Topological Signatures for Population Admixture

Laxmi Parida[1], Filippo Utro[1], Deniz Yorukoglu[2], Anna Paola Carrieri[3], David Kuhn[4], and Saugata Basu[5]

[1] Computational Genomics, IBM T.J. Watson Research Center, Yorktown Heights, NY, USA.
[2] Department of Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA.
[3] Department of Computer Science, University of Milano-Bicocca, Milan, Italy.
[4] USDA-ARS Subtropical Horticultural Research Station, Miami, FL, USA
[5] Department of Mathematics, Purdue University, West Lafayette, IN, USA

Abstract. As populations with multi-linear transmission (i.e., mixing of genetic material from two parents, say) evolve over generations, the genetic transmission lines constitute complicated networks. In contrast, uni-linear transmission leads to simpler network structures (trees). The genetic exchange in multi-linear transmission is further influenced by migration, incubation, mixing and so on. The task we address in the paper is to tease apart subtle admixtures from the usual interrelationships of related populations. We present a combinatorial approach based on persistence in topology to detect admixture in populations. We show, based on controlled simulations, that topological characteristics have the potential for detecting subtle admixture in related populations. We then apply the technique successfully to a set of avocado germplasm data indicating that the approach has the potential for novel characterizations of relatedness in populations. We believe that this approach also has the potential for not only detecting but also discriminating ancient from recent admixture.

1 Background

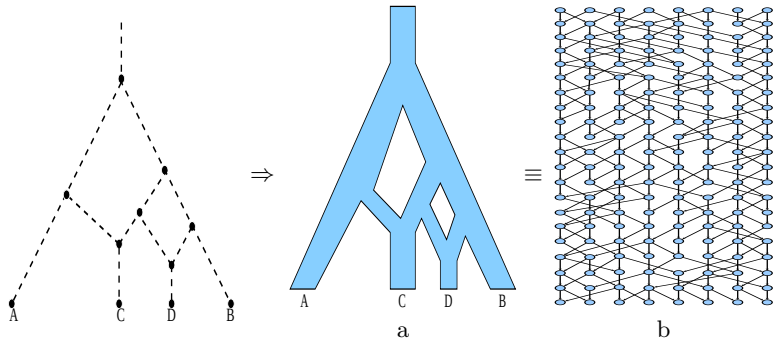
Relatedness of populations is an interesting problem and has been studied extensively in the population genetics community [9, 10]. In the context of plant breeding, this understanding is very important in gauging the diversity in the genetic pool and using it effectively in breeding programs [13]. In the context of humans, admixture mapping of the genome is useful for disease or complex trait association studies [4, 16]. Various statistical models have been proposed in literature [10, 11] to characterize admixture which build mainly on linkage disequilibrium footprints via minimum allele frequencies of the markers. Here, we present a combinatorial model based on persistence to model and study admixture. The authors in [5] have used a similar model to study presence/absence of genetic exchange as recombination or reassortment in viral populations. The

problem we address here is a little more nuanced, i.e., to discern admixture from amongst the ubiquitous recombination events. More precisely, the problem is defined as follows.

1.1 Problem Setting

Ever since Ancestral Recombination Graph (ARG) was introduced by Griffiths and Marjoram [6], it has become a convenient handle to analyze as well as infer evolutionary history of populations. ARG incorporates both recombinations and coalescence in capturing the common history of a set of extant individuals. A combinatorial perspective of this is presented in [12] as $\tilde{G}$, a directed acyclic graph (DAG) with the extant units at the leaf nodes. The internal nodes of $\tilde{G}$ denote ancestors and the edges between nodes denote the transmission of genetic material through them. Each internal node is at some depth d denoted in generations from the leaf nodes. All the leaf nodes are at depth $d = 0$. The nodes and edges are annotated with the portion of the chromosomal segments they transmit. We assume that the populations captured by the ARG are Wright Fisher models [9]. Hence an ARG is a random structure whose topology and annotation is determined by the number of leaf nodes, the recombination rate r, the mutation rate μ and population size at each generation N amongst others. In practice, usually only a portion of the ARG, called the subARG, can be reconstructed [7, 8]. A subARG has a lower resolution of information than $\tilde{G}$ and can be defined as follows: the vertex set V of a subARG is a subset of the vertex set $\tilde{V}$ of $\tilde{G}$. For every directed path in $\tilde{G}$ from v_1 to v_2, $v_1 \neq v_2 \in V \subset \tilde{V}$, there is an edge in G if and only if for every vertex $u(\neq v_1, v_2) \in \tilde{V}$ in a directed path from v_1 to v_2 in $\tilde{G}$, $u \notin V$ holds. In this paper we denote an ARG (or a subARG) as P, where the leaf nodes have an additional population label. Fig 1 (ii) shows an example with four population labels.

Let the relationship between the m populations be defined by a DAG P' with m leaf nodes, called a *scaffold*, as shown in Fig 1 (i). The progress of time is assumed to be from top to bottom and the m leaf nodes are annotated with the population labels. Further, each edge e in P' has three characteristics: the incubation length $\text{len}(e)$, the number of lineages at the bottom of the edge, $l_b(e)$, and the number of lineages at the top of the edge, $l_t(e)$. The length is a time parameter defined in generations. Note that two parameters, an effective population size and a recombination rate, determine the number of lineages $l_t(e)$ for a fixed pair of values of $l_b(e)$ and $\text{len}(e)$. We assume that the scaffold P' is binary (i.e., each internal node in P' has exactly two ascendants or two descendants, but not both). For each internal node, the *junction constraints* are defined as follows. For a node in P' that has two incoming edges e_1 and e_2 and an outgoing edge e_3, the following relationship holds $l_t(e_3) \leq l_b(e_1) + l_b(e_2)$, i.e., the lineages at v is the union of the lineages of the two incoming edges. Similarly if node v has two outgoing edges e_1 and e_2 with one incoming edge e_3, then $l_b(e_3) \leq l_t(e_1) + l_t(e_2)$, i.e., the lineages at v is the union of the lineages of the two outgoing edges. Finally, we say that P' defines *admixture* if there exists a closed path (CP) in P'. Each edge e of P' represents the evolution of a Wright



(i) scaffold P' (ii) ARG P

Fig. 1. An example with four populations A, B, C, D. (i) shows the scaffold P'. (ii) shows a corresponding ARG P. (ii-a) shows the ARG with the "shape" of P' superimposed on it, while the (ii-b) shows some of the details of P of (ii-a). Note that in general the structure of P' is not apparent from P and the ARG P simply looks like the one shown in (ii-b). See text for more details.

