BMC Bioinformatics

CrossMark

Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity

Fabrizio Menardo[1,2]* , Chloé Loiseau[1,2], Daniela Brites[1,2], Mireia Coscolla[1,2], Sebastian M. Gygli[1,2], Liliana K. Rutaihwa[1,2,3], Andrej Trauner[1,2], Christian Beisel[4], Sonia Borrell[1,2] and Sebastien Gagneux[1,2]*

Abstract

Background: Large sequence datasets are difficult to visualize and handle. Additionally, they often do not represent a random subset of the natural diversity, but the result of uncoordinated and convenience sampling. Consequently, they can suffer from redundancy and sampling biases.

Results: Here we present Treemmer, a simple tool to evaluate the redundancy of phylogenetic trees and reduce their complexity by eliminating leaves that contribute the least to the tree diversity.

Conclusions: Treemmer can reduce the size of datasets with different phylogenetic structures and levels of redundancy while maintaining a sub-sample that is representative of the original diversity. Additionally, it is possible to fine-tune the behavior of Treemmer including any kind of meta-information, making Treemmer particularly useful for empirical studies.

Keywords: Representative sample, Large phylogenetic trees, Redundancy reduction, Size reduction, Sampling bias, Clone elimination, Biogeography, Tuberculosis, Influenza

Background

The number of genome sequences deposited into repositories such as NCBI and EBI is increasing rapidly. This wealth of data is at the same time a great opportunity and a challenge for biologists. Large datasets are difficult to visualize and use in downstream analyses. Additionally, being the product of different studies, downloaded datasets are often redundant and suffer from sampling biases. Several software packages such as CD-HIT [1] are available to reduce redundancy in a collection of amino-acid or nucleotide sequences [2]. Essentially, these methods cluster together sequences with a sequence identity higher than a certain threshold (specified by the user), and then select a representative sequence from each cluster for further analysis. While such methods are very efficient and can handle millions of sequences in a short time, they cannot be applied to

whole genome data and do not consider phylogenetic relationships between sequences. To overcome these limitations, methods that reduce the size of datasets based on phylogenies instead of sequence similarity are needed. To date, two software packages have been developed for such purpose:

1) Tree pruner [3] is a tool to manually select and prune leaves/branches from a phylogenetic tree.
2) Treetrimmer [4] automatically reduces the number of leaves in a tree to few representatives for each user-defined operational taxonomical unit (OTU), like genus or species.

While both of these methods address the problem of size reduction in phylogenetic trees, they have some limitations: Tree pruner [3] can be very useful for manual curation, however it is not an automatic method, and it relies on subjective decisions by the users. Treetrimmer [4] is fully automatic, however it is based on user-defined OTU. For some datasets, information on

* Correspondence: fabrizio.menardo@swisstph.ch;
sebastien.gagneux@swisstph.ch
[1]Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Basel, Switzerland
Full list of author information is available at the end of the article

Menardo *et al. BMC Bioinformatics* (2018) 19:164

Page 2 of 8

the taxonomy might not be available and, more importantly, taxonomic categories are only a very rough proxy for genetic diversity.

Here we present Treemmer, a simple tool based on an iterative algorithm to reduce size and evaluate redundancy of phylogenetic datasets. In contrast to previous methods, Treemmer can automatically process any phylogenetic trees with branch lengths and does not require additional information. Treemmer prunes leaves from a phylogenetic tree while minimizing the loss of genetic diversity. At each iteration, all pairs of neighboring leaves are evaluated, the pair with the shortest distance between leaves is selected, and one leaf is pruned off. The user can evaluate the redundancy of the dataset through the plot of the decay of the relative tree length and decide how many leaves to retain, or at what proportion of the original tree length to stop trimming. We applied Treemmer to two datasets (*Mycobacterium tuberculosis* and influenza A virus) and show that it can reduce their size and redundancy while maintaining a subset of samples that are representative of the overall diversity and topology of the original phylogenetic tree.

Methods

Implementation

Treemmer is written in python and uses the ETE library [5] to work with tree structures. Joblib [6] was used to parallelize the search of neighboring leaves (step1).

