bioRxiv preprint doi: https://doi.org/10.1101/2021.03.15.435416; this version posted March 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PHASTSIM: EFFICIENT SIMULATION OF SEQUENCE EVOLUTION FOR PANDEMIC-SCALE DATASETS

A PREPRINT

Nicola De Maio
European Molecular Biology Laboratory
European Bioinformatics Institute
Wellcome Genome Campus, Hinxton
Cambridgeshire, CB10 1SD, UK
demaio@ebi.ac.uk

Lukas Weilguny
European Molecular Biology Laboratory
European Bioinformatics Institute
Wellcome Genome Campus, Hinxton,
Cambridgeshire, CB10 1SD, UK
lukasw@ebi.ac.uk

Conor R. Walker
European Molecular Biology Laboratory
European Bioinformatics Institute
Wellcome Genome Campus, Hinxton,
Cambridgeshire, CB10 1SD, UK
Department of Genetics
University of Cambridge
Cambridge, CB2 3EH, UK
cwalker@ebi.ac.uk

Yatish Turakhia
Department of Biomolecular Engineering
University of California Santa Cruz
Santa Cruz, CA 95064, USA
Genomics Institute
University of California Santa Cruz
Santa Cruz, CA 95064, USA
yturakhi@ucsc.edu

Russell Corbett-Detig
Department of Biomolecular Engineering
University of California Santa Cruz
Santa Cruz, CA 95064, USA
Genomics Institute
University of California Santa Cruz
Santa Cruz, CA 95064, USA
rucorbet@ucsc.edu

Nick Goldman
European Molecular Biology Laboratory
European Bioinformatics Institute
Wellcome Genome Campus, Hinxton,
Cambridgeshire, CB10 1SD, UK
goldman@ebi.ac.uk

March 15, 2021

ABSTRACT

Sequence simulators are fundamental tools in bioinformatics, as they allow us to test data processing and inference tools, as well as being part of some inference methods. The ongoing surge in available sequence data is however testing the limits of our bioinformatics software. One example is the large number of SARS-CoV-2 genomes available, which are beyond the processing power of many methods, and simulating such large datasets is also proving difficult. Here we present a new algorithm and software for efficiently simulating sequence evolution along extremely large trees (e.g. $> 100,000$ tips) when the branches of the tree are short, as is typical in genomic epidemiology. Our algorithm is based on the Gillespie approach, and implements an efficient multi-layered search tree structure that provides high computational efficiency by taking advantage of the fact that only a small proportion of the genome is likely to mutate at each branch of the considered phylogeny. Our open source software is available from `https://github.com/NicolaDM/phastSim` and allows easy integration with other Python packages as well as a variety of evolutionary models, including new ones that we developed to more realistically model SARS-CoV-2 genome evolution.

Keywords SARS-CoV-2 · Simulation · Mutation · Selection · Evolution · Genome

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.15.435416; this version posted March 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 Introduction

Sequence evolution simulation is important for many aspects of bioinformatics [1]. Its most ubiquitous applications are for testing and comparing the performance of various essential tools (such as alignment, phylogenetic, and molecular evolution inference software, see e.g. [2, 3, 4]). However, simulating sequence evolution is also often used for testing hypotheses (e.g. [5]) and for inference, either for example through Approximate Bayesian Computation [6, 7], see e.g. [8, 9], or, more recently, using deep learning see e.g. [10, 11, 12].

Many simulators address the task of simulating gene trees, or ancestral recombination graphs, as well as simulating evolution along these trees (e.g. [13, 14, 15, 16]). Instead, here we focus on the problem of generating sequences given an input tree, as done by "phylogenetic" simulators (e.g. [17, 18, 19]). Realistic simulation of sequence evolution along a phylogenetic tree is essential, for example, for assessing and improving our methods for inference of SARS-CoV-2 phylogenies, which is a largely still open problem [20]. One important factor is the large numbers of available genome sequences for SARS-CoV-2 ($> 500,000$ in the GISAID database [21] as of February 2021). Despite this, there are currently no available simulation frameworks capable of simulating the scale and complex evolutionary features of SARS-CoV-2 and similar genome datasets. For this reason, we focus on the issue of simulating realistic substitution patterns for large datasets of closely related samples, as broadly observed in genomic epidemiology sequence data, and for arbitrarily complex substitution models; for simplicity, we initially ignore the problem of simulating indels.

We show that, indeed, sequence simulation for such large numbers of genomes is often exceedingly computationally demanding for existing software; complex evolutionary models, for example codon substitution models and rate variation, can cause significant further slow-downs. Furthermore, many existing methods do not allow the simulation of mutational patterns realistic for SARS-CoV-2, such as non-stationary and highly variable mutational processes [22, 23, 24]. We also propose a new approach to efficiently simulate the evolution of many closely related genomes along a phylogenetic tree. Our approach simulates one mutation at a time using the Gillespie method [25], and is further tailored to reduce time and memory demand by efficiently representing and storing information regarding non-mutated positions of the genome. Furthermore, we use a multi-layered search tree structure to efficiently sample mutation events along the genome even when each position has its own mutation rate, and to efficiently traverse the phylogenetic tree and avoid redundant operations. Our approach empowers extremely flexible and fine-grained evolutionary models. For example, non-stationary models are specifiable, with each nucleotide position of the genome assigned a distinct mutational profile, and each codon a distinct nonsynonymous/synonymous rate ratio.

2 Methods

We consider the problem of simulating evolution of a DNA sequence along an specified input phylogenetic tree, and under a given evolutionary model. Our simulation approach is based on the Gillespie method [25], as is typically used in molecular evolution simulators [18, 19]. We assume that each position of the genome (either nucleotide or codon) evolves independently of the others, and under a time-homogeneous substitution process; that is, the rates of evolution at each position are initially specified by the user or are sampled randomly by the simulator. We focus on the efficient simulation of sequence evolution for large phylogenetic trees with short branches: we assume that only a few mutations happen on each branch across the genome, which is typical for genomic epidemiology, and in particular for SARS-CoV-2 (needs references here).

2.1 "Vanilla" approach

If we assume that evolutionary rates are homogeneous across the genome, it is simple to use the Gillespie approach efficiently in this scenario by adopting an efficient representation of ancestral genomes in terms of differences with respect to a reference [26]. As a very simplified example, let's consider the case in which there is no selective force at play, mutation rates are constant across the genome, and all bases mutate into all other bases at the same rate (JC69 model [27] with equal nucleotide frequencies). Throughout the manuscript, we will not assume equilibrium or stationarity in sequence evolution, but instead assume that we are given a genome at the root of the phylogeny, which we then evolve down the tree according to given rates.