Fisher population captured in a DAG say P_e. The union of each of these DAGs by appropriately gluing the ends of the edges corresponding to the nodes of P' gives the ARG P that can be written as:

$$P = \bigcup_{e \in P'} P_e.$$

Such a P is shown in Fig 1 (ii) where the leaf nodes correspond to extant units of each population of P': (ii-b) shows some of the typical details of enclosed area of (ii-a). Each row in (ii-b) is a generation and the edges denote the flow of genetic material towards the extant units at the leaf nodes (the arrows are not shown to avoid clutter). A node with two incoming edges in (ii-b) denotes a genetic exchange event such as recombination. Due to space constraints, we refer the reader to [12] for further details of a typical ARG P. Note that a recombination event in the evolution process leads to the occurrence of a CP in P. Now we are ready to define the central problem as a riddle with three actors as follows.

Problem 1. Tom generates a scaffold P' on m populations with the three parameters $\text{len}(e)$, $l_b(e)$ and $l_t(e)$ for each edge $e \in P'$ satisfying the junction constraints. Based on P', Dick constructs an ARG P on m populations. Can Harry detect whether Tom's P' has any CPs i.e., admits admixture, based on the data given to him by Dick:

Scenario I: the ARG P;
Scenario II: a subARG of P that has all leaf nodes of P;
Scenario III: only the leaf nodes of P.

Outline of our approach to the solution. Note that given an ARG or subARG P, its underlying scaffold P' is not immediately computable. Due to recombinations, many CPs exist in P, but they do not necessarily indicate a CP in P'. Fig 2 shows some examples. In this paper, we resort to topology and translate this problem into persistence homology computation in the Vietoris-Rips complex defined by P. Notice that Scenario I is an ideal situation while Scenarios II and III correspond to practical situations, and, we focus on the latter.

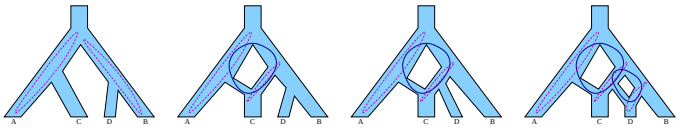


Fig. 2. Examples of CPs shown as solid dark closed paths in the respective ARGs. In contrast, the dashed closed paths cannot correspond to CPs in the underlying scaffolds.

2 Topology Model

In this section we develop a theoretical model that explains the topological signal for the presence or absence of admixtures in the populations being studied in the persistence diagrams that we compute. We model Scenario III of the last section as follows. Denote the leaf nodes of P, by $L(P)$. There exists a notion of distance between nodes v, v' of $L(P)$, denoted $w(v, v')$, obtained by setting

$$w(v, v') = \min_{u \in \mathrm{lca}(v,v')} \mathrm{depth}(u),$$

where $\mathrm{depth}(u)$ denotes the depth of the node u in P (measured in terms of the number of generations), where the depth of any leaf node of P is 0 and $\mathrm{lca}(v, v')$ is the set of least common ancestors of v and v' in P. Recall that the population labels of the leaf nodes (see last section) partitions $L(P)$ into disjoint subsets, where each subset corresponds to a population. Let the set of populations be denoted by $\tilde{L}(P)$. Thus, there exists a surjective map, $\phi : L(P) \to \tilde{L}(P)$. The distance function $w(\cdot, \cdot)$ on L, induces a distance function $\tilde{w}$ on $\tilde{L}(P)$, obtained by setting, for $A, B \in \tilde{L}(P)$ (where A, B are population labels),

$$\tilde{w}(A, B) = \min_{\substack{v \in L(P), \phi(v) = A, \\ v' \in L(P), \phi(v') = B}} w(v, v'). \tag{1}$$

Note that in our method described later, we do not need to know explicitly either the set $\tilde{P}$ or the surjective map ϕ. It is reasonable to assume that w and $\tilde{w}$ defined as above satisfy the following properties. There exists $c > 0$, with $c \ll \mathrm{depth}(P)$, where $\mathrm{depth}(P) = \max_{v \in P} \mathrm{depth}(v)$, and such that

Property 1. For each pair each pair $u, v \in L(P)$,

(a) $\phi(u) = \phi(v)$ implies that $w((u, v)) < c$;
(b) $\phi(u) \neq \phi(v)$ implies that $w((u, v)) > 2c$;
(c) For all $u', v' \in$ with $\phi(u) = \phi(u'), \phi(v) = \phi(v')$, $|w(u, v) - w(u', v')| < c$.

In other word, Property 1 implies that the distance between two leaf nodes of P carrying the same population label is very small, while those carrying different labels is large, and the latter distance depends only slightly on the chosen representatives, u, v, of the respective populations. Property 1 is an ideal property which if satisfied by the data implies a topological result relating the induced Vietoris-Rips complexes on $L(P)$, and on the set of populations $\tilde{L}(P)$ (using the distance measures w and $\tilde{w}$) by virtue of Theorem 1 below. Normally, the data will not satisfy this ideal property exactly – but never-the-less we observe a behavior which is close to what the mathematical theorem suggests.

Before stating the precise topological theorem we first explain the main idea.

The topological framework. Suppose that in a given finite metric space $M = (V, w)$, where $w : V \times V \to \mathbb{R}_{\geq 0}$, the values of w (i.e. the distances) occur in two scales. Suppose also that the points of V form clusters with pairwise distances amongst pairs in each individual cluster belong to the smaller of the two scales – while, the distance between two clusters, measured by taking the minimum of the pairwise distances between the points of the two clusters, belong to the larger scale. We denote the set of clusters by $\tilde{V}$ and denote the induced metric on $\tilde{V}$ by $\tilde{w}$.

