Essential Simplices in Persistent Homology and Subtle Admixture Detection

Saugata Basu[1]
Department of Mathematics, Purdue University, West Lafayette, IN 47906, USA
sbasu@math.purdue.edu
https://orcid.org/0000-0002-2441-0915

Filippo Utro
Computational Biology Center, IBM T. J. Watson Research,
Yorktown Heights, NY 10598, USA
futro@us.ibm.com
https://orcid.org/0000-0003-3226-7642

Laxmi Parida
Computational Biology Center, IBM T. J. Watson Research,
Yorktown Heights, NY 10598, USA
parida@us.ibm.com
https://orcid.org/0000-0002-7872-5074

Abstract

We introduce a robust mathematical definition of the notion of essential elements in a basis of the homology space and prove that these elements are unique. Next we give a novel visualization of the essential elements of the basis of the homology space through a rainfall-like plot (RFL). This plot is data-centric, i.e., is associated with the individual samples of the data, as opposed to the structure-centric barcodes of persistent homology. The proof-of-concept was tested on data generated by SimRA that simulates different admixture scenarios. We show that the barcode analysis can be used not just to detect the presence of admixture but also estimate the number of admixed populations. We also demonstrate that data-centric RFL plots have the potential to further disentangle the common history into admixture events and relative timing of the events, even in very complex scenarios.

2012 ACM Subject Classification Applied computing → Life and medical sciences

Keywords and phrases population admixture, topological data analysis, persistent homology, population evolution

Digital Object Identifier 10.4230/LIPIcs.WABI.2018.14

1 Introduction

A fascinating way to study the relationship between multiple individuals is to understand their potential common history implicated by the genetic signatures of the individuals [6, 7]. If the individuals were bacteria, then their common history is bound to be captured by a tree; while sexually reproducing organisms such as humans have the common history represented as directed acyclic graphs (DAG). A layer of complexity is introduced to the common history through populations. While the common history of individuals *within* a population is captured by a DAG, populations can admix i.e., individuals of two populations can interbreed, at a specific time or time-interval, leaving behind yet a different kind of

[1] The work was partially supported by NSF grant DMS-1620271.

imprint on the genome of the individuals [4, 11]. See Fig 1 for an illustration: A, B, C, and D are extant populations whose common (evolutionary) history is as shown by the structure on the left where the *scaffold* is shown as the DAG with dashed lines. The implicit direction is assumed to flow downwards: the past on the top and the present at the bottom leaf nodes. Note that the scaffold is not a tree; hence there are two admixed populations: C and D. But A and B are not admixed. The figure on the left shows the *macro-level* structure while two pictures on the right show the *micro-level* structure, i.e., transmission of genetic material over time (across generations) in the individuals. The interested reader is directed to [9] for a detailed exposition.

It is easy to appreciate that the DAG due to sexual reproduction within the population (micro-level) is intricately entangled with the DAG due to the admixing populations (macro-level). However, the data available for study is the genome of the individuals of the extant populations only and not necessarily other intermediate (possibly extinct) populations. In fact, in [9] it was shown that given a mixture of populations (possibly without population labels) persistent homology can be used to detect if any admixing event had occurred in their common history. In this paper we extend the analysis to identifying multiple (not just presence or absence of at least one) admixture events. In other words, the question then is whether it is possible to tease apart some essentials of the macro structures (such as the 2 cycles in the structure on the left in the figure) from a collection of individuals which have randomly been drawn from multiple extant populations.

In our experiments we use SimRA [1] to simulate the population data to define the gold-truth. Then we employ persistent homology to address our questions. In literature, barcode diagrams [2, 5, 8] have been widely used to succintly represent persistent homology (see Definition 2 below): in each dimension, the start point of a bar marks the birth and the end point marks the death of a non-zero homology class in that dimension. To detect admixture, not only do we utilize the arrangement (pattern) of the bars in the different dimension, but we also need to associate them to the individuals or the data points.

Each bar in the bar code diagram represents a non-zero homology class – and each such class can be represented by a cycle i.e. by a linear combination of simplices with vanishing boundary. However, this representation is not unique for several reasons – for example, a homology class is defined only up to boundaries (cycles which bound one higher dimensional linear combinations of simplices). We would like to associate to each bar, a unique set of simplices (whose vertices represent individuals). The problem of associating a particular cycle to a bar is an interesting problem in its own right and several approaches has been taken by researchers leading to difficult optimizing problems (for example, computing the shortest length cycle in a homology class). In this paper, we take a different approach. We give a mathematical definition of a well defined (non-empty) set of essential simplices associated to each bar of the persistence diagram that we compute (see Definition 9 below). In this way we resolve the inherent ambiguity of choosing a representative cycle for each homology class. We believe that this new notion of essential simplices can have implications for a host of applications going beyond the one described in this paper.