In this simplified "vanilla" scenario, the total mutation rate across the genome is equal to the mutation rate for one base, $3r$, times the genome length (which we assume constant), L. Starting from the root and its genome, we visit each branch of the tree one at the time in preorder traversal. For each branch of the tree, we consider its length t_b, and we recursively sample a time for the next mutation from an exponential distribution with parameter $\frac{1}{3rL}$. If the sampled time t is over t_b, we move to the next branch. Otherwise, we decrease t_b by t and we sample a mutation event. In the considered scenario, this simply means sampling one position of the genome at random (a random integer number

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.15.435416; this version posted March 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$1 \leq i \leq L$), and then a random allele b to mutate into different from the current allele at position i. Additional steps are also required to keep track of mutations which have already occurred and allow them to further mutate, for example possibly reversing if a mutated allele mutates back to the reference allele. We track each sampled mutation by adding it to a list of mutations for the current branch. It is worth noting however that there are more efficient ways to keep track of mutations that have already occurred, which we discuss in subsequent sections. A pseudocode description of the algorithm is given in Algorithm 1. So overall, the total cost of this "vanilla" algorithm is constant in genome size, and is linear in the number of tips N. It does however scale with the number of mutation events (total tree length) $M = O(Nl)$ where l is the average number of mutations per branch. The initialization step has cost $O(N)$ in order to read the phylogenetic tree, and further $O(L)$ with more complex models in order to keep track of the positions of different alleles. Performing the simulations has cost $O(M \log(N) + M^2 \log(N)/N) = O(l^2 N \log(N))$; the main cost here is to screen previous mutation events at each new mutation, and it can be significantly reduced as explained in the next section. There is a caveat however. The default output of our software phastSim is a text file where each sample name is followed by a list of differences of the simulated sample genome with respect to the reference. We also allow users to print a Newick format tree, annotated with the simulated mutation events. If, however, we want to produce a file containing the full alignment, since the alignment is of size $O(NL)$, so will also be the cost of the algorithm. For this reason, we provide the option for the user to generate a FASTA or PHYLIP alignment output, but by default we only generate the more concise version consisting of a list of differences, which usually leads to a very considerable reduction in time and memory demand.

Algorithm 1 Vanilla algorithm for one phylogenetic branch.

Here evolution on one branch is considered. t_b is initialized as the current branch length. r is the mutation rate from one nucleotide to any other nucleotide. L is genome size. ref$[i]$ is the reference allele at position i. Node is the child node of the currently considered branch.

```
Sample t (the time to next mutation event) from an exponential distribution with parameter 1/(3rL).
while t < t_b do
    Sample a random integer 0 < i ≤ L
    if i is not a position previously mutated in an ancestor of Node then
        a = ref[i].
    else
        a is the current allele for Node
    end if
    Sample a random new allele b ≠ a.
    Add mutation (i, a, b) to the list of mutations of Node.
    Update current allele for Node at position i as b. t_b = t_b − t
    Sample t from an exponential distribution with parameter 1/(3rL).
end while
```

In classical implementations of sequence evolution simulators [17], for each node of the tree we need to update each base of the genome one at the time, therefore incurring in cost $O(NL)$. Therefore, when the number of expected mutations is $M << NL$ we expect a considerable advantage in using this approach.

A considerable limitation of the above "vanilla" approach is that we assume that rates are the same across the genome, and this is hardly realistic [28, 29]. We implemented an extended version of the algorithm above which accounts for both an arbitrary nucleotide mutation model (an UNREST substitution model [30]), and for rate variation across the genome in terms of a finite number of rate categories. phastSim users can therefore specify an arbitrary nucleotide substitution model with any discrete number of rate categories. To achieve this, we extended the algorithm above to keep track of which positions of the genome have which rates. This allows us to efficiently calculate total mutation rates for each class of sites, and to efficiently sample sites within a class.

We also implemented a new model of rate variation in order to better fit the patterns of hypermutability observed in SARS-CoV-2. In this model, small proportions of sites are given a (possibly much) higher mutation rate. For any site that is sampled to be hypermutable, only one specific mutation rate (from one nucleotide to one other nucleotide) is enhanced. For example, one such site with hypermutability might have only the G→T mutation rate increased 100-fold, while all other rates at that site remain the same. This is to model the effects observed in SARS-CoV-2 which are possibly attributable to APOBEC and ROS activity (or other still unclear mechanisms) [22, 23].

However, as the number of site classes increases, and as the number of alleles increase (for example when considering codon models), the extension of the vanilla approach described above is not efficient anymore, in particular in the

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.15.435416; this version posted March 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

extreme case where each site of the genome is given different evolutionary rates. For this reason, we developed a more complex algorithm that remains efficient despite rate variation, although it sacrifices some efficiency relative to the vanilla method in the scenario of no rate variation. We allow phastSim users to choose between the vanilla approach or the more complex one, that we call "hierarchical".

2.2 "Hierarchical" approach

2.2.1 Binary search "genome" tree

We first describe the structure and algorithm that allow us to efficiently sample a mutation event along the genome, accounting for the fact that each position might have distinct mutation rates. This structure needs to be efficiently updatable following a mutation event; in fact, a mutation event changes the allele at a position of the genome, and therefore also its mutation rate. This is very similar to the problem of sampling from a categorical distribution with many elements, where the probabilities can be slightly modified at each sample [31]. A Huffman tree [32] would be close to optimal for this task, however, here we implement a binary search tree, which has a slightly higher expected cost [31] but allows us to more efficiently model blocks of contiguous nucleotides, and therefore will be more easily extendable to indels in the future.

In our search tree, each node corresponds to a contiguous block of nucleotides along the genome. The root node represents the whole genome, and contains a rate value corresponding to the global mutation rate of the whole genome. The two children of the root correspond to the first and the second half of the genome, respectively. There is no overlap between the regions considered by each child node, and their union gives the region considered by the parent node. Consequently, the sum of the rates of the children of a node is equal to the rate of the node. Given this structure, we also refer to this binary search tree as the "genome" tree. A terminal node of the genome tree corresponds to one unit of the genome, either a base or a codon, depending on the model we choose for simulations. The terminal node contains not only information about the position of the unit along the genome, but also the reference allele at this position and the mutation rates associated with it, to allow sampling of a specific mutation event at the given position/node. A graphical representation of an example genome tree is depicted in Figure 1.

Sampling a mutation time is done as in the vanilla approach: sampling from an exponential distribution with parameters determined by the total mutation rate at the root of the genome tree. Then, to sample a specific mutation event at a specific genome position, we first sample a random value uniformly in $[0, 1)$ and multiply it by the total mutation rate. Let's call this parameter R. Then, we traverse the tree from the root to the terminal node corresponding to the mutated position, which takes $\log(L)$ time. Finally, once reaching the corresponding terminal node (genome position) we choose a random mutation event affecting this position and correspondingly a new allele a for this position. An example mutation sampling is depicted in Figure 1. A pseudocode description of this algorithm is given in Algorithm 2. The cost of this approach is also only linear in the number of alleles, instead of quadratic as the classical simulation methods based on matrix exponentiation. However, in practice, the cost for a codon model can be further reduced by considering that typically a codon model only allows a maximum of 9 mutation events from any codon, so at each terminal node we only need to consider a maximum of 9 events and rates. Thanks to this, the cost of running a codon model with this approach is close to the cost of running a nucleotide model.

As mentioned before, once a mutation event is sampled, we need to modify the sampling process so that the change in allele at the mutated position is taken into account, since this change usually affects local and global mutation rates (an exception is for example when mutation rates between any two nucleotides are all equal). Modifying our genome tree following a mutation event is both simple and efficient: we simply need to modify the rates and allele at the mutated terminal node, and then update the rate of all ancestors of this terminal node accordingly. Algorithm 2 for example describes how to sample a mutation event from a genome tree as well as how to update the genome tree accordingly. Again, this can be done in $\log(L)$ time for each new mutation event sampled. However, while this is efficient for simulating evolution along a temporal line, that is, along a single branch of the phylogeny descendant from the root, it becomes inefficient for simulating evolution along a phylogenetic tree. This is because, if we modify the tree, then we cannot use it for sibling nodes. In other words, when we reach a split in the phylogenetic tree, and we have two children of the same phylogenetic node, we need to pass the same genome tree of the phylogenetic parent node to both phylogenetic children. However, we can't only pass a pointer to the same tree to both children, because evolving along one branch leading to one sibling would modify the genome tree also for the other sibling. If we take the approach of duplicating the genome tree at each phylogenetic split, we end up with a cost $O(NL)$, which we are trying to avoid. For this reason, we devise an alternative, hierarchical, multi-layered approach to evolving a genome tree.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.15.435416; this version posted March 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