Given any $d > 0$ (recall that d is "time", in generations, in P), the Vietoris-Rips complex of M with parameter d, which we denote by $\mathbf{Rips}(M, d)$ (see Definition 1), is a certain *simplicial complex* on V (i.e. a family of subsets of V closed under inclusion), and this complex grows with d. For small values of d (i.e. closer to the smaller scale) the Vietoris-Rips complex can have complicated topology (measured by the dimensions of the homology groups or the Betti numbers of the complex $\mathbf{Rips}(M, d)$) which depend only on the induced metric spaces on each of the separate clusters. As d grows, the various Vietoris-Rips subcomplexes corresponding to each cluster become contractible, and all homology groups in dimensions > 0 vanish (and thus the higher Betti numbers which are the dimensions of these homology groups vanish). After the value of d grows even further (i.e. reaches the larger scale), new homology classes in dimensions > 0 might be born and these classes correspond to those of the Vietoris-Rips complex associated to the space $\tilde{M} = (\tilde{V}, \tilde{w})$ obtained from M by clustering.

Persistent homology. A systematic way of understanding the birth and death of homology cycles in the Vietoris-Rips complex is through the persistent homology groups [3] (see Definition 2 for precise definition). Denoting by $\mathbf{Rips}(M, d)$ the Vietoris-Rips complex of M at "time" d, and for all $d' > d$, the inclusion homomorphism $i^{d,d'} : \mathbf{Rips}(M, d) \hookrightarrow \mathbf{Rips}(M, d')$ (which includes $\mathbf{Rips}(M, d)$

in the larger complex $\mathbf{Rips}(M, d'))$ induces a homomorphism

$$i_*^{d,d'} : \mathrm{H}_*(\mathbf{Rips}(M, d)) \to \mathrm{H}_*(\mathbf{Rips}(M, d'))$$

between their respective homology groups. Unlike, the homomorphism $i^{d,d'}$, $i_*^{d,d'}$ is not necessarily injective. A non-zero homology class in $\mathrm{H}_*(\mathbf{Rips}(M, d))$ can map to 0 under $i_*^{d,d'}$. The image of $i_*^{d,d'}$ – whose non-zero elements correspond to non-zero homology classes of $\mathrm{H}_*(\mathbf{Rips}(M, d))$ that *persists* till time d', is called the (d, d')*-th persistent homology group*, which we will denote by $\mathrm{H}_*^{d,d'}(M)$.

One would expect that the the persistent homology groups of the Vietoris-Rips complex associated to M (in dimensions > 0) will also show a separation with respect to the two scales. (The zero-th homology groups will not show such a separation for obvious reasons – and in fact by definition of the Vietoris-Rips complex the zero-th Betti number is just a decreasing function of d.) Moreover, one would expect that the homology classes of the Vietoris-Rips complex associated to $\tilde{M}$ which persists over long periods (which are the ones identified with the larger scale) already appear in the persistent homology of the Vietoris-Rips complex associated to $\tilde{M}$, while those associated to the smaller scale appear much earlier and die earlier.

Theorem 1 assures us that provided $M, \tilde{M}$ satisfy certain conditions (Property 1) any non-zero persistent homology class in $\mathrm{H}_i^{d,d'}(\tilde{M})$, is the image of a class in $\mathrm{H}_i^{d+c,d'}(M)$ (where c is a constant appearing in Property 1) and can be interpreted as an upper bound on the distances of the smaller scale. Thus, even though we do not have direct access to the Vietoris-Rips complexes of $\tilde{M}$, we can obtain information about its persistent homology from those of the Vietoris-Rips complexes of M. In addition, Theorem 1 also assures us of the separation on the time scale, of the homology in the Vietoris-Rips complex of M in the smaller time scale, from the "interesting" homology in the larger time scale which contributes to the homology of the Vietoris-Rips complex of $\tilde{M}$. Together they imply that the persistent homology of the Vietoris-Rips complexes of M contains information allowing us to read the persistent homology of the Vietoris-Rips complex $\tilde{M}$ if the latter is non-zero.

Precise definitions and statement of the topological theorem. To state the topological result alluded to above we first need some definition and notation. We first recall the well known definition of the Vietoris-Rips complex of a finite set V equipped with a distance function $w : V \times V \to \mathbb{R}_{\geq 0}$, satisfying $w(v, v) = 0$ for all $v \in V$.

Definition 1 (Vietoris-Rips Complex). *Let $M = (V, w)$ be a pair, where V is a finite set and $w : V \times V \to \mathbb{R}_{\geq 0}$ is a map (which need not be a metric on V) satisfying $w(v, v) = 0$ for all $v \in V$. Then, for any integer $d > 0$, we define the chain complex of the Vietoris-Rips complex of (M, d), which we will denote by $\mathbf{Rips}_\bullet(M, d) = (\mathbf{C}_\bullet(M, d), \partial_\bullet)$ as follows. Let, $V = \{1, \ldots, n\}$, and for each*

$p \geq 0$, *define*

$$\mathbf{C}_p(M,d) = \bigoplus_{\substack{U \subset V, \\ \mathrm{card}(U)=p+1, \\ \bigwedge_{u,u' \in U} w(u,u') \leq d}} \mathbb{Q} \cdot U.$$

The boundary map ∂_p *is defined by setting for each* $U = \{i_0, \ldots, i_p\} \subset V$, *with* $1 \leq i_0 < \cdots < i_p \leq n$, *where* $U_j = U \setminus \{i_j\}$:

$$\partial_p(U) = \sum_{j=0}^{p} (-1)^j \cdot U_j.$$

Definition 2 (Persistent homology groups of M). *For* $d \leq d'$, *the inclusion map* $i^{d,d'} : \mathbf{Rips}(M,d) \hookrightarrow \mathbf{Rips}(M,d')$ *induces homomorphisms* $i_\bullet^{d,d'} : \mathbf{Rips}_\bullet(M,d) \to \mathbf{Rips}_\bullet(M,d')$ *between the corresponding chain complexes, which in turn induces homomorphisms* $i_*^{d,d'} : \mathrm{H}_*(\mathbf{Rips}_\bullet(M,d)) \to \mathrm{H}_*(\mathbf{Rips}_\bullet(M,d'))$ *in homology. We call the image of* $i_*^{d,d'}$, *the* (d,d')*-th* persistent homology group *of* M *(see for example [3]), and we will denote this group by* $\mathrm{H}_*^{d,d'}(M)$.