In our experiments, we observe that the clustering of the irreducible cycles in the barcode plot capture the admixing events in the population history. This is reinforced by using essential simplices, that further segregates the individuals.

Roadmap. In the next section we give the mathematical underpinnings for the essential simplices and in Section 3 we apply this to simulated population data. In Section 3.1 we describe the visualization of the essential simplices and summarize the results.

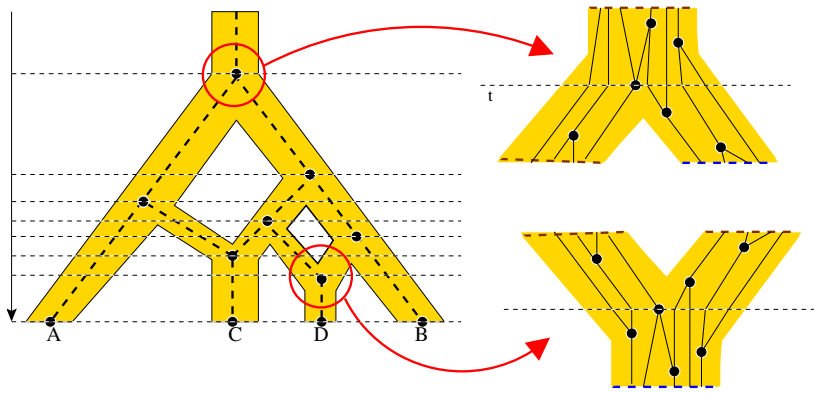


Figure 1 Macro-level structure on the left while two micro-level structures are shown on the right for four populations A, B, C, and D. See text for further details.

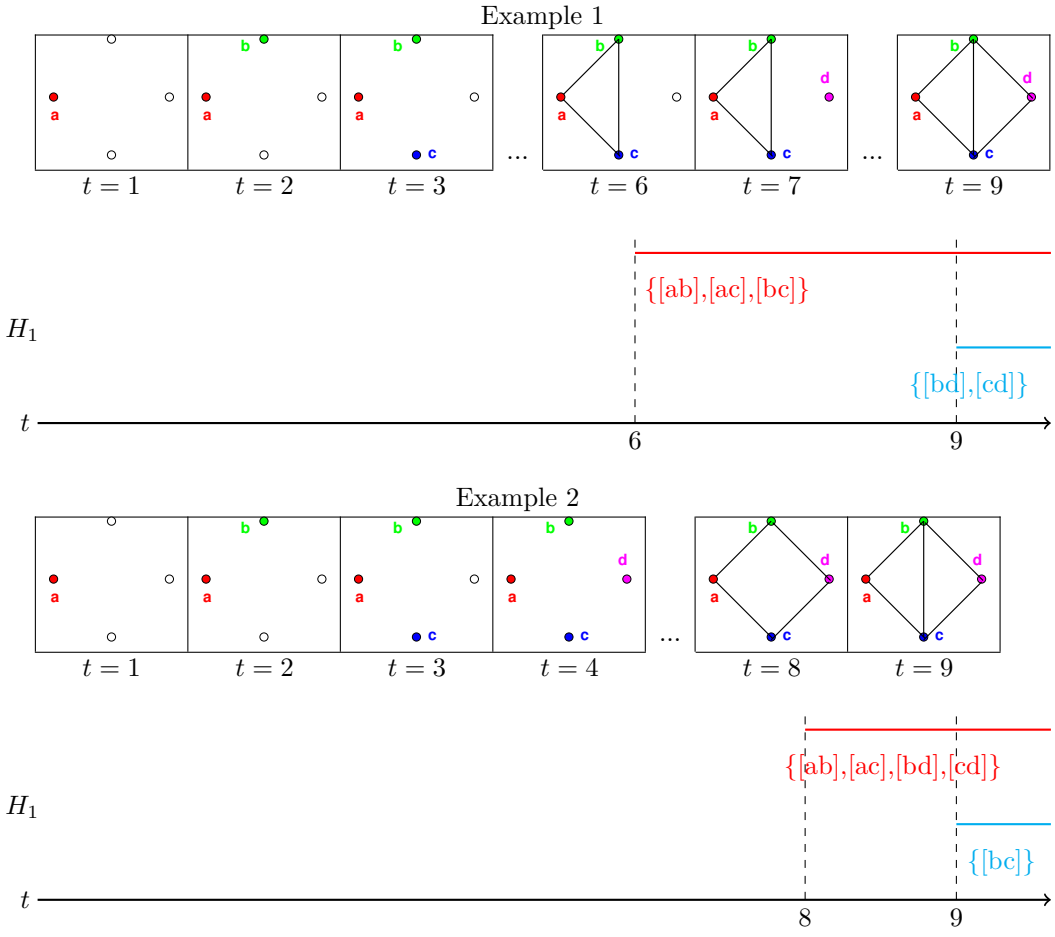


Figure 2 Bar codes for H_1 and their associated essential simplices for the filtrations in Example 10.

