Syst. Biol. XX(XX):1–12, 2023

https://doi.org/10.1093/sysbio/syad028
Advance Access Publication May 19, 2023

PARNAS: Objectively Selecting the Most Representative Taxa on a Phylogeny

Alexey Markin[1,*], Sanket Wagle[2], Siddhant Grover[2], Amy L. Vincent Baker[1], Oliver Eulenstein[2] and Tavis K. Anderson[1]

[1]*Virus and Prion Research Unit, National Animal Disease Center, USDA-ARS, Ames, IA 50010, USA*
[2]*Department of Computer Science, Iowa State University, Ames, IA 50011, USA*
**Correspondence to be sent to: Alexey Markin, Virus and Prion Research Unit, National Animal Disease Center, USDA-ARS, 1920 Dayton ave, Ames, IA 50010, USA; E-mail: alexey.markin@usda.gov.*

Received 13 September 2022; reviews returned 26 April 2022; accepted 28 April 2023
Associate Editor: Matthew Hahn

Abstract.—The use of next-generation sequencing technology has enabled phylogenetic studies with hundreds of thousands of taxa. Such large-scale phylogenies have become a critical component in genomic epidemiology in pathogens such as SARS-CoV-2 and influenza A virus. However, detailed phenotypic characterization of pathogens or generating a computationally tractable dataset for detailed phylogenetic analyses requires objective subsampling of taxa. To address this need, we propose *parnas*, an objective and flexible algorithm to sample and select taxa that best represent observed diversity by solving a generalized k-medoids problem on a phylogenetic tree. *parnas* solves this problem efficiently and exactly by novel optimizations and adapting algorithms from operations research. For more nuanced selections, taxa can be weighted with metadata or genetic sequence parameters, and the pool of potential representatives can be user-constrained. Motivated by influenza A virus genomic surveillance and vaccine design, *parnas* can be applied to identify representative taxa that optimally cover the diversity in a phylogeny within a specified distance radius. We demonstrated that *parnas* is more efficient and flexible than existing approaches. To demonstrate its utility, we applied *parnas* to 1) quantify SARS-CoV-2 genetic diversity over time, 2) select representative influenza A virus in swine genes derived from over 5 years of genomic surveillance data, and 3) identify gaps in H3N2 human influenza A virus vaccine coverage. We suggest that our method, through the objective selection of representatives in a phylogeny, provides criteria for quantifying genetic diversity that has application in the the rational design of multivalent vaccines and genomic epidemiology. PARNAS is available at https://github.com/flu-crew/parnas. [Diversity; epidemiology; influenza A virus; phylogeny; representative sampling; SARS-CoV-2; vaccines.]

Next-generation sequencing technologies are routinely applied to generate thousands to millions of genomes. From the beginning of the COVID-19 pandemic to present, more than 11 million SARS-CoV-2 viruses have been sequenced (Turakhia et al. 2021). Genomic epidemiology and phylogenetic analysis are essential tools to navigate this landscape of genetic data (Hill et al. 2021). However, phylogenetic trees are often insufficient to make informed intervention decisions, especially in public health. For example, it is difficult to select a few representative virus strains from within a phylogeny to include in a polyvalent vaccine, and selecting unique pathogen strains to represent variable genes or genomes for detailed *in vivo* studies on transmission and pathology requires objective criteria that are reproducible. As these assays cannot be performed on the same scale as sequencing, objective subsampling strategies are necessary. Subsampling techniques must satisfy three properties: 1) selected taxa should capture observed genetic diversity; 2) selected taxa should be representative of respective diversity groups; and 3) the method needs to be flexible to allow preferential weighting of taxa and be open to specific constraints, such as a desire to target spatial or temporal metadata due to limited availability of some strains for characterization.

The above "representative sampling" problem can be formulated as the k-medoids problem on a phylogenetic tree. That is, the goal is to select k representative taxa, such that the overall distance from all taxa to the respective closest representative is minimized. This formulation simultaneously partitions the taxa into k (phylo)genetic clusters and chooses best representatives within the clusters, thus satisfying properties (1) and (2) from above. In phylogenetics, this problem was originally considered in the context of biodiversity (Faith 1994) and later as a means to improve phylogenetic inference (Matsen et al. 2013). Independently, a related problem, called *p-medians*, was studied in the context of optimal facility location in Operations Research on general trees (Kariv, Hakimi 1979; Tamir 1996; Benkoczi, Bhattacharya 2005). To date, the best known algorithm to solve the p-medians and, consequently, the k-medoids problem on trees was developed by Tamir (1996). Here, building upon Tamir's algorithm, we designed a fast in-practice algorithm for the k-medoids problem, and implemented our algorithm in the software package *parnas* (https://github.com/flu-crew/parnas). We also developed *parnas* to solve a more general problem and satisfy property 3) from above by allowing taxa to be weighted by metadata so that those with larger weights are better represented within the selection process. Additionally, in *parnas*, users can constrain the pool of potential representatives in multiple ways and indicate previously

selected/employed representative taxa, so that new representatives will optimally complement prior selections.

In virology, the diversity of viruses continually changes through evolutionary processes such as mutation and selection, and it is necessary to monitor and identify possible emerging threats. *parnas* can be applied to identify genetically unique pathogens by allowing users to specify a "coverage radius," such that each potential representative covers all diversity within the specified radius (evolutionary distance) on the phylogenetic tree. Thus, it is possible to choose k representatives so that the total amount of covered diversity is maximized, or choose the minimum number of representatives that cover all the diversity on a tree. The coverage problem also allows the optimal elimination of redundancy from a tree if a small radius is specified. Though linking genetic diversity to phenotype is challenging (Zeller et al. 2021), distance thresholds have been applied to identify discrete genetic clades (Han et al. 2019) that are correlated with antigenic differences (Anderson et al. 2020). Therefore, finding optimal genetic coverage on a phylogenetic tree can identify groups of viruses that may be phenotypically novel and drifted from existing viruses.

In this study, we developed an algorithm that solves the k-medoids problem on a phylogenetic tree. We compared the runtime of *parnas* to prior approaches for the (unweighted) k-medoids problem on a phylogenetic tree proposed by Matsen et al. (2013) as the ADCL algorithms (ADCL-PAM and ADCL-full). Despite ADCL-PAM being an inexact heuristic, we observed *parnas* to be more scalable than both ADCL-PAM and the exact ADCL-full algorithm. We demonstrated that novel optimizations introduced in *parnas* reduce the dynamic programming table size and resulted in 85% performance improvements in terms of runtime and RAM. Further, we demonstrated that *parnas* is faster than Treemmer (Menardo et al. 2018), a popular tool for eliminating redundancy on a phylogeny, while preserving diversity. We showed that Treemmer taxon selections can be 40–50% less representative than the optimal selections by *parnas*. We applied *parnas* to an empirical influenza A virus (IAV) in swine dataset derived from the national USDA IAV in swine surveillance system (Anderson et al. 2013). Our goal was to determine a minimum number of representatives that maximized genetic coverage, and assess the duration that the representatives covered a significant amount of diversity over time. Similarly, we analyzed human H3N2 IAV diversity and incorporated phenotypic prediction models, that is, antigenic advance models (Neher et al. 2016; Hadfield et al. 2018), to identify gaps in human seasonal H3N2 vaccine coverage. Finally, we applied the *parnas* coverage algorithm to quantify the genetic diversity of SARS-CoV-2 genomes and demonstrated how the appearance of novel SARS-CoV-2 clades led to a decrease in observed virus diversity in a process similar to a "selective sweep" (Boyle et al. 2022).

Materials and Methods

Definitions

A *(phylogenetic) tree* over a *taxon set* L is a rooted and full binary tree $T = (V, E)$ with its leaves uniquely labeled (i.e., identified) by the elements of L and *edge lengths* described by the function $l\colon E \to \mathbb{R}^+$. The root of T is denoted by $\rho(T)$. For a node v in T, we denote its parent by $p(v)$ and its children by $v_{(1)}$ and $v_{(2)}$ if such nodes exist. By T_v we denote the subtree of T rooted at v.

The *distance* between two nodes u, v in T, denoted by $d(u, v)$, is defined as the sum of the edge lengths along the simple path between u and v.

To allow for unrooted and/or multifurcating trees, given such a tree, we arbitrarily root it and add the minimum number of edges with length 0 to make it strictly bifurcating and preserve pairwise taxon distances.

Representative Sampling

Our core problem is as follows: given a tree T, identify a set of its leaves that best represents its taxon diversity. Problem 1 formalizes this idea. We allow leaves in T to be weighted according to some real-valued function $w\colon L \to \mathbb{R}^+$ (the default weights are $w(l) = 1$ for all $l \in L$).