We have the following proposition and theorem which relate the persistent homology groups of two pairs $M = (V,w)$ and $\tilde{M} = (\tilde{V}, \tilde{w})$ under certain conditions. [6]

Proposition 1. *Let* $M = (V,w), \tilde{M} = (\tilde{V}, \tilde{w})$ *be as in above with* V, V' *finite,* $c > 0$ *and* $\phi : V \to V'$ *a surjective map, such that for each pair* $u, v \in V$ *satisfies Property 1. Then,*

1. $\mathrm{H}_i(\mathbf{Rips}_\bullet(\tilde{M}, d)) = 0$ *for* $i > 0$, *and* $d < c$.
2. *For all* $d, d' \geq 0$ *satisfying* $d' - d > c$, *the homomorphism* ϕ_* *induced by* ϕ *satisfies*
$$\mathrm{H}_*^{d,d'}(\tilde{M}) \subset \phi_*(\mathrm{H}_*^{d+c,d'}(M)) \subset \mathrm{H}_*^{d+c,d'}(\tilde{M}).$$
Moreover, if $\tilde{i}_*^{d,d'} : \mathrm{H}_*(\mathbf{Rips}_\bullet(\tilde{M}, d)) \longrightarrow \mathrm{H}_*(\mathbf{Rips}_\bullet(\tilde{M}, d'))$ *is an isomorphism, then*
   (a) $\phi_*|_{\mathrm{H}_*^{d+c,d'}(M)} : \mathrm{H}_*^{d+c,d'}(M)$ *is a surjection on to* $\mathrm{H}_*^{d,d'}(\tilde{M})$.
   (b) $\phi_*|_{\mathrm{H}_*^{d,d'}(M)} : \mathrm{H}_*^{d+c,d'}(M)$ *is an injection in to* $\mathrm{H}_*^{d,d'}(\tilde{M})$.

Proof: The first claim immediately follows from Part (c) of Property 1. We now prove the second claim. We first check that for any $d > 0$, the map ϕ induces a simplicial map $\phi : \mathbf{Rips}(M,d) \to \mathbf{Rips}(\tilde{M}, d)$. To see this let $U \subset V$ such that $\bigwedge_{u,u' \in U} w(u,u') \leq d$. We claim that for each $u, u' \in U$, $\tilde{w}(\phi(u), \phi(u')) \leq d$. This follows immediately from the definition of $\tilde{w}$ (see Eqn 1). Notice that

[6] We change slightly the formulation of Proposition 1 and Theorem 1 from the published version. We thank Jose Maria Ibarra Rguez and Victor Prez Abreu from CIMAT, Mexico for bringing an issue with the previous formulation to our attention and helpful discussions.

the min function used in the definition of $\tilde{w}$ is crucial here. This proves that the induced map of ϕ is simplicial i.e. it carries simplices to simplices. Now suppose that $d' - d > c$, and consider a simplex in the Vietoris-Rips complex $\mathbf{Rips}(\tilde{M}, d)$ spanned by $\tilde{U} \subset \tilde{V}$. Since, $\tilde{U}$ is a simplex in the Vietoris-Rips complex, $\mathbf{Rips}(\tilde{M}, d)$, by definition $\bigwedge_{\tilde{u},\tilde{u}' \in \tilde{U}} \tilde{w}(\tilde{u}, \tilde{u}') \leq d$. Then, for all $u \in \phi^{-1}(\tilde{u}), u' \in \phi^{-1}(\tilde{u}')$, $w(u, u') \leq d + c$, (using Parts (a) and (b) of Property 1). Thus, the inverse image of the simplex spanned by $\tilde{U}$ in $\mathbf{Rips}(\tilde{M}, d)$, is contractible inside $\mathbf{Rips}(M, d + c) \hookrightarrow \mathbf{Rips}(M, d')$.

Let $f : \mathbf{Rips}(M, d + c) \to \mathbf{Rips}(\tilde{M}, d')$ denote the simplicial map defined by $f = \phi \circ i$, and let $K = f^{-1}(\mathbf{Rips}(\tilde{M}, d')$ noting that $\mathbf{Rips}(\tilde{M}, d)$ is a subcomplex of $\mathbf{Rips}(\tilde{M}, d')$. Note that since ϕ is surjective, we have the inclusion (of subcomplexes)

$$\mathbf{Rips}(M, d) \hookrightarrow K \hookrightarrow \mathbf{Rips}(M, d + c).$$

We thus have the following commutative diagram of simplicial maps.

$$\begin{array}{ccccccc}
\mathbf{Rips}(M,d) & \xrightarrow{i} & K & \xrightarrow{i} & \mathbf{Rips}(M,d+c) & \xrightarrow{i} & \mathbf{Rips}(M,d') \\
& \searrow^{\phi} & & & \downarrow{\scriptstyle \phi} & & \downarrow{\scriptstyle \phi} \\
& & \mathbf{Rips}(\tilde{M},d) & \xrightarrow{\tilde{i}} & \mathbf{Rips}(\tilde{M},d+c) & \xrightarrow{\tilde{i}} & \mathbf{Rips}(\tilde{M},d')
\end{array}$$

We have the following diagram in homology. The vertical isomorphism is by the Vietoris-Begle theorem (see for example [14, page 344]) and the fact that inverse image under f of each simplex in $\mathbf{Rips}(\tilde{M}, d)$, is contractible inside $\mathbf{Rips}(M, d + c)$ proved above.

$$\begin{array}{ccccccc}
\mathrm{H}_*(\mathbf{Rips}_\bullet(M,d)) & \xrightarrow{i_*} & \mathrm{H}_*(K) & \xrightarrow{i_*} & \mathrm{H}_*(\mathbf{Rips}_\bullet(M,d+c)) & \xrightarrow{i_*} & \mathrm{H}_*(\mathbf{Rips}_\bullet(M,d')) \\
& & \downarrow{\scriptstyle \cong} & & & & \downarrow{\scriptstyle \phi_*} \\
& & \mathrm{H}_*(\mathbf{Rips}_\bullet(\tilde{M},d)) & & \xrightarrow{\tilde{i}_*} & & \mathrm{H}_*(\mathbf{Rips}_\bullet(\tilde{M},d'))
\end{array}$$

From the above diagram it is clear that $\mathrm{H}^{d,d'}(\tilde{M}) \subset \mathrm{Im}(\phi_*|_{\mathrm{H}^{d+c,d'}(M)})$.

