Molecular Ecology (2004) **13**, 1713–1727 doi: 10.1111/j.1365-294X.2004.02177.x

Population Graphs: the graph theoretic shape of genetic structure

RODNEY J. DYER and JOHN D. NASON
Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa 50011, USA

Abstract

Patterns of intraspecific genetic variation result from interactions among both historical and contemporary evolutionary processes. Traditionally, population geneticists have used methods such as *F*-statistics, pairwise isolation by distance models, spatial autocorrelation and coalescent models to analyse this variation and to gain insight about causal evolutionary processes. Here we introduce a novel approach (Population Graphs) that focuses on the analysis of marker-based population genetic data within a graph theoretic framework. This method can be used to estimate traditional population genetic summary statistics, but its primary focus is on characterizing the complex topology resulting from historical and contemporary genetic interactions among populations. We introduce the application of Population Graphs by examining the range-wide population genetic structure of a Sonoran Desert cactus (*Lophocereus schottii*). With this data set, we evaluate hypotheses regarding historical vicariance, isolation by distance, population-level assignment and the importance of specific populations to species-wide genetic connectivity. We close by discussing the applicability of Population Graphs for addressing a wide range of population genetic and phylogeographical problems.

Keywords: genetic structure, graph theory, networks, phylogeography

Received 10 September 2003; revision received 10 January 2004; accepted 13 February 2004

Introduction

The amount and geographical patterning of genetic variation within species is an evolutionary consequence of several historical and contemporary processes including vicariance, range expansion, gene flow and fragmentation (Slatkin 1985; Riddle 1996; Taberlet *et al*. 1998; Hewitt 2001; Sork *et al*. 2001). Quantifying this variation and in turn the extent to which these processes have acted in shaping genetic structure is a chief concern of the field of evolutionary biology. Numerous theoretical models and methodological procedures have been developed to address this goal, many of which are integral to the foundation of population genetic theory. A common conceptual feature of these approaches is an a priori definition of a hierarchically nested population model (for example in which populations are nested within geographical regions) that in turn guides the distillation of genetic differences among populations into single or few measures of average genetic differentiation. Although commonly employed for the analysis of genetic marker data, this general approach suffers from potential several shortcomings that are perhaps not well recognized.

With the goal of stimulating the development of new, hopefully effective, solutions we begin by considering the potential problems associated with an a priori statistical model-based approach characteristic of many traditional population genetic procedures. We then describe the development of a potential solution, a new graph-theoretic approach to the analysis of genetic marker data, which we call Population Graphs.

Population models and summary statistics

Wright's *F*-statistics (Wright 1951), AMOVA (Excoffier *et al*. 1992) and Nei's D (Nei 1972, 1978) are commonly used procedures describing genetic structure in terms of summary statistics. Even approaches relying upon pairwise measures of genetic differentiation, such as isolation by distance (Slatkin 1993; Rousset 1997), summarize relationships in terms of averaging statistics such as a regression slope or a

Correspondence: Rodney J. Dyer. Fax: (515) 294 1337; E-mail: rodney@iastate.edu

correlation coefficient. In contrast, evidence from population genetics, phylogeography and ecology indicates that the processes influencing the evolution of genetic structure vary in both time and space (e.g. Rhodes *et al*. 1996) leading to the expectation that the complexity of interpopulation relationships may not be captured adequately by a single or few averaging statistics. Both simplification and generality often lead to a better understanding of underlying processes, but they do not guarantee that we end up with a more concise interpretation of how evolution has shaped the data.

Even in the presence of a significant summary statistic, it is unclear that the associated a priori statistical model is one that best describes the patterns of variation in the population genetic data and, hence, reveals the correct signature of underlying evolutionary processes. A significant statistic signifies only that the proposed statistical model is sufficiently correct to reject the null hypothesis, often one of no differentiation. However, rejecting the null does not mean that the statistical model is defined precisely or correct. For any given value of a mean measure of differentiation (e.g. F_{ST}, Φ_{ST}) there is an infinite number of ways in which the differentiation between populations may be structured. We cannot distinguish, for example, between all populations being equally differentiated vs. subsets of populations being categorically different. The current repertoire of statistical genetic models lack the ability to inspect both the 'lack of fit' of our models to the observed data and the resulting response surfaces (e.g. Box *et al*. 1978; Box & Draper 1987). As a result, additional information is likely to be present in our genetic data sets that remain quantified inadequately.

Due of the overall complexity of evolutionary processes operating within and among populations of a species, analytical approaches that focus specifically upon the details of genetic interactions and relationships among populations, as opposed to an overall average effect, will provide a more integrated description of observed population genetic structure. Indeed, the extent to which we embrace this, often multivariate, complexity can impact the degree to which we can quantify successfully both intraspecific genetic variability (e.g. Smouse *et al*. 1982; Westfall & Conkle 1992) and the effects of specific evolutionary processes (e.g. Gavrilets 1997). The majority of methodological approaches employed thus far have allowed us to depict evolutionary affects on intraspecific genetic variation with a broad-brush stroke. If we are interested in examining how evolution has structured genetic variation at a finer level of granularity, however, we may often need to look at the problem using alternate perspectives.

Several models have been suggested recently that move beyond averaging summary statistics. These analytical approaches fall, roughly, into two broad categories. In the first category are model-based approaches that either extend or attempt to increase the precision of standard F_{ST}-based approaches. Often these models posit an a priori hierarchical model of population relationships. Examples of these include multivariate ordinations of pairwise differentiation statistics using principal coordinates and multidimensional scaling (e.g. Lessa 1990; Edwards & Sharitz 2000; Zhivotovsky *et al*. 2003) and methods aimed at maximizing among strata genetic variance such as SAMOVA (Dupanloup *et al*. 2002). The second, more recent category includes models based on the coalescence, such as GENETREE (Bahlo & Griffiths 2000) and MIGRATE (Beerli & Felsenstein 2001). In this study we present a third category of models for examining the distribution of intraspecific genetic structure using a multivariate graph-theoretic approach. This method, which we call Population Graphs, is free of an a priori model of population arrangement (e.g. we do not assume that populations are nested within either a hierarchical or bifurcating statistical model). Further, these relationships are not quantified in terms of averaging statistics or coalescence parameters, but in the form of a graphical topology (e.g. the pattern of genetic covariance structures among all populations), which captures the high dimensional genetic covariance relationships among all populations simultaneously rather than in a pairwise fashion. It is with this topology, which is easily visualized, that we address both a priori and *post-hoc* hypotheses regarding the distribution of intraspecific genetic variation.

To present the Population Graph framework, we introduce briefly some graph theoretic notation as well as the statistical methods of conditional independence. We then demonstrate how one utilizes the Population Graph framework by focusing on a set of general population genetic and phylogeographical hypotheses using an empirical example from the Sonoran Desert cactus, *Lophocereus schottii*, presented recently in Nason *et al*. (2002). With this nuclear marker data set, we show specifically how Population Graphs can be used to address hypotheses amenable to traditional analytical approaches, such as the effects of historical vicariance and isolation by distance. Moreover, we show how the graph-theoretic approach on which Population Graphs is based allows us to identify and test hypotheses regarding population genetic history that would not have been apparent from traditional analyses. We close by discussing the utility of Population Graphs for addressing a wide range of population genetic questions focusing on the dynamic nature of co-occurring evolutionary processes.

The model

The essential goal of Population Graphs is to allow the genetic data to describe the statistical relationships among all populations simultaneously. Our approach to defining these sets of relationships relies upon a graph theoretic interpretation of population genetic structure. We begin by

introducing some graph theoretic notations to describe the Population Graph framework.

Graph theory

Graph theory is based upon the association among mathematical sets. A set, S, is simply an unordered collection of objects. The number of objects in S, denoted as $|S|$, is called the *order* of S. A set that has no elements is called a null set and is written as $\varnothing$. We write $x \in S$ and $y \notin S$ to indicate that x is a member of S and y is not a member of S. If all the elements in S are also in another set, T, then S is a subset of T, which is denoted as $S \subset T$. On the other hand, when none of the elements of S are not in T, we write $S \not\subset T$ (i.e. S and T are disjoint). Finally, when examining the relationship between two sets, we can write the *intersection* of the two sets as $S \cap T$, representing the elements shared between both sets.