Problem 1. Given a tree T over a taxon set L, and a positive integer $k < |L|$, find

$$S := \underset{|S|=k, S\subset L}{\arg\min} \left(\sum_{v \in L} d(v, S) \cdot w(v) \right),$$

where $d(v, S) := \min_{c \in S} d(v, c)$.

That is, a set of representatives S should minimize the sum of weighted distances from all leaves to their closest representatives. This problem is a weighted version on the famous k-medoids problem (Kaufman, Rousseeuw 1990) on a phylogenetic tree.

Next, we want to account for potentially pre-selected representatives C and a coverage radius r. The coverage radius implies that a single representative covers all the diversity on the tree within the radius. Therefore, we define the function

$$d_r(v, S) := \begin{cases} 0, & \text{if } d(v, S) \le r \\ d(v, S) - r, & \text{otherwise.} \end{cases}$$

Problem 2 then generalizes Problem 1 and accounts for C and r.

Problem 2. Given a positive integer $k < |L|$, a nonnegative radius r, and a set $C \subset L$ of prior representatives, find

$$S := \underset{|S|=k, S\subset L}{\arg\min} \left(\sum_{v \in L} d_r(v, S \cup C) \cdot w(v) \right). \tag{1}$$

We also define $d_{r,C,w}(v, S) := d_r(v, S \cup C) w(v)$ to simplify notation.

Proposition 1. Problem 2 can be solved in $O(n^2k)$ time for $n = |L|$ using Tamir's algorithm for the p-medians problem on a tree.

Proof: Tamir (1996) solved the following generalization of the p-medians problem on trees:

$$\arg\min_{|S|=p, S\subset V} \left(\sum_{c\in S} c(v) + \sum_{v\in V} f_v(d(v,S)) \right), \tag{2}$$

where $c(v)$ is some cost function on the nodes of the tree and f_v is a non-negative and non-decreasing function specific to node v.

We claim that Problem 2 is a special case of generalized p-medians. We show the reduction by first assigning

$$c(v) = \begin{cases} 0, & \text{if } v \text{ is a leaf,} \\ \infty, & \text{otherwise} \end{cases}$$

to prevent selecting non-leaf nodes as representatives. Further, we set f_v to 0 for all non-leaf v. For $v \in L$, let $m_v := d(v, C)$ if $C \neq \emptyset$ and $m_v := \infty$ otherwise. Then

$$f_v(d) = \begin{cases} 0, & \text{if } \min(d, m_v) \leq r, \\ (\min(d, m_v) - r) \cdot w(v), & \text{otherwise.} \end{cases}$$

Observe that for $v \in L, d_r(v, S\cup C)\cdot w(v) = f_v(d(v,S))$. It is then apparent that under this reduction Equation (2) is equivalent to Equation (1) when $S \subset L$ and $p = k$ (recall that $S \subset L$ is enforced by cost assignments). □

Improving the Dynamic Programming Solution for Problem 2

Tamir's dynamic programming algorithm for generalized p-medians has both best-case and worst-case time and space complexity of $O(kn^2)$. This can be limiting in practice for trees with over 1000 leaves and large k. Therefore, we show how to achieve better (in-practice) time and space efficiency.

Overview of the Tamir algorithm.—We begin by reviewing the original Tamir (1996) algorithm for the p-medians problem (Equation 2). For convenience of connecting the k-medoids and p-medians problems, we set $p = k$.

For each node v_i with $1 \leq i \leq |V|$, let r_i^j be the jth closest to v_i node in V ($1 \leq j \leq |V|$). Ties are resolved, so that nodes in T_{v_i} precede nodes outside the subtree; if two tied nodes are both within T_{v_i} or both outside, then they are placed in order of the post-order traversal of T. Then $[r_i^j]$ lists for each v_i can be computed in $O(n^2)$ time (Tamir 1996).

Tamir then defines two subproblems: G and F. For $v_i \in V, 1 \leq q \leq k$, and $r = r_i^j \in V$, $G[v, q, r]$ is the optimum value of Equation 2 restricted to the subproblem in the subtree T_{v_i}, at most q representatives, and the condition that at least one of the representatives is $r_i^s \in T_{v_i}$, where $s \leq j$. The last condition implies that there is at least one representative at a distance at most $d(v, r)$ from v within the T_{v_i} subtree. Similarly, for $r = r_i^j \in T\backslash T_{v_i}$ (i.e., r outside the T_{v_i} subtree) and $0 \leq q \leq k$, $F[v, q, r]$ is the optimum value of the subproblem restricted to T_{v_i}, q representatives, and the condition that there exists a representative *outside* of T_{v_i} at a distance exactly $d(v, r)$ from v. Intuitively, r in G and F is a radius that controls how far from v we can set a representative. Let $v_0 = \rho(T)$, the solution to the overall problem is then given by $G[v_0, k, v_0^n]$. We can then establish the following relations that allow us to dynamically compute G and F subproblems (for the treatment of edge cases, see Tamir 1996). Assume that v_i has two children u and w, and r_i^j is within the T_u subtree. Then,

$$G[v_i, q, r_i^j] = \min \left(\begin{array}{l} G[v_i, q, r_i^{j-1}], \quad f_{v_i}(d(v_i, r_i^j)) \\ \\ \quad + \min\limits_{\substack{1\leq q_1\leq |V_u| \\ 0\leq q_2\leq |V_w| \\ q_1+q_2=q}} \left(G[u, q_1, r_i^j] + F[w, q_2, r_i^j] \right) \end{array} \right).$$

If r_i^j is outside the T_{v_i} subtree, then

$$G[v_i, q, r_i^j] = G[v_i, q, r_i^{j-1}]; \text{ and}$$

$$F[v_i, q, r_i^j] = \min \left(\begin{array}{l} G[v_i, q, r_i^j], \quad f_{v_i}(d(v_i, r_i^j)) \\ \\ \quad + \min\limits_{\substack{0\leq q_1\leq |V_u| \\ 0\leq q_2\leq |V_w| \\ q_1+q_2=q}} \left(F[u, q_1, r_i^j] + F[w, q_2, r_i^j] \right) \end{array} \right).$$

Reducing the size of the dynamic programming matrix.—First, note that Tamir's p-medians problem considers all nodes in a tree as potential representatives, whereas in Problem 2, we only consider leaf nodes. To save time and space, we restrict the sequences $[r_i^j]$, defined above, to the subsequences $[l_i^j]$, where l_i^j is the jth closest to v_i *leaf* in T. Subsequently, the subproblems G and F are now defined as $G[v_i, q, l_i^j]$ and $F[v_i, q, l_i^j]$, restricting the "radius" parameter in those subproblems to leaves. Lemma 1 then allows us to further reduce the redundancies in the dynamic programming algorithm by Tamir.

Lemma 1. For $q > 0$, let r^{-q} denote the q-th farthest from v leaf in T_v, and $d^{-q} := d(v, r^{-q})$. Then for any $r \in L$ with $d(v, r) > d^{-q}$,

$$G[v, q, r] = G[v, q, r^{-q}] \qquad \text{for } r \in T_v \tag{3}$$

$$F[v, q, r] = G[v, q, r^{-q}] \qquad \text{for } r \in T\backslash T_v. \tag{4}$$

Proof: As there are at least q representatives in T_v, at least one of them has to be no further then d^{-q} from v. Equation (3) then follows from the definition of G.

Further, note that for $r \in T \backslash T_v$ with $d(v,r) > d^{-q}$ and any leaf $l \in T_v$, we have $d(l,r) = d(l,v) + d(v,r)$. Let d_l be the distance from l to the closest representative in T_v. Then $d_l \leq d(l,v) + d^{-q}$. That is, any l is closer to a representative in T_v than to r. Hence, having r as a representative does not affect the subproblem in T_v. □

Lemma 1 implies that we do not need to compute subproblems $G[v,q,r]$ and $F[v,q,r]$ when $d(v,r) > d^{-q}$. This observation significantly improves the best-case complexity of the algorithm as shown in Lemma 2.

Lemma 2. The best-case time and space complexity of the dynamic programming algorithm is $O(n^2 + kn \log n)$, which is achieved when T is a perfectly balanced tree with uniform edge lengths.

In the Results section, we demonstrated that Lemma 1 proved highly effective on simulated data.

Representative Coverage

In addition to the representative sampling, we solve a coverage problem. Similarly to Problem 2, we are given (an optional) set of prior representatives C and a coverage radius r. The goal is to find a minimum set of representatives S, so that $S \cup C$ covers all leaves in the tree (within radius r).