The inclusion $\mathrm{Im}(\phi_*|_{\mathrm{H}_*^{d+c,d'}(M)}) \subset \mathrm{H}_*^{d+c,d'}(\tilde{M})$ is also clear from the diagram

$$\begin{array}{ccc}
\mathrm{H}_*(\mathbf{Rips}_\bullet(M,d+c)) & \xrightarrow{i_*} & \mathrm{H}_*(\mathbf{Rips}_\bullet(M,d')) \\
\downarrow{\scriptstyle \phi_*} & & \downarrow{\scriptstyle \phi_*} \\
\mathrm{H}_*(\mathbf{Rips}_\bullet(\tilde{M},d+c)) & \xrightarrow{i_*} & \mathrm{H}_*(\mathbf{Rips}_\bullet(\tilde{M},d'))
\end{array}$$

Thus, we have

$$\mathrm{H}_*^{d,d'}(\tilde{M}) \subset \mathrm{Im}(\phi_*|_{\mathrm{H}_*^{d+c,d'}(M)}) \subset \mathrm{H}_*^{d+c,d'}(\tilde{M}).$$

Now suppose that $\tilde{i_*}^{d,d'}$ is an isomorphism. Then,

$$\mathrm{H}_*(\mathbf{Rips}_\bullet(\tilde{M},d)) \cong \mathrm{H}_*^{d+c,d'}(\tilde{M}) = \mathrm{H}_*^{d,d'}(\tilde{M}),$$

and it follows that

1. $\phi_*|_{\mathrm{H}_*^{d+c,d'}(M)}$ is a surjection to $\mathrm{H}_*^{d,d'}(\tilde{M})$, and
2. $\phi_*|_{\mathrm{H}^{d,d'}(M)}$ is an injection to $\mathrm{H}_*^{d,d'}(\tilde{M})$.

□

Theorem 1. *Let $M = (V,w), \tilde{M} = (\tilde{V},\tilde{w})$ be as in above with V, V' finite, $c > 0$ and $\phi : V \to V'$ a surjective map, such that for each pair $u, v \in V$ satisfies Property 1.*

Then, for all $d, d' \geq 0$ satisfying $d' - d > 2c$, if $\tilde{i}_*^{d,d'} : \mathrm{H}_*(\mathbf{Rips}_\bullet(\tilde{M},d)) \longrightarrow \mathrm{H}_*(\mathbf{Rips}_\bullet(\tilde{M},d'))$ is an isomorphism, then

$$\phi_*|_{\mathrm{H}_*^{d+c,d'}(M)} : \mathrm{H}_*^{d+c,d'}(M) \to \mathrm{H}_*^{d,d'}(\tilde{M})$$

is an isomorphism.

Proof: Apply Proposition 1 twice, second time with d replaced by $d + c$. □

Theorem 1 is applicable in the context of Scenario III as follows. Take $M = (L(P), w)$ and $\tilde{M} = (\tilde{L}(P), \tilde{w})$. Further suppose that for $A, B \in \tilde{L}(P), A \neq B, |\tilde{w}(A) - \tilde{w}(B)| > 2c$ (say), that is distinct populations are separated by a larger distance than individuals within the same population.

In this case the surjection given in Theorem 1 implies that the presence of persistent homology (i.e. homology cycles that are born after $d = 2c$ and that persists for intervals of length $> c$) in the Vietoris-Rips complex of $\tilde{M}$ can be detected from that of the Vietoris-Rips complex of M. Hence, for all small values of d, d', i.e. $0 < d < d' < c$, the persistent homology groups, $\mathrm{H}_*^{d,d'}(M)$ reflect the topology of the ARG P, created by the recombination events. For $c < d < d + 3c < d'$ by Theorem 1, there is a surjection

$$\mathrm{H}_*^{d+c,d'}(M) \to \mathrm{H}_*^{d,d'}(\tilde{M}).$$

which is an isomorphism if $\tilde{i}_*^{d,d'}$ is an isomorphism, and any persistent homology (in dimension > 0) in this range can be attributed to the cycles in the population graph P' which are caused by admixture.

Topological Signatures. The theorem thus predicts that the presence of admixtures should be detectable from the persistent homology diagrams of the Vietoris-Rips complex of M itself. This is indeed seen in the experimental results. In Figs 3-7, we display the results of computing the homology groups of the Rips complexes obtained from both simulated as well as real data. We take $M = (L(P), w)$ where P is an ARG obtained either from simulated or real data.

Fig 4 shows results for real data while the others are for simulated data. The horizontal axis corresponds to the values of d, and for each fixed d, the number of horizontal lines above is the dimension of homology group of the Vietoris-Rips complex corresponding to this value of d. Thus, each horizontal line depicts the "life" of a non-zero homology cycle. The x-coordinate of its left end point is the time of its "birth" and the right end point the time of its "death". We see a clear separation between persistent cycles in dimensions > 1, in the case of admixed populations – which can be seen as a signal indicating presence of admixture.

3 Experiments

We first describe the simulation experiments. The populations were simulated using SimRA [2]. Once the set of haplotypes were generated for all three populations, we created a distance matrix between all pairs of haplotypes using Hamming distance metric. The Vietoris-Rips complex was constructed on the graph embedding of the distance matrix (a complete graph with each vertex corresponding to an individual haplotype and edge weights corresponding to the Hamming distance between the pair of haplotypes). We computed homology groups on the Vietoris-Rips complex for zero and one dimensions using Javaplex v4.2.0 [15]. Recall that the dimension of the zero-dimensional homology group of a simplicial complex counts the number of connected components of the simplicial complex, while the dimension of the one-dimensional homology group counts the number of independent one-dimensional cycles which do not bound.

In the results, irreducible cycles computed from the simulation experiments are presented as barcode plots, which display when individual cycles representing non-zero one-dimensional homology classes are born and when they disappear. The upper half of each barcode plot for the simulation experiments display the persistence of zero-dimensional homology, while the lower half display barcode line segments indicate the persistence of one-dimensional homology. While short cycles can be due to noise, longer (persistent) cycles represent fundamental topological structures within the genetic distance matrix.

