Outbreaks, Metastases and Homomorphisms: Phylogenetic Inference of Migration Histories of Heterogeneous Populations under Evolutionary and Structural Constraints

Kiril Kuzmin

kkuzmin1@gsu.edu

Georgia State University

Henri Schmidt

Princeton University

Sagi Snir

University of Haifa

Benjamin Raphael

Princeton University

Pavel Skums

University of Connecticut https://orcid.org/0000-0003-4007-5624

Article

Keywords: migration tree, phylogenetic inference, viral transmission, cancer metastasis

Posted Date: September 13th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5040045/v1

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Communications on August 28th, 2025. See the published version at https://doi.org/10.1038/s41467-025-63411-4.

NC

Nature Communications 00 (2024) 1–33

Outbreaks, Metastases and Homomorphisms: Phylogenetic Inference of Migration Histories of Heterogeneous Populations under Evolutionary and Structural Constraints

Kiril Kuzmin[a,*], Henri Schmidt[b], Sagi Snir[c], Ben Raphael[b], Pavel Skums[d,*]

[a]*Department of Computer Science, Georgia State University, 25 Park Place, Atlanta, GA 30303, USA*
[b]*Department of Computer Science, Princeton University, 35 Olden St, Princeton, NJ 08540, USA*
[c]*Department of Evolutionary and Environmental Biology, University of Haifa, 199 Aba Khoushy Ave., Mount Carmel, Haifa 3498838, Israel*
[d]*School of Computing, University of Connecticut, 371 Fairfield Way, Storrs, CT 06269, USA*

Abstract

Many human diseases, including viral infections and cancers, are driven by the evolutionary dynamics of heterogeneous populations of genomic variants. A major type of evolutionary behavior of these populations is migration including viral transmissions and cancer metastatic spread. A common strategy for migration pathways reconstruction involves constructing a phylogenetic tree of observed genotypes and inferring its ancestral states corresponding to migration sites. Key challenges here include determining the conditions when a phylogenetic tree topology reflects the underlying migration tree structure, and balancing computational tractability, flexibility, and biological realism of inference algorithms and models.

In this study, we address these challenges using the powerful machinery of graph homomorphisms, a mathematical concept that describes how one graph can be mapped onto another while preserving its structure. We investigate how structural constraints on migration patterns and migration tree topologies influence the relationship between phylogenies and migration trees, characterize trees compatible with a given phylogeny and propose a series of algorithms to assess whether given phylogenetic and migration trees are compatible under various migration scenarios.

Leveraging our findings, we present a framework for inferring migration trees by sampling potential trees from a prior random tree distribution and identifying a subsample compatible with a given phylogeny. By varying prior tree distributions, this approach expands upon several existing models, offering a versatile strategy applicable to a variety of biological processes. We validate our methodology using simulated datasets and real data from studies of viral outbreaks and cancer metastasis, demonstrating its effectiveness across different contexts.

Keywords: migration tree, phylogenetic inference, viral transmission, cancer metastasis

1. Introduction

Many human diseases are essentially evolutionary processes. This includes viral infections, driven by evolving populations of viral variants [1], as well as cancers associated with diversifying intra-tumor subclonal lineages [2]. Although the biological mechanisms of these diseases are different, both are fueled by highly mutable populations of disease-causing agents, whose extreme genomic diversity originates from error-prone replication processes, whether due to the lack of a proofreading mechanism in RNA-dependent RNA polymerase or retroviral reverse transcriptase in RNA viruses [3, 4, 5], or from the genetic instability of tumor cells manifesting itself in somatic mutations, chromosomal gain/loss/translocation, and aneuploidy [6, 7, 8, 9]. Consequently, the general phylogenetic methodologies applied to these populations exhibit many similarities. From a methodological standpoint, they form a unique segment of phylogenetics and phylodynamics, fostering a mutual exchange of

*Corresponding authors.
Email addresses: kkuzmin1@gsu.edu (Kiril Kuzmin), pavel.skums@uconn.edu (Pavel Skums)

concepts that enhances all areas of application [10, 11, 12].

A major type of evolutionary behaviour of highly mutable populations is *migration*, wherein their members spread from initial sites, seeding new populations at newly invaded sites. For infectious diseases, such migrations equate to pathogen transmissions, whereas in cancer this process is identified with metastatic spread. Thus, accurate inference of migration networks of heterogeneous populations is crucial for public health and medical research [13, 14, 15, 16].

The study of highly mutable population migration has been significantly enhanced by groundbreaking advancements in sequencing technologies. State-of-the-art high-throughput targeted sequencing and single-cell DNA sequencing enable the capture of detailed population snapshots at exceptionally high resolutions, facilitating fine-grained analysis down to the level of individual genotypes [16, 17, 18, 19]. In particular, this allows to examine population migration on the level of individual migration events [20, 17, 21, 22].

A wide array of methods has been developed specifically for reconstructing viral and bacterial transmission trees, reflecting the substantial interest in tracking the spread of infectious diseases. The arsenal of tools available for reconstructing transmission networks is extensive, including but not limited to Outbreaker, Outbreaker 2 [23, 24], SeqTrack [25], SCOTTI [26], SOPHIE [27], Phybreak [28], Bitrugs [29], BadTrIP [30], Phyloscanner [31], StrainHub [32], TransPhylo [33, 34] (along with its extension TransPhyloMulti [35]), STraTUS [36], TreeFix-TP [37], QUENTIN [38], VOICE [39], HIVTrace [40], GHOST [41], MicrobeTrace [42], SharpTNI [43], TiTUS [12], TNeT [44], AutoNet [45], and others [46, 47, 34, 48, 49, 50, 51, 52, 53]. These tools have been instrumental in investigation of outbreaks and monitoring the transmission dynamics of pathogens like HIV, hepatitis C (HCV), SARS, MERS and SARS-CoV-2 [54, 55, 56, 57, 20, 58].

Similarly, the development of methods for deducing metastatic spread histories is burgeoning, driven by advancements in single-cell DNA sequencing and CRISPR-based lineage tracing technologies. Currently, the repertoire of tools in this domain includes MACHINA [22], FitchCount (as part of the Cassiopeia suite) [21, 59], PathFinder [60], TCC, PCC, and PCCH [61]. Notwithstanding the relatively shorter list, the impact of these tools is growing, with several studies published recently leveraging these methods to gain insights into the mechanisms of metastatic spread [21, 62, 63, 64].

Despite significant progress in the field, the wide variety of existing methods underscores that the challenge of accurately inferring heterogeneous population migration remains unresolved. This diversity of approaches indicates both the complexity of the problem and the ongoing efforts to refine and improve upon existing methods. Additionally, a major barrier to advancement in the field is the relative isolation of viral and cancer genomics fields. Some of the aforementioned tools are based on conceptually similar techniques - this applies, for example, to STraTUS (which focus on viral transmission) and FitchCount (which addresses metastatic spread) that, as demonstrated in this paper, yield virtually identical results when applied to the same datasets. This lack of interdisciplinary exchange of ideas often leads researchers to inadvertently duplicate efforts, thereby impeding progress in both fields.

Phylogenetics and phylodynamics provide the most widely used methodological frameworks for migration tree/network reconstruction [35, 22]. However, their application in this context is not straightforward. A phylogenetic tree does not directly equate to a migration tree [35], as the nodes in a phylogenetic tree represent divergences of lineages rather than specific migration events [35, 65]. While some of these divergences may result from migration, others occur within previously invaded sites. Therefore, deriving a migration tree from a phyloge-

netic tree essentially requires solving an ancestral trait inference problem, wherein the internal nodes of a phylogenetic tree are annotated with labels that indicate whether each divergence event occurred within a site or as a result of seeding of a new site upon migration. Furthermore, it is crucial to effectively leverage the full spectrum of intra-site (within-host or within-tumor) population diversity uncovered by high-throughput sequencing, that often provide a strong signal for migration inference. For example, the paraphyletic relationships between populations suggest recent migrations between corresponding sites [66, 67, 39, 68].

The challenges mentioned above highlight several crucial questions that remain unresolved and warrant further exploration. These questions include:

1)*To what extent and under which conditions does the topology of a phylogenetic tree reflect the structure of the underlying migration tree?* This question becomes particularly important when the level of paraphyly in the labeled phylogeny is low, a situation not uncommon as the paraphyletic signal tends to diminish over time and with smaller sample sizes [66]. A number of studies have looked at this subject but their conclusions were mixed. Some studies have found that certain migration patterns, such as super-spreading, migration chains, or, more generally, migrations within networks formed by different models like Erdös-Rényi [69], preferential attachment [70], or Watts-Strogatz models [71], lead to quantitatively distinct phylogenetic tree topologies [72, 73, 74]. Additionally, the spatial structure and dynamics of heterogeneous populations, which are directly related to migrations pathways, have been shown to affect the phylogeny structure of both viral and tumor populations [75, 76, 77, 78]. On the other hand, other studies report that the direct impact of migration patterns on phylogenetic trees ranges from minimal to moderate [79, 80, 81, 82]. Finally, certain studies have drawn mixed conclusions, indicating that while some migration characteristics are reflected in the phylogeny, others are not [83].

2) *How can we limit the solution space to balance computational feasibility, accuracy of inference, generalizability, and biological realism?* The space of migration trees compatible with given phylogenetic trees is often vast, and its properties are not well understood [36, 12]. A sampling-consensus approach is one method to address solution ambiguity, where feasible migration trees are sampled and summarized in a weighted consensus graph, with weights reflecting posterior probabilities of edges [44, 12, 43, 26, 21]. However, the size of solution space may restrict the depth of sampling. As a response, it is common practice to narrow down the solution space to a set of plausible migration trees optimizing a specific objective function under evolutionary-based constraints. Employing constrained models also aids in preventing overfitting in presence of missing data and errors.

Various objectives and constraints have been implemented by existing methods. Limiting number of migration events, sizes of bottlenecks or numbers of back-migrations [31, 12, 44, 22, 59, 21, 61, 36, 25] is more computationally efficient and scalable due to utilization of dynamic programming [84, 85], making such approaches practical in both molecular epidemiology and computational oncology. These can also be formulated as Integer Linear Programming (ILP) problems [86] and solved with reasonable efficiency using existing ILP solvers. Models with more complex Bayesian objectives with constraints regularized as priors [35, 28, 26, 23, 30] offer a richer, biologically nuanced perspective but suffer from scalability issues and usually rely on generic methods like Markov Chain Monte Carlo (MCMC) sampling, which may not yield optimal solutions, in part due to a lack of problem-specific mathematical strategies. Balancing computational efficiency with biological comprehensiveness presents a notable challenge, compounded by the un-

certainty of how much constraints and objectives truly limit the solution space [36, 12].

3) *How to incorporate a variety of models for phylogenetic inference of migrations into a unified modular computational framework?*

Migration inference models draw on varied biological or epidemiological assumptions. For instance, viral transmission inference often incorporates case-specific temporal data like infectious periods, exposure intervals, symptom onset, diagnosis or sample collection dates to establish order of infections and eliminate unlikely transmission links [23, 47, 28, 33, 12, 26, 29]. In some rare cases, contact networks are known and can be used for the same purpose [87, 58]. While effective, such data is often unavailable, non-informative, or sensitive, particularly for endemic and pandemic diseases caused by HIV, Hepatitis C, SARS-CoV-2, or Influenza [27, 38, 53]. In situations where case-specific data cannot be used, genomic epidemiology tools resort to broader assumptions, like the expected degree distribution of transmission networks implied by a structure of a susceptible population [38, 27]. Similarly, methods for inferring metastatic spread use constraints defined by so-called migration patterns that reflect realistic cancer migration scenarios, such as monoclonal, polyclonal or multi-source seeding [22, 61]. A commonality across all these methods is the use of structural constraints on feasible transmission networks that are considered as subgraphs of a larger "pattern" graph. It suggests a need for a versatile, modular migration inference framework that integrates these varied approaches on a unified algorithmic and mathematical basis, akin e.g. to the BEAST framework for Bayesian evolutionary analysis [88].

Addressing these challenges requires a comprehensive investigation into the mathematical properties of migration trees compatible with a given phylogeny, a topic that is not yet fully understood [36]. Our study aims to advance this area by developing a novel methodology based on powerful techniques from graph theory and combinatorics. Additionally, we will use this methodology to introduce novel algorithmic approaches for inferring migration trees.

A number of earlier studies has achieved important progress in this area. Several studies noticed that migration trees compatible with a given phylogeny correspond to partitions of the phylogeny's node set or to coloring of its branches [51, 36, 33, 34]. These observations have informed the development of methods to enumerate and sample these trees, assuming a complete migration bottleneck [36] and a known sequence of migrations [89]. In fact, as we argue in this paper, the relation between a phylogenetic tree and a migration tree is described by the concept of a *graph homomorphism* that generalize both partitions and colorings.

Graph homomorphism is essentially a mapping between the vertices of two graphs that preserve their structure [90]. The theory of graph homomorphisms is well-established area of discrete mathematics, with deep results and rich methodology. All types of migration trees discussed so far can be described by a graph homomorphism with specific constraints. For example, migration trees compatible with a given phylogeny under the assumptions of complete sampling and complete bottleneck, that has been studied in [51, 36] are *minors* [91, 92] of the phylogeny or, equivalently, its homomorphic images such that inverse images of all vertices are connected subtrees.

We use mathematical and algorithmic machinery of graph homomorphism theory to delve into the details of migration inference. Specifically, we provide necessary and sufficient conditions describing trees compatible with a given phylogeny and propose a series of algorithms that evaluate the compatibility of phylogenetic and migration trees under various evolutionary scenarios through the construction of corresponding homomorphisms. We examine particular structural constraints on migra-

tion patterns and migration tree topologies to understand their influence on the relationship between a phylogeny and a migration tree.

Based on aforementioned insights, we propose a general framework for migration inference that samples candidate migration trees from a chosen prior random tree distribution, and identifies a subsample of trees compatible with a given phylogeny. By varying prior tree distributions, this approach expands upon and generalizes several existing models, offering a versatile and computationally efficient strategy applicable to a variety of biological processes associated with heterogeneous population migration. Crucially, the proposed framework is computationally fast, enabling biomedical and public health researchers to quickly test different tree priors that represent common migration models. Beyond migration reconstruction, proposed methods can be used for investigating how phylogenies constrain the space of possible migration trees, for inferring an order of known or suspected migration events, and for determining potential migration events that are definitively ruled out by a phylogeny [36].

Proposed methodology was validated using both simulated datasets and real experimental data gathered from studies of viral outbreaks and cancer metastasis, demonstrating its effectiveness and applicability across different contexts.

2. Methods

2.1. Basic definitions

Throughout this paper, we consider a pair of trees: a phylogenetic tree and a migration tree. For clarity, we refer to elements of a phylogeny as *nodes*, and to elements of a migration network as *vertices*, and denote them by Greek and Latin letters, respectively.

The problem of *migration network inference* is set up as follows. The input is a phylogenetic tree $\Psi = (V(\Psi), E(\Psi))$, with the leaf set $L(\Psi)$ representing genomic variants belonging to different subpopulations (or *demes*), denoted by $\mathcal{L}$. The tree Ψ can be a standard binary phylogeny or non-binary mutation tree used in most cancer studies. Each leaf $\lambda \in L(\Psi)$ has an assigned site label (or *color*) $l_\lambda \in \mathcal{L}$. The aim is to expand this labeling from the leaves to all nodes in the tree, creating a full labeling $f : V(\Psi) \to \mathcal{L}$. In this model, any multi-colored tree edge $\alpha\beta$ represents a migration of genomic variants between demes $f(\alpha)$ and $f(\beta)$. The migration tree $T = T(\Psi, \boldsymbol{l})$ and with vertices $V(T) = \mathcal{L}$, is then formed by contracting the nodes with the same color [51].

As mentioned in the introduction, researchers often seek migration trees satisfying particular constraints restricting types of migration or tree topologies. These constraints can be encoded using a *transition pattern graph* G that describes permissible patterns of migration (specific examples are provided in the following subsection). We will first consider the situation when G is a simple graph; later on, it will be extended to the cases when G is a random graph characterized by some probability distribution. In this model, a migration tree should be isomorphic (i.e. identical up to relabeling of vertices) to a subgraph of the transition pattern. Any corresponding labeling will be called *feasible*.