Problem 3. Given set $C \subset L$ and non-negative $r \in \mathbb{R}$, find minimum set S, s.t.

$$\sum_{v \in L} d_r(v, S \cup C) = 0. \tag{5}$$

We show that this problem can be solved in $O(n^2)$ time using dynamic programming similar to Tamir (1996).

Recall that $[l_i^j]$ lists for each node $v_i \in T$, where l_i^j is the jth closest to v_i leaf in L, can be computed in $O(n^2)$ time. We then define $G(v_i, l_i^j)$ to be the minimum number of representatives required to cover $T_{v_i} \cap L$, so that at least one of these representatives is l_i^k, where $k \leq j$. For $l_i^j \notin T_{v_i}$, $F(v_i, l_i^j)$ is the minimum number of representatives required to cover $T_{v_i} \cap L$, while the closest representative outside of T_{v_i} is l_i^j.

Base case If v_i is a leaf, $G(v_i, l_i^j) = 1$ for all j. If $d(v_i, C) \leq r$ (i.e., v_i is covered by one of the prior centers), then $F(v_i, l_i^j) = 0$ for all j. Otherwise, $F(v_i, l_i^j) = \mathbb{1}[d(v_i, l_i^j) > r]$ for all j, where $\mathbb{1}$ is the indicator function.

Internal nodes For non-leaf v_i and $l_i^j \in T_{v_i}$, let $v_{i(1)}$ and $v_{i(2)}$ denote the children of v_i. For $l_i^j \in T_{v_i}$, WLOG assume that $l_i^j \in T_{v_{i(1)}}$, then

$$G(v_i, l_i^j) = \min\{G(v_i, l_i^{j-1}),\ G(v_{i(1)}, l_i^j) + F(v_{i(2)}, l_i^j)\},$$

where $G(v_i, l_i^{j-1}) = \infty$ for $j = 0$. Further, for $l_i^j \notin T_{v_i}$, we have

$$F(v_i, l_i^j) = \min\{G(v_i, l_i^j),\ F(v_{i(1)}, l_i^j) + F(v_{i(2)}, l_i^j))\}$$

The optimal number of representatives required to cover L is then given by $G(\rho(T), l_{\rho(T)}^{|L|})$. Consequently, the algorithm runs in $O(n^2)$ time and space.

Constraining the Pool of Representatives

We allow users to add constraints of two types to Problems 2 and 3:

1. A subset of taxa A can be chosen as representatives, but do not contribute to the objective function (i.e., excluded from summation in Equations 1 and 5).
2. A subset of taxa B contribute to the objective function, but cannot be chosen as representatives.

For Problem 2, both constraints can be satisfied by appropriately assigning functions c and f_v for the respective taxa. In particular, (1) is satisfied by assigning $f_v = 0$ for $v \in A$ and (2) is satisfied by assigning $c(v) = \infty$ for $v \in B$.

Problem 3 with the added constraints can be solved in a similar fashion. However, it is possible that inclusion of constraint (2) will make the complete coverage infeasible. In that case, *parnas* finds representatives that cover as much diversity as possible by iteratively solving Problem 2 with increasing the number of representatives k, until the objective function cannot be further improved.

Variations of the Coverage Radius

When using a maximum likelihood phylogenetic tree topology constructed from nucleotide sequences, one can be interested in re-scaling the tree so that branch lengths represent the number of amino acid substitutions. This can be needed to appropriately specify a coverage radius in relation to % amino acid sequence divergence rather than nucleotide divergence. *parnas* offers this option by using a user-specified amino acid alignment and TreeTime (Sagulenko et al. 2018) to infer ancestral amino acid substitutions and rescale the tree edges to represent the number of substitutions on that edge. This is motivated by influenza A virus in swine analysis, where a 5% amino acid divergence radius in the HA1 subunit has been used to identify genetically novel clades of viruses that are antigenically novel (Anderson et al. 2020). The tree may also be re-scaled using antigenic prediction models (Neher et al. 2016) so that vaccination coverage of predicted antigenic diversity may be quantified. In addition, we implemented a *binary* coverage problem, where *parnas* selects representatives that cover as much taxon weight as possible (instead of weighted diversity), as follows:

Problem 4. Given a positive integer $k < |L|$, a non-negative radius r, and a set $C \subset L$ of prior representatives, find

$$S = \underset{S \subset L, |S|=k}{\arg\min} \left(\sum_{v \in L} \mathbb{1}[d(v, S \cup C) > r] \cdot w(v) \right).$$

Quantifying Represented Diversity

Given a solution to Problem 2, we may ask how much diversity the selected strains represent. To help answer this question, we adapt ideas from k-means clustering, where an optimal number of clusters is often chosen based on the "variance explained" by the cluster centers—a concept related to an F-ratio statistic (Bock 1985). In particular, given a set $X \in \mathbb{R}^p$ and a partition of X into clusters $X_1, \ldots, X_k$ with cluster means $\bar{X}_1, \ldots, \bar{X}_k$, the variance explained can be calculated as

$$\frac{(\sum_{x \in X} \|x - \bar{X}\|_2^2) - (\sum_{i=1}^{k} \sum_{x \in X_i} \|x - \bar{X}_i\|_2^2)}{\sum_{x \in X} \|x - \bar{X}\|_2^2}.$$

That is, the total sum of squares (TSS) minus within-cluster sum of squares (WSS), divided by TSS.

Our computational problem is defined in terms of representatives. Hence, we define "cluster means" as their respective representatives. That is, let m_0 be a representative for entire set L (a solution to Problem 2 for $k = 1$). Further, let $S = \{m_1, \ldots, m_k\}$ be a solution to Problem 2 for $k > 1$. We partition set L into $L_1, \ldots, L_k$ according to the closest representatives m_i. Then, we say that *S represents P% of overall diversity*, where

$$P := \frac{\sum_{l \in L} d_{r,C,w}(l, m_0) - \sum_{i=1}^{k} \sum_{l \in L_i} d_{r,C,w}(l, m_i)}{\sum_{l \in L} d_{r,C,w}(l, m_0)} \times 100.$$

Note that P accounts for the coverage radius r and prior representatives C. Intuitively, each selected leaf represents the leaves closest to it. Then $d_{r,C,w}(l, m)$ is an "error"-term for leaf l. P measures the reduction in error from 1 to k representatives and hence the increase in represented (weighted) diversity.

Runtime Comparison with ADCL

We implemented *parnas* (Phylogeny-Aware RepreseNtAtive Sampling) using Python 3 and Numba (Lam et al. 2015) for just-in-time compilation of the dynamic programming algorithms. We simulated 80 birth–death trees with the birth rate $\mu = 1$ and death rate $\delta = 0.5$ and the number of leaves, n, varying between 500 and 4000 with step 500 (10 trees per fixed n). We then executed *parnas*, ADCL-full, and ADCL-PAM on each tree with fixed $k = 100$ (i.e., choosing 100 representatives). ADCL-full and ADCL-PAM are methods from (Matsen et al. 2013) which address Problem 1 and are a part of the *pplacer* suite of methods (Matsen et al. 2010). We measured the runtime of each method as well as the memory savings achieved in *parnas*.

In addition, we evaluated how the methods scale as the number of representatives (k) grows. We ran all three methods on the trees with $n = 2000$ leaves and k varying from 40 to 1000 with step 40. This study was conducted on the USDA-ARS SCINet Ceres high-performance computing cluster https://scinet.usda.gov. Each method was given a single 2.4 GHz core with 16GB of RAM per replicate.

Performance Comparison with Treemmer Strain Selections

An existing gene selection approach was introduced by Menardo et al. 2018, and we compared the runtime of *parnas* with this method. Treemmer is a randomized method without an explicit objective function. Therefore, we evaluated the representatives computed by Treemmer based on how close, on average, a taxon was to its closest representative (i.e., the k-medoids objective). For a simulated tree with 1000 leaves, we used 1) Treemmer and 2) random sampling to select $k = 10, 50, 250$ taxa and evaluated the selected taxa using the k-medoids objective. As Treemmer is randomized, for each k we performed random and Treemmer sampling 100 times to obtain a representative distribution.