Fig 3 shows the topological signatures in the context of presence and absence of admixture. The persistent cycles for dimension > 0 clearly separate into two groups. Figs 6-7 in the appendix show the results of experiments with different simulation parameters, including stochasticity of the ARGs.

3.1 Experiments on avocado germplasm

We consider three main avocado cultivars: West Indian (W), Guatamelan (G) and Mexican (M). Moreover, we also consider an F1 population WxG. Each of the group is composed of 19 samples, from which we have 3348 markers. The genotype data were phased using Beagle [1] and both haplotypes are used in our experiments. In particular, using these four groups, we created two datasets to match our simulation study set-up: one composed of W, G and WxG samples

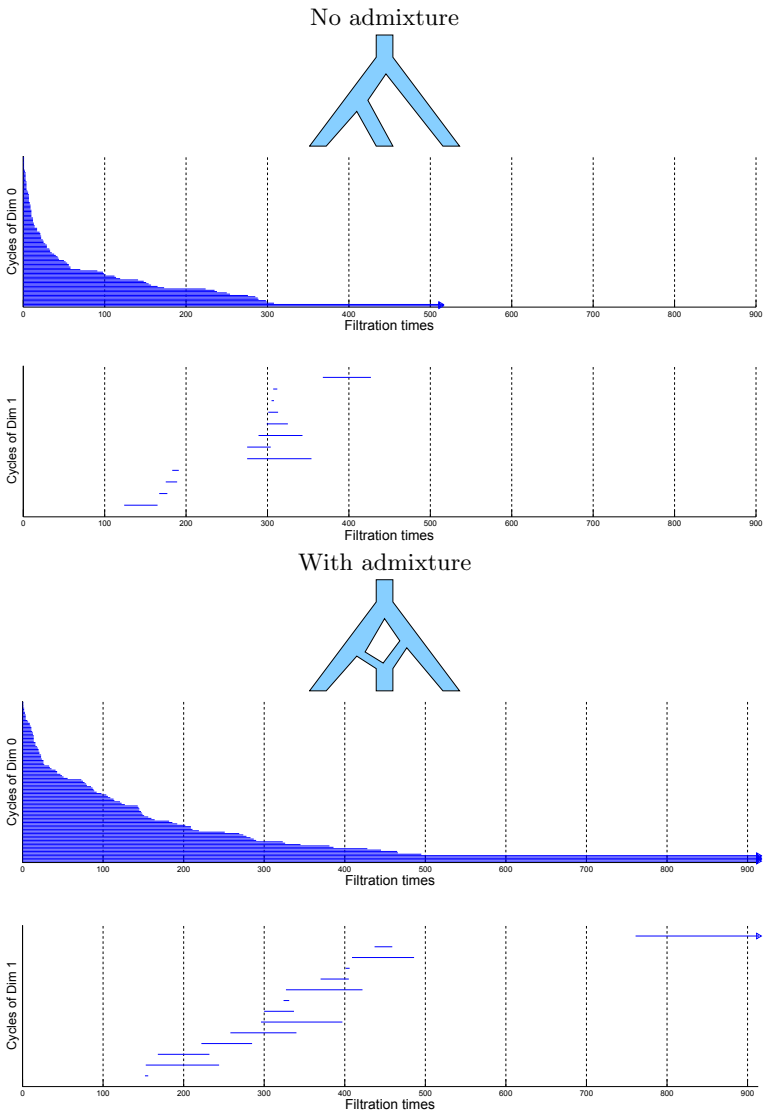


Fig. 3. Topology signatures embedded in the ARGs, on simulated data. There is an absence of admixture in the top while a presence in the bottom panel. This proof-of-concept experimental setting shows that, in ideal scenarios of simulations, topological signatures for recombinations and admixture can be differentiated (notice, in particular, the separation of the persistent cycles of dimension > 0). In the simulations, the effective population size is $N = 10K$. See text for further details.

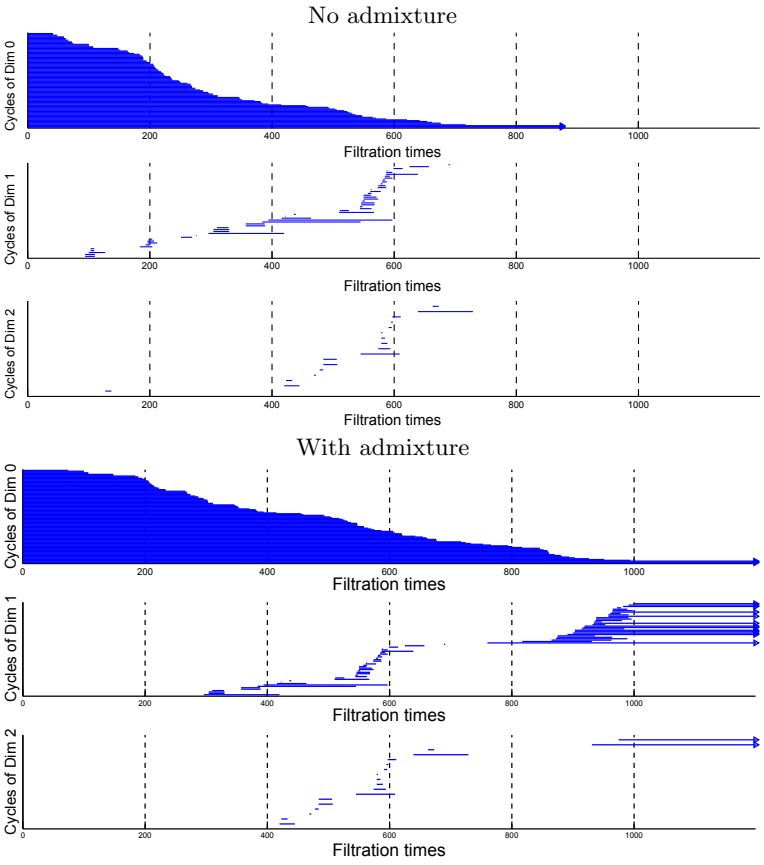


Fig. 4. Haplotypes from three groups of avocado germplasm data: West Indian (W), Mexican (M), and Guatemalan (G). The top plot corresponds to the populations with no admixture, while the bottom admits admixture in the populations. Notice the separation of the persistent cycles in both dimension 1 and 2 for the latter scenario, while the former shows no clear separation.

and the other of G, M and W. The former set admits admixture while the latter does not.