We present here the algorithm implemented in Treemmer:

Step 1) Given a phylogenetic tree, Treemmer iterates through all leaves, for each leaf it identifies the immediate neighboring leaves (separated by one node). If it does not find any immediate neighbor, it extends the search to leaves separated by two nodes. The result of this step is a list of pairs of neighboring leaves and their genetic distances measured as the sum of the lengths of the branches separating the two leaves.

Step 2) Treemmer selects the pair of leaves with the shortest distance among all the pairs of neighboring leaves, then it prunes a random leaf belonging to the pair. In case there are several equidistant pairs, Treemmer selects one at random. After pruning, all pairs of neighboring leaves containing the pruned leaf are eliminated from the list (Fig. 1).

Step 2 can be repeated any number of times before going back to step 1, this behavior can be controlled with the option *-r (–resolution)*. With the default value (*-r = 1*), only one leaf is pruned (step 2) before the set of neighboring leaves is recalculated (step 1). With higher values of *-r = x*, step 2 is repeated x times (x leaves are pruned) before recalculating the set of neighboring leaves (step 1). Since step 1 is the most computationally intensive part of the algorithm, the option *-r* can speed up the running time of Treemmer considerably.

Steps 1 and 2 are repeated until one of the three possible stop criteria is reached:

Stop option 1) The tree is pruned until there are three leaves left (default). At each iteration, the relative tree length (relative to the input tree) and the number of leaves in the tree are stored. Treemmer outputs the plot of the decay of the relative tree length (RTL, compared to the original tree). The RTL decay is a
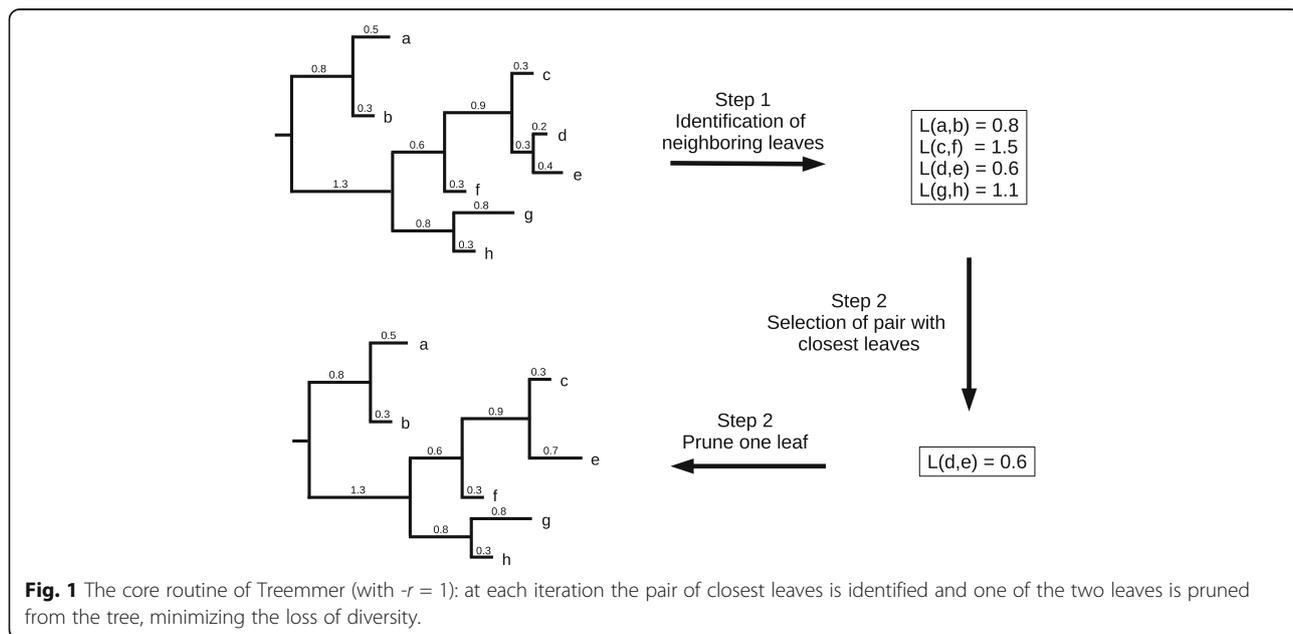
Fig. 1 The core routine of Treemmer (with *-r = 1*): at each iteration the pair of closest leaves is identified and one of the two leaves is pruned from the tree, minimizing the loss of diversity.

