MICROBIAL GENOMICS

Methods paper template

rPinecone: Define sub-lineages of a clonal expansion via a phylogenetic tree.

Alexander M. Wailan[1,*], Francesc Coll[2], Eva Heinz[1], Gerry Tonkin-Hill[1], Jukka Corander[1,3,4], Nicholas A. Feasey[5,6], Nicholas R. Thomson[1,2].

[1]Parasites and Microbes, Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK.

[2]London School of Hygiene & Tropical Medicine, Keppel St, Bloomsbury, London, UK, WC1E 7H

[3]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

[4]Department of Biostatistics, University of Oslo, Oslo, Norway

[5]Liverpool School of Tropical Medicine, Pembroke Pl, Liverpool L3 5QA

, UK

[6]Malawi Liverpool Wellcome Trust Clinical Research Programme, Blantyre, Malawi

*Corresponding author: alex.wailan24@gmail.com

ABSTRACT

The ability to distinguish between pathogens is a fundamental requirement to understand the epidemiology of infectious diseases. Phylogenetic analysis of genomic data can provide a powerful platform to identify lineages within bacterial populations, and thus inform outbreak investigation and transmission dynamics. However, resolving differences between pathogens associated with low variant (LV) populations carrying low median pairwise single nucleotide variant (SNV) distances, remains a major challenge. Here we present rPinecone, an R package designed to define sub-lineages within closely related LV populations. rPinecone uses a root-to-tip directional approach to define sub-lineages within a phylogenetic tree according to SNV distance from the ancestral node. The utility of this program was demonstrated using genomic data of two LV populations: a hospital outbreak of methicillin-resistant *Staphylococcus aureus* and endemic *Salmonella* Typhi from rural Cambodia. rPinecone identified the transmission branches of the hospital outbreak and geographically-confined lineages in Cambodia. Sub-lineages identified by rPinecone in both analyses were phylogenetically robust. It is anticipated that rPinecone can be used to discriminate between lineages of bacteria from LV populations where other methods fail, enabling a deeper understanding of infectious disease epidemiology for public health purposes.

DATA SUMMARY

1. Source code for PlasmidTron is available on GitHub under the open source licence GNU GPL 3; (url: https:// github.com/alexwailan/rpinecone).
2. Newick format files for both phylogenetic trees have been deposited in Figshare; (url: https://doi.org/10.6084/m9.figshare.7022558)
3. Geographical analysis of the *S*. Typhi Dataset using Microreact is available at https://microreact.org/project/r1IqkrN1X.
4. Accession numbers, meta data and sample lineage results of both datasets used in this paper are listed in the supplementary tables.

I/We confirm all supporting data, code and protocols have been provided within the article or through supplementary data files. ☒

IMPACT STATEMENT

Whole genome sequence data from bacterial pathogens is increasingly used in the epidemiological investigation of infectious disease, both in outbreak and endemic situations. However, distinguishing bacterial species which are both very similar and which are likely to come from a small geographical and temporal range presents a major technical challenge for epidemiologists. *rPinecone* was designed to address this challenge and utilises phylogenetic data to define lineages within bacterial populations that have limited variation. This approach is therefore of great interest to epidemiologists as it adds a further level of clarity above and beyond that which is offered by existing approaches which have not been designed to consider bacterial isolates containing variation that only transiently exist, but which is epidemiologically informative. *rPinecone* has the flexibility to be applied to multiple pathogens and has direct application for investigations of clinical outbreaks and endemic disease to understand transmission dynamics or geographical hotspots of disease.

INTRODUCTION

Advances in Whole Genome Sequencing (WGS) permit the study of bacteria at high resolution and this has created the opportunity to use WGS data to reliably discriminate between bacteria that were previously indistinguishable by alternative means, or to “type” them. Using WGS data to understand the epidemiology of infections presents distinct challenges to those faced by evolutionary biologists, more so when aiming to distinguish lineages of bacterial populations that are only transiently extant.

Single nucleotide variants (SNVs) which arise stochastically may offer no evolutionary benefit to a given bacterial isolate, and therefore may only transiently be present in a population. Despite their lack of evolutionary significance, however, they may be epidemiologically informative in the context of localised outbreaks, especially amongst organisms that show low overall population diversity. Traditional genetic clustering algorithms such as Structure(1) and BAPS(2, 3) are not suitable for sub-typing of low-variant (LV) bacterial populations over small timescales such as less than three years. They assume independence between loci and require 10-100's of SNV that are shared between different lineages to confidently cluster populations. Novel bioinformatic approaches to identify sub-lineages within LV bacterial populations are therefore needed for epidemiological investigations, especially when technological advancements have made genomic studies of outbreaks increasingly common.

For genomic data, typing tools are used to identify "clusters" or "partitions" to define a group of isolates as a "lineage" or sub-lineage within a population. These three terms are used synonymously in this context. However, to provide clarity we will use the term "sub-lineage" to refer to a defined group of isolates within a population. These typing tools are non-tree based and fall under two broad approaches, distance-based methods which calculate a pairwise distance matrix often using the SNV distances between isolates, and model-based methods which rely on calculating a population genetic model using either Bayesian or maximum likelihood based methods(1). For distance methods samples are considered to be within the same cluster if their SNV distance is less or equal to a specified SNV threshold. An alternative method is the widely used program hierBAPS(3), which when given a multiple sequence alignment attempts to identify the lineages of the sequences that maximises the posterior probability of the hierBAPS model. The model is based on a Multinomial-Dirichlet distribution and assumes independence between SNV sites. The algorithm is applied first to identify an initial set of lineages, which is then iteratively repeated to generate subsequent sub-lineages within each of the initially defined lineages. While both are suitable for populations where long-term historical evolution is present, neither are appropriate options for short-term clonal expansions such as populations emerging from point source outbreaks or locally endemic disease. In the case of Bayesian approaches, the resulting sub-lineages are not able to reflect the phylogenetic data, whereas methods that use a pairwise SNV distance approach are unable to infer direction of evolution or common ancestry between sub-lineages, which is necessary for source and transmission investigations.