2 Filtrations of finite simplicial complexes, persistent homology, and essential simplices

One important problem in persistent homology theory is to associate to a homology class an actual cycle representing the homology class. Several methods have been developed to choose such a class optimally (with respect to various cost functions). In this paper we choose a different approach. Given a filtered simplicial complex, where the filtration satisfies a certain property (see Definition 3 and Proposition 5 below), we associate to each bar of the bar diagram of the p-th persistent homology of a filtration of a simplicial complex a non-empty set of p-simplices of the simplicial complex in a canonical way – which we call the set of *essential* simplices. Roughly speaking, a p-simplex σ is essential for a bar in the diagram of the p-th persistent homology if and only if σ appears with a non-zero coefficient in any linear combination of p-simplices representing the homology cycle corresponding to the bar. The rest of this section will make this intuitive notion rigorous.

We consider homology with coefficients in $\mathbb{Z}_2$, and which we omit from the notation. We also assume familiarity with basic simplicial homology theory, and denote by $H_*(K)$ the homology groups of a simplicial complex K with coefficients in $\mathbb{Z}_2$. Abusing notation a bit, we will denote by the same letter K the set of simplices of a simplicial complex K, and for $p \geq 0$, we will denote by $K^{(p)}$ the set of simplices of dimension p in K, so that

$$K = \cup_{p \geq 0} K^{(p)}.$$

We now recall the basic definitions pertaining to persistent homology.

Let $\mathcal{F}$ denote a filtration of a finite simplicial complex K given by : $\emptyset = \cdots = K_{-1} = K_0 \subset K_1 \subset \cdots \subset X_s \subset K_{s+1} \subset \cdots \subset K_N = K_{N+1} = \cdots = K$. Here each K_i is a subcomplex of K.

▶ **Notation 1.** *For $s \leq t$, we let $i_n^{s,t} : H_n(K_s) \longrightarrow H_n(K_t)$, denote the homomorphism induced by the inclusion $K_s \hookrightarrow K_t$.*

With the same notation as in the previous section we define:

▶ **Definition 2.** [2] For each triple (n, s, t) with $s \leq t$ the corresponding *persistent homology group*, $H_n^{s,t}(\mathcal{F})$ is defined by

$$H_n^{s,t}(\mathcal{F}) \quad = \quad \mathrm{Im}(i_n^{s,t}).$$

Note that $H_n^{s,t}(\mathcal{F}) \subset H_n(K_t)$, and $H_n^{s,s}(\mathcal{F}) = H_n(K_s)$.

We will consider only a special kind of filtration on simplicial complexes which are induced by orderings on the simplices of the complex satisfying the following property.

▶ **Definition 3** (Admissible ordering)**.** Let K be a finite simplicial complex. We call a total ordering $<$ (of the simplices) of K to be admissible if it satisfies the condition that $\sigma \prec \tau \Rightarrow \sigma < \tau$ for all simplices $\sigma, \tau \in K$. We will denote by $\mathrm{rk}_< : K \to [0, \mathrm{card}(K) - 1]$ the rank function of the ordering $<$.

▶ Remark 4. Note that the rank function, $\mathrm{rk}_<$, corresponding to an admissible ordering $<$ of the simplices of a simplicial complex K, is a discrete Morse function on K in the sense of Forman [3], for which every simplex is a critical simplex (in the sense of discrete Morse theory).

▶ **Proposition 5.** *Let K be a finite simplicial complex, and $<$ an admissible ordering of K. For $s \in [0, \mathrm{card}(K) - 1]$ let $K_s = \{\sigma \in K \mid \mathrm{rk}_<(\sigma) \leq s\}$. Then, $K_0 \subset K_1 \subset \cdots \subset K_{\mathrm{card}(k)-1} = K$ is a filtration of simplicial complexes. (We extend as usual the above filtration by setting $K_i = \emptyset$ for $i < 0$, and $K_j = K$ for all $j \geq \mathrm{card}(K)$, will refer to this filtration as the one induced by the ordering $<$.)*

Proof. The proof is easy and omitted. ◀

▶ **Proposition 6.** *Let $<$ be an admissible ordering of a finite simplicial complex K and let $\mathcal{F}$ denote the induced filtration of K. Then, for each $s, 0 \leq s \leq \text{card}(K) - 1$ and $p \geq 0$,* $\dim H_p(K_{s+1})/i_p^{s,s+1}(H_p(K_s)) \leq 1$.