A graph (G) is a mathematical mapping of two disjoint sets, a set of nodes (V) and a set of edges (E), and is denoted as $G = \{V, E\}$ (Bollobás 2001). Within the Population Graph framework, the set of nodes, $v_i \in V$; $i = \{1, 2, \ldots, N\}$, represent N sample populations and the set of edges, $e_i \in E$; $i = \{1, 2, \ldots, K\}$, the set of multivariate measures of genetic covariance between populations. A saturated graph (Fig. 1A) is one in which all nodes are connected to all other nodes and has $K_{\text{total}} = N(N-1)/2$ edges. Graphs can be represented algebraically by an *incidence matrix*, which describes the *topology* of the graph. The topology, in its most general sense, is simply the pattern of node connections. The following incidence matrix, **A**, describes the graph in Fig. 1B. This incidence matrix has four rows and columns; that is, the order of the graph is $|V| = 4$. Each element of the incidence matrix $\mathbf{A} = \{e_{ij}\}$; $i, j = 1, 2, 3, 4$ denotes the presence (a nonzero value) or absence (a zero value) of an edge, e_{ij}, connecting nodes i and j.

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (1)$$

The strength of edges (i.e. the elements of **A**) can be variable, resulting in a weighted graph, as is used for Population Graphs. Incidence matrices for weighted graphs only differ from eqn 1 in that the presence of an edge is represented by a real number rather than a '1'. Furthermore, the incidence matrix, **A**, in this example is symmetric around the diagonal, although neither graphs in general nor Population Graphs require this. A nonsymmetrical Population Graph incidence matrix would represent anisotropy in genetic covariance, as may arise, for example, under isolation by distance in linear spatial arrangements of populations (such as along a river).

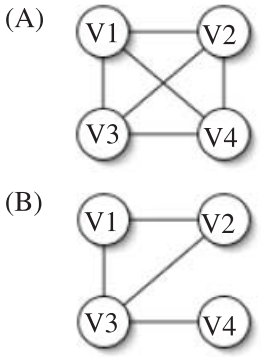

Fig. 1 Two simple graphs. (A) A saturated graph consisting of four nodes (circles) and six edges (lines connecting the nodes). (B) A nonsaturated graph with the same order as (A) but with a reduced edge set.

The topology of a graph has several prominent characteristics salient to the characterization of intraspecific genetic variation. First, a subgraph, G_1, of a larger graph G is one in which the entire edge and node sets of G_1 are included in G; that is, $G_1 = \{V_1, E_1\}$ is a subgraph of G if and only if $V_1 \subset V$ and $E_1 \subset E$. A graph can be partitioned into many subgraphs and we will return to this topic later when discussing the interpretation of Population Graph. Second, the *degree* of a node, $d(v_i)$, is simply the number of edges connected to it. If $d(v_i) = 0$ then the node i is called *isolated*. From a genetic perspective, graphically isolated populations would be independently evolving entities. Next, a graph may contain *cycles*, which are sequences of nonrepeating node–node paths that form a closed loop. Such cycles in Population Graphs may result from reticulate gene flow. The subgraph, G_1, consisting of the nodes $V = \{V_1, V_2, V_3\}$ and the edges $E = \{\{V_1, V_2\}, \{V_2, V_3\}, \{V_3, V_1\}\}$ in Fig. 1B is a cycle. A *tree* is simply a sequence of connected nodes and edges that have no cycles. Again, from Fig. 1B, the subgraph G_2 is a tree when the node set is $V = \{V_1, V_3, V_4\}$ and the edge set $E = \{\{V_1, V_3\}, \{V_3, V_4\}\}$. With this general introduction to graphs, we now describe how one can use graphs to represent population genetic structure using multilocus marker data.

Genetic coding

We assume that genetic markers have been assayed for a relatively large number of individuals spread across several populations. For the purposes of this study we focus specifically on codominant markers, such as allozymes and microsatellites, and for the sake of brevity address the issues of alternate genetic markers and sampling strategy in a subsequent manuscript (Dyer and Nason, in prep). Multilocus genotypes are translated into multivariate coding vectors following the mapping described in Smouse *et al*. (1982). This coding scheme produces a data vector, $\boldsymbol{p}_j$,

representing the multivariate genotype for the *j*th individual. The full data matrix, **X**, will have *m* columns, where *m* is the number of independent genetic variables (e.g. the number of independently assorting alleles across all loci), and n_{total} rows, representing all the individuals within the *N* populations. Even though this method produces a sparse matrix, with a reasonable number of loci and individuals it approaches multivariate normality (see Kempthorne 1969; Westfall & Conkle 1992, for a more complete discussion). The set of individuals within the *i*th population defines a multidimensional population centroid, $\bar{p}_i$, in *m*-dimensional genetic space. The centroid defines a unique *m*-dimensional coordinate representing the average genetic individual within the *i*th population. This distance metric has been chosen as it is identical to that used in the AMOVA framework (Excoffier *et al*. 1992), with which many researchers are familiar.

Once the *N* population centroids are defined from the set of all individuals, we then determine the contribution to the overall genetic variation due to differences among all pairs of populations. These pairwise distances define a distance matrix, **D**, whose off-diagonal elements, d_{ij}, define a statistical distance between populations in *m*-dimensional genetic space. This matrix can be used to describe a saturated Population Graph in which all nodes (populations) are connected to all other nodes by edges of weight d_{ij}. In Appendix I we demonstrate the general equivalence of **D** to the distance matrix in AMOVA (Excoffier *et al*. 1992) and show how Φ-statistics can be estimated directly from the saturated Population Graph. Because all nodes are interconnected, the topology of the saturated graph is not immediately informative as to interpopulation relationships. An informative topology is obtained from the minimal incidence matrix containing the smallest edge set that sufficiently describes the among population genetic covariance structure. The translation of the population distance matrix, **D**, to a minimal incidence matrix (as in eqn 1) relies upon the techniques of conditional independence.

Genetic covariance and conditional independence

Conditional independence forms the foundation of several commonly used parametric statistical models. For example, in multiple regression the inclusion of a variable requires the examination of its conditional independence to those variables already entered into the model (Whittaker 1990). In this case, we examine the Type III sums of squares for the newly entered variable and compare that against the sums of squares accounted for by the set of previously included predictor variables (e.g. Draper & Smith 1981). In addition to multiple regression, conditional independence is also used in such statistical methods as logistic regression and contingency tables (Whittaker 1990). In the context of Population Graphs, we are interested in determining the minimal edge set that sufficiently describes the total among population covariance structure. Following from the regression analogy, each edge in the Population Graph is analogous to a predictor variable. Our goal is to identify edges that do not aid in sufficiently describing the total among population genetic covariance structure. These edges can be 'pruned' from the graph without significantly decreasing the fit of the Population Graph model to the population genetic data.

There are several methods amenable to calculating conditional independence, including edge deviance, covariance selection and vanishing partial correlations (Dempster 1972; Whittaker 1990). For the sake of brevity, we shall use the method of edge deviance in this paper as it has recently been described in an evolutionary context by Magwene (2001) and defer the comparison of alternate methods to a subsequent manuscript. Briefly, the method proceeds as follows. First, we must translate our pairwise population distance matrix, **D**, into a covariance matrix, **C**. Gower (1966) showed the duality of distance and covariance matrices whereby the covariance between the *i*th and *j*th element of **D** is given by $c_{ij} = -0.5(d_{ij} - d_{i.} - d_{.j} + d_{..})$, where the subscripts *i* and *j* index the elements of **D** and the period subscript, '.', indexes the mean of the rows and/or column(s) in **D**. Following directly from Magwene's discussion of edge deviance, we proceed by inverting the covariance matrix producing what is called a *precision matrix* (Cox & Wermuth 1996). The *i*th diagonal element of the precision matrix is equal to $1/(1 - R_i^2)$, where R_i^2 is the multiple correlation coefficient between the *i*th and all remaining populations (Whittaker 1990). R_i^2 is also known as the coefficient of multiple determination (Sokal & Rohlf 1995), which is a measure of the proportion of variation in the *i*th population jointly accounted for by the remaining populations. To aid in interpretation, the precision matrix is standardized to a correlation matrix, **R**, using normal matrix routines (see Johnson & Wichern 1992).

The *ij*th off-diagonal element of the standardized precision matrix, r_{ij}, is simply the partial correlation coefficients between the *i*th and *j*th populations. Absolute values of r_{ij} which are zero denote pairs of populations whose covariance structure are conditionally independent, given all the other populations in the data set. That is, within the edge set *E* there is a sufficient number of alternate paths through the graph, whose presence explain the overall pattern of population covariation such that the removal of e_{ij} does not influence the total genetic covariance.

