingly complex models to extract as much information as possible from the data on the one hand, and on the other hand the recognition that complex models lead to computational difficulties, especially when implemented on datasets of ever-growing sizes. This is reflected by the existence of various methods ranging from description and quantification (estimation of $F$ statistics or factor analysis methods), to complex models based on coalescent theory. In between stand clustering models which achieve a trade-off between biological realism and inference complexity. They borrow to the mainstream statistics literature but also incorporate features of biologically-inspired
models.

**Measuring and testing structure**
This aspect is reviewed by Weir and Hill 2002 and Holsinger and Weir 2009.


**Isolation by distance models**
There is a long tradition of using stochastic models to describe isolation by distance. Probability models of isolation by distance with analytical results include those introduced by Malécot 1975 and Slatkin and Maruyama 1975, Rousset 1997 and Rousset 2000. Epperson 1995 and Epperson and Li 1997 report results based on simulations. Overviews can be found in Hardy and Vekemans 1999 and Charlesworth et al. 2003.

Some of the earliest attempts to compute probabilities of identity of gene lineages or correlations in allele frequencies as a function of distance.

On the basis of simulated data, describes the expected patterns in spatial genetic variation under some models of isolation by distance.

Propose a method to infer dispersal parameters based on the analysis of variation of the pairwise population differentiation with geographic distance.

Investigate theoretically the link between population genetics model of isolation by distance and more descriptive approaches based on spatial auto-correlation functions.

Develops earlier results to make inference of dispersal parameters from data sampled 'continuously' in space instead of in form of sub-populations.

A review article and key reference on the subject.

Detecting demic structure
The most powerful methods to detect demic structure do not attempt to resolve the population history, but seek instead either to estimate admixture proportions or to compute principal components which in both cases synthesizes the population history. The various methods differ by the assumption made about the dependency between loci (with or without linkage equilibrium), the possibility to consider admixed individuals, inbreeding and the technique used to estimate parameters.

The most recent paper to date that reviews and compare methods to perform inference of population structure.

Bayesian mixture models
Bayesian mixture models originate from mainstream statistics. They assume the existence of extant or past clusters at Hardy-Weinberg equilibrium that are characterized by cluster-specific allele frequencies. These models were pioneered by Pritchard et al. 2000, Dawson and Belkhir 2001 and Falush et al. 2003.

A landmark paper that introduces a Bayesian model-based approach to cluster individuals into populations assuming Hardy-Weinberg equilibrium and linkage equilibrium. Method implemented in the widely used computer program Structure.

These two papers tackle the same problem as the paper above by Pritchard et al., with a
slightly different computational strategy and the extra difficulty to estimate the number of clusters directly. Methods implemented in the programs Partition and BAPS.


Demic structure with inbreeding
The methods above have been modified by François et al. 2006 and Gao et al. 2007 to account for departure from Hardy-Weinberg equilibrium due to within-cluster inbreeding.

François O, Ancelet S, Guillot 2006. G. Bayesian clustering using hidden Markov random fields. Genetics, 174, 805-816. Proposes an alternative to the spatial component of the Geneland program and also generalizes the likelihood model to take into account inbreeding.

An alternative approach to the method above (for genetic data only). Method implemented in the Instruct program.

Spatial models
The power of the methods above can be increased by embedding them in a spatially explicit model that makes use of in the information contained in the location of the individuals sampled. The non-admixture case is considered by Guillot et al. 2005 and an admixture model is presented by Durand et al. 2009. The various spatial models are discussed and compared by Safner et al. 2011 and Blair et al. 2012.

A spatially explicit model to make use of geo-referenced genetic data and enable increased power in statistical inferences. Method implemented in the Geneland program.

Introduces a statistical model of spatial variation of admixture coefficients


Two comparisons of the various computer programs used in evolutionary biology to infer demic population genetic structure.

**Likelihood-based inference**

Many of the models for the detection of demic structure discussed above are based on a Bayesian model which is a statistical way to inject information about biological processes by way of probability distributions. This approach proves to be too computer demanding when the number of genetic markers exceeds a few hundreds. An alternative strategy consists in optimizing a likelihood function which is the approach taken by Tang et al. 2005, Tang et al. 2006 and Alexander et al. 2009.