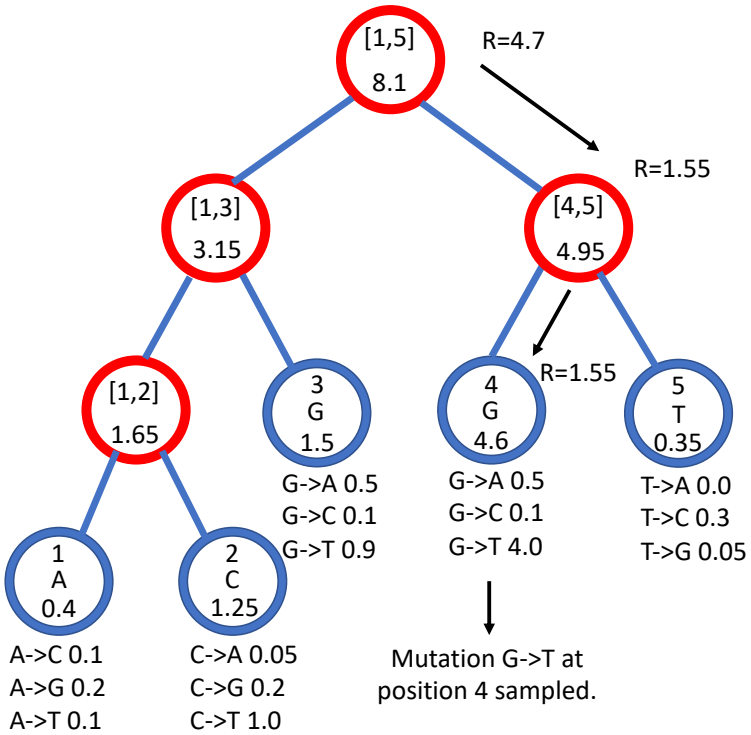

Figure 1: **Example genome tree and genome tree search**. An example genome tree for ancestral genome ACGGT. Blue nodes are terminal and red nodes are internal. Inside each node we represent on top the genome positions represented by the node; at the center inside terminal nodes we show the allele of the node; at the bottom of nodes is their total rate. Under each terminal node we show the example relevant mutation rates. Black arrow represents an example sampling of one mutation event. A parameter R is assigned an initial random number sampled uniformly between 0 and the total rate 8.1, in this case it is $R = 4.7$. As we move downward, the value of R can decrease, and it determines which site will mutate and how. Here, an initial $R = 4.7$ results in the sampling of a G→T mutation at genome position 4.

2.2.2 Hierarchical, multi-layer evolving genome tree

In order to use our genome tree structure to sample mutations along a phylogenetic tree, we add a further "vertical" dimension to it. At each branch of the phylogenetic tree, instead of modifying a genome tree, we take the approach of building on top of it, without modifying the staring genome tree nodes, so that the original genome tree is not lost. By doing this, we can effectively adapt a (multi-layer) genome tree (as new mutations are sampled) without losing the original genome tree. This means that when we need to pass a genome tree to two children of a phylogenetic node, we don't duplicate the genome tree structure, and instead we simply remove (de-allocate, or ignore) the additions to the genome tree made by first child node, and pass the original genome tree structure to the second child node. A graphical representation of an example multi-layer genome tree is given in Figure 2.

We start with a genome tree at the phylogenetic root node; additional nodes are then added at further layers. A genome tree layer n represents the genome nodes specific to a particular depth of the phylogenetic tree; phylogenetic nodes closer to the root (in terms of number of branches that need to be traversed from the root) will be associated with a lower n, and those more distant from the root with higher n. All the initial nodes of the original genome tree belong to layer 0, the one corresponding to the phylogenetic root. Then, as we move from the phylogenetic root to its first child, we add nodes to the tree in layer 1, representing the consequences of mutation events happening along the branch between the phylogenetic root and the first child. Nodes in layer 0 only point to nodes in layer 0, and never to nodes in other layers. More generally, nodes in layer n only point to nodes in layers $m \leq n$. Every time the multi-layer genome tree is passed from phylogenetic parent (layer n) to child (layer $n + 1$), new nodes are added to the corresponding layer $(n + 1)$ if mutation events occur on the corresponding phylogenetic branch.

We traverse the phylogenetic tree in preorder traversal, so, starting from the root, we move to the first child, to which we pass the initial genome tree, add new layers, then do the same for this child's children. For each new mutation occurrring

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.15.435416; this version posted March 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