Menardo *et al. BMC Bioinformatics* (2018) 19:164

Page 3 of 8

function of the redundancy of the dataset and of the phylogenetic structure of the tree. The RTL decay plot is a useful tool to evaluate the redundancy of the datasets and how much of the diversity is lost at each iteration.

Stop option 2) With the stop option -X (–stop_at_X_leaves) the user sets the number of leaves that should be retained in the reduced dataset. Treemmer prunes the input tree until the specified number of leaves is reached and outputs the list of retained leaves and the pruned tree. The output tree is useful for a quick evaluation of the reduced datasets but should not be used for further analysis, a new tree should be inferred with the reduced dataset.

Stop option 3) With the stop option -RTL (–stop_at_RTL) the user sets a threshold on the RTL of the reduced dataset. When the pruned tree RTL falls below the specified value, Treemmer stops pruning and outputs the list of retained leaves and the pruned tree. The output tree is useful for a quick evaluation of the reduced datasets but should not be used for further analysis, a new tree should be inferred with the reduced dataset.

Support values
Treemmer can read and process trees with support values, however these will be ignored by the software. Support values are a measure of confidence in the inferred topology, and they give no information on the relatedness of two leaves, therefore Treemmer does not consider them.

Polytomies
Treemmer can process trees with polytomies, however, very large unresolved polytomies (with thousands of leaves) increase considerably the number of pairs of neighboring leaves, slowing down the calculation. With the option -p / –solve-polytomies all polytomies in the tree are resolved randomly with branch lengths set to zero.

Pruning options
In many studies is often the case that particular clades or geographic locations are known to be oversampled, and it would be desirable to subsample such subset leaving the rest of tree as it is.

To cope with such situations and to increase the flexibility of Treemmer, we implemented three options (-lm, -mc and -lmc) to specify which set of leaves should or

should not be pruned. With these options it is possible to:

- select specific clades that should be pruned (or that should not be pruned)
- prune (or not prune) only leaves originating from (a) specific country(ies) (or any other type of meta information)
- prune the tree maintaining a user-specified number of representatives for each country (or any other type of meta information)
- any combination of the analyses above.

We provide a short tutorial with examples and instructions on how to use these options together with the software.

Mycobacterium tuberculosis dataset
We downloaded Illumina reads of 12,866 isolates of *M. tuberculosis* that we identified in the sequence read archive (SRA) repository. These represent the large majority of *M. tuberculosis* genomes currently available in the public domain.

Illumina adaptors were clipped and low quality reads were trimmed with Trimmomatic v 0.33 (SLIDINGWINDOW:5:20) [7]. Reads shorter than 20 bp were excluded for the downstream analysis. Overlapping paired-end reads were then merged with SeqPrep v 1.2 [8] (overlap size = 15). The resulting reads were mapped to the reconstructed ancestral sequence of the *M. tuberculosis* complex [9] with the mem algorithm of BWA v 0.7.13 [10]. Duplicated reads were marked by the MarkDuplicates module of Picard v 2.9.1 [11] The RealignerTargetCreator and IndelRealigner modules of GATK v 3.4.0 [12] were used to perform local realignment of reads around InDels. Pysam v 0.9.0 [13] was used to exclude reads with alignment score lower than $(0.93*read\_length)-(read\_length*4*0.07))$: this corresponds to more than 7 miss-matches per 100 bp. SNPs were called with Samtools v 1.2 mpileup [14] and VarScan v 2.4.1 [15] using the following thresholds: minimum mapping quality of 20, minimum base quality at a position of 20, minimum read depth at a position of 7X, minimum percentage of reads supporting the call 90%, maximum strand bias for a position 90%.

Strains with average coverage < 20 X were excluded. Additionally, we excluded genomes with more than 50 % of the SNPs excluded due to the strand bias filter. Furthermore, we excluded genomes with more than 50% of SNPs with a percentage of reads supporting the call included between 10% and 90%. Finally, we filtered out genomes with phylogenetic SNPs belonging to different lineages of MTB, as this is an indication that a mix of strains was sequenced.