*A model reminiscent of the approaches above, but based on likelihood than posterior distribution. Method implemented in the Frappe program.*


*A fast likelihood-based approach to infer ancestry blocks. Implemented in the Saber program.*


**Phased data**

Model based clustering algorithms described above work with un-phased data and do not make use of the information about linkage disequilibirum. This is tackled by the recent work by Lawson et al. 2012.


*Makes use of the information available in haplotypes obtained with phasing algorithm. Provides information about fine scale population structure with unprecedented accuracy*
The new standard for massive SNP datasets.

**Factor analysis models**
Bayesian models have been the reference methods for the analysis of micro-satellite data until the advent of SNP datasets for which they showed some limitations due to computing times. This has been addressed by the use of principal component analysis (PCA) methods, for example by Patterson et al. 2006 and Engelhart and Stephens 2010. The decrease in computing is made at the expense of the ease of interpretation of analysis outputs, an aspect discussed by Reich et al. 2008, Novembre and Stephens 2008 and McVean 2009.

*Shows how population structure can be investigated and tested through classical principal components methods at the era of large SNP datasets. Method implemented in the Eigensoft program.*


*Discuss the interpretation of principal components in population genetics in terms of admixture events or population migration events.*

*Develops a framework bridging the gap between cluster models and principal components approaches.*

**Inferring history of populations**

A more statistically oriented review with view to the nested clade analysis method controversy.

Although $\delta_a$ is primarily dedicated to demographic inference, this software can handle multi-population models and estimate from the allele frequency spectrum migration parameters and admixture stages across populations.

Description of distance-based and model-based inference methods for reconstructing the routes of introduction and expansion of invasive species, with reference to examples from the literature.

The first method to date aimed at jointly detecting the presence of sub-population and estimating their common history.

An alternative approach to the method of Choi et al. above suitable for the analysis of large SNP datasets. Method implemented in the Treemix program.

Simulation studies showing the difficulty in disentangling population structure and demographic bottleneck signatures, providing guidelines and solution for correct inference in a conservation context.

Investigates the consequences of range contractions and shifts on pattern of molecular diversity.
diversity. Discovers that the latter will be shaped by the species’ dispersal abilities as well as by the speed of the environmental change.


Introduce allele surfing, a phenomenon that occurs at the wavefront of range expansion and that has consequences on geographic patterns of allele-frequency differentiation.

**Detecting admixture and structure disruption**

Admixture between differentiated populations can occur and disrupt pre-existing population boundaries. Recent interest in admixture has grown considerably with the discovery that non-African human genomes harbour detectable levels of admixture with archaic hominins such as Neanderthal and Denisovans.

Introduces the D-statistics that exploits any excess of shared derived polymorphism between groups as a test for admixture.

gPhoCS is a Bayesian software using Markov Chain Monte Carlo to jointly sample model parameters and genealogies for inferring ancestral population sizes, population divergence times, and migration rates from individual genome sequences.

This article presents the theories underlying the ADMIXTOOLS package and a number of statistical tests for detecting admixture from genome-wide data.

**USING STRUCTURE TO DETECT THE ORIGIN OF INDIVIDUALS**

If some information about allele frequencies of population of known origin is available, one can characterize existing structure and use this information to assign individuals of unknown ancestry. This is the approach taken by Rannala and Mountain 1997, Cornuet et al. 1999 and Anderson and Thompson 2002. Acknowledging that a sample of unknown of
origin may not have direct ancestry in any of the population of known geographic origin, Wasser et al 2004 developed a method for continuous spatial assignment. Yang et al. 2012 proposed an alternative models to deal with large SNP data. A review from an epidemiology perspective is provided by Remais et al. 2013.

The first method to use a formal Bayesian method to assign individuals to some known sub-populations.

A comparison showing that Bayesian methods outperforms other existing methods.

Generalizes existing methods to deal with hybrid individuals

A method to perform continuous spatial assignment of individuals.

Reviews how genetic data can be used in epidemiology to infer the origin of pathogens.

Shows how a simple logistic curve can discriminate between signal and noise in the variation of allele frequencies and enable fast inference of geographical origin without reference samples of known geographic origin.