The relations between the phylogeny Ψ, the migration tree T and the transition pattern G can be captured using the concept of a *graph homomorphism*. A homomorphism $f : \Psi \to G$ [90] is an adjacency-preserving mapping between vertex sets of these graphs, i.e. $f(u)f(v) \in E(G)$ if $uv \in E(\Psi)$. For the sake of mathematical rigor, here and throughout this paper we assume that a transition pattern is *reflexive*, i.e. every vertex is adjacent to itself. With this condition in place, any feasible labeling f is a homomorphism from Ψ to G, making the migration inference problem essentially a problem of finding such a homomorphism. In graph theory, this type of problems is sometimes

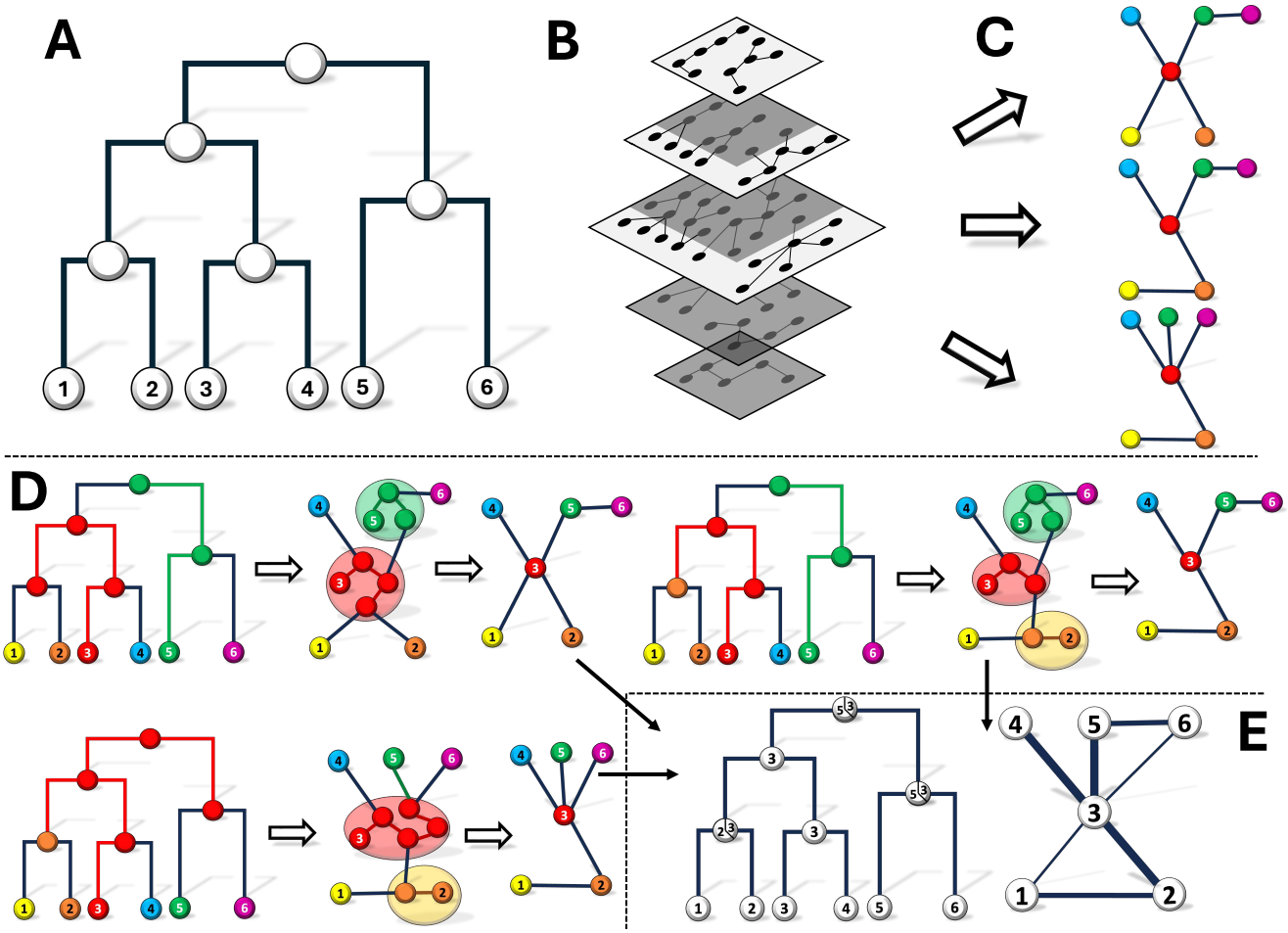

Figure 1. SMiTH: Sampling MIgration Trees with Homomorphisms. **A**: Input phylogenetic tree. **B**: Distribution of possible migration trees. Parallel rectangles depict a probability density function, with each rectangle's width proportional to the corresponding probability. In practice, the distribution is represented either by a random graph model or by a stochastic graph generation procedure. **C**: Candidate migration trees sampled from the distribution B. **D**: Homomorphisms from the phylogeny to three sampled trees. In the phylogenies, nodes are color-coded by their homomorphic images in a migration tree. The phylogeny layouts in the middle of each subfigure showcases how homomorphism transforms them into sampled trees. **E**: consensus solution derived from homomorphisms in D. The solution is shown as potential color distributions for the phylogeny's nodes (left) or as a graph where possible migration edges are weighted according to the number of supporting solutions (right), with the edge thickness indicating weight.

referred to as an *G-coloring* of the tree Ψ [90].

Throughout this paper, we use standard graph theory notations. We denote by Ψ_α a subtree of Ψ rooted at the node α. For any graph G, $G[X]$ represents the subgraph induced by the subset of vertices X. We use $u \sim v$ to indicate that vertices u and v are adjacent. The set of neighbors of a vertex u in G is denoted as $N_G(u)$, $\deg_G(u) = |N_G(u)|$ is the degree of u, $N_G[u] = N_G(u) \cup \{u\}$ and $N_G[X]$ is the union of the neighborhoods of all vertices in a set X. Additionally, the distance between any two vertices u and v in G is denoted by $d_G(u, v)$, and $P_G(u, v)$ refers to the corresponding shortest path between them. When the graph is clear from a context, we may omit the subscripts in these notations. In the case of a phylogeny Ψ, all distances and paths are *undirected*.

Additionally, to simplify the notation, we will apply set-theoretical operations (e.g. intersection and union) directly to subgraph of G, with the understanding that the resulting subgraph is induced by the set obtained from applying these operations to the vertex sets of the original subgraphs.

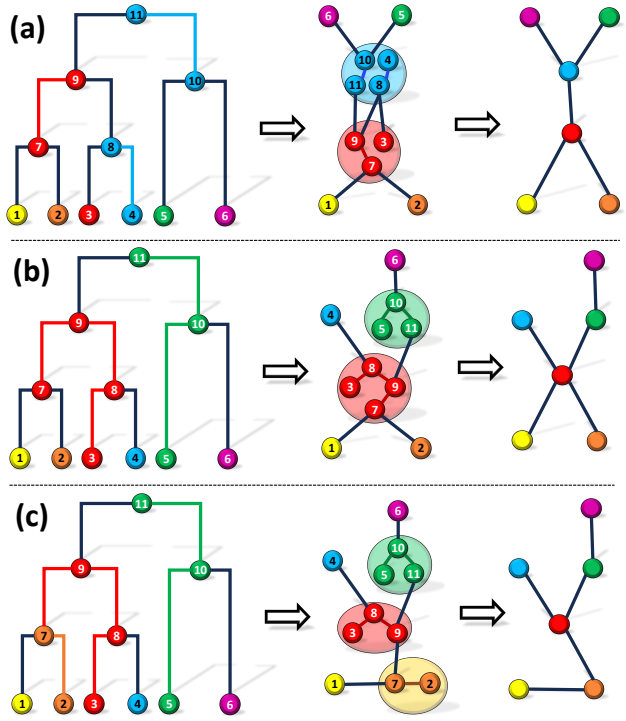

Figure 2. **Solutions for the Migration History Inference Problem under various constraints**. For each scenario, a phylogenetic tree is illustrated in two different layouts on the left and in the middle, with the corresponding migration tree displayed on the right. In the phylogenies, nodes are color-coded by their homomorphic images in the migration tree. The layout in the middle showcases how homomorphism transforms a phylogeny into a migration tree. (a) Unconstrained solution. Subtrees formed by blue and red nodes are not connected, indicating a violation of the convexity constraint. (b) Convex solution. Each color-coded subtree is connected, but compactness is violated in the subtree rooted at node 7. (c) Convex and compact solution.

2.2. *Migration inference under structural constraints*

The simplest form of the migration inference challenge appears when the vertices of the transition pattern graph G directly correspond to the labels of the phylogenetic tree leaves, that is, $V(G) = \mathcal{L}$. This variant is referred to as *labeled inference*. The transition pattern G can be an unweighted graph or a random graph with specified edge probabilities. Several scenarios exemplify this problem:

- For viral transmissions, G could reflect a contact network of potential hosts [87, 58], where an infection spreads through direct interactions within this network.
- Another viral transmission model assumes that transmission is only possible between hosts with overlapping exposure intervals [30, 12]. In this case G is an *interval graph* [93], i.e. a graph with vertices representing time intervals and edges connecting vertices whose intervals intersect.
- For metastatic spread inference, G may represent a *circulatory network* [94, 95, 96], with vertices representing organs and edges reflecting the prior probabilities of cancer spreading between organs.

Problem 1 (labeled migration inference)

Input:

(a1) a phylogenetic tree Ψ with leaf labels $(l_\lambda)_{\lambda \in L(\Psi)}$ forming the label set $\mathcal{L}$;

(b1) a transition pattern G with $V(G) = \mathcal{L}$;

Output:

(c1) a homomorphism $f : \Psi \to G$ such that $f(\lambda) = l_\lambda$ for every $\lambda \in L(\Psi)$.

In practical settings, however, Problem 1 might be too restrictive or not fully reflective of reality. The main issue is the absence of a known mapping between the leaf labels and the vertices of the transition pattern. In other words, the pattern G represents the permissible topology of migration rather than specific allowed migration links. For instance, we may expect that a viral transmission network likely includes a superspreader, but the exact subpopulation associated with this superspreader is not known. The following models to this variant of the problem:

- Viral transmission: The transition pattern G could represent a random graph that describes expected characteristics of transmission trees, like being scale-free or having a particular expected degree distribution. Such models draw on known expected properties of contact networks

essential for infection spread [97, 98, 99, 100] without requiring actual host contact information, and have been explored in several studies [38, 27].

- Metastatic spread inference: In this case, a transition pattern G may describe plausible evolutionary scenarios of cancer migration, such as monoclonal or polyclonal seeding from single or multiple sources [22, 61]. A related idea has been applied in studies analyzing CRISPR-based lineage tracing phylogenies, where G specifies a so-called star homoplasy model [101].

In this version of migration inference problem, is no explicitly given mapping between the set of leaf labels and the vertices of the transition pattern. Instead, a feasible homomorphism should map leaves with distinct labels to distinct vertices of G and vice versa, i.e. for any $\lambda_1, \lambda_2 \in L(\Psi)$ we have $f(\lambda_1) \neq f(\lambda_2)$ whenever $l_{\lambda_1} \neq l_{\lambda_2}$ (Fig. 2 (a)). We will call a homomorphisms satisfying this requirement a *label-distinctive homomorphism*. It should be noted that finding such a homomorphism seems to be a non-standard variant of a graph homomorphism problem, that, to the best of our knowledge, has not been studied previously.

In such context, the first question to be asked is whether a given subtree T of the transition pattern G is compatible with the phylogeny Ψ, i.e. whether there exist a label-distinctive homomorphism $\Psi \to T$.

Problem 2 (unlabeled migration inference)

Input:

(a2) a phylogenetic tree Ψ with leaf labels $(l_\lambda)_{\lambda \in L(\Psi)}$;

(b2) a tree T;

Output:

(c2) a label-distinctive homomorphism $f : \Psi \to T$.

A more restricted version of Problem 3 includes an additional *convexity constraint* [102, 103], where tree nodes mapping to the same vertex of G form a connected subtree of Ψ (Fig. 2 (b)):

Problem 3 (convex unlabeled migration inference)

Input: (a2) and (b2)

Output:

(c3) a label-distinctive homomorphism $f : \Psi \to T$ such that induced subgraphs $\Psi[f^{-1}(v)]$ are connected for all $v \in V(T)$.

We will refer to a homomorphism satisfying (c3) as a *convex homomorphism*. Convex homomorphisms to trees describe migrations with a complete bottleneck; such migration trees were the focus of extensive research in previous studies [51, 36, 89].

To define another type of constraints, consider a labeling l of nodes of the phylogeny Ψ. The labeling l is *compact* if the label of the most recent common ancestor (MRCA) for any group of leaves matches one of the labels within that group.

The rationale for this is based on the understanding that migrations within any given subtree could follow one of the following scenarios: either (i) migrations involve only the demes represented by the leaves of the subtree, or (ii) migrations include external demes, but lineages from these demes were not sampled due to extinction or incomplete collection. Although both scenarios are feasible, the first is more parsimonious, especially when sampling is dense and migration events span a short period of time. Hence, homomorphisms that align with this compact labeling are referred to as compact.

Problem 4 (compact unlabeled migration inference)

Input: (a2) and (b2)

Output:

(c4) a label-distinctive homomorphism $f : \Psi \to T$ such that for any node $\alpha \in V(\Psi)$ we have $f(\alpha) \in f(L(\Psi_\alpha))$.

Finally, it is often desirable to produce a sample of potential solutions rather than a single solution. Such approach offers statistical backing for potential migration network edges and logically shifts the migration inference problem to a Bayesian paradigm. Potential solutions can be sampled from a random graph distribution or, if a transition pattern is a deterministic graph, from the uniform distribution of its subgraphs with specified properties (e.g. subtrees). The sampling-based version of the migration inference problem can be formulated as follows:

Problem 5 (unlabeled migration sampling)

Input:

(a5) a phylogenetic tree Ψ with leaf labels $(l_\lambda)_{\lambda \in L(\Psi)}$;

(b5) a random transition pattern G with edge probabilities $p : E(G) \to [0,1]$ that define a distribution $\mathcal{D}$ of subtrees of G.

Output:

(c5) a sample $S = \{T_1, \ldots, T_m\}$ from the distribution $\mathcal{D}$, where each subtree T_i is a homeomorhic image of Ψ, and the corresponding label-distinctive homomorphisms $f_i : \Psi \to T_i$ are possibly convex and/or compact.

2.3. Labeled migration inference

Problem 1 can be efficiently solved in polynomial time using dynamic programming, as detailed in Algorithm 1. Despite its simplicity, we include the algorithm here due to its relevance for subsequent, more complex methods.

2.4. Unconstrained unlabeled migration inference

A possible strategy to solve Problem 2 involves identifying a bijection $g : \mathcal{L} \to V(T)$ that can be extended to a homomorphism $\Psi \to T$ using Algorithm 1. We will describe such bijections as *feasible*.

Algorithm 1 Unlabeled migration inference

```
Let ρ be the root of Ψ. Perform a post-order traversal of the tree Ψ.
for every node α of the traversal do
    construct a set of potential images I(α) ⊆ ℒ:
    if α is a leaf then
        I(α) ← {l(α)};
    else
        Suppose that β_1,...,β_k are children of α. Then I(α) consists of vertices x ∈ V(G) such that there exist vertices y_i ∈ I(β_i), i = 1,k such that y_i ~ x.
    end if
end for
if I(ρ) = ∅ then
    homomorphism f does not exist
else
    perform a pre-order traversal of Ψ and construct a homomorphism f as follows
    for every node α of the pre-order do
        if α = ρ then
            select any v ∈ I(ρ) and set f(ρ) ← v;
        else
            Let ω be the parent of α. Choose v ∈ I(α) such that v ~ f(ω) and set f(α) ← v.
        end if
    end for
end if
```

Let $L_g(u) = \{\lambda \in L(\Psi) : g(l_\lambda) = u\}$ be the set of leaves in Ψ whose labels map to u. The following theorem establishes a necessary and sufficient condition for the bijection feasibility.

Theorem 1. *The bijection $g : \mathcal{L} \to V(T)$ is feasible if and only if*

$$d_T(u_1, u_2) \le d_\Psi(\lambda_1, \lambda_2) \tag{1}$$

for all $u_1, u_2 \in V(T)$, $\lambda_1 \in L_g(u_1), \lambda_2 \in L_g(u_2)$.

Proof. The necessity of the condition stated in the theorem is a known property of graph homomorphisms [90]. However, it is generally not sufficient [90]. For our specific type of the homomorphism problem, we will demonstrate that it indeed suffices.

Consider a bijection g satisfying the condition (1). Define $C_\alpha \subseteq V(T)$ as the image set of its clade, that is, $C_\alpha = g(l_\lambda) : \lambda \in L(\Psi_\alpha)$. Additionally, for a node $\alpha \in V(\Psi)$ and a vertex $u \in V(T)$, define $B_\alpha(u)$ as a ball centered at u in T, with

radius

$$r_{\alpha,u} = \min_{\lambda \in L_g(u) \cap L(\Psi_\alpha)} d_\Psi(\alpha, \lambda) \quad (2)$$

i.e., $B_\alpha(u) = \{v \in V(T) : d_T(u, v) \leq r_{\alpha,u}\}$.

We proceed by establishing two auxiliary facts. The first follows directly from the properties of a tree:

Lemma 1. *Let $X_1, \ldots, X_k$ be subsets of vertices of the tree T such that each subsets induces a connected subgraph and $\bigcap_{i=1}^k X_i \neq \emptyset$. Then $N[\bigcap_{i=1}^k X_i)] = \bigcap_{i=1}^k N[X_i]$.*

Suppose now that $I(\alpha) : \alpha \in V(\Psi)$ are the sets of potential node images produced by Algorithm 1, given the matching of leaf labels in Ψ and vertices of T through the bijection g. These sets are described by the following lemma.