Determination of how small r_{ij} must be to be considered zero is the final step in estimating the conditional independence structure of the Population Graph. Whittaker (1990) showed that a statistic of information divergence, called edge exclusion deviance (*EED*), could be used to determine if values of r_{ij} are not significantly greater than zero. Edge exclusion deviance is calculated as: $EED = -n_{total} \ln[1 - (r_{ij})^2]$,

where n_{total} is the number of individuals in the entire data set. *EED* has an asymptotic χ^2 distribution with one degree of freedom (Whittaker 1990). Using the *EED* values, we determine the minimal edge set E_{min} describing the hypothesized topology of the Population Graph. The goodness-of-fit for this topology can be evaluated analytically by estimating the model deviance, $D = n_{total} \ln(|\mathbf{\Sigma}| / |\mathbf{S}|)$, where $|\mathbf{\Sigma}|$ is the determinant of a MLE estimate of the covariance matrix (see Edwards 1995 for its calculation) and $|\mathbf{S}|$ is the determinant of observed sample covariance matrix (Whittaker 1990). The model deviance, D, also has an asymptotic χ^2 distribution with the degrees of freedom equal to the number of excluded edges. Significant values of D suggest that the topology does not fit the data. For a more complete discussion of model testing, see Magwene (2001).

Hypothesis testing in Population Graphs

In general, hypotheses are evaluated in Population Graphs by examining either the overall topology or specific topological features of the graph itself. For example, a common a priori hypothesis in many population genetic studies is one specifying a restriction in gene flow between two groups of populations (e.g. vicariance), which is evaluated typically by examining the significance of an average differentiation statistic between groups such as Φ or θ. The same hypothesis can be evaluated using Population Graphs by focusing on topological features of edge connectivity between specified groups of populations. While Population Graphs are not limited to testing only this type of hypothesis, the practice of assessing the significance of topological features within a graph is so common we will focus first on how these hypotheses are evaluated. Later, using the *L. schottii* data set, we will demonstrate methods for testing other categories of hypotheses in Population Graphs.

The null hypothesis regarding the topology of a Population Graph under restricted gene flow, for instance due to vicariance, states that the presence or absence of an edge connecting nodes should be independent of the hypothesized geographical barrier to gene flow. If the null is true then the number of edges connecting nodes across the hypothesized barrier to gene flow should be as numerous as the number of edges in other portions of the graph. As it turns out, as every edge is connected to only two nodes, by definition, the pattern of edge/node connections has a binomial expectation (Bollobás 2001). Given a Population Graph, G, with parameters $K = |E|$ edges and $N = |V|$ nodes the average probability of observing an edge connecting any two nodes can be modelled as a binomial random variable, $B(p, K_{total})$, where the parameter p is:

$$p = \frac{K}{K_{total}} \tag{3}$$

where

$$K_{total} = \frac{N(N-1)}{2} \tag{4}$$

enumerates the number of edges in a saturated graph (following Bollobás 2001).

By way of an example, consider the graph $G = \{V, E\}$ with parameters N and K as above. Further, we are interested in partitioning G into two complete subgraphs, G_1 and G_2, with node sets, V_1 and V_2, where $V_1 \cup V_2 = V$ and $V_1 \cap V_2 = \varnothing$. We assume that these subgraphs are not disconnected (the trivial solution) and share a set of edges, K_{btw}, connecting the two subgraphs. The subgraph G_1 has $N_1 = |V_1|$ nodes and $K_1 = |E_1|$ edges. Similarly, the subgraph G_2 has $N_2 = |V_2| = N - N_1$ nodes and $K_2 = |E_2| = K - K_1 - K_{btw}$ edges. The probability of finding an edge connecting G_1 to G_2 is:

$$p = \frac{N_1 N_2}{K_{total}} \tag{5}$$

and the probability of observing K_{btw} or fewer such edges is:

$$p(x \le K_{btw}) = \sum_{i=0}^{K_{btw}} \binom{K}{i} p^i (1-p)^{K-i} \tag{6}$$

If we are not testing an a priori hypothesis regarding the significance of G_1 and G_2 we need to correct $p(x \le K_{btw})$ to take into account the number of ways we can obtain a subgraph of size N_1 by multiplying eqn 6 by $\binom{N}{N_1}$. This general binomial approach facilitates the testing of a wide range of both a priori as well as *post-hoc* topological hypotheses within the Population Graph framework.

In addition to the binomial approach, one may also use permutation to evaluate the significance of topological features. Following from above, we can easily simulate the null distribution of the number of *between* subgraph edge counts within the set of graphs, $\tilde{G}_{NULL}$, of the same order as G. The node set of each graph in $\tilde{G}_{NULL}$ will be partitioned into populations belonging to subgraphs as in $\tilde{G}_{NULL}$ the observed graph. From the set of graphs in, we extract the null distribution of between subgraph edges, $\tilde{K}_{btw}$, on which we evaluate the significance of the observed value. General permutation approaches are covered in depth in Manly (1997) that can be adapted easily to graph theoretic hypotheses.

Case study: L. schottii

Chief among the historical factors influencing intraspecific genetic variation are deep time geotectonic events over the past several million years and shallow time glacial–interglacial cycles over the most recent hundreds of

thousands of years. The relative importance of geological and climatic factors on the present day distributions of plant and animal taxa has long been contentious (Cracraft 1988; Riddle 1996; Avise 2000). Molecular data are being brought increasingly to bear on this problem, with the strongest inference obtained when testing pre-existing regional hypotheses concerning dispersal and vicariance as opposed to generating *ad hoc* explanations from genetic data alone (Cruzan & Templeton 2000; Knowles & Maddison 2002).

The Sonoran Desert of the southwestern United States and northwestern Mexico is an area of active ecological and evolutionary research due to its diverse habitats, high taxonomic diversity and well-characterized geotectonic and palaeoclimatic history. In particular, three major geotectonic events are predicted sources of vicariance for the region's biota (Case *et al*. 2002): (i) the formation of the Sea of Cortéz (*c*. 5 Mya) separating peninsular Baja from mainland Mexico; (ii) the Isthmus of La Paz separating the southern Cape Region and south-central Baja and resulting from the formation of a trans-peninsular seaway connecting the Sea of Cortéz and the Pacific Ocean (*c*. 3 Mya); and (iii) a geologically cryptic mid-peninsula vicariance event in Central Baja (suggested by vertebrate allozyme and mtDNA data; Riddle *et al*. 2000), resulting from the formation of a trans-peninsular seaway during the mid Pleistocene (*c*. 1 Mya). Layered onto these deep time events are more recent Pleistocene climatic cycles driving changes in geographical distributions that may have eroded or even erased evidence of older vicariance events.

The geographical concordance of phylogenetic topologies with ancient geological events has led Riddle *et al*. (2000) to two general conclusions regarding present day Sonoran Desert vertebrates: first that lineage diversification resulting from vicariance events in deep time persists as important sources of taxic diversity, and second that Quaternary climatic fluctuations have had relatively little impact on biogeographical distributions of taxa in Baja. Sonoran Desert plants, in contrast, are mostly of tropical origin and frost sensitive, suggesting that their populations may be more prone to extinction during climatic fluctuations and less likely to persist on opposite sides of historical sources of vicariance.

Using traditional population genetic tools, Nason *et al*. (2002) tested these hypotheses with respect to a Sonoran Desert plant, the endemic columnar cactus, senita (*L. schottii*). A total of 21 Continental Sonoran and Peninsular Baja populations of *L. schottii* were assayed for 29 polymorphic allozyme loci. The distribution of genetic variation was decomposed using a hierarchical *F*-statistic model with the following levels: individuals nested within populations; populations nested within putative Continental Sonoran and Peninsular Baja phylogroups defined by a Sea of Cortéz vicariance event; and among phylogroups. Nason *et al*. (2002) found genetic differentiation among populations to be exceedingly large for an insect pollinated plant species ($F_{ST} = 0.431$), as was the genetic variance among phylogroups ($F_{PT} = 0.302$). Isolation by distance was significant among Peninsular Baja and among Continental populations (but not between), consistent with the hypothesis of limited gene flow within regions (the lone exception was a Continental population, SenBas, to which we return shortly). A population phenogram of all Peninsular Baja populations, constructed using Nei's genetic distance and neighbour joining, revealed a clear pattern of consecutive nesting of populations from south to north supporting the hypothesis of recent northward range expansion. The phenogram describing Sonoran populations contained less spatial structure.

From the *L. schottii* data set, we examine two distinct sets of hypotheses using Population Graphs. The first set is concerned with more traditional population genetic and phylogeographical questions regarding the historical vicariance between the putative Continental and Peninsular phylogroups (akin to testing for significant differentiation among strata or H_0: $F_{PT} = 0$ in Nason *et al*. 2002) and the presence of isolation by distance. The second set of hypotheses focuses on specific topological characteristics of the Population Graph, examining the placement of a particular continental population (SenBas) within the graph and, more generally, the relative importance of individual populations in maintaining genetic connectivity, or information flow, among all populations.