Menardo *et al. BMC Bioinformatics* (2018) 19:164

Page 4 of 8

After filtering, we obtained 338,553 positions with less than 10% of missing data were polymorphic in at least one strain. After eliminating strains with more than 10% of missing data at these positions, the final dataset comprised 10,303 strains with high-quality genomes (Additional file 1: Table S3).

The phylogenetic tree was inferred with FastTree [16] with options -nocat -nosupport and -fastest.

Influenza A dataset

The influenza A virus tree was download from nextstrain [17] on the 1st of December 2017. In the time calibrated downloaded tree, few branches had small negative values, these were poorly supported branches that were collapsed before running Treemmer.

To run TempEst [18] we used the "divergence tree" corresponding to the same Influenza dataset used in the other analyses.

Results

Analysis of a *Mycobacterium tuberculosis* dataset

To demonstrate one possible application of Treemmer, we analyzed a *Mycobacterium tuberculosis* (MTB) tree built from the variable nucleotide positions of 10,303 isolates. More information on the dataset and on the pipeline used to build the tree is available in the Methods section and in Additional file 2: Table S1. This collection of genome sequences of MTB was not subsampled a priori, but represents all high-quality MTB genome sequences that we were able to retrieve from public repositories. This dataset is the result of several years of sampling by the scientific community, and while it covers most of the known diversity of MTB, it is highly redundant and suffers from sampling bias. One important source of redundancy originates from projects that analyzed individual MTB outbreaks with whole genome sequencing methods. In these projects, several identical or very similar strains were sequenced. Additionally, some phylogenetic lineages of MTB (i.e., L2 and L4) are overrepresented compared to others, mostly because they are predominant in countries where sampling was particularly extensive (Additional file 3: Table S2). We then tested whether Treemmer is able to reduce the redundancy and the sampling bias of this dataset.

We analyzed the decay of the relative tree length with four different values of -*r*: 1, 10, 100 and 1,000. We found that the RTL decay starts slowly and accelerates after about half of the leaves have been pruned. This confirms the high redundancy of this MTB dataset: pruning thousands of leaves affected the tree length only marginally, indicating that there were strains very similar to the pruned ones retained in the tree. Additionally, we found that the trajectories of the decay were overlapping and indistinguishable for -*r* = 1, -*r* = 10 and -*r* = 100.

While for -*r* = 1,000, the decay was comparable to the other values of -*r* until the number of leaves reached about 6,000, and it was slightly faster with fewer leaves (Fig. 2). This finding suggests that the value of -*r* does not influence the results of Treemmer as long as it is two orders of magnitude smaller than the number of leaves. This is important for the analysis of large trees, because increasing the value of the -*r* option can considerably reduce the running time of Treemmer. While this result should be valid in general, we suggest, when possible, to start the analysis of new datasets running Treemmer with several different values of -*r* and comparing the results.

To obtain a reduced tree maintaining 95% of the original tree length, we ran Treemmer with the stop option -*RTL* 0.95. This resulted in a tree with 4,919 leaves, therefore Treemmer reduced the size of the dataset by more than 50 % while retaining 95% of the diversity (measured as tree length). The resulting dataset has the same phylogenetic structure of the full-size original, is easier to handle, and can be used as a starting point for the downstream analysis (Fig. 3).

Analysis of the influenza dataset

We showed that Treemmer can reduce a redundant dataset while maximizing the retained diversity. Next, we tested Treemmer on a dataset with different characteristics and phylogenetic structure. To do this we used a time-calibrated influenza A / H3N2 tree with 2,080 viral sequences downloaded from Nextstrain [17]. In contrast to the MTB dataset, this dataset is already a

Fig. 2 Plot of the relative tree length decay for the MTB dataset. Four different analysis were run with -*r* = 1 (black dots), -*r* = 10 (red dots), -*r* = 100 (blue dots) and -*r* = 1000 (green dots). The slow decay of the RTL is due to the high redundancy of the dataset. The RTL decays for -*r* = 1, 10 and 100 are overlapping and indistinguishable.