**INVESTIGATING ECOLOGICAL FACTORS CAUSING STRUCTURE**
An important question in evolutionary biology consists in understanding the causes of population structure, a question related to the central goal of phylogeography but referred to as landscape genetics when considered at a smaller space and time scale. The topic has been reviewed by Manel et al. 2003, Storfer et al. 2006, Diniz-Filho et al. 2009, Storfer et al. 2010 and Manel and Holderegger. Notable modelling innovation to assess the role of geography have been made by McRae 2006, McRae and Beier 2007

The first paper to review systematically the various questions and methods in the field.

An original model-based approach to go beyond isolation by distance models and assess how landscape affects gene flow.

Storfer, A., Murphy, M. A., Spear, S. F., Holderegger, R., & Waits, L. P. 2010. Landscape genetics: where are we now?. Molecular Ecology, 19(17), 3496-3514.
Review the exponentially growing method and applied literature in the field of landscape genetics.

Show case examples of population structure changes and discusses possible management guidelines in the face of current global warming.

Demonstrates that land clearing following intense agriculture has affected the population structure of the ornate dragon lizard, increasing fragmentation, population differentiation and limiting local genetic diversity.

Investigates which landscape factors amongst climate, land cover and degree of anthropogenic disturbance, affect gene flow in three regions spanning the range of the Cope’s giant salamander, with implication for conservation biology.

Reviews complex models in spatial analyses of genetic variation that account for autocorrelation issues.


Reviews recent conceptual and methodological advances in landscape genomics, with emphasis on future challenges for understanding temporal and geographic patterns of gene-flow and population structure.


Two statistical models for quantifying the effect of distance relatively o that of other barriers on differentiation.

**ADAPTATION AND NON-NEUTRAL CAUSES UNDERLYING STRUCTURE**

Within a given species, populations facing different conditions can become adapted to local environments, creating a geographical structure at the genomic loci underlying adaptive phenotypic traits. This topic goes far beyond the present bibliography but interested readers can refer to Nielsen et al. 2005


Non-mathematical description of the issues involved in detecting selection from DNA sequences and genome-wide genetic datasets.


**SELECTED APPLICATIONS**

Virtually any single study involving population genetics data starts with an attempt to detect structure. Here are only mentioned studies that received a special attention due to the striking results reported or because of their methodological relevance. This includes studies in human genetics or in model species such as *A. thaliana* and also studies in large mammals or marine species for which factors that affect gene flow are poorly understood.

**Humans**

The first application of the Structure program. Provides key insights about the organization of neutral genetic variation in humans at the synoptic scale. Contributed widely to demonstrate the utility of the program. Considered also as a milestone paper in medical genetics for the design of association studies.


A follow-up to the previous reference that investigates the role of the spatial sampling design in the ascertainment of population structure and discusses the interpretation of differences of allele frequencies among populations.


Two papers that received a considerable attention for the striking fine scale pattern of genetic variation observed on a large SNP dataset in Europe. The latter shows also how principal component analysis can be used to extract information about geographic origin of samples.


Two seminal papers presenting draft genomes of archaic hominins and providing the first direct evidence for admixture with ancient modern humans.


Detect signatures of adaptation to life at high-altitude in two independent human populations.

**Apes**


An investigation of the structure of our closest living relative with a study design similar to that of recent studies that have elucidated human structure.


Illustrates the usefulness of continuous assignment methods on genetically structured populations to detect illegal hunting.


Further results on the structure of the chimpanzee population in West Africa.

Unravels the evolutionary history underlying population differentiation among wolves and coyotes.

**Carnivore mammals**

One of the first study to use a model-based clustering method to detect structure and barriers to gene flow in mammals.

A study of the effect of climate and environmental conditions on genetic structure.

Those two studies show the effect of landscape on genetic structure and gene flow in two large mammals.

Describes how a range of landscape barriers has shaped the population structure of brown bears across North-Western Eurasia.

Sacks B, Bannasch DL, Chomel BB, Ernst H 2008 Coyotes demonstrate how habitat specialization by individuals of a generalist species can diversify populations in a heterogeneous ecoregion. Molecular Biology and Evolution, 25, 1354-1395.  
One of the first studies to investigate the effect of habitat specialization at a local scale with model-based clustering methods.

**Herbivore mammals**


Frantz AC, Tigel Pourtois J, Heuertz M et al. 2006 Genetic structure and assignment tests demonstrate illegal translocation of red deer (Cervus elaphus) into a continuous population. Molecular Ecology, 15, 3191–3203  
Shows how model-based approach can be an aid for monitoring individuals with a view to populations management.