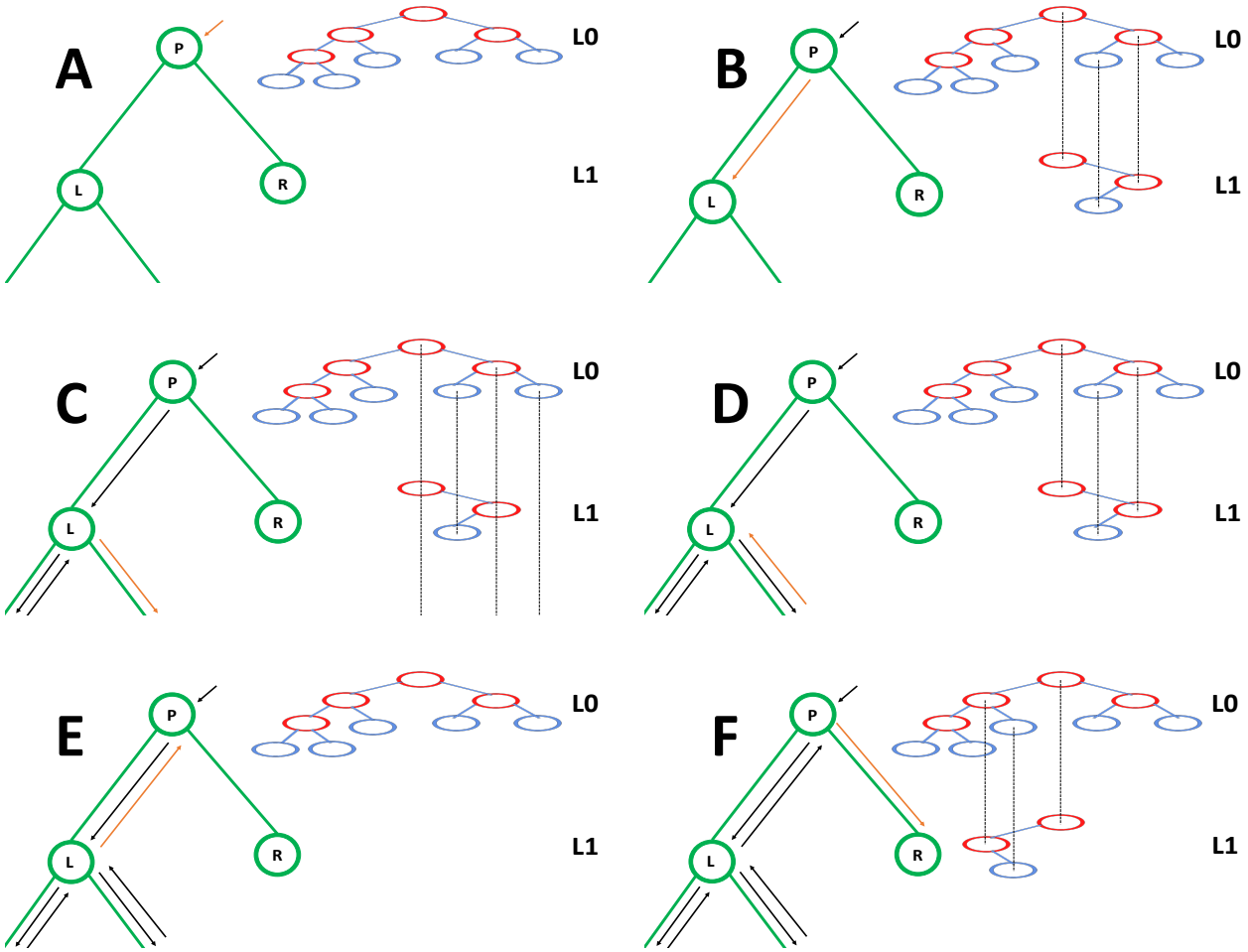

Figure 2: **Example of multi-layer genome tree and its evolution**. We track the evolution of the multi-layer genome tree starting from the genome tree of Figure 1. Colors for the genome tree are the same as in Figure 1. In green, on the left side, we show the phylogenetic tree with three nodes ("P" for parent and "L" and "R" for left and right node). "L" has further descendants, but we don't show the full tree. The orange arrow along the phylogenetic tree shows the current step of the preorder traversal being considered. Black arrows show past steps. Vertical dashed black lines in the multi-layer genome tree connect nodes that represent the same portions of the genome but that are in different layers. "L0" stands for "Layer 0" and "L1" for "Layer 1". **A** At the phylogenetic root "P" we initialize the genome tree for layer 0. **B** As we move to child "L", a new mutation is sampled (as in Figure 1) and 3 corresponding genome nodes are created in layer 1. These nodes correspond to the nodes in the original genome tree whose rate is affected by the new mutation. **C** As we traverse the subtree of the descendants of L, new nodes and mutations might be added in the layers below. **D** We are finished traversing the subtree of the descendants of L, and we return to L, at which point all nodes in layer below 1 have either been removed or are now irrelevant. **E** We return to P, at which point the genome tree nodes previously added layer 1 are also ignored or deleted. **F** We move from P to R, and in doing so new mutation event might be sampled and the corresponding genome nodes might be added to layer 1 (new genome tree nodes corresponding to 1 new mutation are shown in the new layer 1).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.15.435416; this version posted March 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Algorithm 2 Sampling of a mutation event along a genome tree, and updating the genome tree.

We assume we are given node Root, the root of a genome tree structure. For any given genome node, Node, Node.rate represents the corresponding total subtree rate; Node.children represents the list of its child nodes (for now, there are always 2 of them for internal nodes, and 0 for terminal nodes, but in future extensions this might change in order to allow efficient simulation of indels); Node.rates represents the 2-dimensional matrix of mutation rates (from any allele to any other allele) for a terminal node, and for simplicity here we assume that a rate of an allele into itself (a matrix diagonal entry) is 0; Node.parent is its parent node; Node.allele is the current allele for the node. The list "mutations" is used to record the mutation events happening on the considered phylogenetic branch. The function "sample(rates)" samples a rate from a list proportional to its value. We are also given random number $0 \leq R <$Root.rate for which we want to sample the corresponding mutation event.

```
Node=Root
while Node is not terminal do
  for Child in Node.children do
    if Child.rate< R then
      R = R−Child.rate
    else
      Node.rate=Node.rate−Child.rate
      Node=Child
      break
    end if
  end for
end while
oldAllele=Node.allele
Node.allele=sample(Node.rates[oldAllele])
mutations.append([Node.position,oldAllele,Node.allele])
Node.rate= ∑_b Node.rates[Node.allele][b]
NewRate=Node.rate
while Node.parent is not null do
  Node=Node.parent
  Node.rate=Node.rate+NewRate
  NewRate=Node.rate
end while
```

on this phylogenetic branch connecting the root and its first child, we traverse the genome tree, and every time we would modify the genome tree (to update the mutation rates following a change of allele at a position) we instead create new genome tree nodes to a child layer. Once we have traversed the whole phylogenetic subtree of the first child of the root, we have to move to second child of the root. This operation does not incur the cost of duplicating any part of the genome tree, as we only need to pass to the second child the pointer to the root of layer 0 of the hierarchical genome tree. Similarly, at any internal phylogenetic node at layer n, to both children we pass the pointer to the root of layer n of the genome tree. The only additional step which might be required is de-allocation of nodes in layer $n+1$ as we move from one node to its sibling (thanks to our preorder traversal, the nodes currently in and below this layer will not be used again), but this step at most only slows simulations by a small constant factor.

At the start of the simulations for each branch, moving from layer n to $n+1$, we first create a new genome root node for layer $n+1$. This root initially points to the same children as the genome root at layer n, and it also has the same total rate. After creating a new layer root, we sample mutation events for the current phylogenetic branch. To sample mutations, we follow the binary search tree determined by the root of layer $n+1$. As a new mutation event is picked, we either create new layer $n+1$ nodes, or modify existing layer $n+1$ nodes. When sampling a new mutation, every time we reach a node in the genome tree, we either modify the rate of the node, if it's in layer $n+1$, or we create a new layer $n+1$ node, if the original node was in a different layer. The new node is given at first the same children as the original node. When a terminal node is reached, we calculate its new rates (unless they have already been created before for some other node in the phylogenetic tree, in which case we just retrieve them from a dictionary) and total rate, and we pass the new total rate to its parent node, which uses it to update its own total rate, and so on. In total, the cost of sampling a new mutation event and updating the multi-layered structure is $O(\log(L))$. A sketch of the mutation sampling process and multi-layer genome tree update is given in Algorithm 3. The total cost of the algorithm is then $O(L + N + M\log(L))$, where the addendum L is due to the initial creation of the layer 0 genome tree, and N is due to the tree traversing process. In a scenario like SARS-CoV-2 genomic epidemiology, this can lead to dramatic reduction

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.15.435416; this version posted March 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

in computational demand compared to the standard $O(NL)$ in the field, since M appears typically not distant from N [23, 22, 24]. We give a summary of the global hierarchical method in Algorithm 4.

Algorithm 3 SampleMutation(Node,Layer): Sampling of a mutation event along a multi-layer genome tree.

This function is initially run on the root node Root of a genome tree layer. Parameters are as in Algorithm 2. Node.layer represents the layer of the considered node. While below we simplify a few details, in reality we don't recalculate rates at every mutation, but we only calculate the first time they are needed, and then store them in dictionaries.

```
if Node.layer≠Layer then
    create a new node NewNode copy of Node

    NewNode.layer=Layer
    Node=NewNode
end if
if Node is terminal then
    sample mutation event from Node.allele using Node.rates and R.
    expand if needed NewNode.rates, and update Node.rate and Node.allele
    return Node
else
    for c in Length(Node.children) do
        Child=Node.children[c]
        if Child.rate< R then
            R = R−Child.rate
        else
            Node.rate=Node.rate−Child.rate
            NewChild=SampleMutation(Child,Layer) {note that SampleMutation is the current function, so this function
            is called recursively}
            Node.children[c]=NewChild
            Node.rate=Node.rate+NewChild.rate
            return Node
        end if
    end for
end if
```

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.15.435416; this version posted March 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Algorithm 4 SimulatePhyloNode(Node,GenomeNode,Layer): Hierarchical algorithm for simulating sequence evolution along a branch of the phylogenetic tree.

Here evolution on a branch of the phylogenetic tree is considered. The branch is passed through Node, which represents the child node of the branch. The branch length is Node.length. Simulation of the whole phylogeny is performed by calling SimulatePhyloNode(Root,GenomeRoot,0), where Root is the root of the phylogenetic tree (we assume Root.length=0) and GenomeRoot is the root of the initial genome tree for layer 0. This layer 0 genome tree is created by considering the genome of the phylogenetic root, which is typically either sampled at random or read from a reference genome.

