Characterizing SARS-CoV-2 transmission clusters using phylogenies and metadata in DYNAMITE

Description

DYNAMITE stands for DYNAMic Identification of Transmission Epicenters within phylogenies. It is a depth-first search algorithm that is not restricted to monophyletic clades, allowing for improved sensitivity to the detection of dynamic transmission clusters that comprise transmission patterns of interest for public health relevance. A more detailed description of the DYNAMITE algorithm and its applications can be found here. Since the release of this article, we have added a time component to the algorithm, allowing the user to restrain the inclusion of individuals in clusters based on a specified serial interval. This serial interval is defined as the period of time between a primary case-patient (infector) with symptom onset and a secondary case-patient, which has been observed to be 5-6 days for SARS-CoV-2 Rai et al.[1]. Direct transmission chains are intuitively represented by individuals separated by both minimal genetic evolution and time difference within this serial interval. In this tutorial, you will use DYNAMITE in its native R environment to identify putative DARS-CoV-2 Delta direct transmission chains involving the state of Florida and examine the potential underlying risk factors from the availble patient metadata. For additional information on DYNAMITE, refer to this README.md.

Understanding the options

DYNAMITE takes a tree (scaled in substitutions/site) in newick or nexus format with support values at the nodes (e.g., bootstrap). Support values can be provided by any (single) method - i.e., combined result of 2 or more methods will not be accepted. These values can be scaled from 0-1 or 0-100. The tree provided in this tutorial was reconstructed in IQ-TREE and can be vizualized in most standard tree-viewing tools (e.g., Figtree.

Additionally, a metadata file with sequence IDs, sampling dates, and traits of interest is required. Column headers are needed and must contain the strings "ID" and "Date" (specific case format not required). Date information must be consistently either numeric or Date format. Remaining columns can consist of any number of traits of interest.

Step 1. Read the headers of the metadata file in your command line interface (terminal on Mac will be shown from here on) in order to see what metadata information is available for these sequence data.

```
head -n 1 delta_metadata.csv
```

```
Brittanys-MacBook-Air:delta macbook$ head -n 1 delta_metadata.csv
ID,SamplingDate,Location,SampleType,Vaccinated,Sex,Age
```

Step 2. Use the following command to print the available options for DYNAMITE:

```
Rscript dynamite.R --help
```

```
Brittanys-MacBook-Air:delta macbook$ Rscript dynamite.R --help
Usage: dynamite.R [options]


Options:
        -t CHARACTER, --tree=CHARACTER
                tree file name [default= .nwk extension]

        -m CHARACTER, --metadata=CHARACTER
                metadata file name [default= .csv extension]

        -q NUMERIC, --timetree=NUMERIC
                option (Y/N) for molecular clock calibration and time tree output/statistics [default= Y]

        -s NUMERIC, --seqLen=NUMERIC
                sequnce length used in molecular clock calibration [default= 30000]

        -c CHARACTER, --cluster=CHARACTER
                choice of cluster algorithm from c (Phylopart's cladewise) or b (DYNAMITE's branchwise) [default= dynamite]

        -l CHARACTER, --threshold=CHARACTER
                threshold for cluster determination, which can be numeric or median [default= 0.05]. Note that median is very computationally intense

        -i CHARACTER, --serial=CHARACTER
                serial interval for cluster filtering [default= 6]

        -a CHARACTER, --asr=CHARACTER
                option (Y/N) of ancestral state reconstruction for each cluster [default= N]

        -h, --help
                Show this help message and exit
```

Aside from the tree and metadata files, DYNAMITE requires the specification of numerous parameters, or arguments, that dictate how DYNAMITE will be used. For example, DYNAMTE can use the fast tree-dating algorithm (utilizing sampling dates) known as treedater [2] to scale the tree in time and provide time-specific phylogenetic information for each identified cluster (e.g., TMRCA). This option can be specified by *"--timetree=Y"*. Treedater is used by default.

If chosen, treedater requires information on the length of the sequences used to generate the original tree with the *"--seqLen"* parameter. The default is 30,000.