Lemma 2. *For every node $\alpha \in V(\Psi)$, $I(\alpha) = \bigcap_{u \in C_\alpha} B_\alpha(u)$.*

Proof. The proof of the lemma proceeds by induction. Consider a node $\alpha \in V(\Psi)$. The lemma's assertion is trivially true when α is a leaf. Assume now that α is an internal node with children $\beta_1, \ldots, \beta_k$, each set $I(\beta_i)$ is non-empty and, by the inductive assumption,

$$I(\beta_i) = \bigcap_{u \in C_{\beta_i}} B_{\beta_i}(u). \quad (3)$$

Then $C_\alpha = \bigcup_{i=1}^k C_{\beta_i}$; furthermore, according to Algorithm 1 and the equality (3) we have

$$I(\alpha) = \bigcap_{i=1}^k N[I(\beta_i)] = \bigcap_{i=1}^k N\left[\bigcap_{u \in C_{\beta_i}} B_{\beta_i}(u)\right]. \quad (4)$$

Now consider a vertex $u \in C_\alpha$. Let $\beta(u)$ be the child of α such that $r_{\beta(u),u} = \min_{i=1}^k r_{\beta_i,u}$; in cases where there are multiple such children, we pick any of them. Consequently, we have $r_{\alpha,u} = r_{\beta(u),u} + 1$, leading to the relation $B_\alpha(u) = N[B_{\beta(u)}(u)]$. Furthermore, it is obvious that $B_{\beta(u)}(u) \subseteq B_{\beta_i}(u)$ for all nodes β_i such that $u \in B_{\beta_i}(u)$. Together, these observations imply the following sequence of equalities:

$$\bigcap_{u \in C_\alpha} B_\alpha(u) = \bigcap_{u \in C_\alpha} N[B_{\beta(u)}(u)] = \bigcap_{i=1}^k \bigcap_{u:\beta(u)=\beta_i} N[B_{\beta_i}(u)] = \bigcap_{i=1}^k \bigcap_{u \in C_{\beta_i}} N[B_{\beta_i}(u)] \quad (5)$$

By Lemma 1, $N\left[\bigcap_{u \in C_{\beta_i}} B_{\beta_i}(u)\right] = \bigcap_{u \in C_{\beta_i}} N[B_{\beta_i}(u)]$, and thus the expressions (4) and (5) are equal. This completes the proof of Lemma 2. □

According to Lemma 2, Algorithm 1 succeeds whenever $\bigcap_{u \in C_\alpha} B_\alpha(u) \neq \emptyset$ for every node $\alpha \in V(\Psi)$. To establish that this condition holds, we invoke so-called *Helly property* of subtrees. A family of sets $S_1, \ldots, S_k$ has a Helly property [104] if $\bigcap_{i=1}^k S_i \neq \emptyset$ whenever $S_i \cap S_j \neq \emptyset$ for every $i, j \in [k]$; in other words, the existence of non-empty pairwise intersections guarantees a non-empty total intersection.

Subtrees of a given tree are known to have the Helly property [105]. The subsets $B_\alpha(u)$ obviously induce subtrees of T. Therefore, to prove the theorem, it is sufficient to demonstrate that for every node $\alpha \in V(\Psi)$ and for every pair of vertices $u_1, u_2 \in C_\alpha$, the intersection $B_\alpha(u_1) \cap B_\alpha(u_2)$ is non-empty.

Select two leafs $\lambda_1, \lambda_2 \in L(\Psi_\alpha)$ that minimize the distance between nodes of the sets $L_g(u_1) \cap L(\Psi_\alpha)$ and $L_g(u_2) \cap L(\Psi_\alpha)$. According to the theorem's conditions, we have:

$$d_T(u_1, u_2) \leq d_\Psi(\lambda_1, \lambda_2) \leq d_\Psi(\alpha, \lambda_1) + d_\Psi(\alpha, \lambda_2). \quad (6)$$

This implies the existence of a path between u_1 and u_2 in T with a length at most $d_\Psi(\alpha, \lambda_1) + d_\Psi(\alpha, \lambda_2)$. On this path, there is at least one vertex v such that $d_T(u_1, v) \leq d_\Psi(\alpha, \lambda_1)$ and $d_T(u_2, v) \leq d_\Psi(\alpha, \lambda_2)$. Consequently, v belongs to both $B_\alpha(u_1)$ and $B_\alpha(u_2)$, thereby completing the proof. □

Theorem 1 establishes that the Unlabeled Migration Inference problem is algorithmically equivalent to the problem of

finding a bijection that satisfies (1). First of all, this allows us to demonstrate that Problem 2 is $\mathcal{NP}$-hard. We will prove it through a reduction from the following problem:

Graph Bandwidth problem.

Given: A graph G.

Find: The minimal integer $K = bw(G)$ for which there exists is a bijection $f : V(G) \to \{1, \ldots, |V(G)|\}$ such that

$$|f(u) - f(v)| \leq K \text{ for all } u \sim v. \tag{7}$$

The Graph Bandwidth problem is $\mathcal{NP}$-hard [106], and, moreover, it cannot be approximated within any constant factor unless $\mathcal{P} = \mathcal{NP}$, even when the input graph G is a tree [107].

Theorem 2. *The Unlabeled Migration Inference problem is $\mathcal{NP}$-hard, even when all leaf labels in the phylogeny Ψ are unique.*

Proof. For our purposes, it is more convenient to use an equivalent condition for Graph Bandwidth problem:

$$|f(u) - f(v)| \leq K d_G(u, v) \text{ for all } u, v \in V(G). \tag{8}$$

The fact that (8) implies (7) is obvious. To demonstrate that (7) implies (8), consider the shortest (u, v)-path in G ($u = x_0, x_1, \ldots, x_{d-1}, x_d = v$), where $d = d_G(u, v)$. Then we have

$$|f(u) - f(v)| = |\sum_{i=1}^{d} (f(x_{i-1}) - f(x_i))| \leq \leq \sum_{i=1}^{d} |(f(x_{i-1}) - f(x_i))| \leq Kd.$$

Now suppose that T' is an input tree of Graph Bandwidth problem. We can assume that $bw(T') \geq 3$, since graphs where $bw(T') \leq 2$ are recognizable in linear time [108]. To construct an instance for Problem 2, for an integer $K \geq 3$, we proceed as follows:

1) The input phylogeny Ψ_K is constructed by (a) subdividing every edge of T' into K edges; (b) attaching a leaf labeled u' to every node $u \in V(T')$.

2) The input tree T is an n-vertex path P_n with $V(T) = \{1, \ldots, n\}$

For the trees constructed in this manner we have:

(a) $L(\Psi_K) = \{u' : u \in V(T')\}$;

(b) $d_{\Psi_K}(u', v') = K d_S(u, v) + 2$;

(c) $d_T(i, j) = |i - j|$.

We will demonstrate that a polynomial-time algorithm for Problem 2 leads to a $\frac{5}{3}$-approximation algorithm for the Graph Bandwidth problem. Assume the existence of such an algorithm for Problem 2. Let K^* be the smallest integer for which this algorithm produces a sought-for homomorphism $\Psi_{K^*} \to T$. To establish the $\frac{5}{3}$ approximation factor, we need to demonstrate the following relationship:

$$K^* \leq bw(T') \leq \frac{5}{3} K^* \tag{9}$$

To establish an upper bound, let us consider the feasible bijection $g^* : L(\Psi_{K^*}) \to \{1, \ldots, n\}$. We extend this bijection to the nodes of $V(T')$ by setting $g^*(u) = g^*(u')$. Given the inequality (1) and assuming that $K^* \geq 3$, we have:

$$|g^*(u) - g^*(v)| = d_T(g^*(u'), g^*(v')) \leq d_{\Psi_{K^*}}(u', v') = = K^* \cdot d_{T'}(u, v) + 2 \leq \frac{5}{3} K^* d_{T'}(u, v).$$

This implies that $bw(S) \leq \frac{5}{3} K^*$.

Conversely, if $bw(T') = K < K^*$, and the mapping $g : V(T') \to \{1, \ldots, n\}$ is the bijection reflecting this bandwidth, then we can extend it to $L(\Psi_K)$ by setting $g(u') = g(u)$. This yields:

$$d_T(g(u'), g(v')) = |g(u) - g(v)| \le K d_{T'}(u, v) < \\ < K d_{T'}(u, v) + 2 = d_{\Psi_K}(u', v').$$

This inequality contradicts the assumption that K^* is minimal. Therefore, we must have $bw(S) \ge K^*$. This completes the proof. □

Although the Unlabeled Migration Inference problem is $\mathcal{NP}$-hard, we can use Theorem 1 to approach it using Integer Linear Programming (ILP). We define a feasible bijection $g : \mathcal{L} \to V(T)$ using binary variables x_{iu}, where $x_{i,u} = 1$ if $g(i) = u$, for each $i \in \mathcal{L}$ and $u \in V(T)$. The ILP formulation to find such a bijection is as follows:

$$\sum_{i \in \mathcal{L}} \sum_{u \in V(T)} \delta(i) \deg(u) x_{iu} \to \max \quad (10)$$

s.t.

$$\sum_{u \in V(T)} x_{iu} = 1, \quad i \in \mathcal{L}; \quad (11)$$

$$\sum_{i \in \mathcal{L}} x_{iu} = 1, \quad u \in V(T); \quad (12)$$

$$x_{iu} + x_{jv} \le 1, \quad i, j \in \mathcal{L}, u, v \in V(T) \\ \text{and} \min_{\lambda_i \in l^{-1}(i), \lambda_j \in l^{-1}(j)} d_\Psi(\lambda_i, \lambda_j) < d_T(u, v); \quad (13)$$

$$x_{iu} + x_{iv} + x_{ju} + x_{jv} \le y_{ij} + 1 \quad i, j \in \mathcal{L}, uv \in E(T); \quad (14)$$

$$\sum_{i,j \in \mathcal{L}} y_{ij} = n - 1. \quad (15)$$

In this formulation, constraints (11) and (12) ensure that x encodes a bijection, while constraints (13) guarantee that the bijection adheres to the conditions of Theorem 1. The auxiliary variables y_{ij} in constraints (14) indicate whether a pair of leaf labels map to adjacent vertices in T, with constraint (15) ensuring the inferred migration network forms a tree. The objective function (10) facilitates the search for a solution by leveraging the relationship between population diversity and population age [109, 110, 111], that suggests that more diverse populations, which are likely older, are also more probable origins of migration [38, 31, 66]. Consequently, it is more likely that such populations correspond to high-degree vertices in the tree T. Here a coefficient $\delta(i)$ represents the genetic diversity of the ith subpopulation, measured as allelic entropy averaged over all allelic positions.

2.5. Migration inference under convexity constraints

In this section, a convex label-distinctive homomorhism $\Psi \to T$ will be called *feasible*. When such homomorphism exists, T can be obtained from Ψ by a series of edge contractions, making T a *minor* of Ψ. Generally, a graph G_1 is a minor of a graph G_2 if G_1 can be obtained from G_2 by edge contractions, edge removals and node removals [112]. It is known that the problem of detecting whether a given graph is a minor of another graph is $\mathcal{NP}$-hard, even when input graphs are trees [113, 114]. However, minors associated with our problem satisfy a more stringent set of conditions than general graph minors: in our case, only edge contractions are allowed and, in addition, contractions of edges between labeled nodes with different labels are forbidden. We suggest that in practical settings feasible homomorphisms can be efficiently found and enumerated using dynamic programming.

The following simple property of convex homomorphisms will be useful for our subsequent analysis:

Lemma 3. *Suppose that $f : \Psi \to T$ is a convex homomorphism. Then T' with is a subtree of T if and only if $\Psi' = f^{-1}(T')$ is a subtree of Ψ.*

Proof. Homomorphic image of a connected subgraph is connected, as implied by the definition of a homomorphism. The converse is also true, if $|T'| = 1$. Suppose that $|T'| \ge 2$ and the subgraph $f^{-1}(T')$ is not connected. Consider two of its connected components, Ψ'_1 and Ψ'_2, such that the unique path P

between the nodes $\alpha \in \Psi'_1$ and $\beta \in \Psi'_2$ is shortest among all such paths between different components. Then $|P| \geq 3$ and $T'' = f(P \setminus \{\alpha, \beta\}) \cap T' = \emptyset$. Moreover, the vertices $f(\alpha)$ and $f(\beta)$ are adjacent to vertices of T''. This leads to two distinct paths between $f(\alpha)$ and $f(\beta)$ – one in T' and another passing through T''. This contradicts the fact that T is a tree. □

Next, we describe the proposed algorithmic approach. Initially, we simplify the original phylogenetic tree Ψ by collapsing paths between leaves sharing identical labels into a single node, a step made feasible by the convexity constraint. The resulting tree, still referred to as Ψ, may become non-binary and contains uniquely labeled leaves and possibly some labeled internal nodes.

The algorithm performs a post-order traversal of the phylogeny Ψ and, for each node $\alpha \in V(\Psi)$, calculates a set H_α describing possible homomorphisms from the subtree Ψ_α to subtrees of T. At the root node ρ, the set H_ρ thus describes all homomorphisms from Ψ to T. Upon completing the traversal, the algorithm either concludes that no feasible homomorphism exists (when $H_\rho = \emptyset$), or initiates a pre-order traversal of Ψ. During this second traversal, it reconstructs feasible homomorphisms using the information from the sets H_α.

Formally, let Λ_α be the set of labeled nodes in the subtree Ψ_α. A subtree $T[v, X]$ of T is termed *an induced v-subtree* if it includes the vertex v, a subset X of v's neighbors, and all vertices that are connected to v via paths that intersect with X (Fig. 3).

For a vertex $\alpha \in V(\Psi)$, the set H_α consists of triples (v, X, C) called *partial homomorphism tokens* or simply *tokens*. In each token, (i) $v \in V(T)$, (ii) $X \subseteq N_T(v)$, (iii) C is a subset of vertices of an induced v-subtree $T[v, X]$ such that there exists a feasible surjective homomorphism $f : \Psi_\alpha \to T[v, X]$ with $f(\alpha) = v$ and $f(\Lambda_\alpha) = C$.

The algorithm is initialized by setting $H_\lambda = \{(v, \emptyset, \{v\}) : v \in V(T)\}$ for all leafs λ. For an internal node α, the set H_α is constructed based on the sets from its children nodes $\beta_1, \ldots, \beta_k$. The construction utilizes the following lemma:

Lemma 4. *Let $T[v, X]$ be an induced v-subtree. Then there exist a feasible surjective homomorphism $f : \Psi_\alpha \to T[v, X]$ with $f(\alpha) = v$ and $f(\Lambda_\alpha) = C$ if and only if there exist tokens $(v_1, X_1, C_1) \in H_{\beta_1}, \ldots, (v_k, X_k, C_k) \in H_{\beta_k}$ satisfying the following conditions:*

(a1) $v \sim v_i$ *or* $v = v_i$ *for all* $i \in \{1, \ldots, k\}$*;*

(b1) $(V(T[v_i, X_i]) \setminus \{v\}) \cap (V(T[v_j, X_j]) \setminus \{v\}) = \emptyset$ *for all* $i, j \in \{1, \ldots, k\}, i \neq j$*;*

(c1) $v \in V(T[v_i, X_i])$ *if and only if* $v_i = v$*.*

(d1) $v \notin C_i$*, if* α *is labeled.*

(e1) $X_i = N_T(v_i) \setminus \{v\}$*, if* $v_i \neq v$*.*

(f1) $X = \{v_1, \ldots, v_k\} \setminus \{v\}$*;*

(g1) $C = C_1 \cup \cdots \cup C_k$*, if* α *is unlabeled, and* $C = C_1 \cup \cdots \cup C_k \cup \{v\}$*, if* α *is labeled.*

Proof. Let us prove the necessity of conditions (a1) - (g1). Suppose that there exists a feasible surjective homomorphism $f : \Psi_\alpha \to T[v, X]$ such that $f(\alpha) = v$. Define T_i as the image of the subtree Ψ_{β_i} under f, denoted by $f(\Psi_{\beta_i})$, and let $v_i = f(\beta_i)$. Also, let $X_i = N_T(v_i) \cap V(T_i)$ and $C_i = f(\Lambda_{\beta_i})$.

We aim to demonstrate that $T_i = T[v_i, X_i]$. According to Lemma 3, T_i is connected, which suggests that T_i must be a subgraph of $T[v_i, X_i]$. Furthermore, Lemma 3 also indicates that $f^{-1}(T[v_i, X_i])$ is connected. Given the surjectivity of f, it follows that $f^{-1}(T[v_i, X_i]) \subseteq \Psi_{\beta_i}$, leading to the conclusion that $T[v_i, X_i] \subseteq T_i$.