The L. schottii *Population Graph*

A total of 948 individuals with 29 multilocus genotypes were translated into a 21 × 21 population distance matrix, **D** (Table 1; below diagonal). The Population Graph representing the minimal topology of 21 *L. schottii* populations has 50 edges (significant edges shown in Table 1; above the diagonal) with two visually identifiable subgraphs (Fig. 2). This minimal edge set, E_{min}, depicted in Fig. 2, represents the best fit model among several alternate topologies (model deviance: $D = 162.4559$, d.f. = 160, $P = 0.43$). Techniques for exploring alternate topologies are beyond the scope of this study, as it requires a full mathematical treatment (see Magwene 2001 for a general overview). We return to this topic in a later manuscript focusing on global human population genetic structure (Dyer and Nason, in prep).

The smaller of the subgraphs, hereafter G_1 and depicted with lighter coloured nodes in Fig. 2, contains populations PL, LF, CP, Seri, SG, SI, SN and TS, all of which are Continental Sonoran populations. The larger subgraph, G_2, contains the remaining populations that are, with the exception of SenBas from southern Arizona, all Baja Peninsular populations. A single connection, or bridge, connecting SenBas and PL forms the only link between G_1 and G_2. The connectivity of SenBas within Fig. 2 suggests that even though it is geographically proximal to Continental

Table 1 Population-wise multivariate statistical distances (d_{ij}; below diagonal) and standardized inverse correlations (r_{ij}; above diagonal) among peninsular Baja and mainland Sonora populations of *Lophocereus schottii*. Population names in italics denote Continental Populations. Standardized correlations significantly greater than zero are denoted in bold (above diagonal) and index the minimal edge set in the Population Graph shown in Fig. 2. Significance was evaluated using edge exclusion deviance (see text)

	BaC	*CP*	Ctv	LaV	*LF*	Lig	*PL*	PtC	PtP	*SenBas*	*Seri*	*SG*	*SI*	SLG	*SN*	SnE	SnF	SnI	StR	*TS*	TsS
BaC		−0.04	0.05	**0.15**	−0.01	**0.07**	0.00	0.05	**0.10**	**0.07**	0.00	−0.02	0.02	−0.01	0.01	**0.11**	−0.01	**0.07**	**0.23**	0.06	−0.01
CP	10.93		−0.03	0.04	**0.07**	0.02	−0.04	0.01	−0.03	0.01	**0.41**	**0.21**	−0.05	0.04	**0.15**	0.03	0.03	0.00	0.00	**0.20**	−0.01
Ctv	6.84	10.43		−0.02	0.01	0.02	0.00	0.00	**0.25**	**0.10**	0.00	0.00	−0.02	**0.64**	0.01	0.00	**0.20**	0.05	−0.04	0.00	0.00
LaV	9.04	13.23	10.76		0.01	**0.08**	−0.02	0.05	0.02	0.01	−0.02	0.04	0.00	−0.02	−0.01	**0.23**	**0.09**	−0.03	0.03	0.02	**0.38**
LF	10.76	4.29	10.12	13.44		0.02	**0.39**	−0.01	−0.02	−0.01	0.04	**0.28**	**0.17**	0.01	0.06	−0.02	−0.01	0.01	−0.03	0.06	0.01
Lig	9.75	12.35	9.82	12.18	12.40		0.01	**0.09**	0.03	0.05	0.02	0.02	0.01	0.01	0.03	**0.06**	0.04	**0.10**	**0.19**	0.03	0.05
PL	10.77	5.24	10.23	13.67	2.46	12.60		0.00	0.00	**0.07**	0.00	**0.15**	**0.34**	0.03	0.06	0.02	−0.02	−0.02	0.00	0.05	0.04
PtC	10.75	13.84	11.96	10.65	14.01	12.88	14.07		0.01	0.04	0.01	0.03	0.02	−0.01	0.03	**0.09**	0.04	−0.01	**0.12**	0.05	**0.49**
PtP	6.46	10.44	2.65	10.33	10.21	9.62	10.25	11.64		**0.11**	−0.01	0.00	0.02	0.03	0.01	0.01	**0.30**	**0.15**	0.01	0.00	−0.01
SenBas	7.67	9.05	5.73	11.41	8.95	10.40	8.87	12.07	5.82		0.06	0.00	−0.01	0.03	0.05	0.00	0.02	0.05	**0.07**	0.01	0.00
Seri	10.57	2.67	10.23	13.37	3.97	12.13	4.71	13.71	10.16	8.58		**0.19**	0.05	−0.01	**0.25**	0.01	−0.02	0.03	0.01	0.06	−0.02
SG	11.37	3.70	11.04	13.67	2.86	12.85	3.60	14.20	10.95	9.48	3.46		**0.13**	−0.04	0.05	−0.01	0.02	0.00	0.01	−0.01	0.00
SI	10.70	5.29	10.37	13.74	3.19	12.65	2.95	14.12	10.24	9.21	4.67	3.91		0.01	**0.07**	0.00	0.02	0.00	0.02	**0.22**	−0.01
SLG	7.22	10.32	1.39	10.94	10.09	10.02	10.17	12.12	3.26	6.13	10.24	11.07	10.32		−0.03	0.03	**0.18**	**0.10**	0.00	0.03	0.04
SN	10.60	4.08	10.47	13.49	4.49	12.32	4.98	13.73	10.36	8.91	3.57	4.53	4.96	10.56		0.04	−0.01	−0.01	0.02	**0.17**	0.00
SnE	7.76	11.75	8.74	8.44	12.02	10.81	12.05	10.47	8.55	9.68	11.61	12.39	12.18	8.85	11.66		0.03	0.05	**0.19**	−0.05	**0.09**
SnF	7.27	10.69	2.70	10.45	10.63	9.97	10.77	11.99	2.99	6.54	10.62	11.34	10.77	3.01	10.90	8.83		**0.10**	0.03	−0.02	−0.05
SnI	6.88	9.82	4.46	10.68	9.73	9.31	9.93	11.58	4.45	6.66	9.49	10.38	9.93	4.61	9.91	8.50	4.86		**0.09**	0.02	0.05
StR	6.64	11.39	8.58	9.82	11.52	9.25	11.49	10.06	8.07	8.63	11.08	11.84	11.48	8.74	11.18	7.62	8.53	7.74		−0.03	**0.08**
TS	10.22	4.36	9.91	13.31	4.44	12.20	4.81	13.63	9.91	8.91	4.43	4.88	4.34	9.84	4.60	12.05	10.43	9.49	11.37		−0.01
TsS	9.83	13.42	10.86	7.48	13.24	12.08	13.25	6.91	10.70	11.42	13.38	13.67	13.56	10.88	13.39	9.11	11.17	10.40	9.31	13.29	

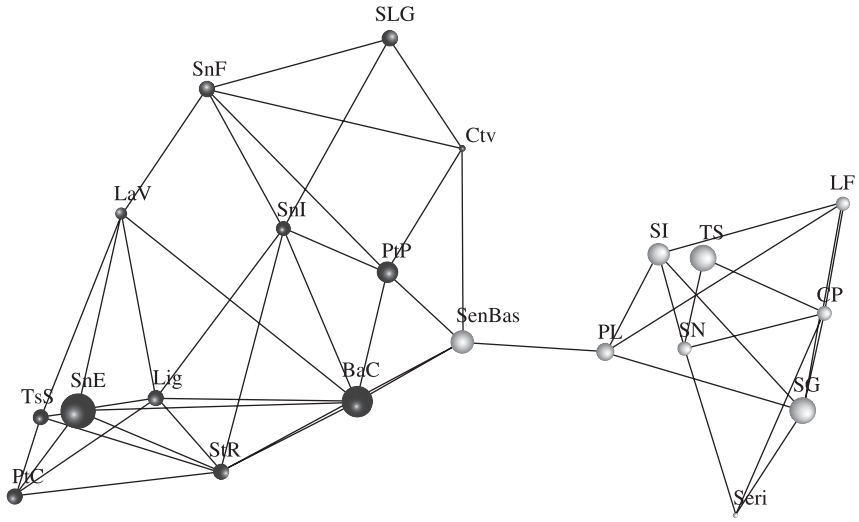

Fig. 2 Population Graph representing the genetic relationships among Peninsular (dark nodes) and Continental (light nodes) populations of *Lophocereus schottii*. The differences in node size reflect differences in within population genetic variability, whereas the edge lengths represent the among population component of genetic variation due to the connecting nodes. Both node sizes and edge lengths are projected within a three-dimensional drawing space.

populations (see Fig. 1 in Nason *et al*. 2002), it shares a higher degree of genetic covariance with Peninsular Baja populations. In contrast, the original hierarchical analysis of Continental vs. Peninsular populations in Nason *et al*. (2002) grouped SenBas with the remaining Continental populations due to its geographical proximity. The authors did single out SenBas as unusual because it alone exhibited no pattern of isolation by distance with other populations (see Fig. 3B in Nason *et al*. 2002). Given palaeoecolgical evidence of the absence of *L. schottii* from northern Baja and Sonora prior to 2000 years ago (Van Devender *et al*. 1994; Peñalba & Van Devender 1998), the Population Graph suggests a new hypothesis for the origin of SenBas, that it is the result of a relatively recent long-distance dispersal event out of Peninsular Baja. This hypothesis would explain the clustering of SenBas within the Peninsular subgraph, G_2, as well as the lack of isolation by distance between this population and the remaining continental populations.