Here we present rPinecone, an R package which evaluates a phylogeny using a root-to-tip approach to define sub-lineages according to SNV distances from ancestral nodes. rPinecone also determines if two or more sub-lineages form part of a larger major lineage if they are related by enough ancestral nodes. To demonstrate this approach, rPinecone was tested on a local endemic dataset of *Salmonella enterica* subsp. *enterica* serovar Typhi (*S*. Typhi) from the H58 haplotype(4) as well as a hospital outbreak of methicillin-resistant *Staphylococcus aureus* (MRSA) ST2371(5).

METHODS

112

113 Phylogenetic tree construction of two case studies

114 Data from two studies were reanalysed to construct phylogenetic trees for input into rPinecone.
115 Mapping of reads was performed using SMALT v0.7.4 (sanger.ac.uk/science/tools/smalt-0/) and SNVs
116 were identified by using samtools mpileup(6). SNV alignments were generated using SNP-sites(7) then
117 used to generate phylogenetic maximum-likelihood (ML) trees with RAxML(8). In the MRSA case
118 study, the alignment and subsequent rooted ML tree was generated by using the *S. aureus* ST22 isolate
119 HO 5096 0412 (accession number HE681097) as a mapping reference and as an outgroup to root the
120 tree. In the case of the locally endemic *S.* Typhi in Cambodia, *S.* Typhi Genotyphi 4.3.1 isolate
121 2010_7898 (accession number GCA_001360555) was used as a mapping reference. *S.* Typhi Genotyphi
122 4.1.1. isolate Mal1017142 (accession number ERR279139)(9) was used as an outgroup to root the ML
123 tree. Methodological details for construction of phylogenetic trees can be found in the supplementary
124 material.

125

126 Ancestral reconstruction of phylogenetic trees

127 The SNV alignment and ML tree from the reanalysis of both datasets was used as an input for the
128 ancestral reconstruction tool "pyjar" (github.com/simonrharris/pyjar)(10) generating a ML joint
129 ancestral reconstruction (JAR) tree. Each JAR tree was used as an input for rPinecone to define the
130 sub-lineages within respective populations.

131

132 hierBAPS as a comparative benchmark

133 The R package for hierBAPS (rhierBAPS, available at github.com/gtonkinhill/rhierbaps) was used as a
134 comparative benchmark for rPinecone when analysing the *S*. Typhi dataset, due to its frequent use in
135 bacterial phylogenetics for lineage identification. Input data for this package was the SNV alignment
136 used to generate the phylogenetic ML tree. A maximum level depth of five and maximum cluster
137 number of 30 was used for the analysis.

138

139 RESULTS

140 rPinecone is a package developed in R and available under the open source licence GNU GPL3 at
141 github.com/alexwailan/rpinecone. Input for rPinecone requires two user-specified integer thresholds
142 for SNV distance and relatability (discussed below), and a rooted phylogenetic tree (Newick format)
143 where the branch lengths are integers corresponding to SNV distance. One type of tree that fits this
144 criterion is a maximum likelihood (ML) joint ancestral reconstruction (JAR) tree.

145

146 The rPinecone package has a primary wrapper function to perform the analysis which is performed in
147 two steps: sub-lineage definition and major-lineage definition (Fig. 1).

148

rPinecone will initially prepare for sub-lineage definition by collapsing zero SNV dichotomies of the tree into polychotomies using the *ape* package(11) for R (v4.1) and performing a depth-first search (DFS)(12) across the resulting tree, to list the tips and ancestral nodes from root-to-tip. Edges are then drawn between nodes where the edge distance is the branch length i.e. the number of SNVs isolates have from an ancestral node on a JAR tree. Using the DPS and SNV branch distance information, rPinecone will then define sub-lineages by traversing the tree in a root-to-tip direction to assess each ancestral node. At each ancestral node, a sub-tree is built with the remaining nodes where the ancestral node becomes the root of the sub-tree. rPinecone uses two methods to define sub-lineages, 1) all root-to-tip paths of the sub-tree are assessed and if the maximum root-to-tip branch distance is equal or below the specified SNV threshold, the remaining tips will be assigned a sub-lineage number and 2) If the maximum distance is over the SNV threshold and there are at least two tips with zero distance from the ancestral node, rPinecone assigns these tips in zero distance a sub-lineage number. The process is repeated throughout the entire tree. Tips that are not assigned a sub-lineage number will be assigned with a singleton number.

After sub-lineages have been declared, their relatability is analysed to determine if they form a larger major lineage. The user specified relatability threshold is an integer used to compare the number of ancestral nodes from the root of the entire tree to each sub-lineage and singleton. Lists of ancestral nodes for each sub-lineage and singleton are compiled. These lists are compared in a pairwise fashion. Sub-lineages and singletons that have the same intersecting ancestral nodes over the relatability threshold are declared to have formed a major lineage within the population, where a major lineage is composed of at least two sub-lineages.