Consequently, the restriction of f to Ψ_{β_i} is a surjective homomorphism from Ψ_{β_i} to $T[v_i, X_i]$. Thus, the token (v_i, X_i, C_i)

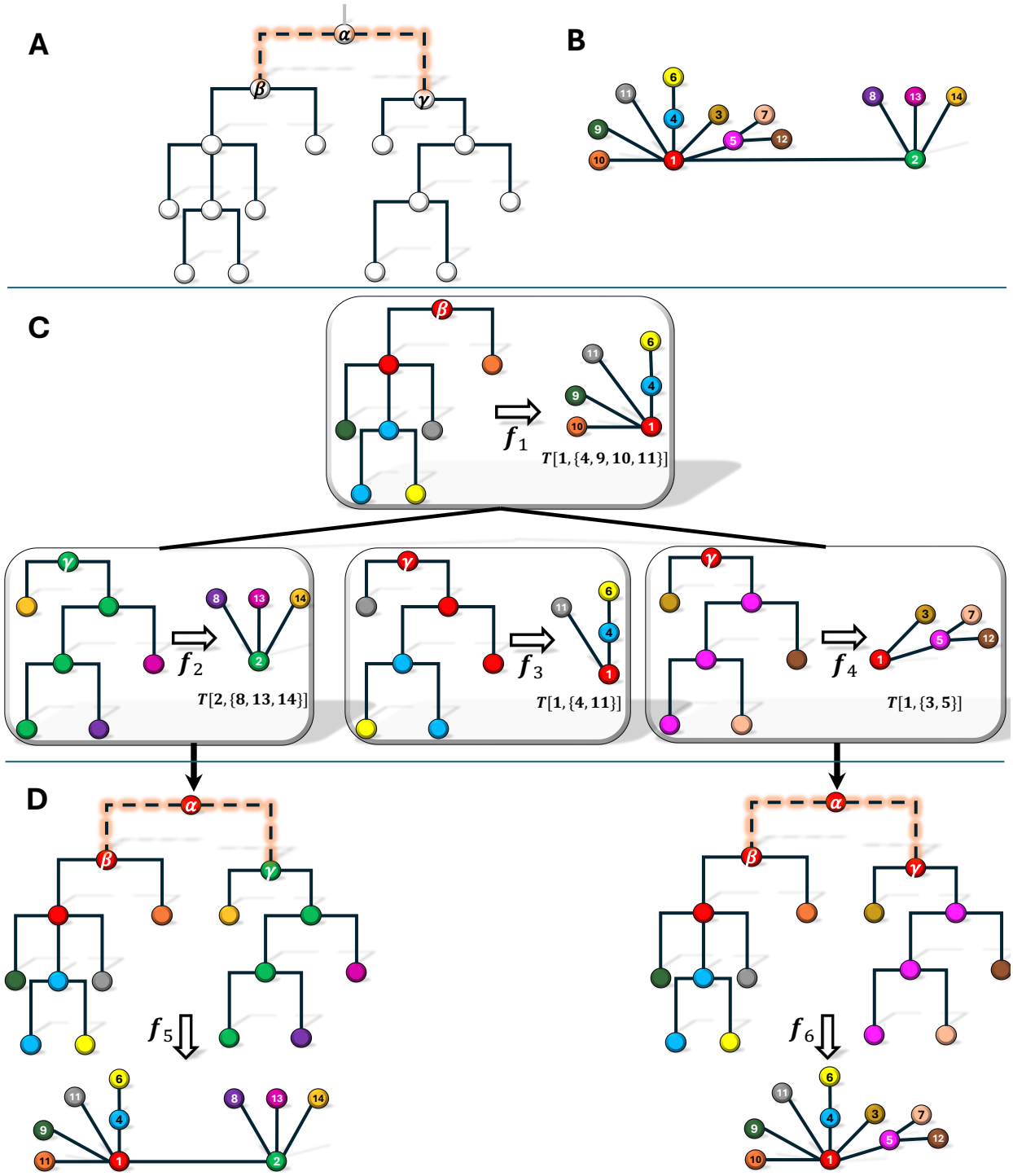

Figure 3. **Overview of the Dynamic Programming Algorithm for Detecting Convex Label-Distinctive Homomorphisms. A.** A phylogenetic subtree, Ψ_α, rooted at node α with two children, β and γ. **B.** Input candidate migration tree T. The goal is to produce convex homomorphisms from Ψ_α to induced subtrees of T from such homomorphisms for Ψ_β and Ψ_γ. **C.** Convex homomorphisms from Ψ_β (top row) and Ψ_γ (bottom row) to induced subtrees of T. For instance, the top figure depicts a homomorphism to an induced 1-tree $T[1,\{4,9,10,11\}]$ that consists of the vertex 1, its neighbors $4, 9, 10, 11$ and all vertices connected to 1 via paths that intersect these neighbors. Nodes of subtrees are colored by their homomorphic images. Homomorphisms are organized into a bipartite graph $\mathcal{G}$, where edges connect homomorphism pairs f_1f_2 and f_1f_4 that are compatible, and there is no edge between homomorphisms f_1 and f_3, that are not compatible. **D.** Homomorphisms f_5 and f_6 obtained by combining compatible homomorphisms f_1, f_2 and f_1, f_4.

belongs to H_{β_i}. The condition (a1) follows from the homomorphism definition, while the convexity of f yields the conditions (b1)-(g1).

Conversely, suppose for each $i \in 1,\ldots,k$, we have feasi-

ble homomorphisms $f_i : \Psi_{\beta_i} \to T[v_i, X_i, C_i]$ that meet conditions (a1) - (e1). We can construct a homomorphism $f : \Psi_\alpha \to T[v, X, C]$ by defining f as follows:

$$f(\gamma) = \begin{cases} f_i(\gamma), \text{ if } \gamma \in V(\Psi_{\beta_i}) \\ v, \text{ if } \gamma = \alpha \end{cases} \quad (16)$$

Condition (a1) implies that f is a homomorphism, condition (d1) ensures that labeled nodes map to distinct vertices of T, and (e1) guarantees that f is surjective.

It remains to show that f is convex. Suppose that $f(\gamma_1) = f(\gamma_2) = w$. If both γ_1 and γ_2 belong to the same subtree Ψ_{β_i}, then the entire path $P_\Psi(\gamma_1, \gamma_2)$ maps to w due to the convexity of f_i. If, on the other hand, $\gamma_1 \in \Psi_{\beta_i}$ and $\gamma_2 \in \Psi_{\beta_j}$, then condition (b1) implies that $w = v$. Furthermore, by the condition (c1) we have $v_i = v_j = v$. Therefore convexity of f_i and f_j, as well as the fact that $f(\alpha) = v$, imply that $f(P_\Psi(\gamma_1, \gamma_2)) = v$. Together, these two facts prove that f is convex. □

To convert Lemma 4 into an algorithm constructing the set of tokens for a node α using the tokens of its children, we first need to identify tokens of children that satisfy conditions (a1)-(g1). This can be achieved through the following steps. For each vertex $v \in V(T)$, we construct a multipartite graph $\mathcal{G} = \mathcal{G}(\alpha, v)$ (i.e. a graph partitioned into k independent sets or *parts*), as described below:

(i) The parts $A_1, \ldots, A_k$ correspond to children of α.

(ii) The vertices of the set A_i are tokens from the set Ψ_{β_i} satisfying the conditions (a1),(c1), (d1) and (e1).

(iii) Two tokens from the sets A_i and A_j are adjacent whenever they satisfy the condition (b1).

In the constructed graph, sets of partial homomorphism tokens that satisfy Lemma 4 can be identified as k-vertex cliques. We employ the Bron-Kerbosch algorithm [115] to generate these cliques. For each identified clique, we use conditions (f1) and (g1) to construct a new token (v, X, C) for the node α.

Additionally, for each token $(v, X, C) \in H_\alpha$, we maintain pointers $p(v, X, C)$ that link to the children tokens used in its construction. These pointers are used in the subsequent phase of the algorithm, which aims to reconstruct full feasible homomorphisms f.

During this phase, the algorithm executes a pre-order traversal of Ψ. As it progresses, it recursively assigns a specific token to each node. When a token $t = (v, X, C)$ is assigned to node α, the algorithm sets $f(\alpha) = v$. It then retrieves tokens for t via the pointers $p(t)$ and assigns them to children, $\beta_1, \ldots, \beta_k$. This ensures that by the end of the traversal, each node in Ψ has been assigned a homomorphic image, completing the construction of the homomorphism f.

In general, the set H_ρ at the root node ρ may include multiple tokens, each representing a feasible homomorphism $f : \Psi \to T$. When multiple feasible homomorphisms are available, the algorithm selects the one that minimizes violations of the compactness constraint (to be discussed in the next subsection). In case of ties, the selection criterion shifts to minimizing the quantity

$$D(f) = \sum_{\alpha \in \Lambda(T)} \delta(\alpha) \cdot d(f(\alpha)), \quad (17)$$

where $\delta(\alpha)$ is the number of labeled children of a node α. This approach prioritizes homomorphisms that map high-degree nodes to high-degree vertices.

The outlined method is formalized in Algorithm 2. Its efficiency can be improved by contracting sibling leaves in both Ψ and T into a single node (vertex). The algorithm maintains a count of copies of contracted nodes used by tokens, and a new token is generated from children tokens only if the total count of each leaf in the children tokens does not exceed its overall count. This modification markedly enhances the dynamic programming algorithm's runtime. The adjustments to Lemma

Algorithm 2 Convex homomorphism minimizing compactness violations

Input: phylogenetic tree Ψ with the root ρ and a candidate migration tree T
Output: a feasible homomorphism $f : \Psi \to T$ or the answer that it does not exist.

modify Ψ by contracting paths between leafs with the same label.
perform a post-order traversal of Ψ.
for every node α of the post-order do
  construct a set of partial homomorphism tokens H_α.
  if α is a leaf then
    $H_\alpha \leftarrow \{(v, \emptyset, \{v\}) : v \in V(T)\}$;
  end if
  if α is an internal node with children $\beta_1, \ldots, \beta_k$ then
    Let $H(\beta_j) = \{(v_i^j, X_i^j, C_i^j) : i = 1, \ldots, l_j\}$, $j = 1, \ldots, k$.
    for $v \in V(T)$ do
      construct the multipartite graph $\mathcal{G}(\alpha, v)$ as described in (i)-(iii);
      Generate the set K of k-vertex cliques of $\mathcal{G}(\alpha, v)$ using Bron-Kerbosch algorithm.
      for each clique $\{(u_{i_1}^1, X_{i_1}^1, C_{i_1}^1), \ldots, (u_{i_k}^k, X_{i_k}^k, C_{i_k}^k)\} \in K$ do
        Construct the sets X and C using formulas (f1) and (g1).
        Set $H_\alpha \leftarrow H_\alpha \cup \{(v, X, C)\}$ and $p(v, X, C) \leftarrow p(v, X, C) \cup \{(i_1, \ldots, i_k)\}$.
      end for
    end for
  end if
end for
if $H_\rho \neq \emptyset$ then
  perform a pre-order traversal of Ψ;
  for each token $t_1, \ldots t_R \in H_\rho$ do
    assign the token t_r to ρ : $AS_\rho \leftarrow t_r$.
    for every node α of the pre-order do
      $f_r(\alpha) \leftarrow v$, where $(v, X, C) = AS_\alpha$.
      if α is an internal node with children $\beta_1, \ldots, \beta_k$ then
        Let $H(\beta_j) = \{(v_i^j, X_i^j, C_i^j) : i = 1, \ldots, l_j\}$, $j = 1, \ldots, k$ and $p(v, X, C) = (i_1, \ldots, i_k)$.
        $AS_{\beta_j} \leftarrow (v_{i_j}^j, X_{i_j}^j, C_{i_j}^j)$, $j = 1, \ldots, k$
      end if
    end for
  end for
  among generated homomorphisms $f_1, \ldots, f_R$, output the homomorphism f_r with the minimal number of compactness violations and, in case of ties, with the minimal $D(f_r)$.
else
  f does not exist
end if

4 and Algorithm 3 in Supplementary Material are straightforward, but involve numerous minor technical details; hence a formal description is omitted. One particular detail, however, should be mentioned: if Ψ' and T' represent the leaf-contracted versions of Ψ and T, respectively, and $f' : \Psi' \to T'$ is a feasible homomorphism, there may be cases where $f(\alpha) = l$ for an internal node α of Ψ and a leaf l in T' that results from the contraction of leaves l_1 and l_2. In such instances, f' can be extended to a homomorphism $f : \Psi \to T$ by designating $f(\alpha)$ as either l_1 or l_2. To resolve this ambiguity, the leaf corresponding to the population with higher diversity is chosen.

Finally, it should be noted that, strictly speaking, Algorithm 2 is not polynomial, since the number of tokens for a node of Ψ theoretically can be exponential. In practical settings, however, the algorithm is extremely fast, and require split seconds to finish.

2.6. *Migration inference with convexity and compactness constraints*

A similar approach to the one outlined in Subsection 2.5 can be employed to identify homomorphisms that are both convex and compact. However, the dynamic programming algorithm can be further optimized by taking advantage of the specific nature

of these constraints.

As with the earlier approach, homomorphisms that are both convex and compact will be referred to as *feasible*. Following a similar methodology to that used in Algorithm 2, we begin by contracting the paths in the phylogeny Ψ. We then construct partial homomorphism tokens similar to those used previously, with the exception that subsets C of images of labeled nodes are not required. Thus, the tokens are simplified to pairs (v, X), where $v \in V(T)$ and $X \subseteq N_T(v)$. A pair (v, X) is included in H_α if there exists a feasible surjective homomorphism $f : \Psi_\alpha \to T[v, X]$ such that $f(\alpha) = v$, and it also satisfies the following condition:

$$|\Lambda_\alpha| = |T[v, X]|. \tag{18}$$

This condition is necessary for a partial homomorphism to be extendable to a full compact homomomorphism $\Psi \to T$.

The algorithm is initialized by setting $H_\lambda = \{(v, \emptyset) : v \in V(T)\}$ for leafs λ. For an internal node $\alpha \in V(\Psi)$, its token set H_α is constructed from the tokens of its children $\beta_1, \dots \beta_k$ using Lemma 5.

Lemma 5. *$(v, X) \in H_\alpha$ if and only if one of the following conditions hold:*

1) *α is not labeled and there exist $w \in X$ such that $(v, X \setminus \{w\}) \in H_{\beta_1}$ and $(w, N_T(w) \setminus \{v\}) \in H_{\beta_2}$.*
2) *α is labeled, $|X| = k$, and there exist a permutation $(v_1, \dots, v_k)$ of elements of X such that $v \sim v_1, \dots, v_k$, and $(v_1, N_T(v_1) \setminus \{v\}) \in H_{\beta_1}, \dots, (v_k, N_T(v_k) \setminus \{v\}) \in H_{\beta_k}$.*

Proof. We present the proof for the case where α is unlabeled; the argument for labeled α follows a similar rationale. In this case, α was not involved in path contraction, and thus $k = 2$.

Suppose that $(v, X) \in H_\alpha$, i.e. $|\Lambda_\alpha| = |T[v, X]|$ and there exists a feasible surjective homomorphism $f : \Psi_\alpha \to T[v, X]$ such that $f(\alpha) = v$. Consequently, $f(\Lambda_\alpha) = V(T[v, X])$, and therefore there must be a labeled node $\gamma \in V(\Psi_{\beta_1})$ such that $f(\gamma) = v$. Given the connectivity constraint, this implies that $f(\beta_1) = v$. Meanwhile, the compactness constraint necessitates $f(\beta_2) = w \neq v$.

Following Lemma 3 and considering the connectivity constraint, it can be shown that:

$$f(\Psi_{\beta_1}) = T[v, X \setminus \{v\}] \text{ and } f(\Psi_{\beta_2}) = T[w, N(w) \setminus \{v\}] \tag{19}$$

Let us now establish the second equality; the method for proving the first is analogous. Let $T_2 = f(\Psi_{\beta_2})$. We know that $v \notin V(T_2)$ and, according to 3, T_2 is connected. These observations imply that $T_2 \subseteq T[w, N(w) \setminus \{v\}]$. Conversely, Lemma 3 suggests that $f^{-1}(T[w, N(w) \setminus \{v\}])$ is connected. Given the surjectivity of f, this implies $f^{-1}(T[w, N(w) \setminus \{v\}]) \subseteq \Psi_{\beta_2}$, leading to $T[w, N(w) \setminus \{v\}] \subseteq T_2$. Thus, both $T_2 \subseteq T[w, N(w) \setminus \{v\}]$ and $T[w, N(w) \setminus \{v\}] \subseteq T_2$ are true, confirming the second equality.

So, the restrictions $f|_{\Psi_{\beta_1}}$ and $f|_{\Psi_{\beta_2}}$ are both feasible surjective homomorphisms. Additionally, $\Lambda_\alpha = \Lambda_{\beta_1} \cup \Lambda_{\beta_2}$, $f(\Lambda_{\beta_1}) = f(\Psi_{\beta_1}) = T[v, X \setminus \{w\}]$ and $f(\Lambda_{\beta_2}) = f(\Psi_{\beta_2}) = T[w, N_T(w) \setminus \{v\}]$, thus confirming that the equality (18) holds for the tokens $(v, X \setminus \{w\})$ and $(w, N_T(w) \setminus \{v\})$. This proves the necessity of condition 1).

To demonstrate the sufficiency of condition 1), we assume that there exist feasible surjective homomorphisms $f_1 : \Psi_{\beta_1} \to T[v, X \setminus \{w\}]$ and $f_2 : \Psi_{\beta_2} \to T[w, N_T(w) \setminus \{v\}]$. By defining f as follows, we can establish a combined feasible homomorphism:

$$f(\chi) = \begin{cases} f_1(\chi) & \text{if } \chi \in V(\Psi_{\beta_1}) \\ f_2(\chi) & \text{if } \chi \in V(\Psi_{\beta_2}) \\ v & \text{if } \chi = \alpha \end{cases} \tag{20}$$

□

Lemma 5 can be directly applied to construct tokens for an unlabeled node α using the tokens from its children. For

a labeled node α, however, the process of finding a permutation $(v_1, \ldots, v_k)$ of the tokens is more complex. The method to achieve this is detailed in the following approach.

Lemma 5 can be straightforwardly use to construct tokens of α from tokens of its children, if α is unlabeled. When α is labeled, then finding a permutation $(v_1, \ldots, v_k)$ required by Lemma 5 is more complicated and can be achieved using the approach described next.