Menardo *et al. BMC Bioinformatics* (2018) 19:164

Page 5 of 8

Fig. 3 Comparison of original (**a**) and reduced (**b**) tree of the MTB dataset, with 10,303 and 4,919 leaves, respectively. The scale bar indicates expected substitution per position (only polymorphic nucleotide positions were included in the alignment). The different lineages of TB are labeled (Maf : *Mycobacterium africanum* (L5 and L6) + animal lineages).

Fig. 4 Comparison of original (**a**) and reduced (**b**) tree of the influenza A virus dataset, with 2,080 and 250 leaves, respectively. The scale bar indicates years.

Menardo *et al. BMC Bioinformatics* (2018) 19:164

Page 6 of 8

sub-sample of all the available influenza A genome sequences, filtered to reduce redundancy and to achieve an equitable temporal and geographic distribution [17]. Additionally, the influenza A tree has a different shape compared to the MTB tree: while the MTB tree comprises well-defined lineages that coalesce close to the root of the tree, the influenza tree is bushy and has a ladder-like structure (Fig. 4).

We found that the trajectory of the RTL decay was steeper compared to the MTB dataset. The greater steepness of the RTL decay compared to the MTB tree is due to the reduced redundancy of the viral dataset: pruning few leaves reduces the tree length considerably. Additionally, we found that the decay of the relative tree length was not very different with three different values of *-r*: 1, 10 and 100 (Fig. 5). This confirms what we already observed during the analysis of the MTB dataset: the value of *-r* does not influence the results of Treemmer as long as it is smaller than two orders of magnitude compared to the number of leaves of the input tree.

We then used Treemmer to subsample 250 leaves from the original influenza tree, corresponding to less than 40% of the original tree length. We found that the reduced tree has the same shape and phylogenetic structure of the original tree, demonstrating that Treemmer can be used to reduce the size of non-redundant datasets maintaining a representative set of samples (Fig. 4).

Fig. 5 Plot of the relative tree length decay for the influenza A virus dataset. Three different analysis were run with *-r* = 1 (black dots), *-r* = 10 (red dots) and *-r* = 100 (blue dots). For this dataset the decay was faster than for the MTB dataset. This is due to the different structure of the phylogenetic trees and to the reduced redundancy of the viral dataset.
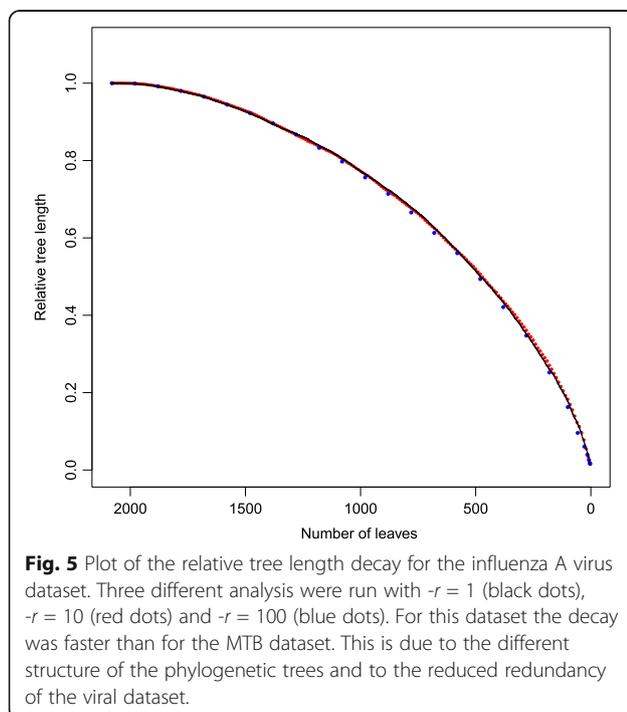
Trimming trees for molecular clock analysis

One of the reasons to reduce the size of a dataset is to analyze it with a molecular clock. Often the main objective of such analyses is to estimate the age of the most common recent ancestor (MRCA) or other internal nodes of the tree. In these cases, it is important to avoid that Treemmer prunes the MRCA node. This can be achieved protecting the outgroup(s), using the *-lm* option or reintroducing it in the reduced dataset. If there is no known outgroup and the samples were collected at different time points, we suggest to check the time structure of the reduced dataset with TempEst [18]. TempEst positions the root in the point of the tree that best fit a strict clock model and performs root-to-tip regression to estimate the age of the MRCA and the substitution rate. In particular, the best-fit root position in the original dataset should not be on a branch that is absent in the reduced dataset, and the root-to-tip regression should give similar estimations of the MRCA age and of the substitution rate.