Vicariance and isolation by distance in Population Graphs

The presence of vicariance and isolation by distance are two hypotheses commonly addressed in population genetic analyses. We begin our examination of the Population Graph in Fig. 2 by evaluating the significance of these two hypotheses. As outlined above, vicariance within a Population Graph should be visually evident by relatively few edges between populations spanning the hypothesized source of vicariance. In Fig. 2, it is immediately obvious that subgraphs G_1 and G_2 may be separated by a potential source of vicariance, which in this case corresponds to the Sea of Cortéz. Nason *et al*. (2002) tested this a priori hypothesis by examining the genetic differentiation between these two putative phylogroups and found a significantly large amount of among phylogroup variation; $F_{PT} = 0.302$; CI 95% = (0.177–0.435). Here we test the same hypothesis within the Population Graph framework using both the binomial representation outlined above as well as a permutation-based approach.

From Fig. 2, we find that subgraph G_1 has $N_1 = 8$ nodes and $K_1 = 16$ edges and subgraph G_2 has $N_2 = 13$ nodes and $K_2 = 33$ edges (assuming for the time being that the SenBas population is part of the Peninsular subgraph). From eqn 5 above, the null hypothesis states that the probability of obtaining an edge connecting G_1 to G_2 in the graph is:

$$p = \frac{N_1 N_2}{K_{\text{total}}} = 0.4952$$

Further, within the entire graph, containing $K = 50$ edges, the probability of observing a single edge (a bridge) connecting G_1 to G_2 is ($K_{\text{btw}} = 1$ from eqn 6 above):

$$p(x \le K_{\text{btw}}) = \sum_{i=0}^{K_{\text{btw}}} \binom{K}{i} p^i (1-p)^{K-1} \approx 7.17e-14$$

a highly significant result. This significant result suggests that the Sea of Cortéz, as represented in the topology of Fig. 2, has indeed acted as a significant source of vicariance between the two subgraphs (or putative phylogroups following Nason *et al*. 2002). Because we are focusing here on demonstrating the use of Population Graphs, we refer the interested reader to the original analysis of these data for discussion of the evolutionary significance of this vicariance.

In addition to the binomial test for vicariance, we also performed a permutation test to highlight the duality of these two analyses. Again, under the null hypothesis, the Sea of Cortéz has no influence on the topology of represented in Fig. 2. As a result, we can simulate a large number of random graphs of the same size and order as that show in Fig. 2. From these graphs, we can build the

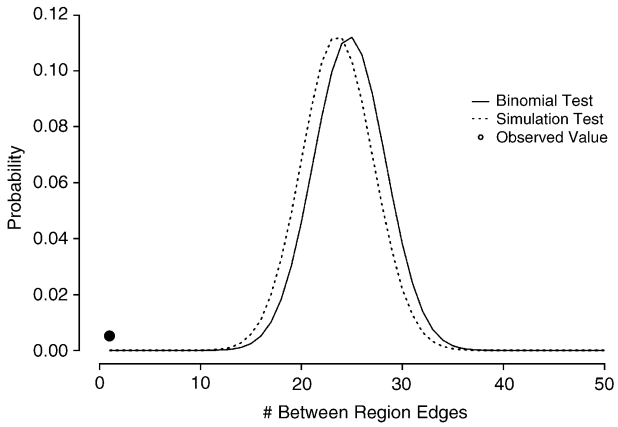

Fig. 3 The probability of observing edges connecting Peninsular and Continental populations of *Lophocereus schottii* using the binomial (solid lines) and permutation (dashed lines) tests. The observed number of interphylogroup edges ($K_{btw} = 1$) is indicated by the filled circle.

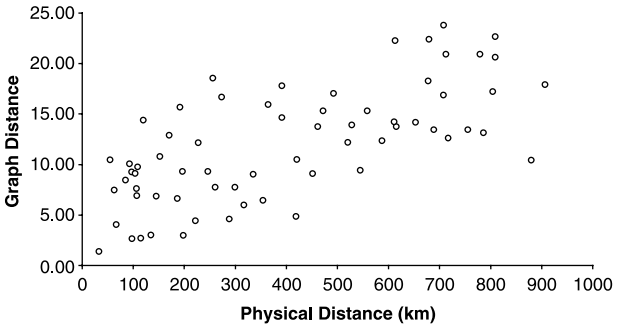

Fig. 4 Relationship between graph distance, *l*, and physical separation between populations (km) as a measure of isolation by distance ($P < 0.001$, $R^2 = 0.48$).

null distribution of how many edges are predicted, under the null, to connect Peninsular and Continental populations. We show in Fig. 3 the probability densities for both the binomial and permutation tests for the graph parameters outlined above. Both approaches yield similar results; namely that the probability of observing a single bridge connecting these two phylogroups is exceedingly unlikely if the Sea of Cortéz is not acting as a source of vicariance. The binomial and permutation approaches are not the only methods to test the significance of topological features; however, they do are relatively intuitive and follow from the work on random graph theory. We return to discussing other analytical approaches in a subsequent manuscript.

The next hypothesis we test concerns the spatial arrangement of populations with respect to their genetic differences under a model of isolation by distance (IBD). IBD is a non-random association of genetic similarity resulting from limited gene exchange among geographically separated populations (e.g. Slatkin 1993). Several approaches have been used to test for isolation by distance, ranging from simple regression of pairwise F_{ST} and genetic distances on physical distance to more sophisticated evolutionary models (e.g. Slatkin 1993; Roussett 1997). Population Graphs can also be used to investigate isolation by distance by examining the relationship between graph and geographical distances.

By definition, the graph distance between nodes *i* and *j* is the shortest path through the graph connecting them, l_{ij}. The length of this minimum path can be measured in terms of the minimum number of edges separating nodes (in unweighted graphs), or the minimum path length (in weighted graphs). In a Population Graph, the graph distance, l_{ij}, corresponds to the among population component of genetic variance, σ^2_A (following Appendix I), for all pairs of populations. As with other IBD methods, these pairwise graph distances are regressed on the geographical separation (using nonparametric methods due to the lack of independence). Under the model of IBD, we expect a positive relationship between pairwise genetic and geographical measures of differentiation. For the set of Peninsular populations, G_2 in Fig. 2 excluding SenBas, we find a significant relationship between graph distance and physical distance (Fig. 4; $P < 0.001$, $R^2 = 0.48$). Similar results were found when testing for IBD among Continental Sonoran populations ($P = 0.017$; $R^2 = 0.15$; figure not shown). Finally, we found no significant relationship between the geographical and graph distances between G_1 and G_2 or between SenBas (a southern Arizona population) and populations from either of these subgraphs. These results concur with Nason *et al*. (2002), who found a significant relationship between a pairwise estimator of gene flow ($\hat{M}$) and physical distance for the same subset of populations.

Population assignment and graph connectivity

Based upon the topology of the Population Graph, we have assumed that the SenBas population in southern Arizona, while geographically associated with the remaining Centennial populations, is genetically more similar to existing Peninsular populations. The connectivity of SenBas (Fig. 2) with a single Sonoran population (PL) and four northern and central Baja populations (Ctv, StR, PtP and BaC) supports the hypothesis of a relatively recent long-distance dispersal event out of Baja into southern Arizona. This interpretation of the data provides an explanation for the apparent lack of isolation by distance between SenBas and all remaining Continental populations as shown above and reported in Nason *et al*. (2002). The improper assignment of SenBas to the Continental phylogroup highlights the differences between an a priori allocation of populations within a hierarchical structure vs. allowing

the data to define their own set of relationships as advocated here.