As DYNAMITE's branchwise algorithm was adapted from the cladewise algorithm used in Phylopart [3], this algorithm is an alternative option. Note that this algorithm restricts clusters to monophyletic clades.

Both algorithms require a genetic distance (i.e., branch length) threshold above which sequences are not considered part of the cluster. This value can usually be defined by studies of the maximum level of genetic divergence within a host leading up to the end of the infectious period and is very well know for HIV but not as

much for SASR-CoV-2. Therefore, the branchwise algorithm, similarly to that of Phylopart, optimizes this value by searching among a sample of values from a distribution of branch lengths for the value that maximizes the number of clusters identified. This threshold does not represent an actual genetic distance, but rather the left-most fraction of branch lengths in the distribution. In other words, the default threshold value is 0.05, which tells DYNAMITE to sample branch lengths from the 5th percentile of the distribution of branch lengths within the tree. The sample size from this distribution is proportional to the threshold value. Another option here is *"median"*, though it is computationally intensive because 100 samples are drawn. It is the recommended value based on simulated datasets; however, for the dataset used in this tutorial, the same results were produced using a threshold of 20%, which is what we will specify here.

DYNAMITE also allows for ancestral state reconstruction of metadata traits, but the results can often be misleading for transmission clusters with minimal evolution, so it is not recommended and will not be used here.

Step 3. Run DYNAMITE (from the infiles folder) using the following command:

```
Rscript ../dynamite.R --tree delta_fb_2021-08-04.fasta.treefile --metadata ./delta_meta
```

DYNAMITE requires numerous additional R libraries, but DYNAMITE will attempt to install them for you. Following the information on the install and loading of libraries, DYNAMITE will print out the parameters used and steps of the process until all output files are saved:

```
[1] "dynamite --tree delta_fb_2021-08-04.fasta.treefile --metadata delta_metadata.csv --timetree Y --seqLen 30000 --cluster b --threshold 0.25 --serial 6 --asr N"
[1] "================================================================================================================================================================"
[1] "Detecting tree file..."
[1] "Newick file detected: delta_fb_2021-08-04.fasta.treefile"
[1] "Searching for metadata file..."
[1] "Converting substitution tree into timed tree using either treedater or lsd2 and dates provided in metadata file. Please make sure the metadata file contains a
column containing the word 'DATE'"
Note: Minimum temporal branch length  (*minblen*) set to 6.18461360615108e-05. Increase *minblen* in the event of convergence failures.
Tree is not rooted. Searching for best root position. Increase searchRoot to try harder.
[1] "The updated most recent sampling date is 2021-07-29"
Warning message:
In dater(sub_tree, decimal_date(sts), s = seqLen, ncpu = numCores,  :
  omega0 provided incompatible with numStartConditions > 0. Setting numStartConditions to zero.
[1] "Defining and extracting well-supported clades within the tree using node labels..."
[1] "Determining branch length limit...."
[1] "DYNAMITE's branchwise algorithm is being used."
[1] "Writing cluster (and background) trees to separate files..."
[1] "ASR not performed - results will rely on sampled data only."
[1] "Performing quick data manipulation... Hold on to your butts!"
[1] "Determining if clusters related by birth..."
[1] "Estimating infection rate for each cluster (Oster, 2018)"
[1] "Annotating trees with cluster assignments..."
Time difference of 8.58115621805191 mins
```

In the meantime (should take approximately 8 minutes on a personal computer), you can navigate to the "Results" folder and visualize the metadata distributions for each cluster over time using the R shiny package, which allows for interactive visualization. This is especially useful when you are interested in certain windows of time (e.g., prior to implementation of public health interventions).