We used TempEst [18] to explore the time structure of the influenza dataset at different levels of reduction. We trimmed the influenza tree inferred under a no-clock model (the "divergence" tree in Nextstrain) at 99%, 90%, 75% and 50% of the relative tree length. Applying the best-fitting root method, we observed that the MRCA of the complete dataset was not pruned in any of the reduced trees, and that the root-to-tip regression yielded similar estimations of the substitution rate and age of MRCA for all trees (Additional file 4: Figure S1 and Additional file 1: Table S3).

Stochastic component of Treemmer

The algorithm implemented in Treemmer is not deterministic: when the pair of leaves with shortest distance is selected, one of the two leaves is pruned at random. Additionally, if there are several pairs with same (shortest) distance, a pair is picked at random. We investigated the effect of the stochastic component of Treemmer on the output of the software, repeating the RTL decay analysis showed in Figs. 2 and 5, 100 times for each datasets. We found that the difference in RTL between runs increase with increasing iterations, and reaches maximum of about 1% of the RTL for the TB dataset, and about 2.5% of the RTL for the influenza dataset (Additional file 5: Figure S2 and Additional file 6: Figure S3).

Although these differences are limited, we recommend running Treemmer several times, and if possible to stop trimming before the point of maximum variability between runs is reached.

Discussion

An important trend in phylogenetic research is the development of methods and software that can handle big

Menardo *et al. BMC Bioinformatics* (2018) 19:164

Page 7 of 8

datasets. Thanks to software such as FastTree [16] and Dendroscope [19], it is now possible to build and visualize trees with hundreds of thousands of sequences.

However, many kinds of analyses, including bootstrap, using a molecular clock or fitting a codon substitution model, are impractical or impossible to perform with very large sequence datasets. To address these limitations, we developed Treemmer, a tool to reduce the size and redundancy of phylogenetic datasets while maintaining a representative diversity. In contrast to methods based on sequence similarity, Treemmer selects a representative subsample considering the phylogenetic relationships between samples and their phylogenetic distance on the tree.

Implicitly, Treemmer adopts a complex definition of diversity that include two basic metrics: genetic distance (branch length) and topological distance (number of internal nodes separating two leaves) (see Vellend et al. 2011 [20] for an overview on methods to measure phylogenetic diversity). At each iteration, Treemmer identifies the pair of leaves with the minimum genetic distance among all pairs of leaves with a topological distance (number of nodes in between) smaller than 3, and prune a random leaf of the selected pair.

With the RTL decay plot, it is possible to evaluate the reduction of the tree length when the size of the dataset is reduced. Treemmer can sample reduced datasets of any size, selecting the balance between size and diversity that best fits the purpose of the user.

The steepness of the RTL decay depends on the level of redundancy of the dataset but also on the structure of the phylogenetic tree. For example, the RTL decay of a tree with few very long branches and several short ones, will have a slower decay compared to a tree with equal number of leaves and all branches with similar length. Additionally, sequences of bad quality, or badly aligned, often produce long branches, and long branches are more likely to be retained by Treemmer, thus enriching bad quality sequences in the dataset. A sensitive approach to bad quality sequences or alignments before inferring the tree is therefore critical.

While the output of Treemmer can be used in many downstream analyses, it should not be considered as a random unbiased sample: the number of leaves belonging to different clades in the reduced dataset depends on the genetic diversity of the different clades and not on the abundance of different clades in nature; highly diverse clades will be represented by more leaves than less diverse ones, irrespectively of the frequency of such clades in natural populations. Additionally, some phenomena, e.g. recent fast speciation or population growth can result in large clades with short branches that would be pruned by Treemmer. Therefore, users interested in such phenomena should be careful when using Treemmer in their pipelines.