We test the hypothesis that SenBas is genetically Continental, despite its high degree of connectivity with several Peninsular populations, using the binomial properties of edge connectivity outlined above. The null hypothesis for this test states that the topology of the Continental subgraph, consisting of SenBas and the remaining Sonoran populations, has a pattern of edge assignment that supports the inclusion of SenBas with the remaining Continental populations. The node set for this subgraph, V_3, contains nine populations (SenBas, PL, LF, CP, Seri, SG, SI, SN and TS), whereas the edge set, E_3, has 17 edges (Fig. 2). We evaluate this subgraph, $G_3 = \{V_3, E_3\}$, in terms of how likely it is to observe a graph with $N_3 = 9$ nodes and $K_3 = 17$ edges that contains a pendant (a node with a single edge as depicted by SenBas).

The binomial nature of edge connectivity leads to an expectation for the degree, or number of edges, for each node. The probability that any two nodes are connected, following eqn 3 above, is $P = 0.472$ for all graphs of the same size as G_3. The expected degree of any node is given by $P(N_3 - 1) = 3.78$. Using the binomial distribution, we find the probability of observing any graph of this size where at least one of the nodes has degree one. This probability is given by:

$$p(pendant \mid N_3, p) = \binom{N_3}{1} p^1 (1 - p)^{N_3 - 1} = 0.0108 \quad (5)$$

Notice here, we used the combinatorial $\binom{N_3}{1}$ in our calculations of the probability. In this case, it is necessary because the hypothesis being tested was not a priori; rather, it tested for the probability of observing a single pendant within a graph defined by the parameters p and N_3. From these results, it appears that within the set of all graphs of this size, it is significantly rare to observe any of them containing a pendant. The high degree of connectivity between SenBas and Peninsular populations combined with the single connection to a Continental Population is consistent with the hypothesized Baja origin of SenBas individuals.

Finally, we focus on the overall topology of the Population Graph in Fig. 2 to address the topic of genetic connectivity among all populations. In so doing, we shift our focus from population genetic and phylogeographical hypotheses and towards ones addressed typically in conservation genetics. However, the issue of genetic connectivity is just as salient to the former fields of inquiry as the latter. From a conservation genetic perspective, populations targeted for preservation are often ones with high genetic variability. Genetic variability is quantifiedtypically with summary statistics including: the proportion of polymorphic loci (P), heterozygosity (H), and the number of alleles per locus (A). The size of the nodes in Fig. 2 summarizes these statistics in multivariate space as it is defined as a measure of the within population genetic variance. As indicated in Fig. 2B, populations BaC and TS have the largest within-population genetic variability for each phylogroup. These populations also have the largest P, H and A (Nason unpubl. data). Nason *et al*. (2002) showed a significant reduction in P, H and A when regressed on latitude ($P < 0.001$ in all cases) consistent with the hypothesis of a recent northward range expansion.

In addition to identifying the differences in within-population genetic variation or heteroscedasticity (a statistical test of which is to be addressed in a later manuscript), we can use the topology of the graph to identify populations important to the flow of genetic material across the landscape. For example, the within-population genetic variability of PL and SenBas is not as large as other populations in Fig. 2, yet they form the only connection between Continental and Peninsular phylogroups; that is, any path originating in one phylogroup and ending in the other must pass through these two populations. Nodes that have this characteristic are called articulation points. In general, any subset of nodes whose removal significantly increases the length of the minimum spanning tree for the graph are of importance to the overall graph connectivity, and in Population Graph terms the overall flow of genetic material across the graph.

Here we focus SenBas and PL because these two populations form a bridge between the two inferred phylogroups and are the graphs most apparent articulation points. This bridge may be present in the graph for one of two reasons. First, this bridge may simply represent the most intermediate point of connectivity between phylogroups in terms of their genetic covariation. In this case, the removal of SenBas or PL from subsequent graph construction would reveal alternative connections between the two phylogroups formed through other population pairs. Conversely, SenBas and PL may represent the only possible link between the two phylogroups. In this case, excluding either population would result in two completely disconnected subgraphs. We can evaluate these alternate hypotheses by removing each one of the two populations in turn and reconstructing the overall topology. To accomplish this we must recalculate the incidence matrix following the removal of either population because the resulting topology is based upon the pattern of genetic covariation among all populations. The Population Graphs resulting from the exclusion of either SenBas or PL reveal two distinct, disconnected subgraphs (not shown). The genetic covariance structure of the remaining populations does not support the hypothesis of robust genetic connectivity between Continental and Peninsular populations without either of the articulation populations. These results suggest that both SenBas and PL represent the sole source of genetic connectivity between Peninsular and Continental populations.

Discussion

The evolution of population genetic structure is a dynamic process influenced by both historical and recurrent evolutionary processes. Vicariance and gene flow, in particular, create a system of interacting populations whose genetic relationships can be readily investigated within a graph theoretic framework. The methods presented in this study provide an introductory treatment of the use of graphical techniques for addressing hypotheses regarding the genetic connectivity of a set of populations within a phylogeographical context. With the example of *L. schottii*, we have shown how the information contained within the topology of a Population Graph can be utilized to address important population genetic questions concerning the nature and significance of genetic separation and isolation by distance. However, if traditional population genetic approaches provide mechanisms to address the same types of questions it becomes relevant to ask why one would specifically use Population Graphs.

There is a fundamental distinction between the models applied to the decomposition of genetic variance using traditional structure statistics (e.g. Wright's *F*-statistics, AMOVA, etc.) and the methods outlined above for Population Graphs. This distinction lies in the fact that when we apply structure statistics to a set of data we impose predefined hierarchical models that we believe reflect the spatial and temporal scales evolutionary processes have operated on to create the observed distribution of genetic structure. The estimation of a significant summary statistic under a predefined model does not mean that the allocation of populations to specific strata is correct, nor does it signify that the hierarchical separation of strata conform to the spatial or temporal scales at which the underlying population genetic processes operate. Rather, it simply implies that the current arrangement of strata is sufficient to assume nonrandom association of genotypes. Short of permuting all strata to find the 'best' fit of the model to the data, there are no methods available to evaluate the assumed hierarchical population model. With the *L. schottii* data set, there are over 2.1×10^6 different ways to allocate 21 populations into two phylogroups. Clearly, evaluating the fit of such population models is not feasible given the amount of time necessary to enumerate all possible arrangements and the effect it would have on the experiment-wise error rates.

By allowing the data to define their own topology, based upon the genetic composition of the entire data set, Population Graphs are not subject to the problem of incorrectly assigning populations to higher-level strata. We specifically used the example of *L. schottii* with respect to the SenBas population to highlight the potential pitfalls of specifying a particular a priori model. Nason *et al*. (2002) had no prior information suggesting that one of their sample populations (SenBas) was the result of a long-distance dispersal event. Only by examining how the all the populations organize themselves within the topology of the Population Graph did the Peninsular origin of this Continental population become apparent. The ability to resolve intrapopulation relationships, within the context of the genetic covariance structure of all populations, is not a strength of traditional population genetic methods. Methods such as pairwise analyses allow the elucidation of pairs of relationships, however, we have no a priori reason to assume that evolutionary processes operate in a pairwise fashion. Methods such as Population Graphs presented here, which focus on the details of population-level relationships within the context of the covariance structures from all populations, are likely to be particularly useful for population genetic structures more complex than that of *L. schottii*.

Population Graphs also provide a single heuristic approach to addressing a variety of general population genetic questions. As shown above with *L. schottii*, questions concerning genetic differentiation, isolation by distance, population assignment and genetic connectivity are all addressed under a single analytical framework. In contrast, traditional population genetic analyses require different analytical procedures to address each of these questions, some of which, such as population assignment and genetic connectivity, are still being developed (e.g. Cornuet *et al*. 1999).

In a more general context, by focusing on an average effect, significant summary statistics indicate only broad trends across an entire data set. The complexity of interacting evolutionary forces shaping intraspecific genetic variation can easily be generalized by summary statistics. Indeed, most of our understanding of how evolutionary processes operate is based on just such statistics. However, if significant details of the evolutionary process lie in the relationships among sets of individual populations rather than their average effects then we must adopt approaches that do not distill the complexity of these relationships into a single or small set of values. Rather, we must adopt methodologies that capture this complexity without undue averaging across strata. A Population Graph is one such method. While it is possible to extract analogues of the traditional summary statistics from this approach, the main focus Population Graphs is directed towards a holistic quantification of population interactions across an entire data set.