Suppose that $\mathcal{S} = (S_1, \ldots, S_k)$ is a collection of sets. A vector $(x_1, \ldots, x_k)$ is termed a *transversal* of $\mathcal{S}$ [116] if $x_i \in S_i$ and all x_i are distinct. Let now $S_i = \{u : (u, Y)) \in H_{\beta_i}$ and $v \sim u\}$. Then the vector $(v_1, \ldots, v_k)$ satisfies the condition 2) of Lemma 5 if and only if it is a transversal of $\mathcal{S}$.

Given this, the set H_α can be obtained by generating all transversals of $\mathcal{S}$. This process involves the following steps:

- Construct a bipartite graph $\mathcal{B}(\mathcal{S})$ with parts $\mathcal{I}$ and $\mathcal{J}$, where $\mathcal{I} = 1, \ldots, k$ represents the set indices, and $\mathcal{J} = \bigcup_{i=1}^{k} S_i$, represents all elements in the sets. In this graph, a vertex $i \in \mathcal{I}$ is adjacent to a vertex $j \in \mathcal{J}$ if j belongs to S_i.
- In $\mathcal{B}$, each transversal corresponds to a maximal matching of size k. To generate these matchings, construct a *line graph* $L(\mathcal{B})$, where each vertex represents an edge of $\mathcal{B}$, and two vertices are adjacent if their corresponding edges in $\mathcal{B}$ share a common vertex. Maximal matchings of $\mathcal{B}$ correspond to maximal independent sets in $L(\mathcal{B})$, that can be produced using the Bron–Kerbosch algorithm [115].

The entire method is detailed in Supplementary Material, Algorithm 3. Like Algorithm 2, efficiency can be significantly improved by contracting sibling leaves in both Ψ and T.

2.7. SMiTH: Sampling Migration Trees via Homomorphisms

The algorithms discussed can be effectively integrated into an Unlabeled Migration Sampling framework (Problem 5). It allows for the identification of homeomorphic images of a given phylogeny Ψ within a collection of candidate migration trees sampled from a given migration pattern represented by a specified tree distribution.

The obtained sample of migration trees can be directly analyzed to estimate the probabilities of specific migration routes or to obtain summary statistics and confidence intervals for derivative evolutionary parameters. It can be also synthesized into a single weighted *consensus graph*, where each edge is weighted by the number of candidate trees that support it. When the homomorphism reconstruction includes an objective function, the consensus graph is constructed from a subsample comprising the top κ% of trees ranked by their objective values. For applications requiring a specific output tree – such as for benchmarking and comparison with other methods described in Subsection 3.1 – the tree is determined by calculating the maximum-weight spanning tree of the consensus graph. The entire algorithmic pipeline, named SMiTH (Sampling Migration Trees via Homomorphisms), is illustrated in Figure 1.

3. Results

3.1. Simulated data

To generate synthetic data, we used FAVITES [117], a tool capable of simulating genomes, phylogenies, and migration networks under various evolutionary scenarios. Although originally designed to simulate viral outbreaks, FAVITES supports general phylogenetic and population genetics models, making it suitable for simulating migrations of heterogeneous populations besides viruses. It also should be noted that, to the best of our knowledge, specialized simulation tools for metastatic

spread with capabilities comparable to FAVITES are currently not available.

We simulated the migration of a heterogeneous population over a network of sites formed according to the Barabasi-Albert model [70]. This assumption can be valid for both viral [98, 118] and cancer [119, 95] spread. Migrations occur at a constant rate along each network edge (in viral context, this corresponds to the network-based Susceptible-Infected (SI) transmission model). Within each site, phylogenies evolved under the exponential coalescent, a model previously used to simulate intra-host [66] and intra-tumor [120] evolution. Genotypes were assumed to evolve under the GTR+Γ substitution model, and were sampled simultaneously at the end of the simulation. In total, 275 simulated datasets were generated, encompassing 5–30 demes with 100 sequences sampled per deme.

In the first series of experiments, we sampled candidate migration trees from 3 distinct prior distributions of tree topologies and evaluated their compatibility with simulated phylogenies under three different types of constraints. The prior distributions included:

T1) Degenerate distribution consisting of the true topology. Even though this scenario is unrealistic for actual migration inference, it serves as a test to determine if migration links can be accurately reconstructed when the topology is known but sites need to be correctly mapped to migration network vertices. To avoid bias linked to correlations between vertex IDs and migration times, that can be potentially introduced by the simulation methods, we produced multiple samples with randomly permuted vertex IDs.

T2) Random scale-free trees produced by the preferential attachment procedure.

T3) Uniformly distributed trees of a given size. Sampling was performed by generating random Prufer codes [121], integer sequences of length $n-2$ that uniquely define n-vertex trees.

The following types of constraints were used:

H1) unconstrained homomorphism;

H2) convex homomorphism minimizing the number of compactness constraint violations;

H3) convex and compact homomorphism.

For each simulated dataset, we produced 9 samples of candidate migration trees corresponding to all combinations of conditions T1)–T3) and H1)–H3). The sample sizes ranged from 1,000 for the degenerate distribution without constraints, up to 1,000,000 for the uniform distribution with convexity and compactness constraints. This variation in sample size was necessary because stricter constraints require larger samples to ensure that a sufficient number of feasible trees are produced. We assessed the compatibility of these sampled trees with the given phylogenies under the respective constraints, using methods detailed in Subsections 2.4–2.6. Using these assessments, subsamples of compatible tree were extracted.

Sampled trees were compared with true migration trees produced by FAVITES. Individual trees were compared by measuring recall, defined as the fraction of inferred transmission edges among true transmission edges; precision, the fraction of true transmission edges among inferred transmission edges; and the f-score, i.e., the harmonic mean of precision and recall.

Additionally, we summarized each subsample of compatible migration trees using a *consensus graph*, where each edge is weighted by the proportion of candidate trees that support that edge [26, 34, 122]. A solution can be extracted from the consensus graph by discarding edges with support below a predefined threshold. For each graph, we estimated the area under

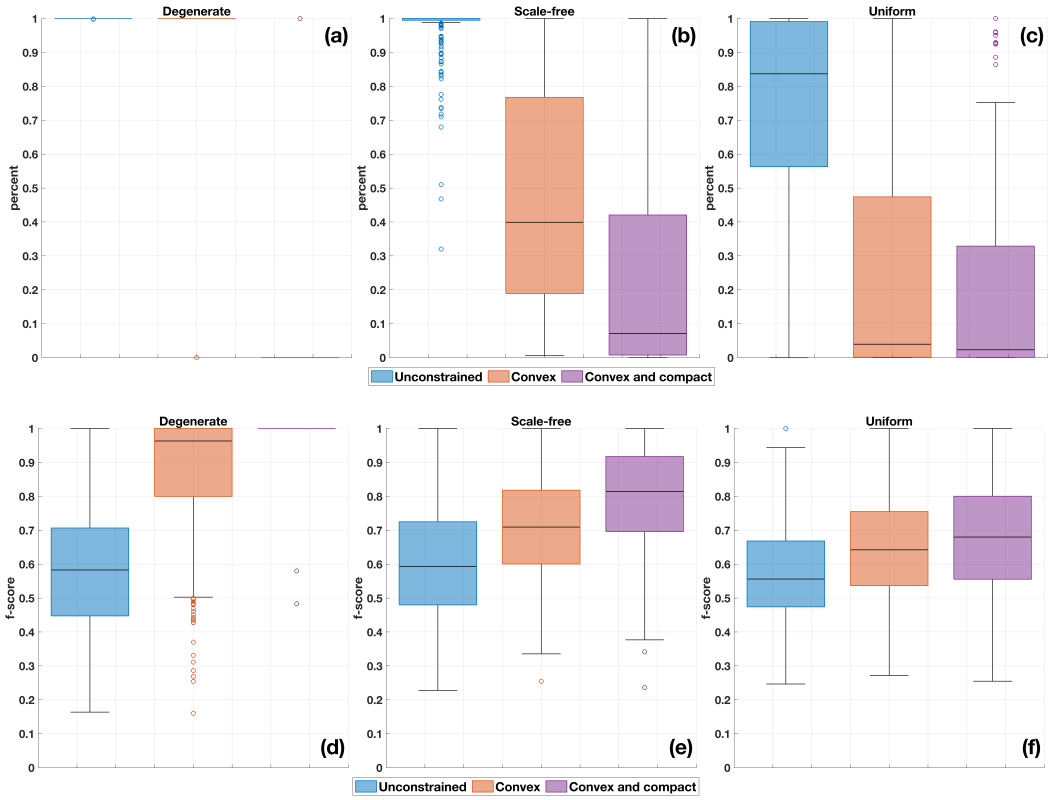

Figure 4. (a)-(c) Percent of sampled trees that are compatible with given phylogenies. (d-f) Area under the precision-recall curve (AUC) calculated by varying the support threshold.

the precision-recall curve, which was calculated by varying the support threshold. We opted for the precision-recall curve instead of the more common ROC curve due to the imbalance between the classes of true and false migration edges.

We found that the relationship between phylogenetic trees and migration trees is heavily influenced by structural constraints. In the absence of constraints, there is considerable ambiguity in the possible migration histories that align with a given phylogenetic tree topology, even when prior knowledge about true migration tree topology is available. Notably, almost every sampled scale-free network proved compatible with the given phylogenies, with a median fraction of compatible trees at $\mu = 1$ (Fig. 4b). It is not entirely unexpected in light of Theorem 1, which suggests that trees with a low diameter – a common feature of scale-free networks – are more likely to be compatible with a given phylogenetic tree. However, even among uniformly sampled trees, a high compatibility rate was observed when no constraints are applied (median $\mu = 0.84$, Fig. 4c).

Furthermore, without constraints individual compatible trees display only marginal agreement with true migration trees. This holds not only for scale-free and uniformly sampled trees, but even for the degenerate distribution, with median f-score within the range 0.33–0.36 for all three tree priors (Supplementary Material, Fig. 8). In other words, even if the topology of a true migration tree is known, numerous labelings of that topology are compatible with the original phylogeny, most of which substantially diverge from the true labeling. Consequently, the true labeling is not immediately distinguishable among the alternatives without additional information. Combining a compatible tree subsample into a consensus graph, however, brings it closer to the true migration tree, with the median AUC values ranging from 0.56 to 0.59 for all three tree priors (see Fig. 4).

The introduction of convexity and compactness constraints significantly reduces the percentage of trees deemed compati-

ble, as can be expected (Fig. 4abc). This reduction primarily eliminates incidental solutions, enhancing the alignment of the trees that meet these constraints with the true migration trees (Fig. 4def). In particular, for the degenerate distribution, introducing constraints effectively filters out ambiguous vertex mappings, nearly always recovering the true mapping, provided that the solutions satisfying the constraints exist.

Interestingly, without constraints, the use of tree priors does not enhance accuracy, as demonstrated by the lack of significant differences in AUC distributions among the tree priors ($p = 0.083$, Kruskal-Wallis test). In contrast, when constraints are applied, prior knowledge of the migration tree structure becomes beneficial, with AUCs improving as the tree prior becomes tighter. In particular, under constraints, AUCs for the scale-free prior are significantly higher than those for the uniform prior ($p = 4.4 \cdot 10^{-6}$ and $p = 1.76 \cdot 10^{-16}$ for convex and both convex and compact cases, respectively, Kruskal–Wallis test). Given that true migration trees are generated by the preferential attachment, this suggests that under constraints, the phylogeny to a certain degree reflects the properties of the underlying migration network.

Taken together, these observations indicate that phylogeny topologies do indeed reflect underlying migration tree structures, but the extent of this reflection is influenced by evolutionary constraints. Moreover, the correspondence between phylogenies and migration trees is primarily discernible when analyzed statistically across a large sample of feasible migration trees that are compatible with the phylogeny. A single compatible tree may be arbitrary, and thus relying on a single solution may lead to misleading conclusions.

Based on those observations, we have developed a method named SMiTH (Sampling MIgration Trees using Homomorphisms) for the constrained inference of migration trees with expected general properties. This method involves sampling candidate migration trees from a designated random tree distribution, identifying convex homomorphisms from the given phylogeny to these sampled trees while minimizing an objective function defined by the number of compactness constraint violations, constructing a consensus graph from trees with top objective values, and ultimately inferring the final migration tree as the minimal spanning tree of this consensus graph.

We benchmarked SMiTH against several existing tools designed to infer migration networks from phylogenetic tree topologies. For the sake of fairness, tools that use dated phylogenies and/or case-specific epidemiological information were not considered. The tools selected for this comparison include Cassiopeia [21, 59], MACHINA [22], Phyloscanner [31], STraTUS [36], and TNet [44]. For MACHINA, we ran all four migration models provided by the tool and report the best result, which was achieved using the single-source seeding model. STraTUS generates a sample of migration trees rather than a single tree; thus, similarly to SMiTH, we used the minimum spanning tree of the consensus graph for benchmarking purposes.

The results of the algorithm comparison are shown in Fig. 5. It was found that both variants of SMiTH – with uniform and scale-free tree priors – allow for a statistically significant improvement over other tools ($p < 10^{-9}$, multiple comparison of f-score distributions by Kruskal–Wallis test). SMiTH is followed by Cassiopeia and STraTUS – two other sampling-based methods, whose accuracies were statistically indiscernible ($p = 0.58$, Kruskal–Wallis test). These tools indeed both produce samples of convex solutions, albeit using different algorithms. While STraTUS is doing it directly, while Cassiopeia's module FitchCount samples most parsimonious solutions that in our examples were almost always convex. These tools were followed by MACHINA that, similarly to our approach, imposes structural constraints on plausible migration trees by considering them as subgraphs of so-called transition patterns. However,

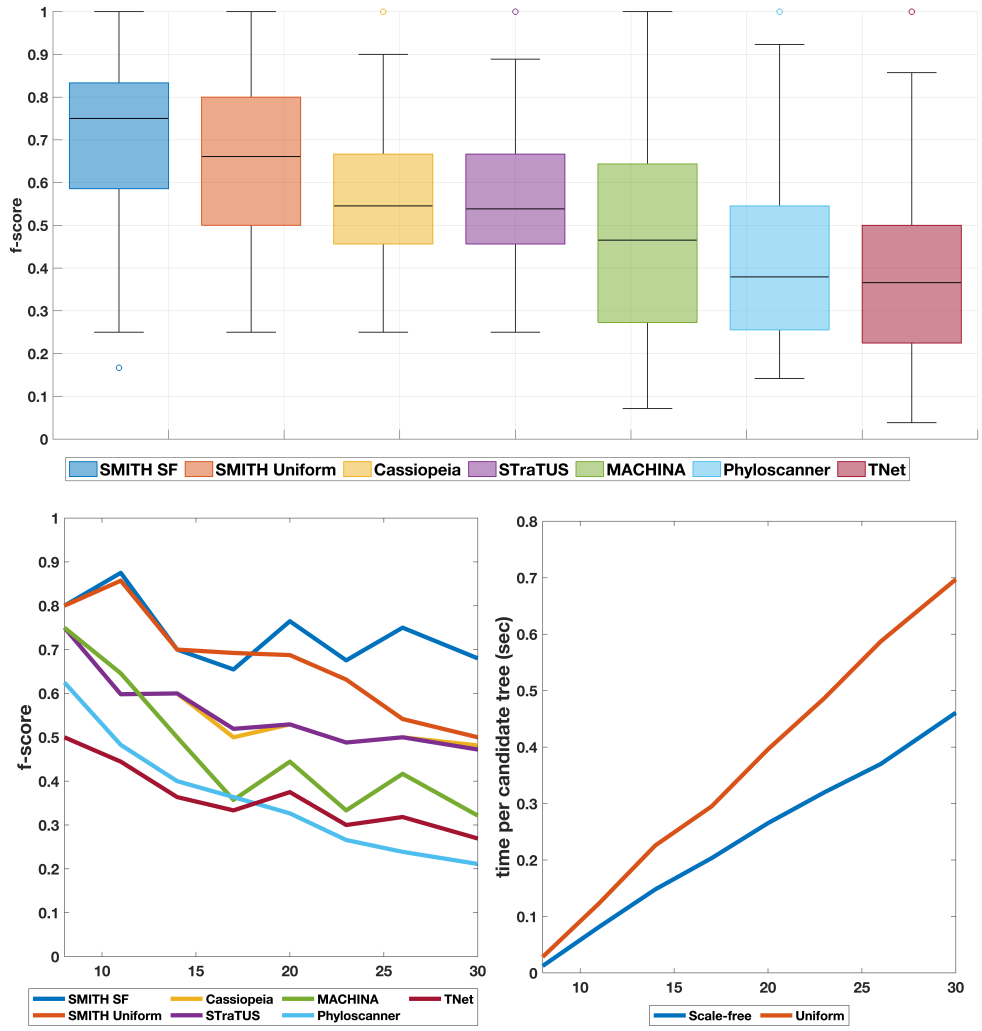

Figure 5. (a) Summary statistics of methods performance on all simulated datasets. (b) Median f-scores of different methods on simulated datasets with varying numbers of populations. (c) Median running times of SMiTH for construction of a constrained homomorphism for a given phylogenetic tree and candidate migration tree as a function of number of migration sites

MACHINA produces a single solution optimizing particular objectives rather than summarizes a sample of such solutions; this, in light of the observations described above, likely hampered its performance vis-à-vis sampling-based methods. Similar reasoning can be applied to Phyloscanner, that also produces a single most parsimonious solution. In addition, Phyloscanner is specifically designed to make use of paraphyly, that usually provides a strong signal for migration [66] when present; consequently, its accuracy can be affected when, as in analyzed test cases, the number of paraphyletic clades is limited. Furthermore, Phyloscanner usually assumes that when the populations sampled from different cites are monophyletic, then they all have a common source [66] - the assumption that is opposite to compactness and that seems to be not always valid. The relation between algorithms' performances is mostly stable with regard to the number of migration sites (Fig. 5b).