Proof. Note that the rank function $\text{rk}_<$ is a discrete Morse function for which every simplex is critical (cf. Remark 4). The proposition is a consequence of the basic results of discrete Morse theory. ◀

▶ Remark 7. As a consequence of Proposition 6, we have that for each $p \geq 0$, and $s \geq 0$, the bar diagram for the p-dimensional persistent homology corresponding to an admissible filtration has at most one bar starting at time s.

Moreover, if $\sigma \in K^{(p)}$ with $\text{rk}_<(\sigma) = s$, then if there exists a cycle of the form $\sigma + \sum_{\tau \neq \sigma, \text{rk}_<(\tau) < \text{rk}_<(\sigma)} n_\tau \cdot \tau, \in Z_p(K_s)$, then and $H_p(K_s)/i_*^{s-1,s}(H_p(K_s)) \neq 0$, and this cycle represents the unique non-zero class in $H_p(K_s)/i_*^{s-1,s}(H_p(K_s)) \neq 0$.

▶ **Notation 8.** *For any filtration $\mathcal{F}$ and $p \geq 0$, we will denote by $\text{Bar}_p(\mathcal{F})$ to be the set of pairs (s,t) where each pair (s,t) corresponds to a bar starting at time s and ending at time t, in the bar diagram of the p-dimensional persistent homology of $\mathcal{F}$. For $c = (s,t) \in \text{Bar}_p(\mathcal{F})$, we denote $s(c) = s$.*

We are now in a position to define the set of essential simplices associated to a bar of a filtration induced by an admissible ordering.

▶ **Definition 9** (Essential simplices)**.** Let $\mathcal{F}$ be a filtration of a finite simplicial complex K induced by an admissible ordering. Suppose that for some $s \geq 0$, $H_p(K_s)/i_*^{s-1,s}(H_p(K_s)) \neq 0$, and suppose that $\dim H_p(K_s)/i_*^{s-1,s}(H_p(K_{s-1})) = 1$. Then, there is a unique bar $c \in \text{Bar}_p(\mathcal{F})$, with $s(c) = s$, in the bar diagram of $\mathcal{F}$. We call a p-simplex σ_0 to be *essential* with respect to c, if σ_0 satisfies the following condition.

For all $z = \sum_\sigma n_\sigma \cdot \sigma \in Z_p(K_s)$ (where the sum if taken over all p-simplices σ in the complex K_s), such that the image of z in $H_p(K_s)/i_p^{s-1,s}(H_p(K_{s-1}))$ under the canonical homomorphism is not zero, the coefficient n_{σ_0} of σ_0 is not equal to 0.

We will denote the set of essential simplices corresponding to c by Σ_c.

Before proceeding further we consider some examples.

▶ **Example 10** (Example of filtrations induced by admissible orderings and essential simplices)**.** Consider a graph on 4 vertices labelled a, b, c, d with 5 edges $[ab], [ac], [bc], [bd], [cd]$ (see Figure 2). Let $<$ be the admissible order

$$[a] < [b] < [c] < [ab] < [ac] < [bc] < [d] < [bd] < [cd].$$

Consder the bar diagram of the 1-dimensional homology for the induced filtration. It has two bars, $c = (6, \infty), c' = (9, \infty)$. It is easy to check that

$$\begin{aligned} \Sigma_c &= \{[ab], [ac], [bc]\}, \\ \Sigma_{c'} &= \{[bd], [cd]\}. \end{aligned}$$

Now consider the same graph but with a different admissible ordering. Let $<'$ be the order defined by

$$[a] <' [b] <' [c] <' [d] <' [ab] <' [ac] <' [bd] <' [cd] <' [bc].$$

The bar diagram of the 1-dimensional homology for the induced filtration again has two bars, $d = (8, \infty), d' = (9, \infty)$, and in this case we have

$$\begin{aligned} \Sigma_d &= \{[ab], [ac], [bd], [cd]\}, \\ \Sigma_{d'} &= \{[bc]\}. \end{aligned}$$

▶ Remark 11. As discussed earlier the set of essential simplices associated to a bar in the bar diagram corresponding to the filtration induced by an admissible ordering intuitively consists of simplices that must be present in any cycle representation of the homology cycle being born at that moment (the start time of the bar). In applications, this set of simplices can thus be considered as essential for the existence of the bar.