Once sub-lineages and major lineages are defined, rPinecone outputs a list of six variables each able to be used for downstream processes, including; the tree used for analysis, number of sub-lineages, major lineages, singletons identified and number of isolates within the tree. The final variable is a three-column table listing each “Taxa” label of the tree and their respective sub-lineage and major lineage. Furthermore, phylogenetic trees can be displayed using the online display tool iTOL(13). The rPinecone package includes three functions taking the output from the primary rPinecone function. These functions parse the rPinecone output to a file in a format to display in iTOL. Three files can be generated, a “LABELS” file to change the tip labels according to their sub-lineage or singleton number as well as two “DATASET_COLORSTRIP” format files to display the sub-lineages and major lineages as a colour strip/ block (Example displayed in the Fig. 2). To show the utility of rPinecone we perform two case studies on a MRSA hospital outbreak and local endemic dataset of *S.* Typhi.

HOSPITAL OUTBREAK OF MRSA ST2371 – HEALTHCARE WORKERS WERE COLONISED BY A MAJOR LINEAGE

In 2011 over a 6-month period, the National Health Service Foundation Trust in Cambridge, UK, investigated transmission of MRSA within the hospital neonatal unit and community and described a complex transmission network(5). We reanalysed the data of this study and generated a phylogenetic

ML tree. From this analysis the population had a median pairwise distance of five SNVs. The phylogenetic tree (Fig. 2) of the outbreak had a structure that begins with a root isolate (Patient 5), then expands outwards in a "star-burst"-like fashion where each branch of the tree represents a transmission pathway of MRSA between infants and their mothers, other mothers on the ward, and to partners of affected mothers as described in the original study(5). rPinecone identified nine sub-lineages and a major lineage composed of three sub-lineages (25 isolates) using SNV threshold of four and a relatability threshold of three. These sub-lineages corresponded to the transmission pathways determined in the original investigation using epidemiological data. The major lineage represents the sub-lineages of MRSA (rPinecone sub-lineages 4, 5 and 6) carried by colonised healthcare workers which were associated with subsequent episodes of MRSA 64 days after the last MRSA-positive patient left the unit.

The major lineages identified by rPinecone highlight sub-lineages related by ancestral nodes and potentially lineages that have separated from the remaining clonal population. This type of analysis was not performed during the original analysis, however we successfully identified MRSA transmission pathways using rPinecone and provided an additional layer of sub-lineage data to complement the epidemiological data when associating sub-lineages with likely transmission events.

LOCALLY ENDEMIC *S.* TYPHI IN RURAL CAMBODIA – HIGHER RESOLUTION TO IDENTIFY CIRCULATING SUB-LINEAGES

In the second example, rPinecone distinguished lineages of an *S*. Typhi population circulating in a defined geographical area causing locally endemic disease, as opposed to a clonal outbreak. LV bacterial pathogens responsible for localised, endemic disease present a different challenge; in this context, it is highly likely that multiple, genetically similar sub-lineages will be in circulation, and in order to be epidemiologically informative, phylogenetic analysis must be able to distinguish these lineages despite the low variation between them. Whilst hierBAPS is unable to utilise the low variant SNV information to define lineages, rPinecone can be used in such situations. We demonstrate this by defining sub-lineages in a LV population of *S.* Typhi H58 responsible for endemic typhoid fever in a rural part of Cambodia(4). Reanalysis of these data determined the population to have a median pairwise SNV distance of two. The subsequent phylogenetic tree of the reanalysis can best be described from root-to-tip (Fig. 3). At the root of the tree resides a "root group" of isolates, then a main branch composing of identical isolates which can be referred to as the "primary group" of the clonal expansion, followed by "diverging groups" of isolates and singletons which have diverged from this main branch. Using the reanalysed data, rPinecone identified 14 sub-lineages and a major lineage composed of two sub-lineages (seven isolates) of *S.* Typhi, using a SNV threshold of two and a relatability threshold of three (Fig. 3). 27 isolates were also identified as singletons. In comparison, rhierBAPS identified a maximum of nine sub-lineages at level three and offered no further resolution after further analysis (2, 7, 9, 9 and 9 sub-lineages were identified at levels one through to five [Supplementary Fig.1]).

229 rPinecone sub-lineages were nested within the sub-lineages defined by rhierBAPS. rPinecone also
230 acknowledges singleton isolates which have accumulated their own "private" SNVs. Therefore,
231 singletons that have individual SNVs may be excluded in the definition of sub-lineages. In contrast,
232 rhierBAPS cannot detect the private SNVs of singletons. This is because rhierBAPs seeks to maximise
233 the posterior probability of a sub-lineage assuming independence between SNV sites, and is less
234 suitable for the identification of singletons. In addition, rhierBAPS also identifies isolates situated
235 across the tree to be members of the same sub-lineage. hierBAPS and other similar programs like
236 Structure are not designed to separate population data that would be observed towards the tips of a
237 phylogenetic tree including those observed with a median pairwise SNV of two as observed here.
238 Rather they aim to cluster isolates into groups that are likely to come from similar source populations.
239 When there are few isolates and few SNVs there is little information to estimate the likely distribution
240 of allele frequencies in these source populations.

241

242 In this example, rPinecone provided greater resolution to distinguish sub-lineages than across five
243 different levels of rhierBAPS and distinguished additional sub-lineages. The original study noted their
244 sub-lineages to have significant geographical variation. The rPinecone sub-lineages provided a cross-
245 sectional view of the bacterial population, where some sub-lineages were geographically confined to
246 districts within one province and others span multiple provinces (Supplementary Fig. 2).

247

248 CONCLUSION

249 Investigators must be able to distinguish isolates to identify epidemiologically informative sub-
250 lineages, however this is difficult for bacterial populations associated with both outbreaks and
251 endemic disease over small temporal or geographical distances when the pathogen has a slow
252 mutation rate and appears largely "clonal". The ability to make this distinction is critical if WGS data
253 is to inform epidemiological investigation of such pathogens, whether in epidemic, or locally endemic
254 disease. When aiming to geographically locate hotspots of disease transmission, it is critical to be able
255 to incorporate associated geospatial data associated with samples.