In order to compute the persistent homology groups on the avocado germplasm data, we concatenated SNP loci from all 12 chromosomes into a single sequence for each haplotype and computed the distance matrix based on the Hamming distance metric as described above. For the two avocado germplasm datasets, we computed zero, one and two-dimensional cycles representing non-zero elements of the persistent homology groups on the Vietoris-Rips complex using Javaplex. Fig 4 shows barcode plots describing zero, one and two-dimensional topological signatures on these two avocado germplasm data sets with and without admixture present. Further analysis of the persistent cycles in terms of their mean length and variances again shows distinguishing characteristics: see Fig 5 in the appendix.

4 Conclusion

We present the first combinatorial approach to characterizing admixture in populations, based on ARGs. Traditionally admixture has been addressed by studying linkage disequilibrium distributions. In this study, we show through controlled simulations that it is feasible to detect admixture by topological signatures. Moreover, when the model was applied on avocado germplasm data, we observed similar signatures of the persistent cycles, as was seen in the simulation experiments. Due to noise and other unknown factors in real data, the signatures may be require to be calibrated (i.e., values of c in Section 2) based on training data. This preliminary work is promising and in our future work, we plan to explore more complex admixture models, both in terms of complex topology of P' as well as complex characterizations of admixture. We believe that the topological signatures have the potential for not only detecting but also discriminating ancient from recent admixture in multiple populations.

Acknowledgments

Some of the work was done while DY was a summer intern at IBM T. J. Watson Research Center and APC was visiting the Center for her doctoral work.

Authors' contributions

LP and SB defined the mathematical model. LP and FU designed the study; FU and DY carried out the experiments; APC carried out the scaffold and population simulations. DK provided the real data and the result interpretation.

References

1. Browning, S., Browning, B.: Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics 81(5), 1084–1097 (2007)

2. Carrieri, A.P., Utro, F., Parida, L.: Accurate and efficient sampling of underlying ARG of multiple populations under subdivision and admixture. Bioinformatics (2016)
3. Edelsbrunner, H., Harer, J.L.: Computational topology. American Mathematical Society, Providence, RI (2010)
4. Freedman, M.L., Haiman, C.A., Patterson, N., McDonald, G.J., Tandon, A., Waliszewska, A., Penney, K., Steen, R.G., Ardlie, K., John, E.M., Oakley-Girvan, I., Whittemore, A.S., Cooney, K.A., Ingles, S.A., Altshuler, D., Henderson, B.E., Reich, D.: Admixture mapping identifies 8q24 as a prostate cancer risk locus in african-american men. Proceedings of the National Academy of Sciences 103(38), 14068–14073 (2006)
5. Greenbaum, B.D., Li, O.T., Poon, L.L., Levine, A.J., Rabadan, R.: Viral reassortment as an information exchange between viral segments. Proceedings of the National Academy of Sciences 109(9), 3341–3346 (2012)
6. Griffiths, R.C., Marjoram, P.: Ancestral inference from samples of dna sequences with recombination. Journal of Computational Biology 3(4), 479–502 (1996)
7. Javed, A., Pybus, M., Melè, M., Utro, F., Bertranpetit, J., Calafell, F., Parida, L.: IRiS: Construction of ARG network at genomic scales. Bioinformatics 27, 2448–2450 (2011)
8. Javed, A., Melé, M., Pybus, M., Zalloua, P., Haber, M., Comas, D., Netea, M.G., Balanovsky, O., Balanovska, E., Jin, L., et al.: Recombination networks as genetic markers in a human variation study of the old world. Human genetics 131(4), 601–613 (2012)
9. Jobling, M., Hollox, E., Hurles, M., Kivisild, T., Tyler-Smith, C.: Human evolutionary genetics. Garland Science, UK (2013)
10. Kearsey, M., Pooni, H.: The Genetical Analysis of Quantitative Traits. Stanley Thornes, UK (2004)
11. Loh, P.R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J.K., Reich, D., Berger, B.: Inferring admixture histories of human populations using linkage disequilibrium. Genetics 193(4), 1233–1254 (2013)
12. Parida, L.: Ancestral Recombinations Graph: A reconstructability perspective using random-graphs framework. Journal of Computational Biology 17, 1345–1350 (2010)
13. Semon, M., Nielsen, R., Jones, M.P., McCouch, S.R.: The population structure of african cultivated rice oryza glaberrima (steud.) evidence for elevated levels of linkage disequilibrium caused by admixture with o. sativa and ecological adaptation. Genetics 169(3), 1639–1647 (2005)
14. Spanier, E.H.: Algebraic topology. McGraw-Hill Book Co., New York (1966)
15. Tausz, A., Vejdemo-Johansson, M., Adams, H.: Javaplex: A research software package for persistent (co)homology. Software available at `http://javaplex.github.io/` (2011)
16. Wall, J., Hammer, M.: Archaic admixture in the human genome. Current opinion in genetics & development 16(6), 606–610 (2006)

Appendix: Additional Experiments

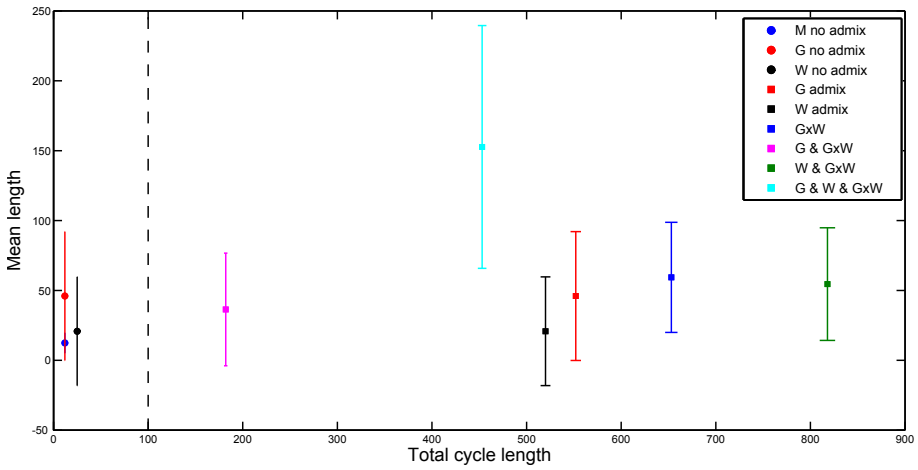


Fig. 5. Analysis of the persistent cycles of avocado germplasm data: It shows that the admixed samples have larger cycle lengths (> 100). The mean length for the 3 admixed populations is larger than the other cases. Also, the individuals for the G and W cultivars are the same in both the experiments, and they have comparable mean length (the red and black lines in the plot), while the total cycle lengths are different.