▶ **Theorem 12.** *Let $\mathcal{F}$ be a filtration induced by a perfect function, $p \geq 0$ and $c \in \mathrm{Bar}_p(\mathcal{F})$. Suppose that $z = \sum_{\sigma \in \Sigma} n_\sigma \cdot \sigma \in Z_p(K_s)$ (where the sum if taken over all p-simplices σ in the complex K_s), is such that its image in $H_p(K_s)/i_*^{s-1,s}(H_p(K_{s-1}))$ under the canonical homomorphism is not zero, and $n_\sigma \neq 0$ for each $\sigma \in \Sigma$.*

Then,

$$\Sigma_c = \Sigma \setminus \bigcup_{c' \in \mathrm{Bar}_p(\mathcal{F}), s(c') < s(c)} \Sigma_{c'}. \tag{1}$$

Proof. Let $\sigma \in \Sigma_c$. We prove that $\sigma \in \Sigma \setminus \bigcup_{c' \in \mathrm{Bar}_p(\mathcal{F}), s(c') < s(c)} \Sigma_{c'}$. By the defining property of Σ_c, it is clear that $\sigma \in \Sigma$. Now suppose that $\sigma \in \Sigma_{c'}$ for some c' with $s' = s(c') < s(c)$. Then, there exists a cycle,

$$\sum_\tau m_\tau \cdot \tau, \tag{2}$$

representing the class $H_p(K_{s'})/i_*^{s'-1,s'}(H_p(K_{s'-1}))$, with $m_\sigma \neq 0$, and $\mathrm{rk}_<(\tau) < s(c)$ for each τ with $m_\tau \neq 0$. Thus, there exists a relation

$$\sigma = \sum_{\tau \neq \sigma} m_\tau \cdot \tau \mod i_*^{s'-1,s'}(H_p(K_{s'-1})) \tag{3}$$

with $\partial_p \sigma = \partial_p \left(\sum_{\tau \neq \sigma} m_\tau \cdot \tau \right)$, and $\mathrm{rk}_<(\tau) < s(c)$ for each $\tau \neq \sigma$ with $m_\tau \neq 0$. Moreover, it is clear from the definition of persistent homology and the fact that $s > s'$, that, $\sigma = \sum_{\tau \neq \sigma} m_\tau \cdot \tau$ mod $i_*^{s-1,s}(H_p(K_{s-1}))$ as well.

Thus, we can substitute for σ in (2) by the right hand side of (2) and thus obtain an equivalent expression for the cycle representing the non-zero class in $H_p(K_s)/i_*^{s-1,s}(H_p(K_{s-1}))$ which does not contain σ, thus contradicting the fact that $\sigma \in \Sigma_c$. This proves that $\sigma \notin \bigcup_{c' \in \mathrm{Bar}_p(\mathcal{F}), s(c') < s(c)} \Sigma_{c'}$, proving that

$$\sigma \in \Sigma \setminus \bigcup_{c' \in \mathrm{Bar}_p(\mathcal{F}), s(c') < s(c)} \Sigma_{c'}.$$

This proves the inclusion $\Sigma_c \subset \Sigma \setminus \bigcup_{c' \in \mathrm{Bar}_p(\mathcal{F}), s(c') < s(c)} \Sigma_{c'}$.

We now prove the reverse inclusion.

Suppose that $\Sigma \setminus \bigcup_{c' \in \mathrm{Bar}_p(\mathcal{F}), s(c') < s(c)} \Sigma_{c'} \not\subset \Sigma_c$. Let

$$\sigma \in \Sigma \setminus \bigcup_{c' \in \mathrm{Bar}_p(\mathcal{F}), s(c') < s(c)} \Sigma_{c'} \setminus \Sigma_c$$

having the maximal rank. Clearly, $\mathrm{rk}_<(\sigma) \leq s(c)$. Now clearly, if $\mathrm{rk}_<(\sigma) = s(c)$, then $\sigma \in \Sigma_c$. Otherwise, if $\sigma \notin \Sigma_c$, there exists an expression for a cycle

$$\sum_\tau m'_\tau \cdot \tau, \tag{4}$$

congruent to the cycle in (4), with $m'_\sigma = 0$, and $m'_\tau = m_\tau$ for all τ satisfying $m_\tau \neq 0$, and $\mathrm{rk}_<(\tau) > \mathrm{rk}_<(\sigma)$. Subtracting the expression (4) from that in (2) we get a cycle,

$$\sigma + \sum_\tau m''_\tau \cdot \tau$$

with $\mathrm{rk}_<(\tau) < \mathrm{rk}_<(\sigma)$ for all τ with $m''_\tau \neq 0$. This implies (using the second part of Remark 7) that $\sigma \in \Sigma_{c''}$, where $c'' \in \mathrm{Bar}_p(\mathcal{F})$ with $s(c'') = \mathrm{rk}_<(\sigma)$. This contradicts the fact that $\sigma \notin \bigcup_{c' \in \mathrm{Bar}_p(\mathcal{F}), s(c') < s(c)} \Sigma_{c'}$. This finishes the proof of the reverse inclusion. ◀