256

257 rPinecone was designed to analyse LV populations and to be applied to any bacterial species with at
258 least a median SNV distance of two within the population. Using the genomic data of a MRSA hospital
259 outbreak and endemic *S.* Typhi, we demonstrate that rPinecone can be used to identify sub-lineages
260 within LV populations that are reflective of the phylogenetic data. Furthermore, when compared to
261 rhierBAPS, rPinecone identified additional sub-lineages. We highly recommend using heirBAPs for
262 initial analysis of phylogenetic data to understand the bacterial population. Once this has been done,
263 rPinecone will define sub-lineages of LV bacterial populations to set the platform for understanding
264 the transmission dynamics or geographical hotspots of disease.

265

266 ---

267 AUTHOR STATEMENTS

Funding information

This project and A.M.W (Postdoctoral Fellow) was supported by a grant from the Bill and Melinda Gates Foundation (OPP1128444). The Wellcome Trust Sanger Institute is core funded by Wellcome Trust grant 206194. JC was funded by ERC grant no. 742158. F.C. is funded by a Wellcome Trust Sir Henry Postdoctoral Fellowship (201344/Z/16/Z).

ABBREVIATIONS

LV: Low-Variant

SNV: Single Nucleotide Variant

WGS: whole genome sequencing

ML: Maximum-Likelihood

JAR: Joint Ancestral Reconstruction

REFERENCES

1. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945-59.

2. Corander J, Waldmann P, Sillanpaa MJ. Bayesian analysis of genetic differentiation between populations. Genetics. 2003;163(1):367-74.

3. Cheng L, Connor TR, Siren J, Aanensen DM, Corander J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Mol Biol Evol. 2013;30(5):1224-8.

4. Pham Thanh D, Thompson CN, Rabaa MA, Sona S, Sopheary S, Kumar V, et al. The Molecular and Spatial Epidemiology of Typhoid Fever in Rural Cambodia. PLoS Negl Trop Dis. 2016;10(6):e0004785.

5. Harris SR, Cartwright EJ, Torok ME, Holden MT, Brown NM, Ogilvy-Stuart AL, et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. Lancet Infect Dis. 2013;13(2):130-6.

6. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

7. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom. 2016;2(4):e000056.

8. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-3.

9. Wong VK, Baker S, Connor TR, Pickard D, Page AJ, Dave J, et al. An extended genotyping framework for *Salmonella enterica* serovar Typhi, the cause of human typhoid. Nat Commun. 2016;7:12827.

10. Pupko T, Pe'er I, Shamir R, Graur D. A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol Biol Evol. 2000;17(6):890-6.

11. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20(2):289-90.

12. Prosperi MC, Ciccozzi M, Fanti I, Saladini F, Pecorari M, Borghi V, et al. A novel methodology for large-scale phylogeny partition. Nat Commun. 2011;2:321.

13. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Research. 2016;44(W1):W242-W5.

DATA BIBLIOGRAPHY

1. Holden, M.T. et al., *Staphylococcus aureus* subsp. *aureus* str. HO 5096 0412, 2015, accession number HE681097.

2. Pham Thanh, D et al., *Salmonella enterica* subsp. *enterica* serovar Typhi str. Mal1017142, 2016, accession number ERR279139.

3. Pham Thanh, D et al., *Salmonella enterica* subsp. *enterica* serovar Typhi str. 2010_7898, 2016, accession number GCA_001360555.