The median running time of our method for construction of a constrained homomorphism for a given phylogenetic tree and candidate migration tree was within 0.7 seconds for all tests and tree priors (Fig. 5c), even though theoretically our algorithms can be exponential in the worst case. It allows us, using straightforward parallelization, to produce and process samples consisting in hundreds of thousands of candidate trees in reasonable time.

3.2. Experimental viral data

Analysis of simulated data highlights the role of structural constraints in migration tree inference, particularly when paraphyly is limited. Interestingly, these constraints prove just as essential in scenarios with a high degree of paraphyly, albeit for different reasons. This is evidenced by the analysis of data of Hepatitis C (HCV) outbreaks, which have been considered in previous studies [38, 39, 44, 27]. The data comprises intra-host HCV populations from several outbreaks investigated by the Centers for Disease Control and Prevention, each population consisting of sequences covering Hypervariable Region 1 (HVR1) of the HCV genome. In each outbreak, a single primary host infected all other hosts, rendering the migration tree in graph-theoretical terms a *star*.

We analyzed two largest outbreaks involving 15 and 19 infected hosts. Phylogenetic trees for each outbreak were constructed using RAxML [123]. All transmissions occurred within a short time frame and, as a result, intra-host populations are highly intermixed (Fig. 6a). This makes paraphylytic signal strong, but oversaturated, thus impeding its use to reconstruct true transmission history.

This effect can be demonstrated by examining internal node labels generated by Fitch algorithm, that serves as a basis for several methods considered in the previous section. In the trees analyzed, 32–34% of internal nodes were assigned a single provisional label during the post-order traversal step of the dynamic programming algorithm, indicating that these labels appear in all most parsimonious solutions (or solutions with the minimal migration number in terms of [22]). Many of these nodes are adjacent, suggesting that the transmission links they represent will be identified by any parsimony-based sampling approach similar to those employed by existing tools [44, 12, 21]. For one outbreak, these links form a connected graph, whereas in the other, only one host does not integrate into this single connected component (see Fig. 6b). These resulting graphs are relatively dense and include not only true edges but also a significant number of false positives (see Fig. 6b). Consequently, even if all true positive edges are correctly identified using nodes with multiple Fitch labels, the f-scores would not exceed 0.54 and 0.51, respectively. Enhancing the accuracy of this approach requires the filtering out of false positive edges, achievable only through the integration of additional prior information or constraints.

In contrast, sampling unconstrained candidate transmission trees from the scale-free tree distribution produce the results that are significantly closer to true transmission histories (Fig. 6c). For consensus networks derived from these samples, areas under precision-recall curve are estimated at 0.76 and 0.70. Furthermore, f-scores of solutions obtained as minimal spanning trees of consensus networks produced from top 1% of sampled trees according to the objective (10) are 0.86 and 0.94. Comparable results – $f = 0.79$ and $f = 0.89$ – are obtained if we use top 1% of sampled trees based on the parsimony score.

3.3. Experimental cancer data

We employed SMiTH to analyze the migration history of metastatic ovarian cancer using the data published in [124]. The dataset comprised whole-genome and targeted sequencing data from samples collected at various anatomical sites, including the left ovary (LOv), the right ovary (ROv), and several metastases. In the original study, migration networks were inferred using hierarchical clustering trees and a Dollo parsimony model. Subsequent re-analysis using MACHINA [22] revealed several additional, more parsimonious migration histories.

We focused on the data from Patients 1, 3, and 7, that included the highest number of anatomical sites (7-8 sites) and that were thoroughly analyzed in [22]. We used clone trees shared by the authors of [22]. For each patient, we sampled candidate migration trees from a uniform distribution using an

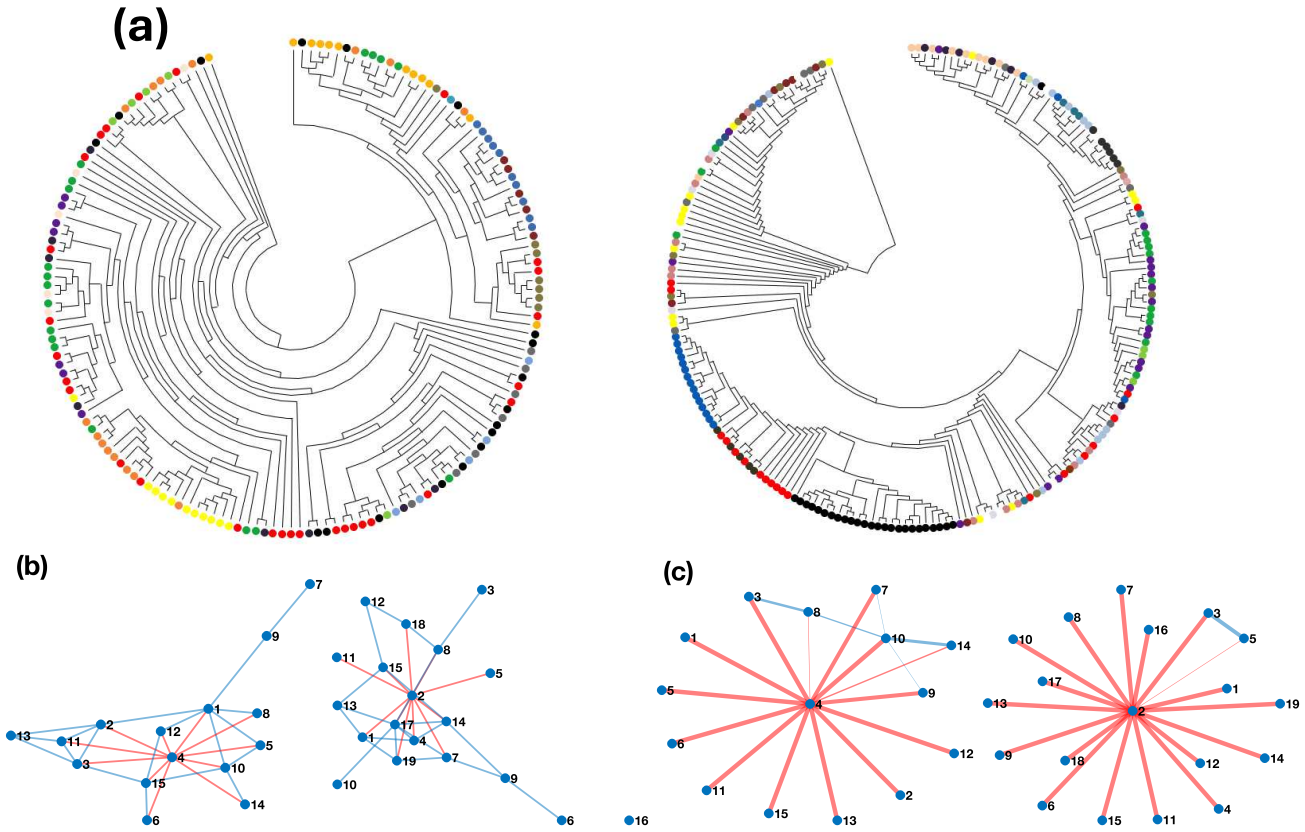

Figure 6. (a) Phylogenetic trees of HCV variants from two outbreaks. Variants sampled from different hosts are highlighted in different colors. (b) Graphs formed by edges corresponding to adjacent tree nodes with unique Fitch labels. True edges are highlighted in red. (c) Consensus networks of top 1% of sampled trees with respect to the objective (10). Edge thicknesses are proportional to their frequencies, true edges are highlighted in red.

unconstrained model. Following the methodology described in [22], we resolved polytomies in clone trees to match the solutions reported there. In instances where the resolution of polytomies was ambiguous, we applied a random resolution, generating a new random resolution for each sampled candidate migration tree.

For Patient 1, [124] identified a complex migration history designating ROv as the primary tumor site. MACHINA was able to find several more parsimonious histories that suggested either LOv or ROv as the primary tumor location, but was not able to distinguish between them. These histories shared parsimony scores (referred to as "migration numbers" [22]) and/or the same number of migration events (named "co-migration numbers" [22]), which left the primary tumor's location ambiguous. In contrast, SMiTH enabled the comparison of migration number distributions for different potential primary tumor sources (Fig. 7) rather than making the decision based on single most parsimonious solutions. Migration numbers associated with LOv and ROv were significantly lower than those of other potential sources ($p < 10^{-65}$, Mann–Whitney U test). Among these two, the lowest numbers were observed for ROv ($p < 10^{-38}$, Mann–Whitney U test), indicating a stronger statistical support for ROv as the primary tumor source.

A similar situation was observed for Patient 7. Here, [124] suggested the right uterosacral ligament (RUt) as the primary tumor location, while MACHINA identified several alternative migration histories with either the left ovary (LOv) or the right ovary (ROv) as the source, each sharing identical migration and

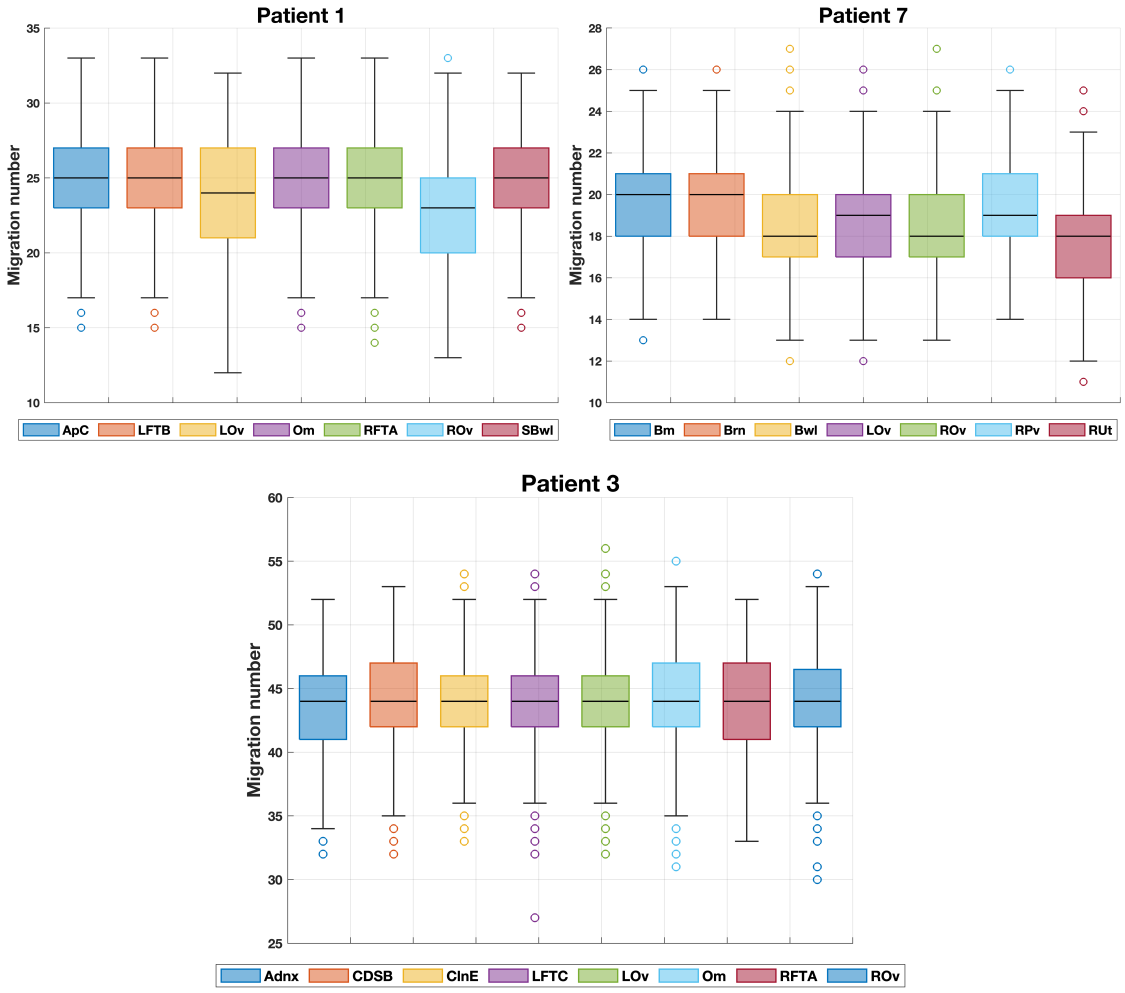

Figure 7. Distributions of migration numbers for trees with different primary tumor sites.

co-migration numbers. Based on these results, [22] argued that available data provides no evidence for the assertion that the primary tumor is located in the RUt as opposed to the ovaries. However, SMiTH provided such statistical evidence (Fig. 7), showing that migration trees with RUt as the source generally exhibited lower migration numbers ($p < 10^{-154}$, Mann–Whitney U test).

Patient 3 presents a different scenario. Here, MACHINA identified several migration histories with LOv or ROv as primary tumor sources. There is also an alternative history with the omentum (Om) as the source, which has a lower migration number. The latter hypothesis appears preferable if judged solely by the single most parsimonious solution. Yet, this conclusion is not reliable due to the highly symmetric distribution of clones from different sites in the clone tree. Many clones form polytomies, offering insufficient data to clearly differentiate between the corresponding sites (e.g., all clones from LOv and LFTC are siblings, rendering these sites indistinguishable; see Supplementary Material, Fig. 9). The apparently lower migration number for Om, compared to other sites, is simply due to its representation by five clones, versus four for several other sites – a difference that could stem from sampling bias given the small number of clones involved.

SMiTH was able to capture and quantify this uncertainty. Specifically, it did not find significant differences in the distribution of migration numbers among potential primary sources,

with p-values ranging from 0.09 to 0.96 in pairwise Mann–Whitney U tests and a $p = 0.65$ in a joint Kruskal–Wallis test (Fig. 7). This provides a statistical support for suggestion that the existing data is insufficient to draw reliable conclusions about the migration history.

In total, these examples illustrate how SMiTH can be used to provide statistical support for hypotheses regarding metastatic spread pathways.

4. Discussion

This study is dedicated to in-depth mathematical exploration of the relationships between phylogenies and migration trees of heterogeneous genomic populations. Although it is established that phylogenetic trees impose some restrictions on migration pathways [36], the exact nature and extent of these constraints are still not well understood, despite a considerable amount of research dedicated to this problem [73, 72, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 36]. Our approach adds both depth and rigor to this area by utilizing the powerful theoretical and algorithmic framework of theory of graph homomorphisms. This framework allowed us to derive necessary and sufficient conditions for the compatibility of phylogenetic and migration trees, and to develop efficient algorithms for analyzing this compatibility through numerical experiments.

Based on our findings, we propose a general and flexible computational framework that can be used to infer migration networks under various assumptions, quantitatively assess competing hypotheses about migration dynamics, investigate the influence of phylogenies on the migration tree space, and to determine whether a potential migration history is definitively contradicted by a phylogeny or set of phylogenies.

Methodologically, our approach balances the advantages of probabilistic and parsimony methods. It incorporates scalability and the use of advanced combinatorial optimization techniques from the latter, along with the biological plausibility of the former that comes from employing appropriate prior random tree distributions.

This study aligns well with the context of previous research. Several earlier studies have conceptualized migration inference as a coloring problem [35, 34, 36], employing this framework both to develop efficient inference algorithms and to explore the structure of migration tree space. The methodology introduced in this paper advances these prior approaches by incorporating a more comprehensive mathematical model and applying more sophisticated mathematical techniques. Similarly, the concept of constraining the tree space to random trees from a specified distribution was first introduced in our earlier studies on viral transmissions [38, 27], while a related concept of restricting the tree space to subgraphs of specific migration patterns has been utilized in computational cancer genomics [22, 61]. This paper not only refines and extends these methodologies but also integrates them into a cohesive modeling and computational framework.

The proposed approach certainly has both advantages and disadvantages. We recognize that uniform sampling from the space of candidate migration trees may not be the most optimal tool for migration dynamics inference, and employing more precise optimization or sampling techniques to navigate the tree space could substantially enhance the method's accuracy and efficiency. This work lays a foundation for further theoretical and algorithmic development in this direction, equipping researchers with the tools needed to expand the use of graph homomorphism methodologies. On the other hand, uniform sampling may be more suitable for hypothesis testing and comparison. It also should be noted that a key aspect of our methodology is that it involves sampling from the space of subtrees of a transition pattern, rather than from the space of phylogeny node labelings as was done in other studies [36, 44, 12].

The former space is considerably smaller in practical scenarios which often makes the uniform sampling computationally feasible.