▶ Remark 13. Theorem 12 furnishes us with an algorithm for computing the set Σ_c of essential simplices for each bar $c \in \mathrm{Bar}_p(\mathcal{F}), p \geq 0$, once we have an algorithm for computing a representative cycle for each such bar. Let $c \in \mathrm{Bar}_p(\mathcal{F})$ with $s(c) = s_0$, and we have computed (using an algorithm for computing persistent homology) the set $\Sigma \subset K^{(p)}$ of p-simplices appearing in a cycle representing a homology class corresponding to c. Assuming by induction that we have computed $\Sigma_{c'}$ for bars $c' \in \mathrm{Bar}_p(\mathcal{F})$ with $s(c') < s_0$, we can compute Σ_c using (1).

One filtered simplicial complex that plays an important role in applications of persistent homology theory, including the one in this paper, is the so called Vietoris-Rips complex associated to a weighted graph or equivalently a finite set V equipped with a distance function $w : V \times V \to \mathbb{R}$, satisfying $w(v, v) = 0$ for all $v \in V$.

▶ **Definition 14** (Vietoris-Rips filtration)**.** 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 $v \in V$.

Let $K = 2^V$ denote the simplicial complex corresponding to the simplex with vertices elements of the set V. For any real number $d \geq 0$, we denote by K_d the sub-complex of K, defined by setting for each $0 \leq p \leq \mathrm{card}(V) - 1$, $K^{(p)} = \{[v_0, \ldots, v_p] \mid w(v_i, v_j) \leq d, 0 \leq i, j \leq p\}$.

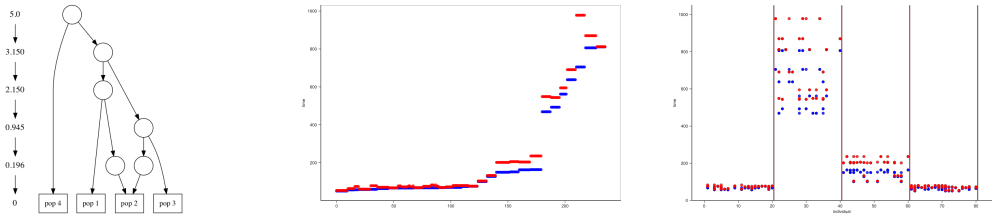
Clearly, if $d \leq d'$, $K_d \subset K_{d'}$, and there exists a a finite set of $0 = d_0 < d_1 < \cdots < d_N$ such that, $K_{d_i} \neq K_{d_{i+1}}$, and $K_{d_i} = K_{d_i + t}$ for $0 \leq t < d_{i+1} - d_i$, for all $i, 0 \leq i < N$.

We call the above filtration the *Vietoris-Rips filtration associated to* M.

▶ Remark 15. Note that for a generic weight function w on $V \times V$, the corresponding Vietoris-Rips filtration is identical to that induced by an admissible ordering (up to breaking ties) in an obvious manner.

3 Experiments

Population simulation. We specify the scaffold to the simulator for a variety of configurations that ranges form zero to three admixed populations. To avoid any unintended bias each population was simulated with the same number of individuals and similar population parameters: mutation rate 4×10^{-8} mut/bp/gen, recombination rate 0.3×10^{-8} cM/Mb/gen and segment length 150Kb and effective population size of 10^4. Each simulated scenario is repeated at least five times. SimRA generates the set of individual haplotypes for each of the populations which is fed to detection algorithm for processing as follows.



(a) Populations scaffold

(b) RFL plot of essential simplices of 1-D cycles

(c) Sorted RFL plot

Figure 3 (a) The scaffold of the evolution scenario of four populations. Only populations 2 is admixed. (b) The corresponding RFL plot of the essential simplices of the one-dimensional cycles. The blue dot marks the birth and the red dot marks the death of the corresponding irreducible cycle. Notice the obvious kink in the plot. (c) Here the known population labels of the individuals are used to put the data in the four buckets.

Persistent Homology. Note that all the individuals of all the populations are put in a single group, i.e., we keep the population label aside and do not use them in any of the computations. Next we create a distance matrix between all pairs of individuals using the Hamming distance metric. The graph embedding of this distance matrix is the complete graph with each vertex corresponding to an individual and edge weight is the distance between the pair of vertices.