Influenza A Virus Dataset Collection and Curation

Influenza A viruses (IAV) are the causative agents of an important viral respiratory disease in pigs and humans. In pigs, subtypes of H1N1, H1N2, and H3N2 are endemic in swine around the world. Despite only three circulating subtypes, the genes encoding the surface glycoproteins, hemagglutinin (HA), and neuraminidase (NA), exhibit significant diversity due to two-way transmission of IAV between swine and humans (Nelson et al. 2012). Globally, approximately 30 phylogenetic clades of H1 and H3 genes were detected worldwide in the past 3 years in swine (Anderson et al. 2020), and across the same time period 16 H1 and H3 clades were detected in the United States (Arendsee et al. 2021). We downloaded 4090 H1 and 1572 H3 IAV in swine hemagglutinin (HA) genes from the Influenza Research Database (Zhang et al. 2017) [accessed 13 June 2022]. We restricted analyses to sequences within the USDA influenza A virus in swine surveillance system (indicated by a nine digit alpha-numeric "A0" code in the strain name) collected between 2016 and 2021. Phylogenetic clade classifications were determined using the H1 swine influenza H1 HA clade tool on IRD (Anderson et al. 2016) and H3 clades were assigned using octoFLU (Chang et al. 2019). The sequences were subsequently aligned using MAFFT v7.475 (Katoh, Standley 2013) and we inferred H1 and H3 phylogenetic trees using FastTree v2.1.11 (Price et al. 2010) that were then rooted using TreeTime v.0.8.4 (Sagulenko et al. 2018). We extracted the HA1 subunit amino-acid sequences using flutile v0.13.1 (https://github.com/flu-crew/flutile) since the HA1 subunit is a major target for protective antibody immunity and divergence in HA1 may

act as a proxy for the antigenic distance between HA genes (Zeller et al. 2021).

To demonstrate how *parnas* can re-scale trees to include antigenic phenotype predictions, a process not yet accomplished for IAV in swine (Zeller et al. 2021), we also applied *parnas* to analyze the diversity of human seasonal H3N2 IAV. Using Nextstrain (Hadfield et al. 2018; Sagulenko et al. 2018) [accessed 2 March 2023], we downloaded a time-scaled phylogenetic tree covering human H3N2 viruses in a 6-month window up to 3 February 2023 that included $n = 2011$ taxa and edges annotated by the tree antigenic advance model (Neher et al. 2016). We rescaled the input tree for *parnas* so that the edge lengths represented their respective antigenic advance and applied *parnas*'s coverage algorithm to find representatives that were required to cover all diversity on the tree within a 2 antigenic unit (AU) radius. The 2 AU threshold is based upon empirical hemagglutination inhibition (HI) data (Lapedes, Farber 2001) where the distance between viruses is quantified in antigenic space and 1 AU is equivalent to a 2-fold dilution in the HI assay. Currently, an 8-fold HI difference (equivalent to 3 AU) has been sufficient to consider updating the human seasonal vaccine strain (Anderson et al. 2020), and greater than 2 AU is considered to be a biologically significant difference between viruses.

Quantifying Changes in SARS-CoV-2 Genomic Diversity

To demonstrate how *parnas* can identify changes in genomic diversity on a broader and larger dataset, we downloaded the maximum parsimony SARS-CoV-2 Audacity v1.32 tree from GISAID [accessed 2 January 2023] (Shu, McCauley 2017; Lanfear 2020). This phylogenetic tree contained 11,390,555 high-quality SARS-CoV-2 genomes as leaves and was built using UShER (Turakhia et al. 2021).

We grouped all genomes according to their date of collection into quarters, starting with 2020Q1 (January to March 2020, inclusively) and up to 2022Q4. The number of genomes within each quarter ranged from 43,000 to 2,673,000. We subsequently randomly sampled 10,000 genomes (100 times) from each quarter and extracted the corresponding subtree from the global Audacity phylogeny. We then analyzed the resulting 12×100 subtrees using *parnas*. For each subtree, we calculated the number of genomes required to cover all leaves in the subtree within a five nucleotide radius. That is, *parnas* finds the minimum number of unique genomes so that all other genomes are at most five substitutions away from a representative. We chose the five substitution threshold, because the commonly applied lineage classification pipeline, pangolin, frequently separates genetic clades by four to five substitutions across the genome (see Rambaut et al. 2020). By repeating the process 100 times for each quarter, we were able to estimate the number of unique strains per 10,000 genomes per quarter providing a robust measure of SARS-CoV-2 genomic diversity over time.

Results

PARNAS Outperforms Existing Representative Sampling Methods

Comparison with ADCL.—We compared the scalability of *parnas* with two core algorithms from Matsen et al. (2013). These algorithms solve the unweighted version of Problem 1 (representative sampling). The first algorithm, *ADCL-full*, is an exact algorithm for Problem 1, wheras the second algorithm, *ADCL-PAM*, is a heuristic and an adaptation of the classic Partition Around Medoids (PAM) k-medoids heuristic (Kaufman, Rousseeuw 1990). *parnas* solves a more general problem (Problem 2) than ADCL methods in polynomial-time. Specifically, we measured both computational and memory savings in terms of the reduction of the size of the dynamic programming (DP) table.

parnas was significantly more scalable than ADCL methods on simulated data (Fig. 1). This is particularly encouraging in the case of ADCL-PAM, as it is an inexact heuristic, whereas *parnas* is guaranteed to compute optimal representatives. Notably, *parnas* showed nearly linear runtime increase under both fixed n (number of

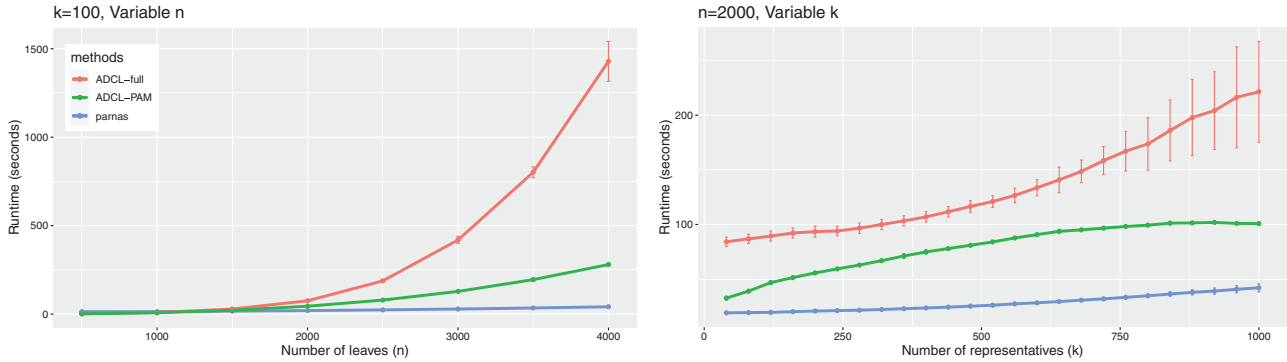

Figure 1. Runtime of *parnas* and ADCL methods with increasing number of leaves (left) and increasing number of representatives (right). The vertical error bars show standard deviation (across 10 replicates) around the mean.

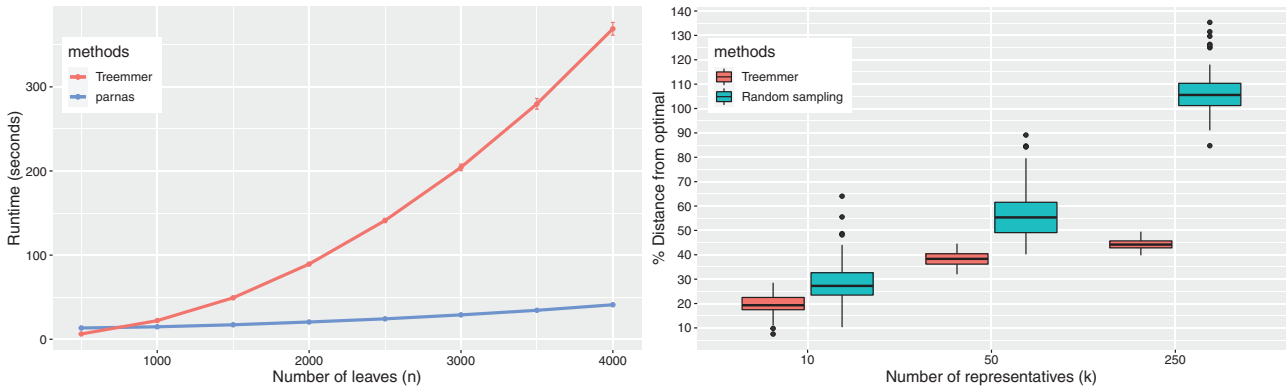

FIGURE 2. Runtime of *parnas* and Treemmer with increasing number of leaves and fixed $k = 100$ (left); and the assessment of Treemmer representatives in comparison to the optimal representatives computed by *parnas* and random sampling (right).

taxa) and k (number of representatives). We further evaluated the DP table reduction due to Lemma 1 above. For fixed $k = 100$, the average memory savings were 85% across 80 replicates, supporting the effectiveness of our approach.