Sample t (the time to next mutation event) from an exponential distribution with parameter $1/$GenomeNode.rate
CurrentTime$= t$
while CurrentTime<Node.length **do**
  Sample a random uniform vaule $0 \leq R <$GenomeNode.rate
  GenomeNode=SampleMutation(GenomeNode,Layer) {note that this is the function defined by Algorithm 3}
  Sample t (the time to next mutation event) from an exponential distribution with parameter $1/$GenomeNode.rate
  CurrentTime=CurrentTime+t
end while
for Child in Node.children **do**
  run SimulatePhyloNode(Node,GenomeNode,Layer+1) {note that this function is the one defined in the current Algorithm, which is therefore recursive in nature}
end for
if needed, de-allocate all nodes of layer Layer from GenomeNode down to its descendants. {Because gnome nodes with layer Layer are descendant from nodes with layer Layer, we do not need to explore the whole multi-layer genome tree, but only its layer Layer.}

2.3 Further details of the implementation

In this section we describe several features of our software implementation. See Table 1 for a comparison with some other popular sequence evolution simulators.

	phastSim	Seq-Gen [17]	INDELible [18]	pyvolve [33]
Codon models	Yes	No	Yes	Yes
Non-stationary models	Yes	No	Yes	Yes
Indels	No	No	Yes	No

Table 1: Evolutionary models implemented by different simulators

2.3.1 Substitution models

Thanks to our algorithm, we can allow any substitution model without incurring a dramatic increase in computational demand, and without risking numerical instability (which can sometimes be a problem with classical matrix exponentiation approaches). In fact, we allow the most general nucleotide substitution model, UNREST [30]. By default mutation rates are assumed to be those estimated from SARS-CoV-2 [22], however, any rates can be specified by the user, including those of simpler, more restricted models (e.g. JC [27], HKY [34], or GTR [6]).

We also implemented codon models, which, with our hierarchical approach, come at only a small additional computational demand compared to nucleotide models. For the definition of the substitution rates of codon models, we use an extension of the GY94 [35] model. We separately model the nucleotide mutation process and the amino acid selection one. Unlike GY94 (which assumes an HKY nucleotide mutation process), we allow any general nucleotide mutation process as defined by an UNREST matrix. Then, nonsynonymous mutations rates are modified by a single factor ω (see next section for variation of ω across codons). Under this model, a substitution from codon c_1 to codon c_2 therefore has rate:

$$r_{c_1 \to c_2} = \begin{cases} m_{n_1 \to n_2}\,, & \text{if } c_1 \text{ and } c_2 \text{ are synonymous and differ only by nucleotides } n_1 \text{ and } n_2 \text{ at a position ,} \\ \omega m_{n_1 \to n_2}\,, & \text{if } c_1 \text{ and } c_2 \text{ are non-synonymous and differ only by nucleotides } n_1 \text{ and } n_2 \text{ at a position ,} \\ 0\,, & \text{if } c_1 \text{ and } c_2 \text{ differ by more than one nucleotide ,} \end{cases} \tag{1}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.15.435416; this version posted March 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

where $m_{n_1 \to n_2}$ is the mutation rate from nucleotide n_1 to nucleotide n_2.

We don't allow, at this stage, instantaneous multi-nucleotide mutation events, or amino acid substitution models, but we plan to address them in future extensions.

2.3.2 Models of rate variation

We consider four types of variation in rates across the genome. These types can be used in combination, or separately, as required.

The first type of variation is changes in the position-specific mutation rate across the genome. Every nucleotide position i in the genome (even using a codon model) is assigned its own mutation rate scaling factor γ_i. This means that, at position i, the mutation rate from any nucleotide n_1 to any other nucleotide n_2 becomes $\gamma_i m_{n_1 \to n_2}$. We allow two ways to sample values of γ_i for each i. One way is to sample them from a continuous Gamma distribution with parameters $\Gamma(\alpha, \alpha)$, with α specified by the user; this results in each genome position having a distinct γ_i. Alternatively, we allow the definition of district categories, with a finite number of categories, each with its own frequency and γ rate.

The second type of variation we model is variation in ω, with codon position i across the genome being given its own ω_i. As with γ_i, values of ω_i can either be sampled from a continuous Gamma or a finite categorical distribution.

Lastly, to accommodate the strong variation in mutation rates observed in SARS-CoV-2 [23, 22], and attributable to APOBEC, ADAR, or ROS activity, we introduce a new model of rate variation. This model allows, for a certain position, to have one specific mutation rate (from one specific nucleotide to another specific nucleotide) enhanced by a certain amount μ. In this case we only allow a categorical distribution, with the first category having no enhancement ($\mu = 1$) and the other categories having $\mu > 1$. For any nucleotide position i that is assigned a hypermutable category and therefore has $\mu_i > 1$, we then sample uniformly a start nucleotide n_s and a destination nucleotide n_d. The mutation rates $m^i_{n_1 \to n_2}$ for position i then become:

$$m^i_{n_1 \to n_2} = \begin{cases} \gamma_i m_{n_1 \to n_2} \ , & \text{if } n_1 \neq n_s \text{ or } n_2 \neq n_d \ , \\ \gamma_i \mu_i m_{n_1 \to n_2} \ , & \text{otherwise} \ . \end{cases} \tag{2}$$

2.3.3 Rate normalization

We assume that, for the given input phylogenetic tress, branch lengths represent expected numbers of substitutions per nucleotide - no matter if a nucleotide or a codon model is used. However, as mutations accumulate across the phylogeny, the total mutation rate of the genome might slightly change. This is because we allow substitution models that are not reversible and not at equilibrium. Therefore, we always consider the mutation rates at the root genome to normalize the mutation rates. This means that while branch lengths near the root accurately represent the expected numbers of substitutions per nucleotide, lower down the tree this might slightly change.

2.3.4 Output formats

As default, our software creates an output file where it stores information about which genome position evolved under which rate. It also creates a file where each tip name is listed together with the mutations it contains that distinguish its genome from the reference genome. In scenarios similar to SARS-CoV-2 datasets (where each genome is very similar to the reference), this format requires much less space and time to generate than FASTA or PHYLIP formats (see Section 2.1).

An optional output format that our software can create is a tree in Newick format, where each branch of the input phylogeny is annotated with a list of mutation events that occurred on that branch. This format is richer than the others, as it provides information regarding each mutation event, even those that might be over-written by other mutations at the same position; it is also more efficient than multiple sequence alignment formats in the scenario of short branch lengths considered here.

Finally, we also allow the creation of FASTA and PHYLIP format alignment output. However, note that the creation of these files costs $O(NL)$ in time and space.