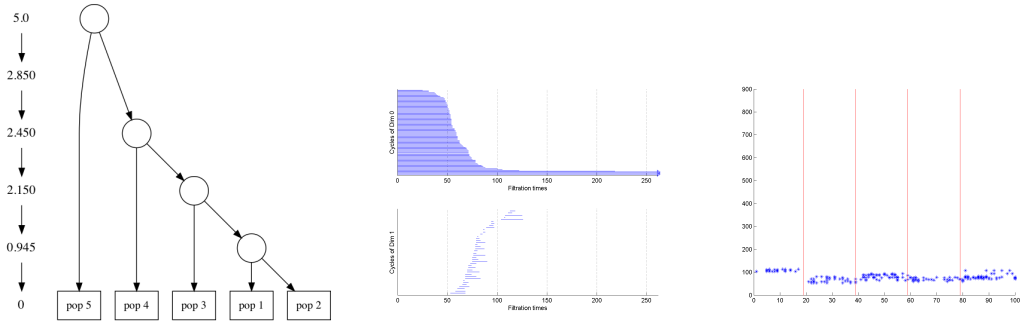
Next the Vietoris-Rips filtration (cf. Definition 14) is constructed on this embedding graph. The zero and one-dimensional persistent homology bar diagrams of the Vietoris-Rips filtration are computed using JavaPlex V 4.3.1 [10]. The set of essential simplices associated to the bars in the bar diagram are then computed using Remark 13.

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. The barcode plots display individual cycles representing non-zero one-dimensional homology classes are born and when they disappear. The top half of each barcode plot for the simulation experiments shows the persistence of zero-dimensional homology and the bottom half shows that of one-dimensional homology. While short cycles can be due to noise, longer (persistent) cycles represent fundamental topological structures within the genetic distance matrix.

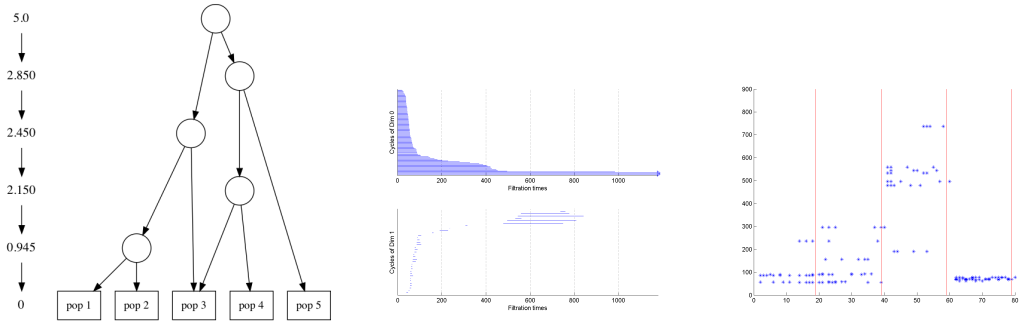
3.1 Visualization of the essential simplices

Note that the bar plot helps in visualizing the cycles representing non-zero homology classes each dimension. Similarly the RFL plot is meant to visualize the set of essential simplices associated to the different bars in the bar diagram (cf. Definition 9). The individual data plots are on the x-axis, in any order natural to the given problem. The y-axis represents the filtration time and shows the birth and death time of each individual as two dots in two different colors. See Fig 3 for an example with four populations. The RFL plot is shown in Fig 3 (b). The individuals here cluster into two groups with the esssential simplices of the late bars corresponding to the admixed population in this example.

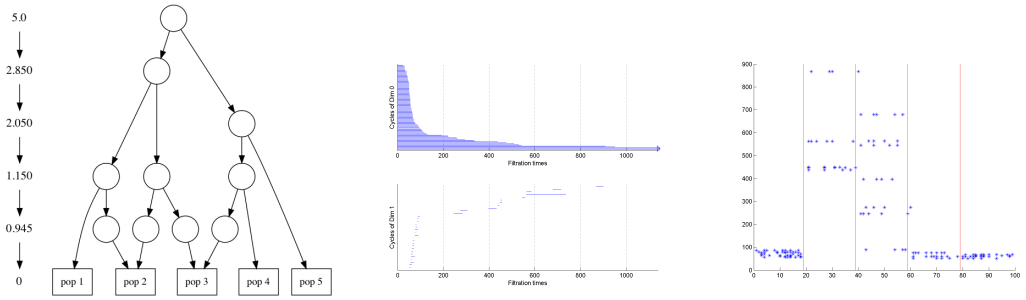
Fig 4 shows four distinct scenarios of admixed populations. The persistent homology of the Vietoris-Rips filtration are shown as bars. Notice that the 1-dimensional bars cluster into different groups whose transitions roughly correspond to the number of admixed populations. This is an empirical observation which we are currently continuing to probe. Further, the