Comparison with Treemmer strain selections.—We evaluated *parnas* against Treemmer in terms of the runtime and the representativity of generated selections. Treemmer uses a randomized procedure to select representative taxa across a phylogeny with a goal to reduce branch redundancy and does not explicitly solve the k-medoids problem like *parnas* and ADCL. We evaluated Treemmer in terms of the average distance of taxa to their closest Treemmer-selected representative (i.e., the k-medoids objective), and we computed how far Treemmer-selected representatives were from the optimal representatives by *parnas* in terms of that objective. *parnas* is significantly faster than Treemmer on large phylogenetic trees (Fig. 2). We observed that Treemmer representatives were generally better than randomly selected representatives, but they were 10–50% away from the optimum computed by *parnas*. That is, Treemmer-selected representatives are 10–50% further away from the taxa than the optimal representatives. The difference was particularly large for $k = 250$, where Treemmer representatives were 40–50% divergent from the optimum.

PARNAS-Selected Representatives in Influenza A Virus in Swine

A central question in the development of a polyvalent influenza vaccine is determining how many HA components are sufficient to cover observed diversity (Anderson et al. 2012). We addressed this using the last 6 years of USDA IAV in swine surveillance data, and evaluated how many H1 and H3 HA genes were required to cover the genetic diversity of IAV in swine. Though the link between genetic diversity and antigenic phenotype varies (Bolton et al. 2019), in general antigenic and genetic distances are well-correlated (Smith et al. 2004). This correlation can be further strengthened by increasing the weight of mutations in epitope sites or locations that have been identified as important in determining antigenic phenotype (Abente et al. 2016; Rajão et al. 2018; Zeller et al. 2021). Here, we applied the conservative assumption that a hemagglutinin (HA) gene would retain some cross-reactivity against another HA gene that is within 5% amino-acid divergence in the HA1 subunit (Anderson et al. 2020). A second consideration in vaccine design is the determination of when to update the components, and our study evaluated how long *parnas*-selected strains remained adequate representatives of the more contemporary viruses detected in the USDA IAV in swine surveillance system.

For the H1 subtype, we used *parnas* to select 2, 4, and 6 of the most representative HA genes for the surveillance data collected in 2016, 2017, and 2018. Subsequently, we calculated how many HA genes in each year were within 5% divergence from the *parnas* representatives, that is, for the 2016 selections, we determined how much diversity they covered in each year between 2016 and 2021. Similarly, for the H3 subtype, where fewer genetic clades cocirculate than in the H1 subtype (Arendsee et al. 2021), we used *parnas* to select 1, 2, and 3 of the most representative HA genes for 2016, 2017, and 2018. We executed *parnas* with the option to rescale the tree edges with the number of HA1 amino acid substitutions and specified a 5% divergence radius (16 amino acid substitutions).

We observed that the *parnas* selected representatives came from the most frequently detected circulating HA clades. For example, four 2017 selections from the H1 tree came from clades 1A.3.3.3 (γ), 1A.1.1 (α), 1B.2.2.1 (δ_{1a}), and 1B.2.1 (δ_2), which were the most frequently detected H1 clades in the United States that year (Arendsee et al. 2021). Similarly, selecting two representatives in the H3 tree consistently produced strains from the 1990.4.a and 2010.1 clades—the two most frequently detected H3 clades in the United States since 2015 (Zeller et al. 2018). Selecting four H1 HA genes were sufficient to cover over 70% overall diversity for

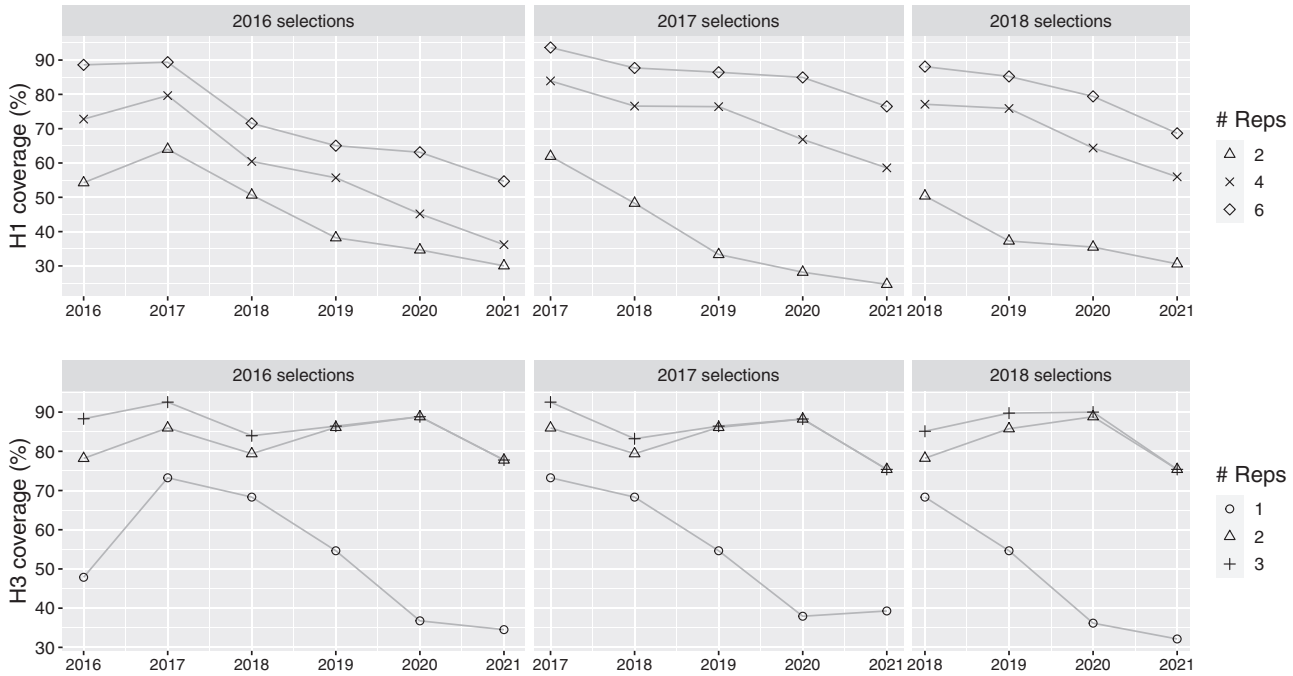

FIGURE 3. Percent of circulating influenza A virus in swine hemagglutinin genes that are within 5% divergence of the *parnas*-selected representatives. For H1 (top row), we selected 2, 4, and 6 representative strains with *parnas* for 2016, 2017, and 2018, and tracked how these representative HA genes covered H1 HA genes that circulated in the following years. Similarly, for H3 (bottom row), we selected 1, 2, and 3 representatives, and tracked how these representative HA genes covered H3 HA genes that circulated in the following years.

the first year and the subsequent year, that is, selections were adequate to cover the majority of observed diversity across 2 years (Fig. 3). Increasing the number of representative strains to six guaranteed over 85% coverage in the first 2 years, whereas decreasing the number of selections to two reduced coverage to less than 50%. For the H3 subtype, three HA genes were sufficient to cover more than 85% of diversity in the same year, and the coverage remained consistently high for each subsequent year (Fig. 3). The 2016 representatives provided close to 90% coverage until 2020 (i.e., for 5 years straight). In H1s, the decrease in coverage was more pronounced over the years.

PARNAS Identified a Gap in Human H3 Influenza A Virus Vaccine Coverage

For human H3N2 IAVs it is possible to integrate *antigenic advance* prediction models (Neher et al. 2016; Huddleston et al. 2020) contrasting the 5% threshold we applied as a proxy for antigenic difference with IAV in swine genomic data. This approach can improve the sensitivity of *parnas* to single mutations in the HA gene that may have a disproportionate impact on phenotype. We downloaded a phylogenetic tree of $n = 2011$ H3N2 viruses with edges annotated by the tree model of antigenic advance (Neher et al. 2016; Hadfield et al. 2018; Sagulenko et al. 2018). One thousand nine hundred and thirty-eight of the viruses were contemporary, that is, collected between August 2022 and January 2023. Using *parnas* we determined that three representatives were sufficient to cover the observed antigenic diversity of contemporary viruses within a 2 AU distance. The *parnas* selection of three representatives agrees with the Nextstrain genomic nomenclature that identifies three major contemporary H3N2 clades (2a, 2b, and 1a). We then specified the most recent WHO recommended vaccine strains (A/Darwin/06/2021 and A/Cambodia/e0826360/2020) as existing representatives and used *parnas* to identify whether these two strains alone were sufficient for complete coverage. This analysis (Fig. 4) shows that two of the contemporary H3N2 clades are sufficiently represented by existing vaccine strains; however, a component of observed antigenic diversity is not represented, and an additional strain from the 2b clade would be required to complete the coverage of the antigenic diversity of the observed H3 genes. Of the 804 H3 2b HA genes detected, 52 were outside the 2 AU distance and these strains could form a persistent clade of viruses into future influenza seasons.