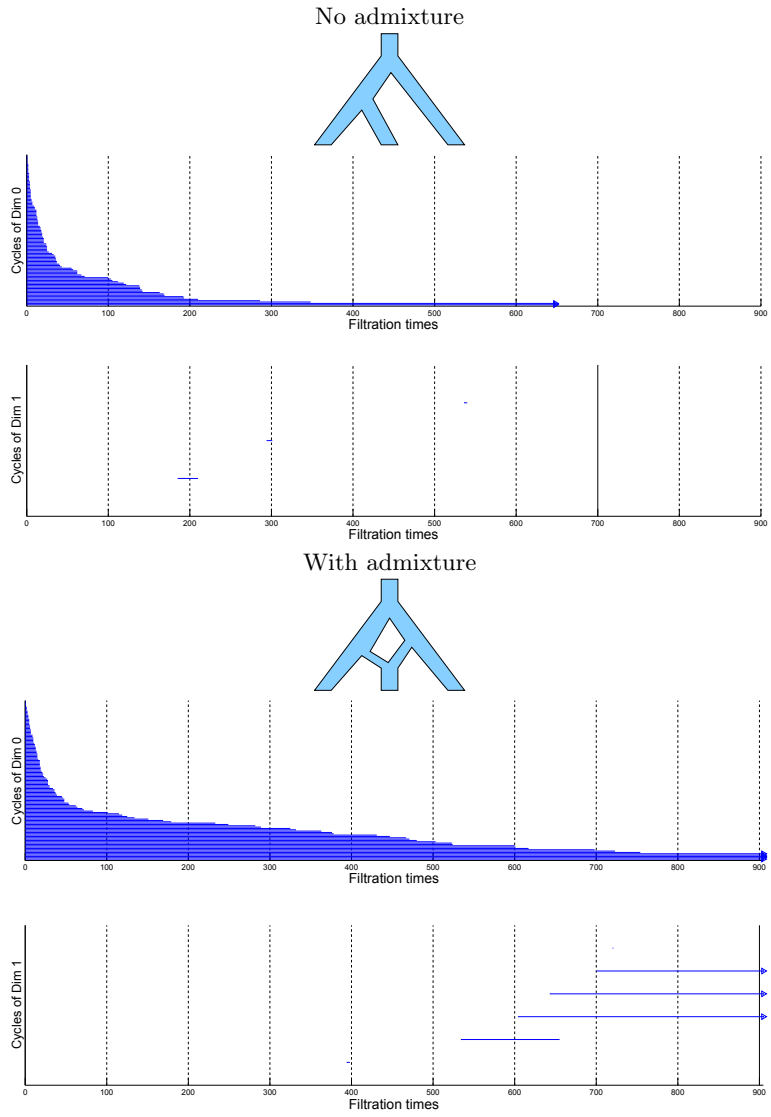


Fig. 6. In the simulations recombination rate $r = 0.1 \times 10^{-8}$. Notice that in the absence of recombinations, no particular separation of persistent cycles is observed. In the simulations, the effective population size is $N = 10K$.

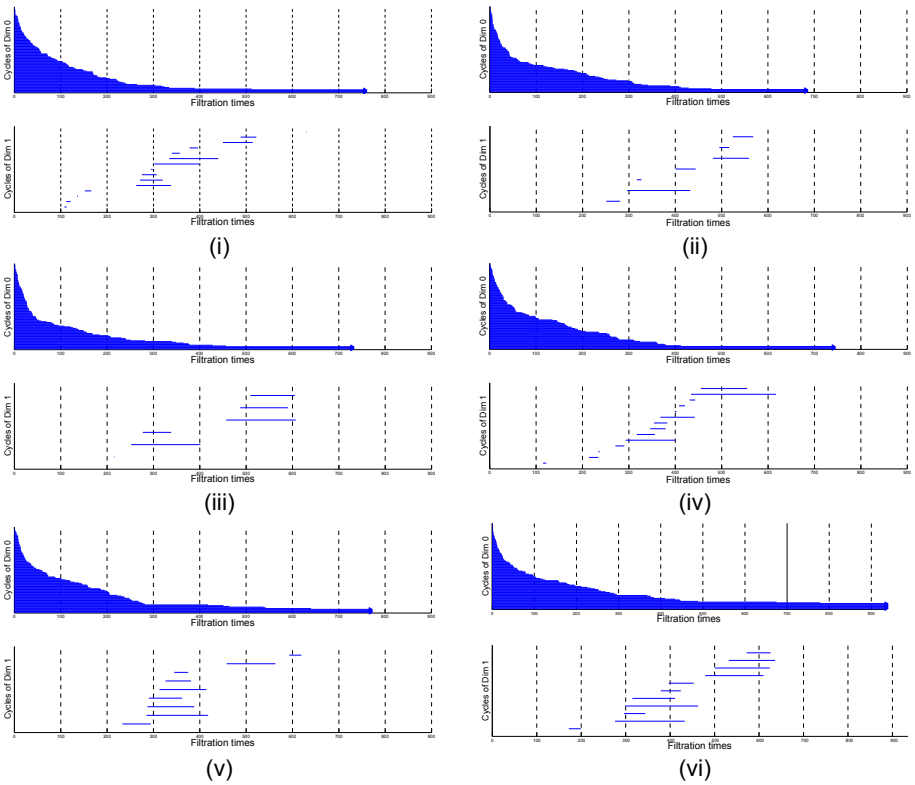


Fig. 7. Six simulations, each with effective population size $N = 10K$; with recombination ($r = 0.3 \times 10^{-8}$) as well as admixture to show that stochasticity does not affect the topological signature, i.e. the separation of the persistent cycles into roughly two groups.