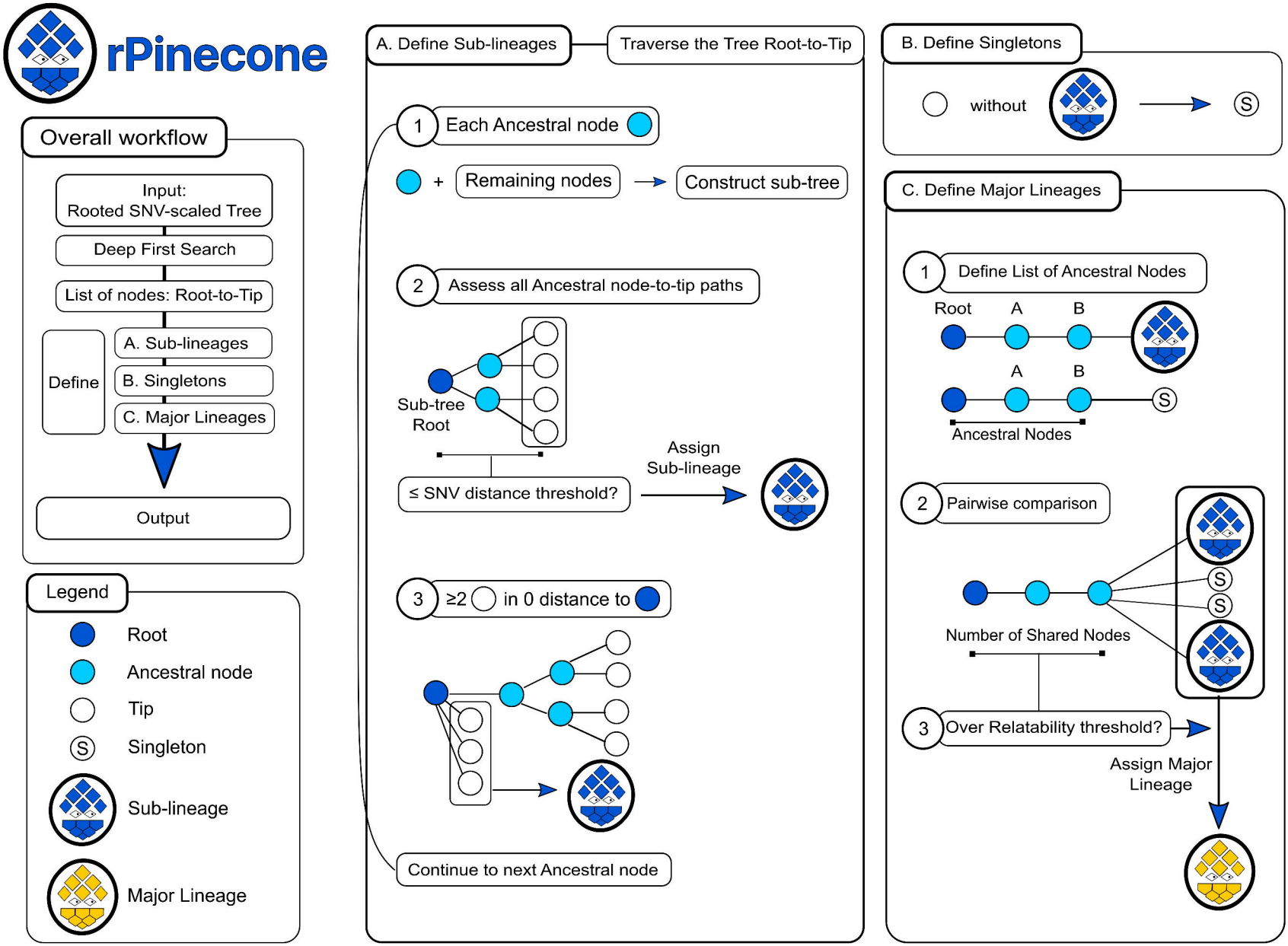
FIGURES AND TABLES

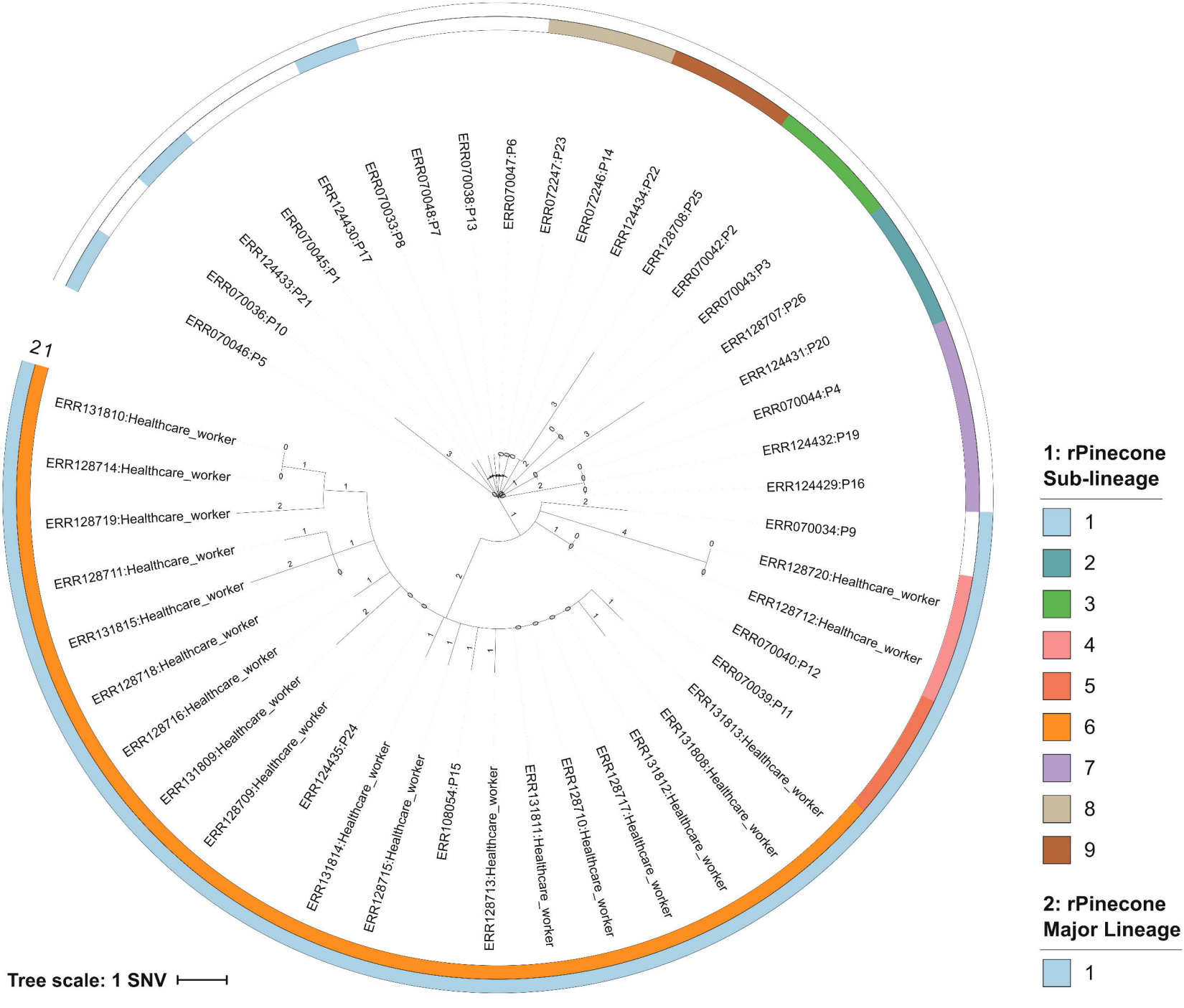
Figure 1. The rPinecone package evaluates a phylogenetic tree in a root-to-tip approach and defines sub-lineages according to single nucleotide variants (SNV) distances from ancestral nodes. (A) Define Sub-lineages. The rPinecone algorithm traverses the tree from root-to-tip in the order of nodes listed by the depth-first search. At each ancestral node, a sub-tree is constructed containing the ancestral and remaining nodes. The ancestral node becomes the root of the sub-tree and all root-to-tip paths are assessed to identify the maximum root-to-tip branch distance. If this distance is below or equal to