Reduced Genetic Diversity Following the Emergence of SARS-CoV-2 Variants

We applied *parnas* to quantify the genetic diversity of SARS-CoV-2 viruses over time. We used the *parnas* coverage algorithm to identify the number of unique SARS-CoV-2 genomes—up to a five nucleotide substitution

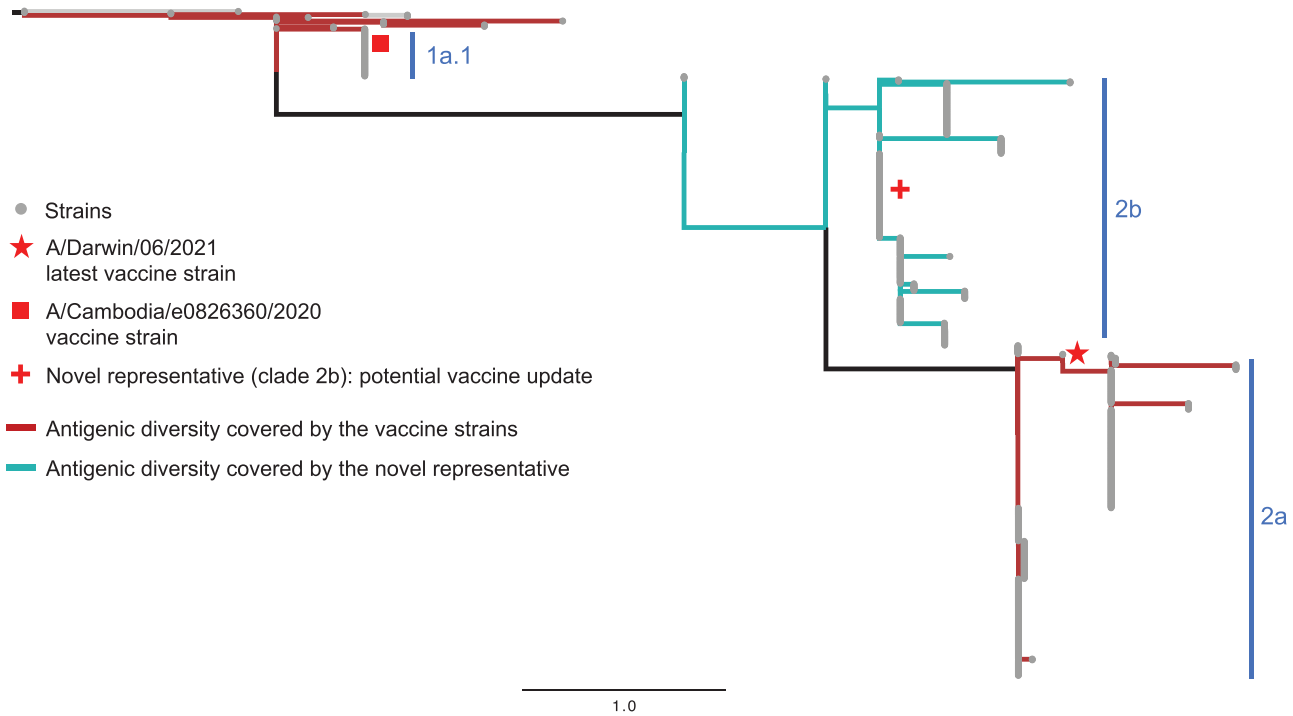

FIGURE 4. A phylogenetic tree of $n = 2011$ human H3N2 hemagglutinin genes with edge lengths representing antigenic advance (1 U of antigenic advance correlates with a 2-fold dilution in the hemagglutination inhibition assay). The *parnas* coverage algorithm identified that an additional strain from the 2b clade would ensure complete antigenic coverage of the tree, complementing the previously used vaccine antigens.

radius—that circulated globally between January 2020 and December 2022. We grouped 11,390,555 genomes from the Audacity maximum parsimony tree into yearly quarters (2020Q1 to 2022Q4), and determined the number of unique strains per 10,000 genomes within each quarter. Thus, we estimated general fluctuations in the genetic diversity of SARS-CoV-2 genomes over time (Fig. 5).

We observed that the genetic diversity of SARS-CoV-2 viruses rapidly increased in 2020 following the emergence of the virus (Van Dorp et al. 2020). The overall diversity, however, decreased in 2021Q2 and Q3, coinciding with the global spread of Alpha and Delta variant lineages. Peak diversity was reached in 2021Q4, during the later phase of the Delta wave and the emergence of the Omicron variant lineage (Fig. 5b). Later, as Omicron rapidly became a dominant clade globally, we observed a significant drop in overall genetic diversity in 2022Q1 with a gradual increase in diversity over the remainder of 2022. These data (Fig. 5) suggest that the emergence of a "successful" variant may be associated with the amount of genetic diversity in the virus population, that is, a virus that can escape vaccine and infection driven population immunity is higher when there is more genetic diversity (a sampling effect). These data also show how novel (invasive) SARS-CoV-2 variant lineages can reduce observed genetic diversity through a mechanism similar to a selective sweep (Kang et al. 2021; Boyle et al. 2022).

Discussion

Given the rapid growth of sequence data in genetic databases, representative subsampling techniques are essential for computation-intensive bioinformatics studies, objective selection of pathogen strains for phenotypic characterization, and for genomic epidemiology (Hill et al. 2021). There has been a surge in the number of tools that can parse sequence data or phylogenetic trees. One group of methods identify clusters of related sequences, for example, TreeCluster (Balaban et al. 2019) or PhyCLIP (Han et al. 2019), but these are unable to objectively select strains within the identified clusters. A second group of methods select or remove single strains: such as TARDiS (Marini et al. 2021) that can perform time-aware sampling of genetic sequences, or Treemmer (Menardo et al. 2018) that reduces taxa on a phylogeny through pruning redundant branches. These selection approaches either do not account for evolution or are not able to objectively select representatives across tens of thousands of taxa in a reasonable time. *parnas* allows researchers to identify diversity groups within their data and objectively choose representatives. Furthermore, *parnas* provides wide flexibility in constraining the pool of potential representatives through the specification of prior representatives or through the use of an optional coverage radius. Consequently, the method can achieve time-representative and geo-representative sampling by individually sampling from