Furthermore, it is likely that migration models tailored to the specific characteristics of underlying populations could potentially yield more accurate insights into the biological processes involved. In particular, the biological mechanisms driving viral and cancer migrations are certainly very different. Nonetheless, employing a unified phylogenetic approach to study highly mutable populations offers several advantages. First, it decouples the initial phylogenetic reconstruction from its biological interpretation, thereby minimizing the risk of overfitting and ensuring that the results are less biased by underlying models [35]. Additionally, such methods are significantly more computationally efficient and scalable compared to parameter-rich models [35, 27]. Finally, more general phylogenetic models offer greater flexibility and versatility, and usually can be readily extended to more specific settings through the integration of suitable priors [35]. These features make them an excellent foundation for more detailed analyses.

5. Acknowledgements

K.K. has been supported by a Georgia State University Molecular Basis of Disease fellowship. P.S. has been supported by the NSF grants 2047828 and 2212508. The authors thank the organizers of the Computational Genomics Summer Institute (UCLA, July 12 – August 4, 2023 and July 10 - August 2, 2024), where several ideas of this paper were conceived.

6. Code availability

The code developed and used in this study is available at https://github.com/compbel/SMiTH.

References

[1] E. Domingo, J. Sheldon, C. Perales, Viral quasispecies evolution, Microbiology and Molecular Biology Reviews 76 (2) (2012) 159–216.

[2] R. A. Burrell, N. McGranahan, J. Bartek, C. Swanton, The causes and consequences of genetic heterogeneity in cancer evolution, Nature 501 (7467) (2013) 338–345.

[3] E. C. Smith, The not-so-infinite malleability of rna viruses: Viral and cellular determinants of rna virus mutation rates, PLoS pathogens 13 (4) (2017) e1006254.

[4] S. Duffy, Why are rna virus mutation rates so damn high?, PLoS biology 16 (8) (2018) e3000003.

[5] R. Sanjuán, P. Domingo-Calap, Mechanisms of viral mutation, Cellular and molecular life sciences 73 (2016) 4433–4448.

[6] I. P. Tomlinson, M. Novelli, W. Bodmer, The mutation rate and cancer, Proceedings of the National Academy of Sciences 93 (25) (1996) 14800–14803.

[7] M. Greaves, Nothing in cancer makes sense except…, BMC biology 16 (1) (2018) 1–8.

[8] W. M. Grady, Genomic instability and colon cancer, Cancer and metastasis reviews 23 (1-2) (2004) 11–27.

[9] G. S. Charames, B. Bapat, Genomic instability and cancer, Current molecular medicine 3 (7) (2003) 589–596.

[10] F. Utro, C. Levovitz, K. Rhrissorrakrai, L. Parida, A common methodological phylogenomics framework for intra-patient heteroplasmies to infer sars-cov-2 sublineages and tumor clones, BMC genomics 22 (5) (2021) 1–13.

[11] T. Stadler, O. G. Pybus, M. P. Stumpf, Phylodynamics for cell biologists, Science 371 (6526) (2021) eaah6266.

[12] P. Sashittal, M. El-Kebir, Sampling and summarizing transmission trees with multi-strain infections, Bioinformatics 36 (Supplement_1) (2020) i362–i370.

[13] J. Z. Sanborn, J. Chung, E. Purdom, N. J. Wang, H. Kakavand, J. S. Wilmott, T. Butler, J. F. Thompson, G. J. Mann, L. E. Haydu, et al., Phylogenetic analyses of melanoma reveal complex patterns of metastatic dissemination, Proceedings of the National Academy of Sciences 112 (35) (2015) 10995–11000.

[14] D. X. Nguyen, J. Massagué, Genetic determinants of cancer metastasis, Nature Reviews Genetics 8 (5) (2007) 341–352.

[15] N. D. Grubaugh, J. T. Ladner, P. Lemey, O. G. Pybus, A. Rambaut, E. C. Holmes, K. G. Andersen, Tracking virus outbreaks in the twenty-first century, Nature microbiology 4 (1) (2019) 10–19.

[16] G. L. Armstrong, D. R. MacCannell, J. Taylor, H. A. Carleton, E. B. Neuhaus, R. S. Bradbury, J. E. Posey, M. Gwinn, Pathogen genomics in

public health, New England Journal of Medicine 381 (26) (2019) 2569–2580.

[17] A. Black, D. R. MacCannell, T. R. Sibley, T. Bedford, Ten recommendations for supporting open pathogen genomic analysis in public health, Nature Medicine (2020) 1–10.

[18] D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz, et al., Eleven grand challenges in single-cell data science, Genome biology 21 (1) (2020) 1–35.

[19] N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, et al., Tumour evolution inferred by single-cell sequencing, Nature 472 (7341) (2011) 90–94.

[20] S. Knyazev, L. Hughes, P. Skums, A. Zelikovsky, Epidemiological data analysis of viral quasispecies in the next-generation sequencing era, Briefings in bioinformatics 22 (1) (2021) 96–108.

[21] J. J. Quinn, M. G. Jones, R. A. Okimoto, S. Nanjo, M. M. Chan, N. Yosef, T. G. Bivona, J. S. Weissman, Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts, Science 371 (6532) (2021) eabc1944.

[22] M. El-Kebir, G. Satas, B. J. Raphael, Inferring parsimonious migration histories for metastatic cancers, Nature Genetics 2 (2018) 5.

[23] Jombart, A. Cori, X. Didelot, S. Cauchemez, C. Fraser, N. Ferguson, Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data, PLoS Comput Biol 10 (1) (2014) e1003457.

[24] F. Campbell, X. Didelot, R. Fitzjohn, N. Ferguson, A. Cori, T. Jombart, outbreaker2: a modular platform for outbreak reconstruction, BMC bioinformatics 19 (11) (2018) 1–8.

[25] T. Jombart, R. Eggo, P. Dodd, F. Balloux, Reconstructing disease outbreaks from genetic data: a graph approach, Heredity 106 (2) (2011) 383–390.

[26] N. De Maio, C.-H. Wu, D. J. Wilson, Scotti: efficient reconstruction of transmission within outbreaks with the structured coalescent, PLoS computational biology 12 (9) (2016) e1005130.

[27] P. Skums, F. Mohebbi, V. Tsyvina, P. Icer, S. Ramachandran, Y. Khudyakov, Sophie: viral outbreak investigation and transmission history reconstruction in a joint phylogenetic and network theory framework, in: International Conference on Research in Computational Molecular Biology, Springer, 2022, pp. 369–370.

[28] D. Klinkenberg, J. A. Backer, X. Didelot, C. Colijn, J. Wallinga, Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks, PLoS computational biology 13 (5) (2017) e1005495.

[29] C. J. Worby, P. D. O'Neill, T. Kypraios, J. V. Robotham, D. De Angelis, E. J. Cartwright, S. J. Peacock, B. S. Cooper, Reconstructing transmission trees for communicable diseases using densely sampled genetic data, The annals of applied statistics 10 (1) (2016) 395.

[30] N. De Maio, C. J. Worby, D. J. Wilson, N. Stoesser, Bayesian reconstruction of transmission within outbreaks using genomic variants, PLoS computational biology 14 (4) (2018) e1006117.

[31] C. Wymant, M. Hall, O. Ratmann, D. Bonsall, T. Golubchik, M. de Cesare, A. Gall, M. Cornelissen, C. Fraser, T. M. P. C. STOP-HCV Consortium, T. B. Collaboration, Phyloscanner: inferring transmission from within-and between-host pathogen genetic diversity, Molecular biology and evolution 35 (3) (2017) 719–733.

[32] A. de Bernardi Schneider, C. T. Ford, R. Hostager, J. Williams, M. Cioce, Ü. V. Çatalyürek, J. O. Wertheim, D. Janies, Strainhub: A phylogenetic tool to construct pathogen transmission networks, Bioinformatics 36 (3) (2020) 945–947.

[33] X. Didelot, C. Fraser, J. Gardy, C. Colijn, Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks, Molecular biology and evolution 34 (4) (2017) 997–1007.

[34] X. Didelot, J. Gardy, C. Colijn, Bayesian inference of infectious disease transmission from whole-genome sequence data, Molecular biology and evolution 31 (7) (2014) 1869–1879.

[35] J. Carson, M. Keeling, D. Wyllie, P. Ribeca, X. Didelot, Inference of infectious disease transmission through a relaxed bottleneck using multiple genomes per host, Molecular Biology and Evolution (2024) msad288.

[36] M. D. Hall, C. Colijn, Transmission trees on a known pathogen phylogeny: Enumeration and sampling, Molecular biology and evolution 36 (6) (2019) 1333–1343.

[37] S. Sledzieski, C. Zhang, I. Mandoiu, M. S. Bansal, Treefix-tp: Phylogenetic error-correction for infectious disease transmission network inference, bioRxiv (2019) 813931.

[38] P. Skums, A. Zelikovsky, R. Singh, W. Gussler, Z. Dimitrova, S. Knyazev, I. Mandric, S. Ramachandran, D. Campo, D. Jha, et al., Quentin: reconstruction of disease transmissions from viral quasispecies genomic data, Bioinformatics 34 (1) (2017) 163–170.

[39] O. Glebova, S. Knyazev, A. Melnyk, A. Artyomenko, Y. Khudyakov, A. Zelikovsky, P. Skums, Inference of genetic relatedness between viral quasispecies from sequencing data, BMC genomics 18 (10) (2017) 918.

[40] S. L. Kosakovsky Pond, S. Weaver, A. J. Leigh Brown, J. O. Wertheim, Hiv-trace (transmission cluster engine): a tool for large scale molecular epidemiology of hiv-1 and other rapidly evolving pathogens, Molecular biology and evolution 35 (7) (2018) 1812–1819.

[41] A. G. Longmire, S. Sims, I. Rytsareva, D. S. Campo, P. Skums, Z. Dimitrova, S. Ramachandran, M. Medrzycki, H. Thai, L. Ganova-Raeva,

et al., Ghost: global hepatitis outbreak and surveillance technology, BMC genomics 18 (10) (2017) 916.

[42] E. M. Campbell, A. Boyles, A. Shankar, J. Kim, S. Knyazev, R. Cintron, W. M. Switzer, Microbetrace: retooling molecular epidemiology for rapid public health response, PLoS computational biology 17 (9) (2021) e1009300.

[43] P. Sashittal, M. El-Kebir, Sharptni: counting and sampling parsimonious transmission networks under a weak bottleneck, bioRxiv (2019) 842237.

[44] S. Dhar, C. Zhang, I. Mandoiu, M. S. Bansal, Tnet: Transmission network inference using within-host strain diversity and its application to geographical tracking of covid-19 spread, IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[45] Z. Ke, H. Vikalo, Graph-based reconstruction and analysis of disease transmission networks using viral genomic data, Journal of Computational Biology.

[46] R. J. Ypma, W. M. van Ballegooijen, J. Wallinga, Relating phylogenetic trees to transmission trees of infectious disease outbreaks, Genetics 195 (3) (2013) 1055–1062.

[47] N. Mollentze, L. H. Nel, S. Townsend, K. Le Roux, K. Hampson, D. T. Haydon, S. Soubeyrand, A bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data, Proceedings of the Royal Society of London B: Biological Sciences 281 (1782) (2014) 20133251.

[48] M. J. Morelli, G. Thébaud, J. Chadœuf, D. P. King, D. T. Haydon, S. Soubeyrand, A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data, PLoS Comput Biol 8 (11) (2012) e1002768.

[49] E. M. Cottam, G. Thébaud, J. Wadsworth, J. Gloster, L. Mansley, D. J. Paton, D. P. King, D. T. Haydon, Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus, Proceedings of the Royal Society of London B: Biological Sciences 275 (1637) (2008) 887–895.

[50] F. Campbell, A. Cori, N. Ferguson, T. Jombart, Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data, PLoS computational biology 15 (3) (2019) e1006930.

[51] M. Hall, M. Woolhouse, A. Rambaut, Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set, PLoS computational biology 11 (12) (2015) e1004613.

[52] M. Senghore, H. Read, P. Oza, S. Johnson, H. Passarelli-Araujo, B. P. Taylor, S. Ashley, A. Grey, A. Callendrello, R. Lee, et al., Inferring bacterial transmission dynamics using deep sequencing genomic surveillance data, Nature Communications 14 (1) (2023) 6397.

[53] D. S. Campo, G.-L. Xia, Z. Dimitrova, Y. Lin, J. C. Forbi, L. Ganova, Raeva, L. Punkova, S. Ramachandran, H. Thai, P. Skums, et al., Accurate genetic detection of hepatitis c virus transmissions in outbreak settings, The Journal of infectious diseases 213 (6) (2016) 957–965.

[54] J. O. Wertheim, S. L. Kosakovsky Pond, L. A. Forgione, S. R. Mehta, B. Murrell, S. Shah, D. M. Smith, K. Scheffler, L. V. Torian, Social and genetic networks of hiv-1 transmission in new york city, PLoS pathogens 13 (1) (2017) e1006000.

[55] O. Ratmann, M. K. Grabowski, M. Hall, T. Golubchik, C. Wymant, L. Abeler-Dörner, D. Bonsall, A. Hoppe, A. L. Brown, T. de Oliveira, et al., Inferring hiv-1 transmission networks and sources of epidemic spread in africa with deep-sequence phylogenetic analysis, Nature communications 10 (1) (2019) 1–13.

[56] Y. Zhang, C. Wymant, O. Laeyendecker, M. K. Grabowski, M. Hall, S. Hudelson, E. Piwowar-Manning, M. McCauley, T. Gamble, M. C. Hosseinipour, et al., Evaluation of phylogenetic methods for inferring the direction of human immunodeficiency virus (hiv) transmission: Hiv prevention trials network (hptn) 052, Clinical Infectious Diseases.

[57] S. Ramachandran, H. Thai, J. C. Forbi, R. R. Galang, Z. Dimitrova, G.-l. Xia, Y. Lin, L. T. Punkova, P. R. Pontones, J. Gentry, et al., A large hcv transmission network enabled a fast-growing hiv outbreak in rural indiana, 2015, EBioMedicine 37 (2018) 374–381.

[58] E. M. Campbell, H. Jia, A. Shankar, D. Hanson, W. Luo, S. Masciotra, S. M. Owen, A. M. Oster, R. R. Galang, M. W. Spiller, et al., Detailed transmission network analysis of a large opiate-driven outbreak of hiv infection in the united states, The Journal of infectious diseases 216 (9) (2017) 1053–1062.

[59] M. G. Jones, A. Khodaverdian, J. J. Quinn, M. M. Chan, J. A. Hussmann, R. Wang, C. Xu, J. S. Weissman, N. Yosef, Inference of single-cell phylogenies from lineage tracing data using cassiopeia, Genome biology 21 (1) (2020) 1–27.

[60] S. Kumar, A. Chroni, K. Tamura, M. Sanderford, O. Oladeinde, V. Aly, T. Vu, S. Miura, Pathfinder: Bayesian inference of clone migration histories in cancer, Bioinformatics 36 (Supplement_2) (2020) i675–i683.

[61] M. S. Roddur, S. Snir, M. El-Kebir, Inferring temporally consistent migration histories, in: 23rd International Workshop on Algorithms in Bioinformatics (WABI 2023), Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2023, pp. 9.1–9.22.

[62] S. Hessey, P. Fessas, S. Zaccaria, M. Jamal-Hanjani, C. Swanton, Insights into the metastatic cascade through research autopsies, Trends in Cancer.

[63] M. Al Bakir, A. Huebner, C. Martínez-Ruiz, K. Grigoriadis, T. B. Watkins, O. Pich, D. A. Moore, S. Veeriah, S. Ward, J. Laycock, et al., The evolution of non-small cell lung cancer metastases in tracerx, Na-

ture (2023) 1–10.

[64] H.-N. Chen, Y. Shu, F. Liao, X. Liao, H. Zhang, Y. Qin, Z. Wang, M. Luo, Q. Liu, Z. Xue, et al., Genomic evolution and diverse models of systemic metastases in colorectal cancer, Gut 71 (2) (2022) 322–332.

[65] R. Schwartz, A. A. Schäffer, The evolution of tumour phylogenetics: principles and practice, Nature Reviews Genetics 18 (4) (2017) 213–229.

[66] E. O. Romero-Severson, I. Bulla, T. Leitner, Phylogenetically resolving epidemiologic linkage, Proceedings of the National Academy of Sciences (2016) 201522930.

[67] T. Leitner, W. Fitch, The phylogenetics of known transmission histories, The evolution of HIV. Johns Hopkins University Press, Baltimore, Md (1999) 315–345.

[68] G. E. Fischer, M. K. Schaefer, B. J. Labus, L. Sands, P. Rowley, I. A. Azzam, P. Armour, Y. E. Khudyakov, Y. Lin, G. Xia, et al., Hepatitis c virus infections from unsafe injection practices at an endoscopy clinic in las vegas, nevada, 2007–2008, Clinical Infectious Diseases 51 (3) (2010) 267–273.

[69] P. Erdős, A. Rényi, et al., On the evolution of random graphs, Publ. math. inst. hung. acad. sci 5 (1) (1960) 17–60.

[70] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, science 286 (5439) (1999) 509–512.

[71] D. J. Watts, S. H. Strogatz, Collective dynamics of 'small-world'networks, nature 393 (6684) (1998) 440.

[72] C. Colijn, J. Gardy, Phylogenetic tree shapes resolve disease transmission patterns, Evolution, medicine, and public health 2014 (1) (2014) 96–108.