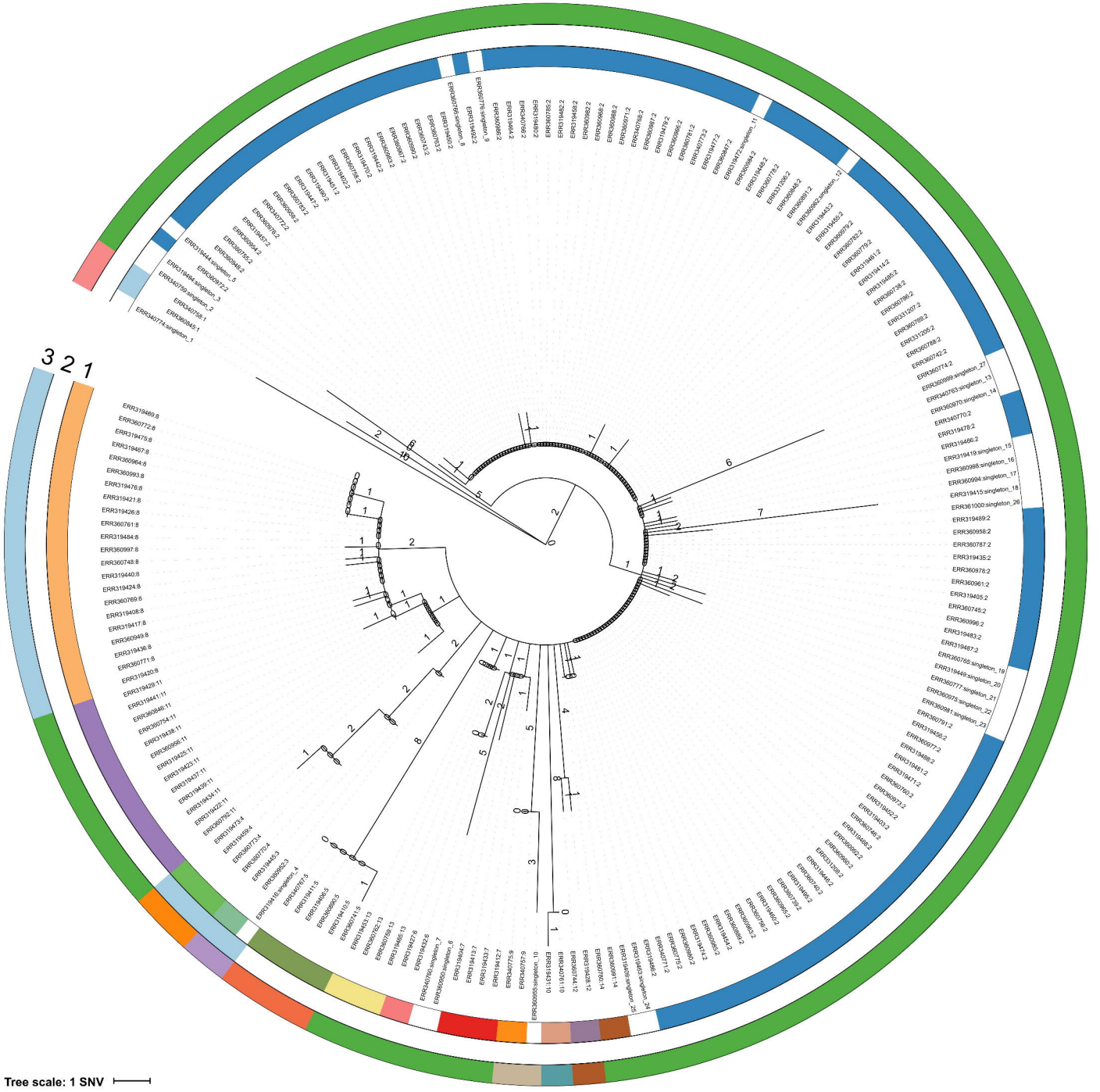
337 the user specified SNV distance threshold, the tips of the sub-tree are assigned to a sub-lineage
338 number. If the distance exceeds the threshold and two or more tips are in zero distance to the root of
339 the sub-tree, only the tips in zero distance will be assigned a sub-lineage number. The process is
340 repeated throughout the entire tree. (B) Define Singletons. Tips which have not been assigned a sub-
341 lineage number are assigned a singleton number. (C) Define Major lineages. Using the entire tree, the
342 list of ancestral nodes from the root to each sub-lineage and singleton is compiled. A pairwise
343 comparison of these lists is performed. Sub-lineages and singletons which have the same number of
344 ancestral nodes over the user specified relatability threshold will be defined as part of a major lineage.
345 The root, ancestral nodes, and tips of the tree are denoted by navy, cyan and white circles respectively.
346 Singletons, sub-lineages, major lineages are represented by a 'S' labelled circle, a blue pinecone, a gold
347 pinecone.

348 Figure 2. Rooted accelerated transformation tree based on whole genome sequencing of 45 MRSA
349 ST2371 isolates. Outgroup rooted using *S. aureus* ST22 isolate HO 5096. Tips are labelled by the
350 sample's ERR number followed by the origin of sample, patient (P) or Healthcare worker. Branch
351 lengths are single nucleotide variants scaled. The inner and outer coloured rings correspond to the
352 sub-lineages and, major lineage identified by rPinecone.