2.3.5 Python package

Our software phastSim is implemented as a Python package, and can be found at `https://github.com/NicolaDM/phastSim` or `https://pypi.org/project/phastSim/`. phastSim uses the ETE3 library [36] to robustly read input trees in different variants of the Newick format.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.15.435416; this version posted March 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3 Results

3.1 SARS-CoV-2 scenario

To assess the performance of our approach compared to existing evolutionary simulators, we consider different scenarios typical for genomic epidemiology. First, we consider the simulation of a scenario similar to SARS-CoV-2 evolution. We simulate trees with a custom script and under a Yule process with birth rate equal to genome length (29,903), so to have in the order of one mutation per branch. Evolution is simulated under an UNREST model [30] with rates inferred from SARS-CoV-2 data [22] where possible (for phastSim, pyvolve [33] and INDELible [18]) and a GTR model [6] otherwise (for Seq-Gen [17]). We run INDELible both using method 1 ("INDELible-m1", which uses matrix exponentiation to model substitutions) and method 2 ("INDELible-m2", which instead uses the Gillespie approach for the same task). For now we ignore sequence rate variation. While phastSim and pyvolve are both Python implementations, therefore sharing similar benefits (high compatibility with other packages and ease of extensions) and draw-backs (reduced efficiency compared to some other languages), we see that the two approaches have dramatically different computational demands (Figure 3): simulating 50 sequences under pyvolve requires on average more time than simulating 500,000 in phastSim. We can also see that INDELible-m2 is marginally more efficient than INDELible-m1 in this scenario, due to the low number of mutations per branch. However, while phastSim and INDELible-m2 are both similar Gillespie approaches, simulating 5,000 sequences with INDELible-m2 requires slightly more time than simulating 500,000 sequences in phastSim (Figure 3), despite the fact that INDELible is coded in C++. Seq-Gen appears to be very efficient, but it's still more than one order of magnitude slower than phastSim on large phylogenetic trees in this scenario. Also note that, for large trees considered here, we can reduce computational demand in phastSim by more than 5-fold by not producing a FASTA output alignment; this way we can also save very significant amounts of memory demand. Regarding small trees ($< 10^4$ tips) most of the demand in phastSim is associated with initializing the simulations (loading packages and initializing the genome tree structure); these initialization costs do not depend on tree size, and instead depend on genome size, and they are why phastSim is relatively less efficient on small trees. If simulation on small trees are indeed of interest, these initialization costs could be reduced by re-using the same genome tree structure over multiple replicates, or, in the case of simple evolutionary models, by using our "vanilla" simulation approach.

3.2 Bacterial scenario

To demonstrate a scenario in which we are interested in simulating bacterial genome evolution within one outbreak, we use the *E. Coli* reference genome (`https://www.ncbi.nlm.nih.gov/nuccore/U00096.3` [37], 4,641,652 nucleotides) as our root genome sequence. For now we still focus on the simple scenario of a nucleotide model without rate variation. We again assume a scenario typical for genomic epidemiology, that the birth rate of the simulated tree is equal to the genome length. The number of mutations simulated is therefore comparable to the number of branches in the tree.

As genome size increases, time and memory demand of traditional simulators is expected to grow linearly. Indeed, we now see that Seq-Gen takes considerably more time to simulate the same number of genomes than in the SARS-CoV-2 scenario (Figure 4). phastSim also has an increased computational demand, but only in terms of the initial step of generating an initial genome tree. This initial cost is linear with respect to genome length, but does not increase with the number of samples or with the number of mutations simulated. In total, in this scenario phastSim can simulate sequence evolution along trees with more than 1000 times more samples than Seq-Gen. A further reduction in computational demand, in particular in terms of the initial cost of generating a genome tree, can be obtained by using the "vanilla" algorithm (Figure 4), which however comes at the cost of narrowing the choice of evolutionary models to less complex ones.

3.3 Evolutionary model

One of the advantages of the approach we present here is that simulating evolution under increasingly complex models comes at almost no additional computational cost (Figure 5). It can be seen, for example, that INDELible-m1 and Seq-Gen incur a significantly higher cost when using a continuous variation in mutation rate, and that INDELible-m2 is more demanding when simulating discrete rate categories. Surprisingly, running INDELible with a codon model appears to come with no additional computational demand, similarly to phastSim (Figure 5). For these comparisons we have considered the SARS-CoV-2 simulation scenario.

3.4 Branch lengths

Probably the main limiting factor in the applicability of the approach presented here are tree branch lengths. Since the demand of our approach is affected linearly by the number of mutation events, and as we scale up the length of the tree

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.15.435416; this version posted March 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

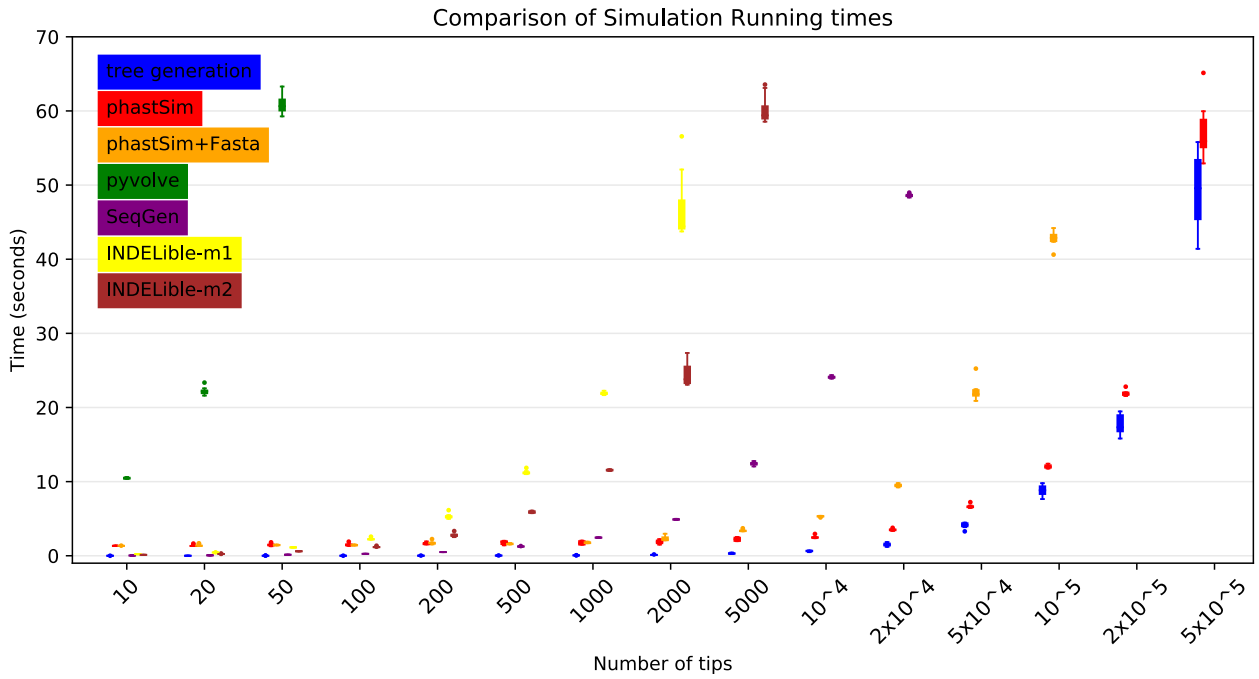

Figure 3: **Comparison of running times of different simulators in a scenario similar to SARS-CoV-2 data**. On the Y axis we show the number of seconds it takes to perform simulations using different software. On the X axis is the number of tips simulated. Each boxplot represents ten replicates. We do not run the most demanding simulators when each replicate would take substantially more than 1 minute to run. In blue is the computational demand for generating the random tree. In red is the time to run phastSim with a concise output, and in orange is the time for phastSim with additionally generating a FASTA format output. In green is the demand of pyvolve, and in purple of Seq-Gen. In yellow and brown are respectively the time for running INDELible with method 1 (matrix exponentiation) and method 2 (Gillespie approach).