[73] G. E. Leventhal, R. Kouyos, T. Stadler, V. Von Wyl, S. Yerly, J. Böni, C. Cellerai, T. Klimkait, H. F. Günthard, S. Bonhoeffer, Inferring epidemic contact structure from phylogenetic trees, PLoS computational biology 8 (3) (2012) e1002413.

[74] N. B. Carnegie, Effects of contact network structure on epidemic transmission trees: implications for data required to estimate network structure, Statistics in medicine 37 (2) (2018) 236–248.

[75] B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. Wood, J. M. Daly, J. A. Mumford, E. C. Holmes, Unifying the epidemiological and evolutionary dynamics of pathogens, science 303 (5656) (2004) 327–332.

[76] E. M. Volz, K. Koelle, T. Bedford, Viral phylodynamics, PLoS computational biology 9 (3) (2013) e1002947.

[77] R. Noble, D. Burri, C. Le Sueur, J. Lemant, Y. Viossat, J. N. Kather, N. Beerenwinkel, Spatial structure governs the mode of tumour evolution, Nature ecology & evolution 6 (2) (2022) 207–217.

[78] M. A. Lewinsohn, T. Bedford, N. F. Müller, A. F. Feder, State-dependent evolutionary models reveal modes of solid tumour growth, Nature Ecology & Evolution 7 (4) (2023) 581–596.

[79] K. Robinson, N. Fyson, T. Cohen, C. Fraser, C. Colijn, How the dynamics and structure of sexual contact networks shape pathogen phylogenies, PLoS computational biology 9 (6) (2013) e1003105.

[80] L. Villandre, D. A. Stephens, A. Labbe, H. F. Günthard, R. Kouyos, T. Stadler, S. H. C. Study, et al., Assessment of overlap of phylogenetic transmission clusters and communities in simple sexual contact networks: applications to hiv-1, PloS one 11 (2) (2016) e0148459.

[81] D. Welch, Is network clustering detectable in transmission trees?, Viruses 3 (6) (2011) 659–676.

[82] F. Giardina, E. O. Romero-Severson, J. Albert, T. Britton, T. Leitner, Inference of transmission network structure from hiv phylogenetic trees, PLoS computational biology 13 (1) (2017) e1005316.

[83] R. M. McCloskey, R. H. Liang, A. F. Poon, Reconstructing contact network parameters from viral phylogenies, Virus evolution 2 (2) (2016) vew029.

[84] D. Sankoff, Minimal mutation trees of sequences, SIAM Journal on Applied Mathematics 28 (1) (1975) 35–42.

[85] W. Fitch, Towards defining the course of evolution: minimum change for a specified tree topology, Systematic Zoology 20 (1971) 406–416.

[86] D. Gusfield, Integer linear programming in computational and systems biology: an entry-level text and course, Cambridge University Press, 2019.

[87] W. M. Switzer, A. Shankar, H. Jia, S. Knyazev, F. Ambrosio, R. Kelly, H. Zheng, E. M. Campbell, R. Cintron, Y. Pan, et al., High hiv diversity, recombination, and superinfection revealed in a large outbreak among persons who inject drugs in kentucky and ohio, usa, Virus Evolution (2024) veae015.

[88] A. J. Drummond, A. Rambaut, Beast: Bayesian evolutionary analysis by sampling trees, BMC evolutionary biology 7 (1) (2007) 214.

[89] E. Kenah, T. Britton, M. E. Halloran, I. M. Longini Jr, Molecular infectious disease epidemiology: survival analysis and algorithms linking phylogenies to transmission trees, PLoS computational biology 12 (4) (2016) e1004869.

[90] P. Hell, J. Nešetril, Graphs and homomorphisms, volume 28 of Oxford Lecture Series in Mathematics and its Applications, Oxford University Press, 2004.

[91] L. Lovász, Graph minor theory, Bulletin of the American Mathematical Society 43 (1) (2006) 75–86.

[92] N. Robertson, P. D. Seymour, Graph minors. xx. wagner's conjecture, Journal of Combinatorial Theory, Series B 92 (2) (2004) 325–357.

[93] A. Brandstädt, V. B. Le, J. P. Spinrad, Graph classes: a survey, SIAM,

1999.

[94] J. Scott, P. Kuhn, A. R. Anderson, Unifying metastasis—integrating intravasation, circulation and end-organ colonization, Nature Reviews Cancer 12 (7) (2012) 445–446.

[95] P. K. Newton, J. Mason, K. Bethel, L. Bazhenova, J. Nieva, L. Norton, P. Kuhn, Spreaders and sponges define metastasis in lung cancer: a markov chain monte carlo mathematical model, Cancer research 73 (9) (2013) 2760–2769.

[96] P. Gerlee, M. Johansson, Inferring rates of metastatic dissemination using stochastic network models, PLoS Computational Biology 15 (4) (2019) e1006868.

[97] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, Y. Åberg, The web of human sexual contacts, Nature 411 (6840) (2001) 907–908.

[98] J. O. Wertheim, A. J. Leigh Brown, N. L. Hepler, S. R. Mehta, D. D. Richman, D. M. Smith, S. L. Kosakovsky Pond, The global transmission network of hiv-1, The Journal of infectious diseases 209 (2) (2014) 304–313.

[99] G. J. Hughes, E. Fearnhill, D. Dunn, S. J. Lycett, A. Rambaut, A. J. L. Brown, U. H. D. R. Collaboration, Molecular phylodynamics of the heterosexual hiv epidemic in the united kingdom, PLoS pathogens 5 (9) (2009) e1000590.

[100] C. M. Romano, I. M. G. de Carvalho-Mello, L. F. Jamal, F. L. de Melo, A. Iamarino, M. Motoki, J. R. R. Pinho, E. C. Holmes, P. M. de Andrade Zanotto, V. Consortium, et al., Social networks shape the transmission dynamics of hepatitis c virus, PLoS One 5 (6) (2010) e11170.

[101] P. Sashittal, H. Schmidt, M. M. Chan, B. J. Raphael, Startle: a star homoplasy approach for crispr-cas9 lineage tracing, bioRxiv (2022) 2022–12.

[102] C. A. Meacham, T. Duncan, The necessity of convex groups in biological classification, Systematic Botany (1987) 78–90.

[103] S. Moran, S. Snir, Convex recolorings of strings and trees: Definitions, hardness results and algorithms, Journal of Computer and System Sciences 74 (5) (2008) 850–869.

[104] B. Bollobás, Combinatorics: set systems, hypergraphs, families of vectors, and combinatorial probability, Cambridge University Press, 1986.

[105] M. C. Golumbic, Algorithmic graph theory and perfect graphs, Vol. 57, Elsevier, 2004.

[106] M. R. Garey, R. L. Graham, D. S. Johnson, D. E. Knuth, Complexity results for bandwidth minimization, SIAM Journal on Applied Mathematics 34 (3) (1978) 477–495.

[107] C. Dubey, U. Feige, W. Unger, Hardness results for approximating the bandwidth, Journal of Computer and System Sciences 77 (1) (2011) 62–90.

[108] A. Caprara, F. Malucelli, D. Pretolani, On bandwidth-2 graphs, Discrete Applied Mathematics 117 (1-3) (2002) 1–13.

[109] V. Puller, R. Neher, J. Albert, Estimating time of hiv-1 infection from next-generation sequence diversity, PLOS Computational Biology 13 (10) (2017) e1005775.

[110] M. L. Russell, C. S. Fish, S. Drescher, N. A. Cassidy, P. Chanana, S. Benki-Nugent, J. Slyker, D. Mbori-Ngacha, R. Bosire, B. Richardson, et al., Using viral sequence diversity to estimate time of hiv infection in infants, PLoS Pathogens 19 (12) (2023) e1011861.

[111] P. B. I. Baykal, J. Lara, Y. Khudyakov, A. Zelikovsky, P. Skums, Quantitative differences between intra-host hcv populations from persons with recently established and persistent infections, Virus Evolution 6 (2) (2021) veaa103.

[112] R. Diestel, Graph theory, volume 173 of, Graduate texts in mathematics (2012) 7.

[113] P. Kilpeläinen, H. Mannila, Ordered and unordered tree inclusion, SIAM Journal on Computing 24 (2) (1995) 340–356.

[114] J. Matoušek, R. Thomas, On the complexity of finding iso-and other morphisms for partial k-trees, Discrete Mathematics 108 (1-3) (1992) 343–364.

[115] C. Bron, J. Kerbosch, Algorithm 457: finding all cliques of an undirected graph, Communications of the ACM 16 (9) (1973) 575–577.

[116] L. Lovász, M. D. Plummer, Matching theory, Vol. 367, American Mathematical Soc., 2009.

[117] N. Moshiri, M. Ragonnet-Cronin, J. O. Wertheim, S. Mirarab, Favites: simultaneous simulation of transmission networks, phylogenetic trees and sequences, Bioinformatics 35 (11) (2019) 1852–1861.

[118] A. J. L. Brown, S. J. Lycett, L. Weinert, G. J. Hughes, E. Fearnhill, D. T. Dunn, Transmission network parameters estimated from hiv sequences for a nationwide epidemic, Journal of Infectious Diseases (2011) jir550.

[119] S. P. Castillo, R. A. Rebolledo, M. Arim, M. E. Hochberg, P. A. Marquet, Metastatic cells exploit their stoichiometric niche in the network of cancer ecosystems, Science Advances 9 (50) (2023) eadi7902.

[120] D. Posada, Cellcoal: coalescent simulation of single-cell sequencing samples, Molecular biology and evolution 37 (5) (2020) 1535–1542.

[121] S. Caminiti, I. Finocchi, R. Petreschi, On coding labeled trees, Theoretical computer science 382 (2) (2007) 97–108.

[122] C. Wymant, F. Blanquart, T. Golubchik, A. Gall, M. Bakker, D. Bezemer, N. J. Croucher, M. Hall, M. Hillebregt, S. H. Ong, et al., Easy and accurate reconstruction of whole hiv genomes from short-read sequence data with shiver, Virus evolution 4 (1) (2018) vey007.

[123] A. Stamatakis, Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 30 (9) (2014) 1312–1313.

[124] A. McPherson, A. Roth, E. Laks, T. Masud, A. Bashashati, A. W. Zhang, G. Ha, J. Biele, D. Yap, A. Wan, et al., Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer, Nature genetics 48 (7) (2016) 758–767.

7. Supplementary Material

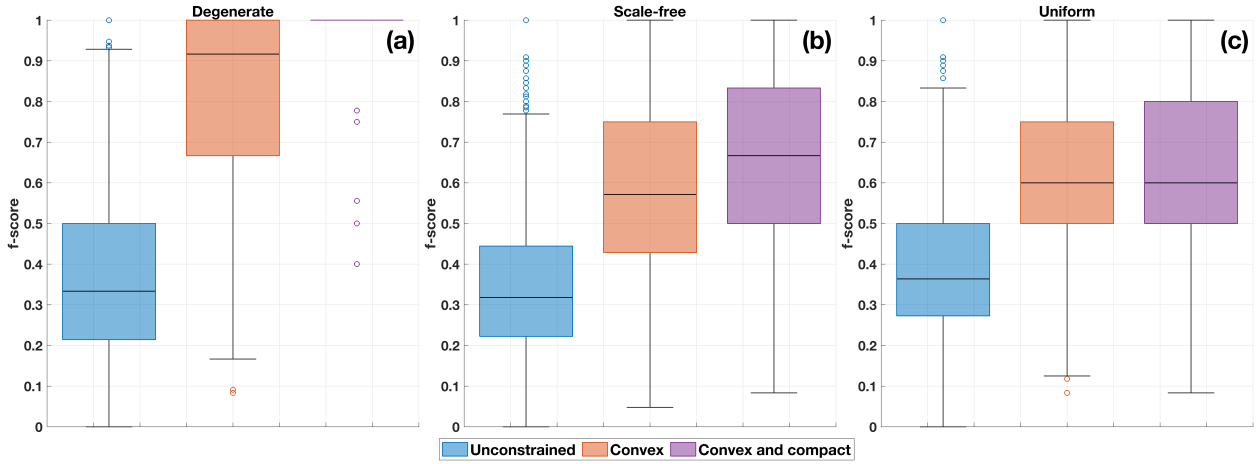

Figure 8. Comparison of sampled compatible trees with true trees under different constraints.

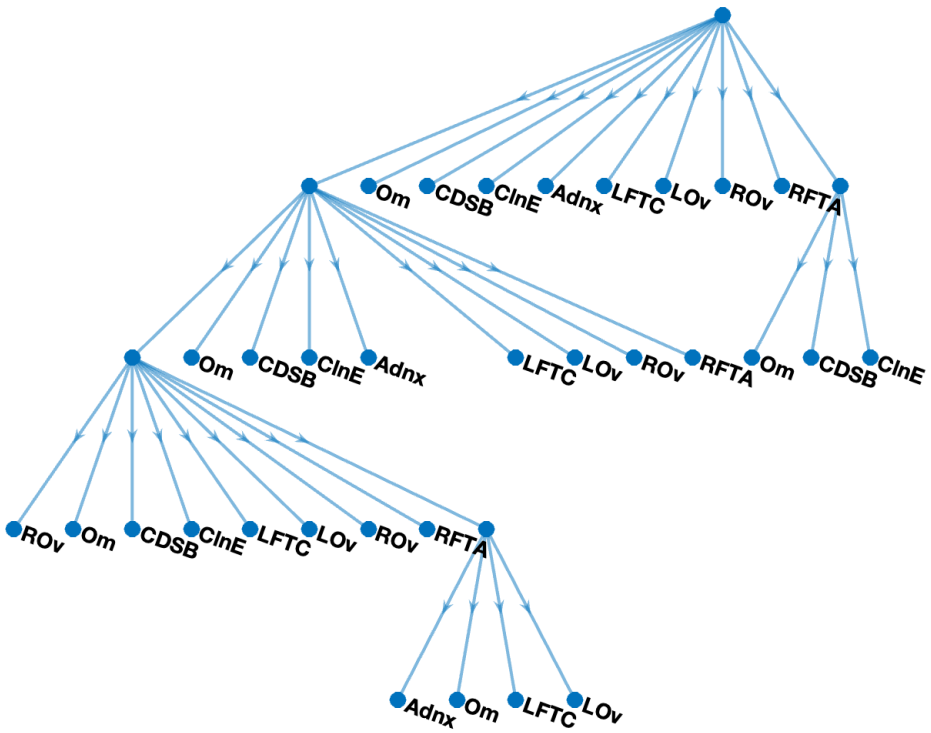

Figure 9. Clone tree for Patient 3

Algorithm 3 Homomorphism under convexity and sampling parsimony constraints

Input: phylogenetic tree Ψ with the root ρ and a candidate migration tree T
Output: a feasible homomorphism $f : \Psi \to T$ or the answer that it does not exist.

modify Ψ by contracting paths between leafs with the same label.
perform a post-order traversal of the tree Ψ.
for every node α of the post-order do
  construct a set of partial homomorphism tokens H_α.
  if α is a leaf then
    $H_\alpha \leftarrow \{(v, \emptyset) : v \in V(T)\}$;
  end if
  if α is an unlabeled internal node with children β_1 and β_2 then
    for all $(v, X) \in H_{\beta_1}$ and $(u, Y) \in H_{\beta_2}$ do
      if $v \sim u$, $u \notin X$ and $N_T(u) = Y \cup \{v\}$ then
        Add $(v, X \cup \{u\})$ to H_α
      end if
      if $v \sim u$, $v \notin Y$ and $N_T(v) = X \cup \{u\}$ then
        Add $(u, Y \cup \{v\})$ to H_α
      end if
    end for
  end if
  if α is a labeled internal node with children $\beta_1, \ldots, \beta_k$ then
    $S_i \leftarrow \{v : (v, X) \in H_{\beta_i}$ and $|X| = \deg(v) - 1\}$, $i = 1, k$;
    Generate the set Tr of transversals of the set system $(S_1, \ldots, S_k)$;
    for transversals $V = (v_1, \ldots, v_k) \in Tr$ do
      if there exists $v \sim v_1, \ldots, v_k$ then
        Add (v, V) to H_α
      end if
    end for
  end if
end for
if $H_\rho \neq \emptyset$ then
  perform a pre-order traversal of Ψ;
  for each token $t_1, \ldots t_R \in H_\rho$ do
    assign the token t_r to ρ : $AS_\rho \leftarrow t_r$.
    for every node α of the pre-order do
      $f(\alpha) \leftarrow v$, where $(v, X) = AS_\alpha$.
      if α is an unlabeled internal node with children β_1 and β_2 then
        select a vertex $u \in X$ such that $(v, X \setminus \{u\}) \in H_{\beta_1}$ and $(u, N_T(u) \setminus \{v\}) \in H_{\beta_2}$.
        $AS_{\beta_1} \leftarrow (v, X \setminus \{u\})$ and $AS_{\beta_2} \leftarrow (u, N_T(u) \setminus \{v\})$
      end if
      if α is a labeled internal node with children $\beta_1, \ldots, \beta_k$ then
        $AS_{\beta_i} \leftarrow (v_i, N_T(v_i) \setminus \{v\})$, $i = 1, \ldots, k$, where $(v, \{v_1, \ldots, v_k\}) = AS_\alpha$
      end if
    end for
    among generated homomorphisms $f_1, \ldots, f_R$, output the homomorphism f_r with the minimal $D(f_r)$.
  end for
else
  f does not exist
end if