It must be pointed out that the Population Graph framework does not increase the precision of a summary statistics. In fact, in some cases it appears to decrease the precision resulting in a larger variance around the test statistic (Dyer unpubl. data). The degree to which the variance surrounding the test statistic increases depends upon the order and size of the graph, as well as how among-population components of genetic variation (e.g. the edge lengths) are distributed across the topology. We address

this issue in greater depth as well as provide an index relating to how minimizing the topology influences the variance around summary statistics in a subsequent manuscript.

Finally, we believe that the intrinsic visual nature of Population Graphs facilitates the interpretation of population genetic data. Several aspects of population genetic structure are immediately apparent in the *L. schottii* graph (Fig. 2) including the unique placement of SenBas, the separation of Continental Sonoran and Peninsular populations, and the difference in within population genetic variability. While we have chosen to illustrate Population Graphs using the *L. schottii* data set because of its relative simplicity and obvious patterns, in other species the patterns in the data will often be more complex. In these cases especially, being able to visually inspect the relationships among all populations may prove to be immensely important for exploratory data analysis, data interpretation and the development of subsequent hypotheses.

Conclusions

The primary focus of Population Graphs is concerned with examining specific topological characteristics of the graph in terms of the relationships among individual populations. The focus on topological characteristics is essential as the structure of the graph represents a model of how evolutionary processes have acted on interacting populations. Albert & Barabási (2002) argue that when the time scales governing the dynamics on *any* type of graph are comparable to those characterizing the graph assembly (e.g. the process governing the connectivity of nodes), the dynamical processes influence the resulting topology. In the case of Population Graphs, it is the interaction of historical and recurrent evolutionary processes that structure genetic variation within and among populations and as a result each of these processes are expected to leave specific topological features within the overall graph.

We are currently examining several aspects of the evolution of topological features within Population Graphs. What we have presented here is a simply an introduction to the utility of graphical methods for describing population genetic structure. Currently we are actively expanding upon the framework presented herein. Some of the most important directions include: (i) quantifying the directionality of genetic covariation as would arise from periods of range expansion, anisotropic gene flow or source–sink dynamics; (ii) a more complete analysis of the relative importance of particular populations (articulation points) or edges (bridges) to the graphs overall connectivity; (iii) topological clustering algorithms for objectively partitioning a graph into completely induced subgraphs; and (iv) the set of topological features characteristic of particular models of evolution (e.g. stepping-stone, *N*-island). By characterizing the topological features that are expected to emerge due to identifiable evolutionary processes, we are in a better position to interpret correctly the topologies we observe in natural populations.

Acknowledgements

The authors thank L. Excoffier, J. Fernandez, P. Smouse, V. Sork, R. Westfall and two anonymous reviewers for insightful comments on this manuscript. The authors would also like to thank the Office of Biotechnology and the Plant Sciences Institute at Iowa State University for their support of this research.

References

Albert R, Barabási AL (2002) Statistical mechanics of complex networks. *Review of Modern Physics*, **74**, 47–97.

Avise JC (2000) *PhyloGeographyraphy: the History and Formation of Species*. Harvard University Press, Cambridge, MA.

Bahlo M, Griffiths RC (2000) Inference from gene trees in a subdivided population. *Theoretical Population Biology*, **57**, 79–95.

Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences USA*, **98**, 4563–4568.

Bollobás B (2001) *Random Graphs*, 2nd edn. Cambridge University Press, Cambridge, UK.

Box GEP, Draper NR (1987) *Empirical Model-Building and Response Surfaces*. J. Wiley and Son, New York.

Box GEP, Hunter WG, Hunter JS (1978) *Statistics for Experimenters: an Introduction to Design, Data Analysis, and Model Building*. J. Wiley and Son, New York.

Case TJ, Cody ML, Ezcurra E (2002) *A New Island Biogeography of the Sea of Cortez*. Oxford University Press, New York.

Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*, **153**, 1989–2000.

Cox JM, N.Wermuth (1996) *Multivariate Dependencies: Models, Analysis, and Interpretation*. Chapman & Hall, New York.

Cracraft J (1988) Deep-history biogeography: retrieving the historical pattern of evolving continental biotas. *Systematic Zoology*, **37**, 221–236.

Cruzan MB, Templeton AR (2000) Paleoecology and coalescence: phylogeographic analysis of hypotheses from the fossil record. *Trends in Ecology and Evolution*, **15**, 491–496.

Dempster AP (1972) Covariance selection. *Biometrics*, **28**, 157–175.

Draper NR, Smith H (1981) *Applied Regression Analysis*. J. Wiley and Sons, New York.

Dupanloup I, Schneider S, Excoffier L (2002) A simulated annealing approach to define the genetic structure of populations. *Molecular Ecology*, **11**, 2571–2581.

Edwards D (1995) *Introduction to Graphical Modeling*. Springer-Verlag, New York.

Edwards AL, Sharitz RR (2000) Population genetics of two rare perennials in isolated wetlands: *Sagittaria isoetiformis* and *S. teres* (Alismataceae). *American Journal of Botany*, **87**, 1147–1158.

Excoffier 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.

Gavrilets S (1997) Evolution and speciation on holey adaptive landscapes. *Trends in Ecology and Evolution*, **12**, 307–312.

Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.
Hewitt GM (2001) Speciation, hybrid zones and phylogeography — or seeing genes in space and time. *Molecular Ecology*, **10**, 537–550.
Johnson RA, Wichern DW (1992) *Applied Multivariate Statistical Analysis*. Prentice Hall, Upper Saddle River, NJ.
Kempthorne O (1969) *An Introduction to Genetic Statistics*. Iowa State University Press, Ames, IA.
Knowles LL, Maddison WP (2002) Statistical phylogeography. *Molecular Ecology*, **11**, 2623–2635.
Lessa EP (1990) Multidimensional analysis of geographic genetic structure. *Systematic Zoology*, **39**, 242–252.
Long JC, Smouse PE, Wood JW (1987) The allelic correlation structure of Gainj- and Kalam-speaking people. II. The genetic distance between population subdivisions. *Genetics*, **117**, 273–283.
Magwene PM (2001) New tools for studying integration and modularity. *Evolution*, **55**, 1734–1745.
Manly BFJ (1997) *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd edn. Chapman & Hall, New York.
Nason JD, Hamrick JL, Fleming TH (2002) Gene migration, range expansion, and phylogeography of the columnar cactus genus Lophocereus. *Evolution*, **56**, 2214–2226.
Nei M (1972) Genetic distance between populations. *American Naturalist*, **106**, 283–292.
Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, **89**, 583–590.
Peñalba MC, Van Devender TR (1998) Cambios de vegetacíon y clima en Baja California, México, durante los últimos 20 000 años. *Geología Del Noroeste*, **2**, 21–23.
Rhodes OEJ, Chesser RK, Smith MH (1996) *Population Dynamics in Ecological Space and Time*. University of Chicago Press, Chicago.
Riddle BR (1996) The molecular phylogeographic bridge between deep and shallow history in continental biotas. *Trends in Ecology and Evolution*, **11**, 207–211.
Riddle BR, Hafner DJ, Alexander LF, Jaeger JR (2000) Cryptic vicariance in the historical assembly of a Baja California peninsular desert biota. *Proceedings of the National Academy of Sciences USA*, **97**, 14438–14443.
Rousset F (1997) Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. *Genetics*, **145**, 1219–1228.
Slatkin M (1985) Gene flow in natural populations. *Annual Review of Ecology and Systematics*, **16**, 393–430.
Slatkin M (1993) Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, **47**, 264–279.
Smouse PE, Spielman RS, Park MH (1982) Multiple-locus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. *American Naturalist*, **119**, 445–463.
Sokal RR, Rohlf FJ (1995) *Biometry*. W.H. Freeman, New York.
Sork VL, Nason J, Campbell DR, Fernandez JF (2001) Landscape approaches to historical and contemporary gene flow in plants. *Trends in Ecology and Evolution*, **14**, 219–224.
Taberlet P, Fumagalli L, Wust-Saucy AG, Cosson JF (1998) Comparative phylogeography and post-glacial colonization routes in Europe. *Molecular Ecology*, **7**, 453–464.
Van Devender TR, Burgess TL, Piper JC, Turner RM (1994) Paleoclimatic implications of Holocene plant remains from the Sierra Bacha, Sonora, Mexico. *Quaternary Research*, **41**, 99–108.
Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.
Westfall RD, Conkle MT (1992) Allozyme markers in breeding zone designations. *New Forests*, **6**, 279–309.
Whittaker J (1990) *Graphical Methods in Applied Multivariate Statistics*. J. Wiley and Son, New York.
Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.
Zhivotovsky LA, Rosenberg NA, Feldman MW (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *American Journal of Human Genetics*, **72**, 1171–1186.