Figure 4: **Comparison of running times of different simulators in a scenario similar to *E. Coli* outbreak data**. On the Y axis we show the number of seconds it takes to perform simulations using different software. On the X axis is the number of tips simulated. Each boxplot represents ten replicates. We do not run Seq-Gen for more than 1000 tips due to high computational demand. In red is the time to run phastSim, and in orange is the time for phastSim with the "vanilla" approach. In purple is the time demand of Seq-Gen.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.15.435416; this version posted March 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

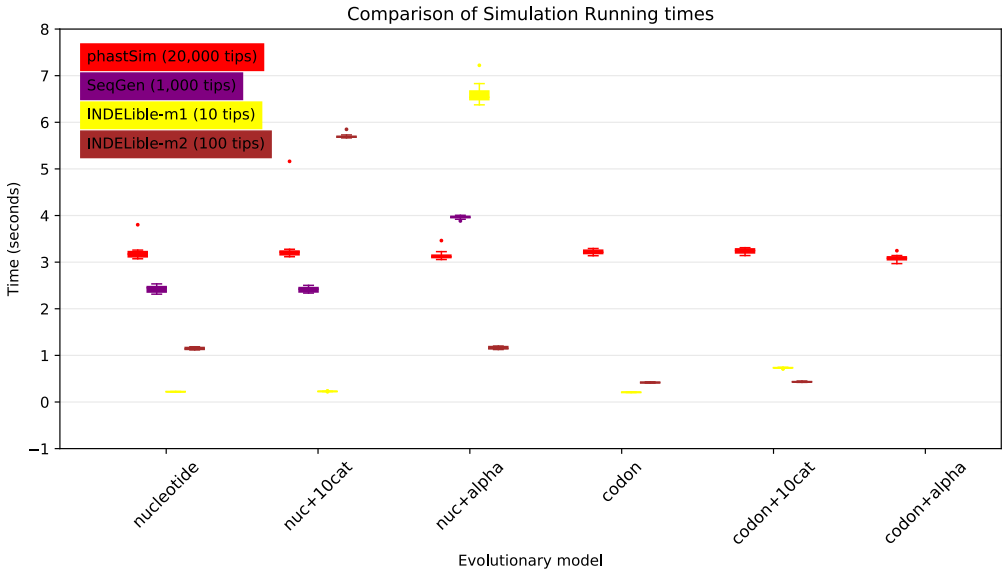

Figure 5: **Comparison of running times of different simulators in a SARS-CoV-2 scenario using different evolutionary models**. On the Y axis we show the number of seconds it takes to perform simulations using different software. On the X axis is the model used for simulations: "nucleotide" is a nucleotide substitution model without variation; "nuc+10cat" is a nucleotide model with 10 rate categories; "nuc+alpha" is a nucleotide model with continuous variation in rate (each site has a distinct rate sampled from a Gamma distribution); "codon" represents a codon substitution model; "codon+10cat" represents a codon substitution model with 10 categories for ω; "codon+alpha" is a codon model with continuous rate variation in mutation rate and in ω (only allowed in phastSim). Each boxplot represents ten replicates. Seq-Gen does not allow codon models. Colors are as in Figure 3. To aid the visual comparison, we use trees of different sizes for different simulators: 10 tips for INDELible-m1; 100 tips for INDELible-m2; 1,000 tips for SEq-Gen; 20,000 tips for phastSim.

we need to simulate more mutation events, then the length of the phylogenetic branches will significantly affect the performance of our approach. We can see that, in the SARS-CoV-2 scenario, the impact is not strictly linear (Figure 6). This is because there are additional factors which contribute to phastSim demand in addition to the number of mutation events. For example, one also has to consider the time to initialize the genome tree, which is linear in genome size, as well as the time to read, initialize, and traverse the input phylogenetic tree, which are linear in the number of tips. Expectedly, the computational demands of Seq-Gen and INDELible-m1 seem not affected by the length of the branches. It is instead surprising to see that the computational demand of INDELible-m2 seems also not affected by the branch lengths, despite it using a Gillespie approach; the reason is that probably other factors, independent of the number of mutations, cause the bulk of the demand in this scenario.

4 Conclusions

We have introduced a new approach to simulating sequence evolution that is particularly efficient when used on phylogenies with many tips and with short branches. Our software phastSim implements this new algorithm and is implemented in Python, allowing it to be easily extended and combined with other Python packages. phastSim relies on the ETE 3 tree phylogenetic structure, and in particular it uses ETE 3 to read input phylogenetic trees. This allows flexibility in the phylogenetic tree input format. Furthermore, thanks to the fact that the efficiency of the algorithm is not affected by the complexity of the substitution model used, we allow a broad choice of evolutionary models, such as codon models with position-specific mutation rates and nonsynonymous/synonymous rates. We also implement a new model of hypermutability to more realistically describe the mutational process in SARS-CoV-2.

We show that, compared with other simulators, phastSim is more efficient in the scenarios common to genomic epidemiology, that is, when simulating many closely related bacterial or viral genomes. Its particular efficiency with bacterial genomes means that it ideally matches the needs of software that simulate bacterial ancestral recombination graphs (e.g. [38, 9]).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.15.435416; this version posted March 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

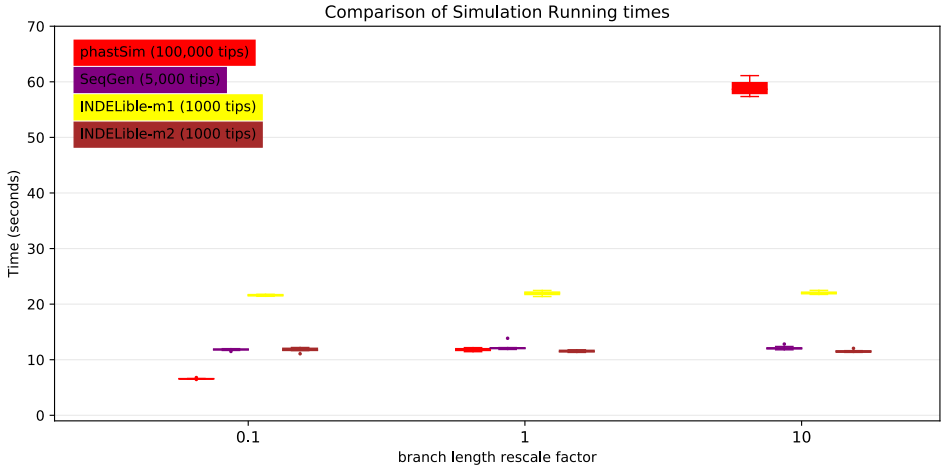

Figure 6: **Comparison of running times of different simulators in a SARS-CoV-2 scenario after rescaling the tree branch lengths by different factors**. On the Y axis we show the number of seconds it takes to perform simulations using different software. On the X axis is the rescaling factor we use to make the phylogenetic tree branch lengths longer or shorter. Colors are as in Figure 3. To aid the visual comparison, we use trees of different sizes for different simulators: 1000 tips for INDELible; 5,000 tips for SEq-Gen; 100,000 tips for phastSim.

phastSim is implemented as a Python package, which allows for easy integration into other Python pipelines. However, Python comes typically at additional computational demand compared to packages written in more efficient programming languages.

In the future, we plan to expand the features of phastSim, in particular we intend to:

- Allow the efficient simulation of indels.
- Allow the modeling of the effect of mutations on the reproductive fitness of a lineage (i.e. its transmissibility in the context of genomic epidemiology)
- Allow a broader spectrum of models, for example column-specific amino acid fitness profiles.
- Include the algorithm within efficient implementations, for example within Seq-Gen.
- Extend our work on DNA data format, for example allowing popular phylogenetic inference methods to process alignment formats that, in the context of genomic epidemiology, are more efficient alignment formats than FASTA or PHYLIP (e.g. `https://github.com/lh3/minimap2#the-cs-optional-tag`).

In conclusion, we have presented a novel algorithm, and corresponding software implementation phastSim, to efficiently simulate sequence evolution along large trees of closely related sequences. This new approach considerably outperforms other methods in the scenarios of genomic epidemiology, for example when simulating SARS-CoV-2 genome sequence datasets. This approach also allows for more realistic models of sequence evolution, and paves the way for future further extensions and more efficient and realistic sequence data simulation and inference.