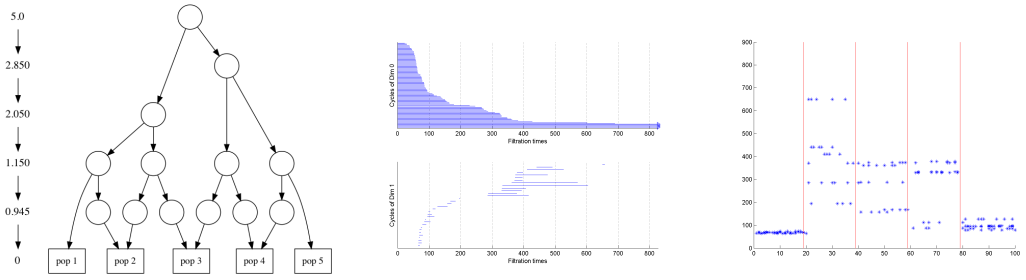
(a) No admixed populations.



(b) 1 population admixed.



(c) 2 populations admixed.



(d) 3 populations admixed.

Figure 4 Each row corresponds to a distinct scenario, whose scaffold is shown on the left. Each scenario has 100 individuals distributed equally amongst the populations. The bar plot of the persistent homology groups of the Vietoris-Rips filtration (cf. Definition 14) of the graph embedding of the distance matrix of the data points (of each scaffold) is shown in the center. The top half shows the persistence of the zero-dimensional while the bottom half shows that of the one-dimensional homologies. The corresponding RFL plot of the essential simplices of the irreducible one-dimensional cycles is shown on the right. For the latter only the birth points are shown. Also, in the RFL plot the individuals of the populations (x-axis) are separated based on their input labels, for convenience.

birth points on the RFL plot cluster at different filtration time points in the reverse order (recall that the time on the scaffold goes from present to past since the reference is 0 at present, for convenience). Thus the number of admixing events and their relative timing can be deduced from the RFL plots in combination with bar diagram. See Fig 4 for the details of the four scenarios.

4 Conclusion

We introduced the notion of essential simplices. This enables us to study the role of individuals in the persistent homology space in an unambiguous manner. Through simulations we show that the clustering of the bars in the bar diagram captures some of the admixing events in the population history and the relative timing of the events. This is reinforced by using essential simplices that further segregate the individuals. We believe that the notion of essential simplices is general enough to be of use in other applications as well.

References

1 Anna P. Carrieri, Filippo Utro, and Laxmi Parida. Sampling arg of multiple populations under complex configurations of subdivision and admixture. *Bioinformatics*, 32(7):1048–1056, 2016.

2 Herbert Edelsbrunner and John L. Harer. *Computational topology: An Introduction*. American Mathematical Society, Providence, RI, 2010.

3 Robin Forman. A user's guide to discrete Morse theory. *Sém. Lothar. Combin.*, 48:Art. B48c, 35, 2002.

4 Matthew L. Freedman, Christopher A. Haiman, Nick Patterson, Gavin J. McDonald, Arti Tandon, Alicja Waliszewska, Kathryn Penney, Robert G. Steen, Kristin Ardlie, Esther M. John, Ingrid Oakley-Girvan, Alice S. Whittemore, Kathleen A. Cooney, Sue A. Ingles, David Altshuler, Brian E. Henderson, and David Reich. 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 Robert Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45:61–75, 2008.

6 Mark Jobling, Edward Hollox, Matthew Hurles, Toomas Kivisild, and Chris Tyler-Smith. *Human evolutionary genetics*. Garland Science, UK, 2013.

7 M.J. Kearsey and H.S. Pooni. *The Genetical Analysis of Quantitative Traits*. Stanley Thornes, UK, 2004.

8 P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Alagappan, J. Carlsson, G. Carlsson, and Mikael Vilhelm Vejdemo Johansson. Extracting insights from the shape of complex data using topology. *Scientific Reports*, 3, 2013.

9 Lammi Parida, Filippo Utro, Deniz Yorukoglu, Anna Paola Carrieri, David Kuhn, and Saugata Basu. Topological signatures for population admixture. In Teresa M. Przytycka, editor, *Research in Computational Molecular Biology*, pages 261–275. Springer International Publishing, 2015.

10 Andrew Tausz, Mikael Vejdemo-Johansson, and Henry Adams. JavaPlex: A research software package for persistent (co)homology. In Han Hong and Chee Yap, editors, *Proceedings of ICMS 2014*, Lecture Notes in Computer Science 8592, pages 129–136, 2014. Software available at `http://appliedtopology.github.io/javaplex/`.

11 J.D. Wall and M.F. Hammer. Archaic admixture in the human genome. *Current opinion in genetics & development*, 16(6):606–610, 2006.