Rodney Dyer is currently a postdoctoral researcher working with John Nason at Iowa State University. Rodney's interests are in theoretical population genetics with particular emphasis in issues relating to quantifying the complexity of population genetic processes. John Nason researches the population genetics of plants and their associated insects with a special interest in obligate pollination mutualisms and host-race formation in phytophagous insects. This study reflects our shared interest in developing analytical tools for the analysis of population genetic data.

Appendix I

Corollary between Population Graphs and the AMOVA *analysis*

In this Appendix, we show how the a saturated graph is identical to the genetic variance decomposition under the AMOVA framework. Our overall goal is to show how the genetic variance can be partitioned into population wise components in a manner similar to Long *et al*. (1987). The AMOVA analysis is a random effects, multivariate analogue of Weir and Cockerham's θ (Weir & Cockerham 1984) that relies upon an elegant use of squared genetic distances, δ^2_{ij}, measured in a pairwise fashion between all individuals. For a single strata analysis, the sums of squared distances (SSD) allow the partitioning of the total genetic variance, σ^2_T, into within- (or error) and among-population components of genetic variation, σ^2_W and σ^2_A, respectively. The ratio of σ^2_A to σ^2_T supplies the statistic of differentiation, Φ_{ST}.

The error variance, σ^2_W, in the AMOVA model is the sum, across populations, of the average genetic distance among all individuals within populations. From Excoffier *et al*. (1992), with some simplification due to considering a non-nested model, it is given by (notation changed from Excoffier *et al*. 1992 for consistency):

$$\begin{aligned}\sigma^2_W = MSW &= \frac{SSW}{N-1} \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\frac{\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\delta^2_{jk}}{2n_i} \\ &= \frac{1}{N-1}(SSW_1 + SSW_2 + \ldots + SSW_K)\end{aligned} \tag{A1}$$

where N is the number of populations (or nodes), n_i is the number of individuals within the ith population, and δ^2_{ij} is the pairwise squared genetic distance between the jth and kth individuals in population i. The construction of Population Graphs partitions the within population sums of squares as:

$$\begin{aligned}\sigma^2_W = MSW &= \frac{SSW}{N-1} \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\sum_{j=1}^{n_i}[\boldsymbol{p}_{ij} - \bar{\boldsymbol{p}}_i]^T[\boldsymbol{p}_{ij} - \bar{\boldsymbol{p}}_i]\right) \\ &= \frac{1}{N-1}(SSW_1 + SSW_2 + \ldots + SSW_K)\end{aligned} \tag{A2}$$

where $|N|$ is the number of nodes (populations) in the graph, $\boldsymbol{p}_{ij}$ is the jth individual within the ith population and is the centroid of the ith population in m-dimensional space. Therefore, the average distance between an individual and its population centroid is defined as the error variance attributable to that particular population. We define the volume of a node within the graph as $v_i = SSW_i$. The sum of all node volumes would equal the total sums of squares within populations.

The decomposition of the among-population component of variance, MSA in the AMOVA framework, is given by:

$$\begin{aligned}MSA &= \frac{SSA}{N_{total} - N} \\ &= \frac{1}{N_{total} - N}\sum_{i=1}^{N}\sum_{j\neq l}^{N}\frac{\sum_{k=1}^{n_i}\sum_{l=l}^{n_i}\delta^2_{kl}}{n_i + n_j} \\ &= \frac{1}{N_{total} - N}\left(\frac{SSA_{1vs.2}}{2} + \frac{SSA_{1vs.3}}{2} + \ldots + \frac{SSA_{N-1vs.N}}{2}\right)\end{aligned} \tag{A3}$$

where the notation is as above with the addition that N_{total} is the total number of individuals and $SSA_{ivs.j}$ represents the contribution to the overall among population sums of squares due to the differences between populations i and j. The quantity $SSA_{ivs.j}$ is divided by two because the among-population sums of squares due to individual pairs of populations is calculated twice, once in terms of the distances between populations i and j and again for the calculation of the distance between j and i. Within the distance-based approach of the AMOVA framework, $SSA_{ivs.j}$ is the distance between the centroids of the ith and jth population, exactly how we have defined the off-diagonal components of the incidence matrix within Population Graphs. Therefore, the among-population component of genetic variance in the AMOVA model is equal to the sum of the edge lengths in Population Graphs.

From a graph theoretic perspective we can define a graph, G, consisting of the node set $V = \{ A, B, C \}$ and the edge set $E = \{e_1, e_2, e_3\}$ (Fig. A1A). Again, the sum of the volumes of all nodes is defined as σ^2_W, and the sums of the edges are similarly defined as σ^2_A. Therefore, an overall Φ_{ST} for all three populations is estimated, by the definitions of how we construct Population Graphs, as:

$$\begin{aligned}\Phi_{ST} &= \frac{\sum_{i=1}^{|E|} e_i}{\sum_{i=1}^{|E|} e_i + \sum_{i=1}^{|V|} v_i} \\ &= \frac{\sigma^2_A}{\sigma^2_A + \sigma^2_W}\end{aligned} \tag{A5}$$

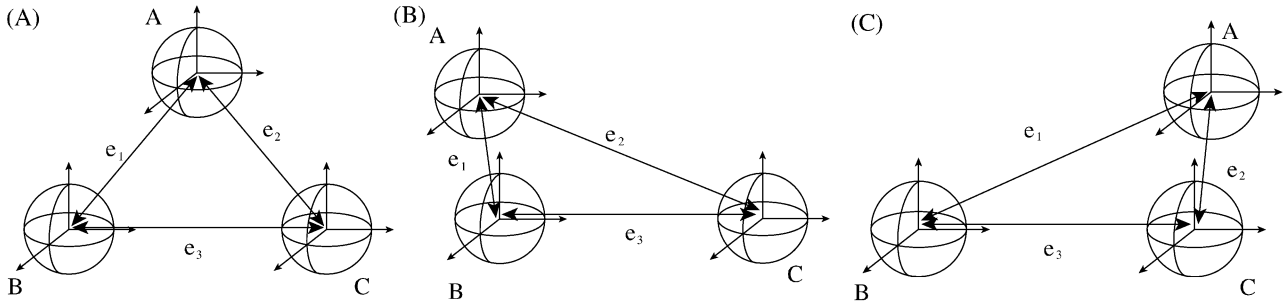

Fig. A1 Graphical representation of population genetic differentiation among three populations (A, B and C). Differentiation among populations quantified by edges e_1, e_2 and e_3. All graphs have the same Φ_{ST}. (A) All populations equally differentiated from each other. (B) Nonsymmetrical differentiation among three populations under the constraint that the sum of edge lengths is equal to the sum of the edge lengths in (A). (C) Permutation of populations within a graph where the sum of the edge lengths equals the previous two examples.

where $|E|$ is the number of edges, $|V|$ is the number of nodes, v_i is the volume of the ith node and the σ^2_A and σ^2_W terms follow the definition of random effects estimates of differentiation of Weir & Cockerham (1984) and Excoffier *et al*. (1992).

There are two salient points to make with respect to the geometry of averaging statistics such as Φ_{ST}. First, the centroid of any population is, by definition, determined entirely by the genetic identity of the individuals within the population. This means that the coordinates of the population centroid are independent of how different a particular population is from others. The distances from the centroids to the individuals within a population define σ^2_W. So, σ^2_W is independent of where populations are located in m-space. The only thing that affects an average statistic of genetic differentiation such as Φ_{ST}, or its univariate analogs such as F_{ST} or Φ_{ST}, is σ^2_A.

Second, σ^2_A, defined as the sum of the edge lengths, has no unique solution. The Population Graph in Fig. A1A is portrayed as an equilateral triangle, that is $e_1 = e_2 = e_3$, and σ^2_A is defined as the sum of these edge lengths. However, there is an infinite number of other ways one could draw a triangle with the restriction that the sum of the lengths of all edges equals that shown in Fig. A1A. Figure A1B and A1C also show the arrangement of three populations within three unique topologies. However, if the sums of the edges all equal the same σ^2_A, then all three will estimate the exact same Φ_{ST} value. In general, for any set of populations, the nature of the pairwise genetic relationships among populations will not be revealed in statistics of average differentiation. However, heterogeneity in the nature of these relationships will be represented explicitly in the topology of a Population Graph.