5 Code Availability

The code and data used for this project (except for the SARS-CoV-2 phylogenetic tree which falls under the restrictions of the GISAID terms of use) are available at `https://github.com/NicolaDM/phastSim`. phastSim can be easily installed across most platforms (see PyPI repository `https://pypi.org/project/phastSim/`) using the pip installer.

6 Acknowledgments

NG, LW, CRW and NDM were supported by the European Molecular Biology Laboratory (EMBL). R.C.-D. was supported by R35GM128932, by an Alfred P. Sloan foundation fellowship, and by funding from the Schmidt Futures Foundation. We thank William Boulton for comments on the early versions of the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.15.435416; this version posted March 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

[1] Miguel Arenas. Simulation of molecular data under diverse evolutionary scenarios. *PLoS Comput Biol*, 8(5):e1002495, 2012.

[2] William Fletcher and Ziheng Yang. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Molecular biology and evolution*, 27(10):2257–2267, 2010.

[3] Gregory Jordan and Nick Goldman. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Molecular biology and evolution*, 29(4):1125–1139, 2012.

[4] Ricardo Assunção Vialle, Asif U Tamuri, and Nick Goldman. Alignment modulates ancestral sequence reconstruction accuracy. *Molecular biology and evolution*, 35(7):1783–1797, 2018.

[5] Michael Worobey, Jonathan Pekar, Brendan B Larsen, Martha I Nelson, Verity Hill, Jeffrey B Joy, Andrew Rambaut, Marc A Suchard, Joel O Wertheim, and Philippe Lemey. The emergence of sars-cov-2 in europe and north america. *Science*, 370(6516):564–570, 2020.

[6] Simon Tavaré et al. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, 17(2):57–86, 1986.

[7] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[8] Daniel J Wilson, Edith Gabriel, Andrew JH Leatherbarrow, John Cheesbrough, Steven Gee, Eric Bolton, Andrew Fox, C Anthony Hart, Peter J Diggle, and Paul Fearnhead. Rapid evolution and the importance of recombination to the gastroenteric pathogen campylobacter jejuni. *Molecular biology and evolution*, 26(2):385–397, 2009.

[9] Nicola De Maio and Daniel J Wilson. The bacterial sequential markov coalescent. *Genetics*, 206(1):333–343, 2017.

[10] Zhengting Zou, Hongjiu Zhang, Yuanfang Guan, and Jianzhi Zhang. Deep residual neural networks resolve quartet molecular phylogenies. *Molecular Biology and Evolution*, 37(5):1495–1507, 12 2019.

[11] Anton Suvorov, Joshua Hochuli, and Daniel R Schrider. Accurate inference of tree topologies from multiple sequence alignments using deep learning. *Systematic biology*, 69(2):221–233, 2020.

[12] Alina F Leuchtenberger, Stephen M Crotty, Tamara Drucks, Heiko A Schmidt, Sebastian Burgstaller-Muehlbacher, and Arndt von Haeseler. Distinguishing Felsenstein zone from Farris zone using neural networks. *Molecular Biology and Evolution*, 37(12):3632–3641, 07 2020.

[13] Robert G Beiko and Robert L Charlebois. A simulation test bed for hypotheses of genome evolution. *Bioinformatics*, 23(7):825–831, 2007.

[14] Richard R Hudson. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

[15] Guillaume Laval and Laurent Excoffier. Simcoal 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*, 20(15):2485–2487, 2004.

[16] Gregory Ewing and Joachim Hermisson. Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065, 2010.

[17] Andrew Rambaut and Nicholas C Grass. Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238, 1997.

[18] William Fletcher and Ziheng Yang. Indelible: a flexible simulator of biological sequence evolution. *Molecular biology and evolution*, 26(8):1879–1888, 2009.

[19] Botond Sipos, Tim Massingham, Gregory E Jordan, and Nick Goldman. Phylosim-monte carlo simulation of sequence evolution in the r statistical computing environment. *BMC bioinformatics*, 12(1):1–6, 2011.

[20] Benoit Morel, Pierre Barbera, Lucas Czech, Ben Bettisworth, Lukas Hübner, Sarah Lutteropp, Dora Serdari, Evangelia-Georgia Kostaki, Ioannis Mamais, Alexey Kozlov, et al. Phylogenetic analysis of sars-cov-2 data is difficult. *bioRxiv*, 2020.

[21] Yuelong Shu and John McCauley. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13):30494, 2017.

[22] Nicola De Maio, Conor R Walker, Yatish Turakhia, Robert Lanfear, Russell Corbett-Detig, and Nick Goldman. Mutation rates and selection on synonymous mutations in sars-cov-2. *bioRxiv*, 2021.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.15.435416; this version posted March 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[23] Yatish Turakhia, Nicola De Maio, Bryan Thornlow, Landen Gozashti, Robert Lanfear, Conor R. Walker, Angie S. Hinrichs, Jason D. Fernandes, Rui Borges, Greg Slodkowicz, Lukas Weilguny, David Haussler, Nick Goldman, and Russell Corbett-Detig. Stability of SARS-CoV-2 phylogenies. *PLOS Genetics*, 16(11):e1009175, November 2020.

[24] Alan M Rice, Atahualpa Castillo Morales, Alexander T Ho, Christine Mordstein, Stefanie Mühlhausen, Samir Watson, Laura Cano, Bethan Young, Grzegorz Kudla, and Laurence D Hurst. Evidence for strong mutation bias towards, and selection against, u content in sars-cov-2: implications for vaccine design. *Molecular Biology and Evolution*, 2020.

[25] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

[26] Yatish Turakhia, Bryan Thornlow, Angie S Hinrichs, Nicola De Maio, Landen Gozashti, Robert Lanfear, David Haussler, and Russ Corbett-Detig. Ultrafast sample placement on existing trees (usher) empowers real-time phylogenetics for the sars-cov-2 pandemic. *bioRxiv*, 2020.

[27] Thomas H Jukes, Charles R Cantor, et al. Evolution of protein molecules. *Mammalian protein metabolism*, 3:21–132, 1969.

[28] Hans Ellegren, Nick GC Smith, and Matthew T Webster. Mutation rate variation in the mammalian genome. *Current opinion in genetics & development*, 13(6):562–568, 2003.

[29] Ziheng Yang. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11(9):367–372, 1996.

[30] Ziheng Yang. Estimating the pattern of nucleotide substitution. *Journal of molecular evolution*, 39(1):105–111, 1994.

[31] Daniel Tang. Efficient algorithms for modifying and sampling from a categorical distribution. *arXiv preprint arXiv:1906.11700*, 2019.

[32] David A Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

[33] Stephanie J Spielman and Claus O Wilke. Pyvolve: a flexible python module for simulating sequences along phylogenies. *PloS one*, 10(9):e0139047, 2015.

[34] Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of molecular evolution*, 22(2):160–174, 1985.

[35] Nick Goldman and Ziheng Yang. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular biology and evolution*, 11(5):725–736, 1994.

[36] Jaime Huerta-Cepas, François Serra, and Peer Bork. Ete 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution*, 33(6):1635–1638, 2016.

[37] Frederick R Blattner, Guy Plunkett, Craig A Bloch, Nicole T Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D Glasner, Christopher K Rode, George F Mayhew, et al. The complete genome sequence of escherichia coli k-12. *science*, 277(5331):1453–1462, 1997.

[38] Thomas Brown, Xavier Didelot, Daniel J Wilson, and Nicola De Maio. Simbac: simulation of whole bacterial genomes with homologous recombination. *Microbial genomics*, 2(1), 2016.