353 Figure 3. Rooted joint ancestral reconstruction tree based on whole genome sequencing of 203 *S*.
354 Typhi H58 isolates. Outgroup rooted using *S.* Typhi Genotyphi 4.1.1 isolate Mal1017142. Tips are
355 labelled by the sample's ERR number and rPinecone sub-lineage. Branch lengths are single
356 nucleotide variants scaled. Colour rings one and two correspond to the sub-lineages and major
357 lineage defined by rPinecone. The colour ring three corresponds to clusters defined by rhierBAPs
358 level three.


rPinecone
Overall workflow
Input:
Rooted SNV-scaled Tree
Deep First Search
List of nodes: Root-to-Tip
Define
A. Sub-lineages
B. Singletons
C. Major Lineages
Output
Legend
Root
Ancestral node
Tip
Singleton
Sub-lineage
Major Lineage
A. Define Sub-lineages
Traverse the Tree Root-to-Tip
1
Each Ancestral node
+
Remaining nodes
Construct sub-tree
2
Assess all Ancestral node-to-tip paths
Sub-tree
Root
≤ SNV distance threshold?
Assign
Sub-lineage
3
≥2
in 0 distance to
Continue to next Ancestral node
B. Define Singletons
without
S
C. Define Major Lineages
1
Define List of Ancestral Nodes
Root
A
B
A
B
S
Ancestral Nodes
2
Pairwise comparison
S
S
Number of Shared Nodes
3
Over Relatability threshold?
Assign Major
Lineage


21
ERR131810:Healthcare_worker
ERR128714:Healthcare_worker
ERR128719:Healthcare_worker
ERR128711:Healthcare_worker
ERR131815:Healthcare_worker
ERR128718:Healthcare_worker
ERR128716:Healthcare_worker
ERR131809:Healthcare_worker
ERR128709:Healthcare_worker
ERR124435:P24
ERR131814:Healthcare_worker
ERR128715:Healthcare_worker
ERR108054:P15
ERR128713:Healthcare_worker
ERR131811:Healthcare_worker
ERR128710:Healthcare_worker
ERR128717:Healthcare_worker
ERR131812:Healthcare_worker
ERR131808:Healthcare_worker
ERR131813:Healthcare_worker
ERR070039:P11
ERR070040:P12
ERR128712:Healthcare_worker
ERR128720:Healthcare_worker
ERR070034:P9
ERR124429:P16
ERR124432:P19
ERR070044:P4
ERR124431:P20
ERR128707:P26
ERR070043:P3
ERR070042:P2
ERR128708:P25
ERR124434:P22
ERR072246:P14
ERR072247:P23
ERR070047:P6
ERR070038:P13
ERR070048:P7
ERR070033:P8
ERR124430:P17
ERR070045:P1
ERR124433:P21
ERR070036:P10
ERR070046:P5
Tree scale: 1 SNV
1: rPinecone Sub-lineage
1
2
3
4
5
6
7
8
9
2: rPinecone Major Lineage
1



1: rPinecone Sub-lineage

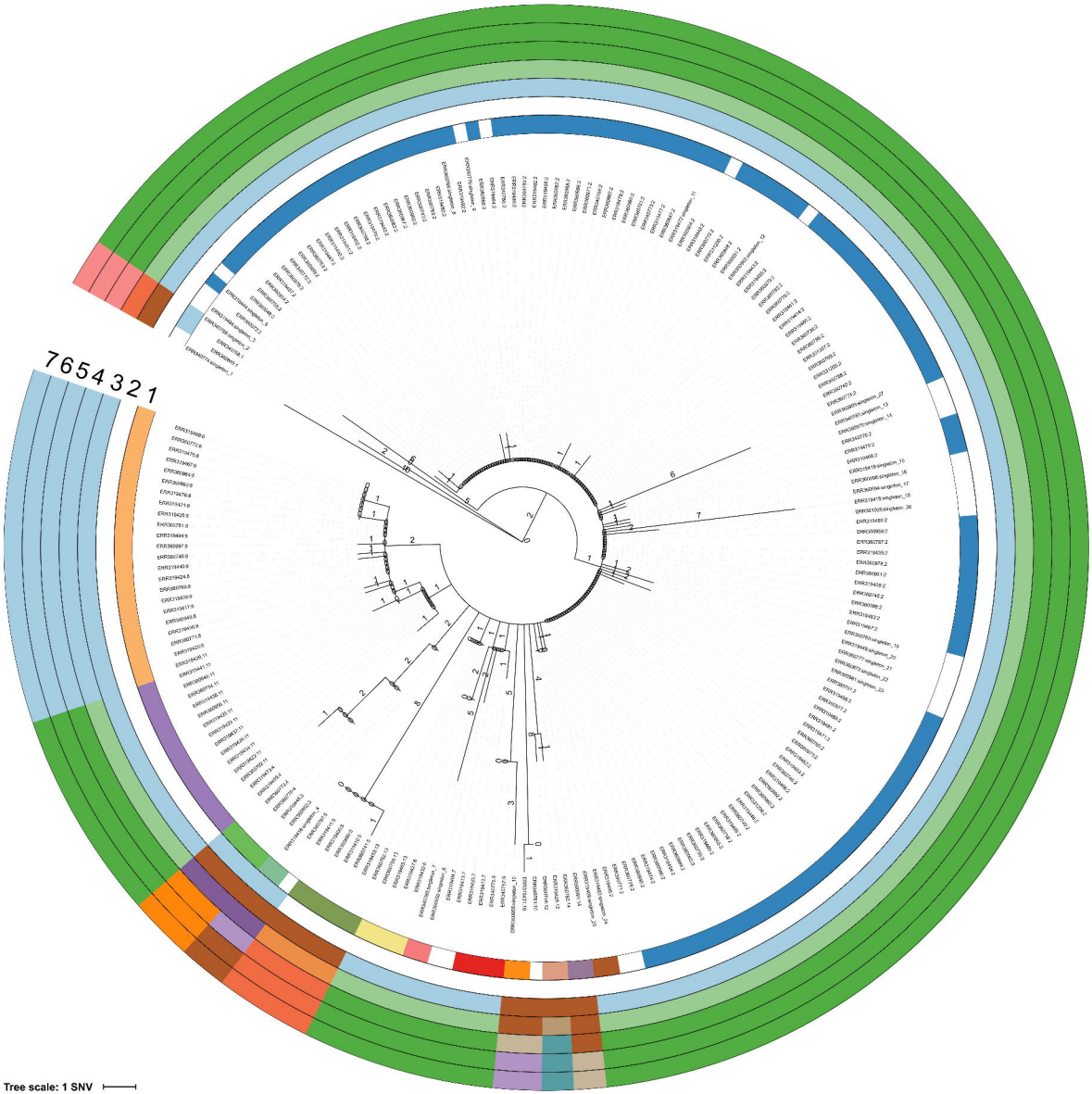
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14