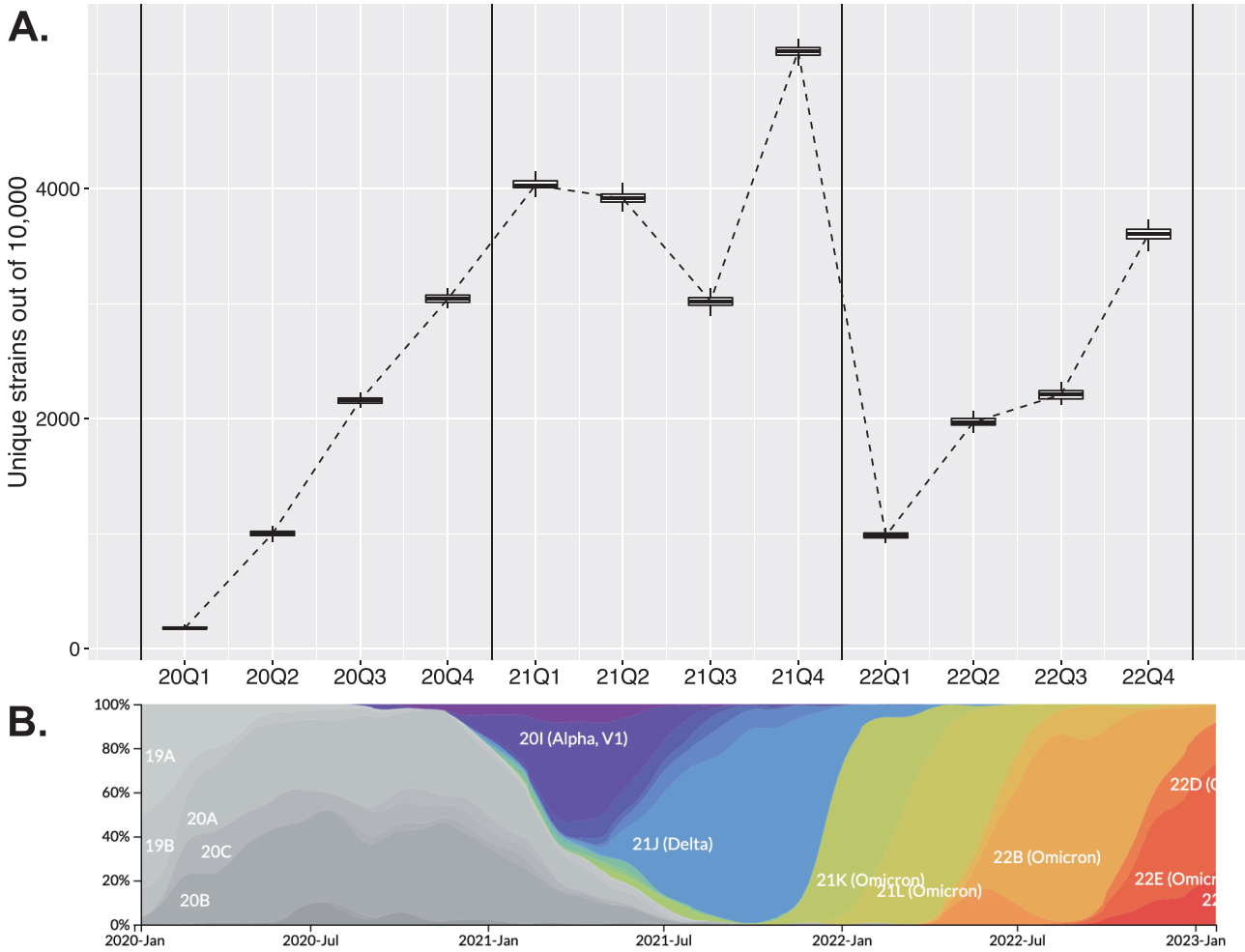

FIGURE 5. a) The number of unique strains per 10,000 genomes (up to five substitutions across the genome) over 3 years following the emergence of the SARS-CoV-2 virus: all genomes were assessed according to their date of collection and grouped into calendar quarters. b) The frequency of detection of SARS-CoV-2 genetic clades across the same time interval generated by Nextstrain (Hadfield et al. 2018) (accessed 27 January 2023). The clade names follow the Nextstrain SARS-CoV-2 clade nomenclature that is derived from the frequency of clade detection and stability.

different time-periods and geographic regions across thousands of taxa.

We demonstrated that *parnas* is faster and broader in scope than ADCL (Matsen et al. 2013). Apart from subsampling and reference strain selection applications, ADCL was suggested to be used for genotype imputation (Kang et al. 2015; Ye et al. 2019). Therefore, *parnas* can be applied in a similar manner in the genotype imputation pipelines with large reference datasets. Matsen et al. also noted that "the computational complexity class of the ADCL optimization problem is not yet clear." In this work, we resolve this question and demonstrate that *parnas* solves the ADCL optimization problem, and a significantly more general problem, in polynomial time.

A primary motivation for developing *parnas* was the objective and representative selection of IAV in swine for phenotypic characterization. To this end, we demonstrated that *parnas*, unsupervised, selects representative strains from the most frequently detected IAV sequences in swine HA clades. We showed that as few as 6 HA genes (4 H1 and 2 H3) may be sufficient for effective representation of circulating IAV in the United States for approximately 2 years. A consistent challenge in vaccine design is to determine what antigenic components are required, and *parnas* provides an objective approach to select HA genes that represent circulating diversity. Similarly, the algorithm provides a metric to determine when genetic coverage is reduced. Alternative pipelines for selection of representative strains in IAV research have previously involved 1) clustering of genes/genomes under a fixed divergence threshold and 2) selecting consensus or random strains within the clusters (cf. Jones et al. 2021). The advantage of *parnas* is that it automatically achieves the same goal as these manual approaches while also using an objective criterion

that accounts for the interaction between the selected strains.

An additional application of *parnas* is the identification of genes that are not within a prescribed radius of prior representatives. For IAV in swine, the utility is the automated identification of HA genes that are genetically, and potentially antigenically novel. There are as many as 30 genetic and antigenically distinct clades of IAV in swine globally (Anderson et al. 2020), and more than 1000 IAV in swine HA genes are sequenced every year in the United States (Arendsee et al. 2021). *parnas* provides a rapid, reproducible, and objective approach to determine which of these viruses should be characterized using *in vivo* and *in vitro* methods. Linking computational assessments of circulating viruses with antigenic characterization can provide empirical data for use in pandemic risk assessments of viruses circulating in animal hosts (e.g., Souza et al. 2022). The representative viruses identified by *parnas* can then be screened for mutations in antigenic epitopes (Bush et al. 1999; Plotkin et al. 2002; Koelle et al. 2006) relative to existing viruses and vaccines as these changes may be associated with antibody-binding and IAV fitness (Łuksza and Lässig 2014). Furthermore, we demonstrated how the input tree to *parnas* may be rescaled using phenotypic information: in our empirical human seasonal H3N2 example, these data were used to 1) identify antigenically distinct groups of viruses, 2) determine whether existing vaccine strains provided adequate coverage across all observed diversity, and 3) we were able to identify a subset of viruses that are likely antigenically distinct from other circulating H3 viruses. In general, *parnas* can provide a rational and reproducible approach for parsing genomic surveillance data and developing a prioritization of strains to be comprehensively evaluated with a goal to detect novel viruses that may impact animal and human health.

Funding

This project was funded in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. 75N93021C00015, the USDA Agricultural Research Service (5030-32000-231-000-D and 5030-32000-231-095-S), and used resources provided by the SCINet project of the USDA Agricultural Research Service (6500-00093-001-00-D). The funding sources had no role in study design, data collection, and interpretation, or the decision to submit the work for publication. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. USDA is an equal opportunity provider and employer.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments that helped improve the manuscript. The authors would like to thank Yiyan Yang for pointing us toward Treemmer. The authors gratefully acknowledge pork producers, swine veterinarians, and laboratories for participating in the USDA Influenza A Virus in Swine Surveillance System and publicly sharing sequences in NCBI GenBank.

Supplemenatry Material

Data and additional scripts are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.sbcc2fr9m.

References

Abente E.J., Santos J., Lewis N.S, Gauger P.C., Stratton J., Skepner E., Anderson T.K., Rajao D.S., Perez D.R., Vincent A.L. 2016. The molecular determinants of antibody recognition and antigenic drift in the H3 hemagglutinin of swine influenza A virus. J. Virol. 90(18):8266–8280.

Anderson T.K., Chang J., Arendsee Z.W., Venkatesh D., Souza C.K., Kimble J.B., Lewis N.S., Davis C.T., Vincent A.L. 2020. Swine influenza A viruses and the tangled relationship with humans. Cold Spring Harb. Perspect. Med. 11(3):a038737.

Anderson T.K., Laegreid William W., Cerutti F., Osorio F.A., Nelson E.A., Christopher-Hennings J., Goldberg T.L. 2012. Ranking viruses: measures of positional importance within networks-define core viruses for rational polyvalent vaccine development. Bioinformatics. 28(12):1624–1632.

Anderson T.K., Macken C.A., Lewis N.S., Scheuermann R.H., Van Reeth K., Brown I.H., Swenson S.L., Simon G., Saito T., Berhane Y., Ciacci-Zanella J., Pereda A., Davis C.T., Donis R.O., Webby R.J., Vincent A.L. 2016. A phylogeny-based global nomenclature system and automated annotation tool for H1 hemagglutinin genes from Swine influenza A viruses. mSphere. 1(6):e00275–16.

Anderson T.K., Nelson M.I., Kitikoon P., Swenson S.L., Korslund J.A., Vincent A.L. 2013. Population dynamics of cocirculating swine influenza A viruses in the United States from 2009 to 2012. Influenza Other Respir. Vir. 7(s4):42–51.

Arendsee Z.W., Chang J., Hufnagel D.E., Markin A., Janas-Martindale A., Vincent A.L., Anderson T.K. 2021. octoFLUshow: an interactive tool describing spatial and temporal trends in the genetic diversity of influenza A virus in US Swine. Microbiol. Res. Announc. 10(50):e01081-21.

Balaban M., Moshiri N., Mai U., Jia X., Mirarab S. 2019. TreeCluster: clustering biological sequences using phylogenetic trees. PloS One. 14(8):e0221068.

Benkoczi R., Bhattacharya B. 2005. A new template for solving p-median problems for trees in sub-quadratic time. In: Brodal G.S., Leonardi S., editors. Algorithms–ESA 2005. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 271–282.

Bock H.-H. 1985. On some significance tests in cluster analysis. J. Class. 2(1):77–108.