Step 4. Run the following command to utilize the R script for interactive

visualization:

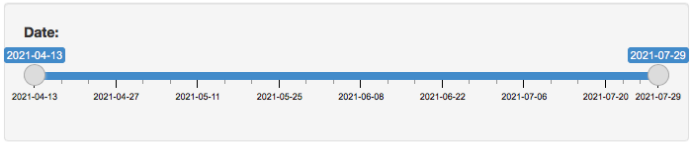
```
Rscript ../dynaviz.R
```

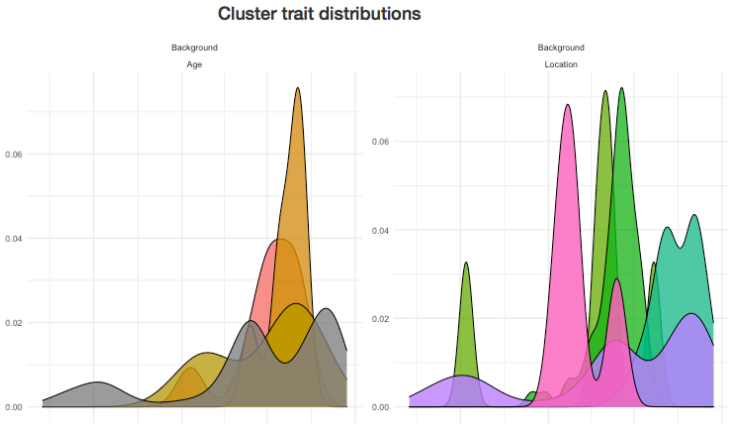
```
Brittanys-MacBook-Air:Results macbook$ Rscript ../cluster_traits_shiny.R

Listening on http://127.0.0.1:6677
```

Step 5. Copy the URL from the output into your browser

You should see kernel density estmates for each of the metadata traits (columns) for each of the four identified clusters and background population (rows), as well as a bar in the upper left-hand corner for specifying the window of time you are interested in:





Play around with the window bar to focus on specific months. What do these plots tell you about transmission in Gainesville, Florida? E.g.,

- How many transmission clusters were identified?
- What were their geographical origins?
- When did they originate?
- Were age or sex risk factors?
- Were vaccinated individuals included?

This shiny R script outputs a PDF copy of the trait distribution plots but also the PDF plot with the timed tree and heatmap used in the presentation accompanying this tutorial. As the legend is not very intuitive using this R package, a copy of the legend is also saved for manual editing for publication figure quality (if you have any other suggestions, please let me know!)

There are also a number of other output files from DYNAMITE, including a summary file detailing tree statistics such as the estimated infection rate, described by Oster et al. [4] and overall phylogenetic diversity (sum of

branch lengths) [5]. These statistics are calculated for each cluster as well as for the background population.

Step 6. Open tree statistics ("dynamite_tree_stats") file in Excel or other text-editing program:

cluster_id	PD	timespan	size	tmrca	R_Oster
c2	0.0005669	60.4126183	7	5/27/21	3.9300832
c3	0.00171351	120.848775	20	3/28/21	2.34099509
c6	0.00056755	99.3611814	6	4/17/21	1.94403201
c10	0.00037094	64.043665	8	5/20/21	3.95848265
Background	0.03387862	192.066239	433	1/17/21	1.20240691

How do cluster infection rates compare to the overall rate? Phylogenetic diversity? Do these represent groups at potentially high risk for transmission?

References

[1] Rai B, Shukla A, Kant Dwivedi L (2021). Estimates of serial interval for COVID-19: A systematic review and meta-analysis. Clin Epidemiol Glob Health. 9: 157-61.

[2] Erik Volz (2020). treedater: Fast Molecular Clock Dating of Phylogenetic Trees with Rate Variation. R package version 0.5.0. https://CRAN.R-project.org/package=treedater

[3] Prosperi MCF (2011). A novel methodology for large-scale phylogeny partition. Nat Commun. 2: 321.

[4] Oster AM, France AM, Panneer N, Bañez Ocfemia MC, Campbell E, Dasgupta S, Switzer WM, Wertheim JO, Hernandez AL (2018). Identifying Clusters of Recent and Rapid HIV Transmission Through Analysis of Molecular Surveillance Data. J Acquir Immune Defic Syndr. 79(5):543-550.

[5] Faith DP (1992). Conservation evaluation and phylogenetic diversity. Biological Conservation. 61 (1): 1-10