Pope L, Pope BR, Wilson G et al. 2007 Genetic evidence that culling increases badger
Shows how the analysis of population genetics structure can give insight into the spread of pathogens.

Other terrestrial animals
Investigates the (absence of) effect of a potential barrier to gene flow and all suggests a general methodology to validate the output of clustering programs.

A timely discussion of the interpretation of the outputs of clustering programs in light of isolation by distance.

Amphibians
Illustrates a common difficulty associated MCMC-based Bayesian clustering model, namely the frequent incongruence between programs outputs.

Birds
The first paper to evaluate the accuracy of the Structure program on real data in a controlled set-up (20 chicken breeds).

Shows how the analysis of 20 microsatellite markers for 1000 individuals can provide insight into the effect of habitat fragmentation of an endangered bird species.

Illustrates the power of continuous assignment methods for the study of bird migration and also shows how they can be improved by combining genetic data to isotopic data.
Aquatic species
Discusses the use of genetic markers for the study of marine organisms in a conservation or management setting.

A key paper in 'seascape' genetics litterature as it provides a method to predict gene flow in a marine environment.

One of the first studies to undertake the analysis of a large dataset of a marine mammal. Provides also key insight into the interpretation of inferred clusters in presence of clinal variation.

Explores the effect of variation in water density on seven fish species.

Investigates population differentiation of harbour porpoise in relation to morphology, habitat and resources.

Uses a RAD-seq approach to investigate the population structure of the threespine stickleback fish, a species that has repeatedly undergone parallel phenotypic and genetic differentiation following the colonization of freshwater habitats from oceanic sources.

Insects
Shows how the study of population structure can highlight the mode of propagation of a pathogen.

Utilization of RAD-sequencing for unraveling the population structure and phylogeographic history of the pitcher plant mosquito.


Investigates one Columbian hybrid zone between two diverging Heliconius butterflies subspecies exhibiting different wing colour patterns.

Plants

One of the first examples in plant genetics of how a clustering program can be used in an association study to control for the confounding effect of population structure.


A careful study of the important question of pollen dispersal investigated with molecular markers.


Analyzes how kinship systems among farmers in Gabon impinge on the dynamics of crop genetic diversity by investigating the relationships between marriage exchanges and seed exchange networks.


Detailed studies of the population structure of A. thaliana at global and local scales.
Reviews plant studies in landscape genetics, with emphasis on selection-driven geographical patterns of genetic variation.

Identifies Scandinavia as a postglacial contact zone between divergent lineages of Populus tremula, the widespread forest common aspen tree, and reveals signatures of adaptation within admixed populations.

Micro-organisms
The first application of the Structure program to bacteria. Relates inferred genetic structure in bacteria to routes of human migration.

Suggests a method to gain insights about the spatial scales involved in the dispersal process.

STATISTICAL AND COMPUTATIONNAL TOOLS

Statistical background
The modelling of population structure is a technical subject. In addition to probability theory (Durrett 2008), it has relied extensively on several recent statistical techniques such as Bayesian modelling (Beamont and Rannala 2004), Monte Carlo simulation (Gilks et al. 1996) and approximate Bayesian computations (Beaumont 2010, Csillery et al. 2010).

A reference book on the most widely used computational technique in the study of population structure. Does not contain specific material in genetics or evolutionary biology but addresses most of the questions encountered by users of MCMC-based clustering programs.

Reviews Genetics, 5(4), 251-261.
Introduces all the key concepts underlying the increasingly important field of Bayesian modelling, from a geneticist perspective.

The mathematical background to many models and computational statistical methods in population genetics. Includes a chapter on spatial models of structure known as stepping stone models.

These three review papers introduce ABC as the most important alternative to MCMC computation with mathematical background and connection to the evolutionary biology litterature.

Coalescent theory
Theoretical results on populations genetics models are sometimes easier to obtain if one thin backward in time, which is the key idea behind coalescent theory (Schierup and Wiuf 2004, Wakeley 2009).

An excellent introduction to coalescent theory, showing clear mathematical and theoretical derivations as well as a diversity of examples.

One of the most authoritative references on the models and computational methods used to infer evolutionary scenarios.

Computer resources
A well-informed catalogue of computer programs.