Bolton M.J., Abente E.J., Venkatesh D., Stratton J.A., Zeller M., Anderson T.K., Lewis N.S., Vincent A.L. 2019. Antigenic evolution of H3N2 influenza A viruses in swine in the United States from 2012 to 2016. Influenza and other respiratory viruses. 13(1):83–90.

Boyle L., Hletko S., Huang J., Lee J., Pallod G., Tung H.-R., Durrett R. 2022. Selective sweeps in SARS-CoV-2 variant competition. Proc. Natl. Acad. Sci. 119(47): e2213879119.

Bush R.M., Bender C.A., Subbarao K., Cox N.J., Fitch W.M. 1999. Predicting the evolution of human influenza A. Science. 286(5446): 1921–1925.

Chang J., Anderson T.K., Zeller M.A., Gauger P.C., Vincent A.L. 2019. octoFLU: automated classification for the evolutionary origin of influenza A virus gene sequences detected in US Swine. Microbiol. Res. Announ. 8, 32.

Faith D. P. 1994. Phylogenetic pattern and the quantification of organismal biodiversity. Philosophical Trans. R. Soc. Lond. Ser. B: Biol. Sci. 345(1311): 45–58.

Hadfield J., Megill C., Bell S.M., Huddleston J., Potter B., Callender C., Sagulenko P., Bedford T., Neher R.A. 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 34(23): 4121–4123.

Han A.X., Parker E., Scholer F., Maurer-Stroh S., Russell C.A. 2019. Phylogenetic clustering by linear integer programming (PhyCLIP). Mol. Biol. Evolut. 36(7): 1580–1595.

Hill V. Ruis C., Bajaj S., Pybus O.G., Kraemer M.U.G. 2021. Progress and challenges in virus genomic epidemiology. Trend. Parasitol. 37(12): 1038–1049.

Huddleston J., Barnes J.R., Rowe T., Xu X., Kondor R., Wentworth D.E., Whittaker L., Ermetal B., Daniels R.S., McCauley J.W., Fujisaki S., Nakamura K., Kishida N., Watanabe S., Hasegawa H., Barr I., Subbarao K., Barrat-Charlaix P., Neher R.A., Bedford T. 2020. Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution. Elife. 9. e60067.

Jones J.E., Le Sage V., Padovani G.H., Calderon M., Wright E.S., Lakdawala S.S. 2021. Parallel evolution between genomic segments of seasonal human influenza viruses reveals RNA–RNA relationships. Elife. 10. e66525.

Kang J.T.L., Zhang P., Zöllner S., Rosenberg N.A. 2015. Choosing subsamples for sequencing studies by minimizing the average distance to the closest leaf. Genetics. 201(2): 499–511.

Kang L., He G., Sharp A.K., Wang X., Brown A.M., Michalak P., Weger-Lucarelli J. 2021. A selective sweep in the spike gene has driven SARS-CoV-2 human adaptation. Cell. 184(17): 4392–4400.

Kariv O., Hakimi S.L. 1979. An algorithmic approach to network location problems. II: the p-median. SIAM J. Appl. Math. 37(3): 539–560.

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evolut. 30(4): 772–780.

Kaufman L., Rousseeuw P.J. 1990. Finding groups in data: an introduction to cluster analysis. 344.

Koelle K., Cobey S., Grenfell B., Pascual M. 2006. Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. Science. 314(5807): 1898–1903.

Lam S.K., Pitrou A., Seibert S. 2015. Numba: a LLVM-based Python JIT compiler. Proceedings of the second workshop on the LLVM compiler infrastructure in HPC. New York, NY: Association for Computing Machinery. (LLVM '15).

Lanfear R. 2020. A global phylogeny of hCoV-19 sequences from GISAID.

Lapedes A., Farber R. 2001. The geometry of shape space: application to influenza. J. Theor. Biol. 212(1): 57–69.

Łuksza M., Lässig M. 2014. A predictive fitness model for influenza. Nature. 507(7490): 57–61.

Marini S., Mavian C., Riva A., Prosperi M., Salemi M., Rife Magalis B. 2021. Optimizing viral genome subsampling by genetic diversity and temporal distribution (TARDiS) for phylogenetics. Bioinformatics. 38(3): 856–860.

Matsen F.A., Kodner R.B., Armbrust E. 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics. 11(1): 1–16.

Matsen IV F.A., Gallagher A., McCoy C.O. 2013. Minimizing the average distance to a closest leaf in a phylogenetic tree. Syst. Biol. 62(6): 824–836.

Menardo F., Loiseau C., Brites D., Coscolla M., Gygli S.M., Rutaihwa L.K., Trauner A., Beisel C., Borrell S., Gagneux S. 2018. Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. BMC Bioinformatics. 19(1): 1–8.

Neher R.A., Bedford T., Daniels R.S., Russell C.A., Shraiman B.I. 2016. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. Proc. Natl. Acad. Sci. 113(12): E1701–E1709.

Nelson M.I., Gramer M.R., Vincent A.L., Holmes E.C. 2012. Global transmission of influenza viruses from humans to swine. J. General Virol. 93(10): 2195–2203.

Plotkin J.B., Dushoff J., Levin S.A. 2002. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. Proc. Natl. Acad. Sci. 99(9): 6263–6268.

Price M.N., Dehal P.S., Arkin A.P. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS One. 5(3): e9490.

Rajão D.S., Anderson T.K., Kitikoon P., Stratton J., Lewis N.S., Vincent A.L. 2018. Antigenic and genetic evolution of contemporary swine H1 influenza viruses in the United States. Virology. 518: 45–54.

Rambaut A., Holmes E.C., O'Toole Á., Hill V., McCrone J.T., Ruis C., du Plessis L., Pybus O.G. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. 5(11): 1403–1407.

Sagulenko P., Puller V., Neher R.A. 2018. TreeTime: maximum-likelihood phylodynamic analysis. Virus Evolut. 4(1): vex042.

Shu Y., McCauley J. 2017. GISAID: Global initiative on sharing all influenza data–from vision to reality. Eurosurveillance. 22(13): 30494.

Smith D.J., Lapedes A.S., De Jong J.C., Bestebroer T.M., Rimmelzwaan G.F., Osterhaus A.D.M.E., Fouchier R.A.M. 2004. Mapping the antigenic and genetic evolution of influenza virus. Science. 305(5682): 371–376.

Souza C.K., Anderson T.K., Chang J., Venkatesh D., Lewis N.S., Pekosz A., Shaw-Saliba K., Rothman R.E., Chen K.-F., Vincent A.L. 2022. Antigenic distance between North American swine and human seasonal H3N2 influenza A viruses as an indication of zoonotic risk to humans. J. Virol. 96(2): e01374–21.

Tamir A. 1996. An $O(pn^2)$ algorithm for the p-median and related problems on tree graphs. Oper. Res. Lett. 19(2): 59–64.

Turakhia Y., Thornlow B., Hinrichs A.S., De Maio N., Gozashti L., Lanfear R., Haussler D., Corbett-Detig R. 2021. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nat. Genet. 53(6): 809–816.

Van Dorp L., Acman M., Richard D., Shaw L.P., Ford C.E., Ormond L., Owen C.J., Pang J., Tan C.C.S., Boshier F.A.T., Ortiz A.T., Balloux F. 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect. Genet. Evolut. 83: 104351.

Ye S., Yuan X., Huang S., Zhang H., Chen Z., Li J., Zhang X., Zhang Z. 2019. Comparison of genotype imputation strategies using a combined reference panel for chicken population. Animal. 13(6): 1119–1126.

Zeller M.A., Anderson T.K., Walia R.W., Vincent A.L., Gauger P.C. 2018. ISU FLUture: a veterinary diagnostic laboratory web-based platform to monitor the temporal genetic patterns of Influenza A virus in swine. BMC Bioinformat. 19(1): 397.

Zeller M.A., Gauger P.C., Arendsee Z.W., Souza C.K., Vincent A.L., Anderson T.K. 2021. Machine learning prediction and experimental validation of antigenic drift in H3 influenza A viruses in swine. Msphere. 6(2): e00920–20.

Zhang Y., Aevermann B.D., Anderson T.K., Burke D.F., Dauphin G., Gu Z., He S., Kumar S., Larsen C.N., Lee A.J., Li X., Macken C., Mahaffey C., Pickett B.E., Reardon B., Smith T., Stewart L., Suloway C., Sun G., Tong L., Vincent A.L., Walters B., Zaremba S., Zhao H., Zhou L., Zmasek C., Klem E.B., Scheuermann R.H. 2017. Influenza Research Database: an integrated bioinformatics resource for influenza virus research. Nucl. Acid. Res. 45(D1): D466.