Conclusions

We developed Treemmer, a tool to reduce the size of large phylogenetic datasets maintaining a sub-sample that is representative of the original diversity. With Treemmer it is possible to reduce the size of datasets that are too large for specific analysis, additionally the possibility of including many kind of meta-information makes Treemmer particularly flexible and useful for empirical studies.

Availability and requirements

Project name: Treemmer
    Project home page: https://git.scicore.unibas.ch/TBRU/Treemmer
    Operating system: platform independent
    Programming language: Python
    Other requirements: ETE3 and Joblib
    License: GPL

Additional files

Additional file 1: Table S3. Results of the root-to-tip regression analysis performed with TempEst onto the influenza tree at different levels of reduction. (TXT 269 bytes)

Additional file 2: Table S1. Accession numbers used for the MTB tree. (TXT 341 kb)

Additional file 3: Table S2. Number of leaves per lineage in the complete and reduced MTB tree showed in Fig. 3 (TXT 174 bytes)

Additional file 4: Figure S1. Root-to-tip regression for the complete influenza dataset (2063 leaves, black dots and black regression line, in gray the 99% confidence interval) and four reduced trees (orange: 99% or RTL , red: 90% of RTL, green: 75% of RTL, blue: 50% of RTL. All trees were re-rooted with the best-fit method implemented in TempEst. (PDF 12 kb)

Additional file 5: Figure S2. Plot of 100 RTL decays (-r =100) for the TB dataset (black dots). All decays are similar. Red dots indicates the range of RTL among the different runs for the corresponding iteration, the variability among runs increases slowly at first, it reaches a maximum of 1% when the tree is reduced to 20% of the RTL. (PDF 25 kb)

Additional file 6: Figure S3. Plot of 100 RTL decays (-r =10) for the Influenza dataset (black dots). All decays are similar, but there is a larger variability compared to the TB dataset. Red dots indicate the range of RTL among the different runs for the corresponding iteration, the variability among runs increases steadily and it reaches a maximum of 2.5% when the tree is reduced to 40% of the RTL. (PDF 56 kb)

Menardo *et al. BMC Bioinformatics* (2018) 19:164

Page 8 of 8

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Basel, Switzerland. [2]University of Basel, Basel, Switzerland. [3]Ifakara Health Institute, Bagamoyo, Dar es Salaam, Tanzania. [4]Department of Biosystems Science and Engineering, ETH Zürich, 4058 Basel, Switzerland.

References
1. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*. 2001;*17*(3):282–3.
2. Sikic K, Carugo O. Protein sequence redundancy reduction: comparison of various method. *Bioinformation*. 2010;*5*(6):234.
3. Krishnamoorthy M, Patel P, Dimitrijevic M, Dietrich J, Green M, Macken C. Tree pruner: An efficient tool for selecting data from a biased genetic database. *BMC bioinformatics*. 2011;*12*(1):51.
4. Maruyama S, Eveleigh RJ, Archibald JM. Treetrimmer: a method for phylogenetic dataset size reduction. *BMC research notes*. 2013;*6*(1):145.
5. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution*. 2016;*33*(6):1635–8.
6. Joblib: https://pythonhosted.org/joblib.
7. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*. 2014;*30*(15):2114–20.
8. SeqPrep: https://github.com/jstjohn/SeqPrep.
9. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nature genetics*. 2013;*45*(10):1176–82.
10. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*. 2009;25:1754–60.
11. Picard: https://github.com/broadinstitute/picard.
12. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010;20:1297–303.
13. Pysam: https://github.com/pysam-developers/pysam.
14. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987–93.
15. Koboldt D, Zhang Q, Larson D, Shen D, McLellan M, Lin L, Miller C, Mardis E, Ding L, Wilson R. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research. 2012;22(3):568–76.
16. Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE. 2010;5(3):e9490.
17. Neher RA, Bedford T. nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics*. 2015;31(21):3546–8.
18. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus evolution*. 2016;*2*(1)
19. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC bioinformatics*. 2007;*8*(1):460.
20. Vellend, M., Cornwell, W. K., Magnuson-Ford, K., & Mooers, A. Ø. (2011). Measuring phylogenetic biodiversity. *Biological diversity: frontiers in measurement and assessment. Oxford University Press, Oxford*, 194-207.