2: rPinecone Major Lineage

- 1

3: rhierBAPS level 3

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9



1: rPinecone Sub-lineage

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14

2: rPinecone Major Lineage

- 1

3: rhierBAPS level 1

- 1
- 2

4: rhierBAPS level 2

- 1
- 2
- 3
- 4
- 5
- 6
- 7

5: rhierBAPS level 3

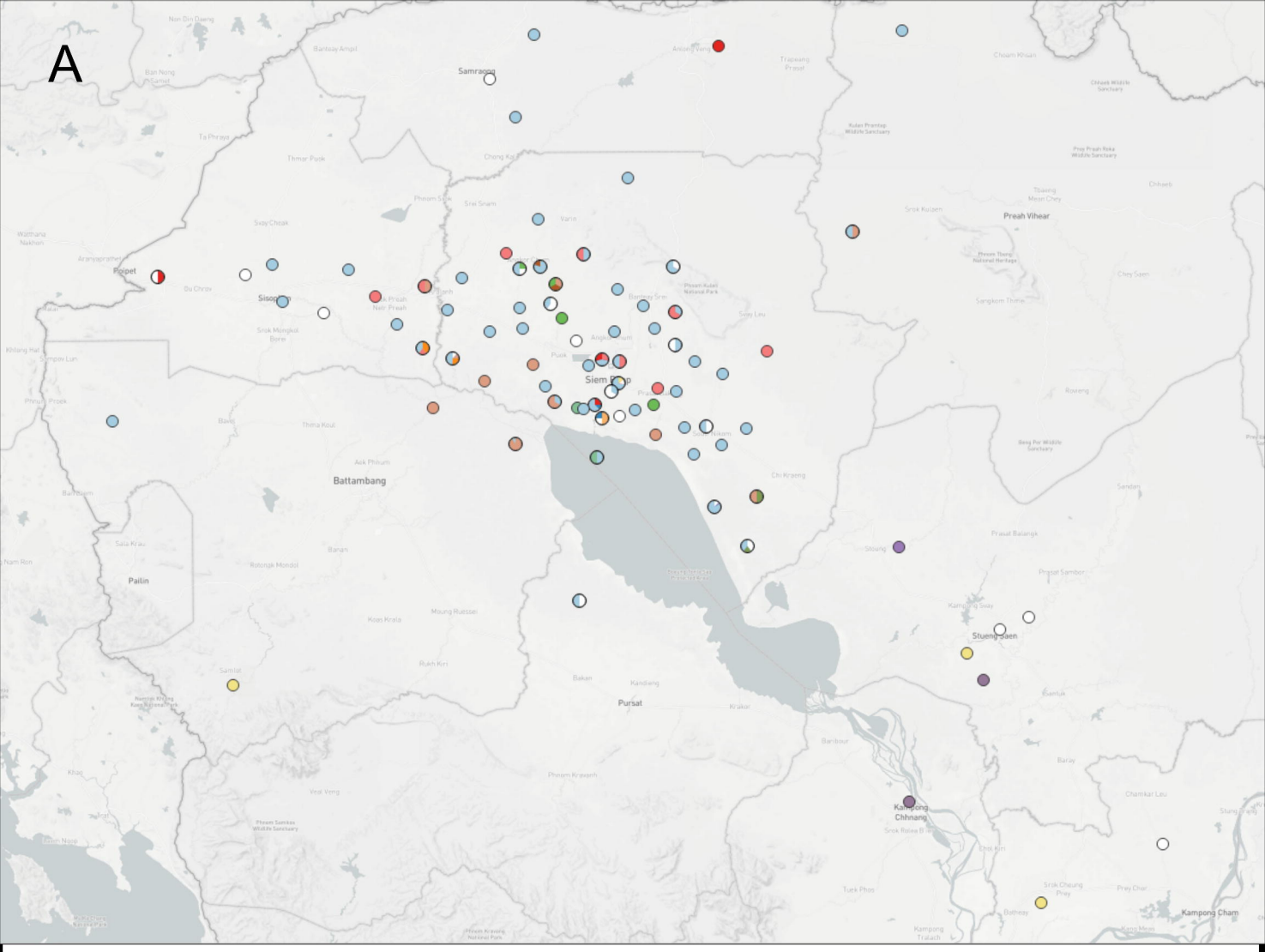
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9

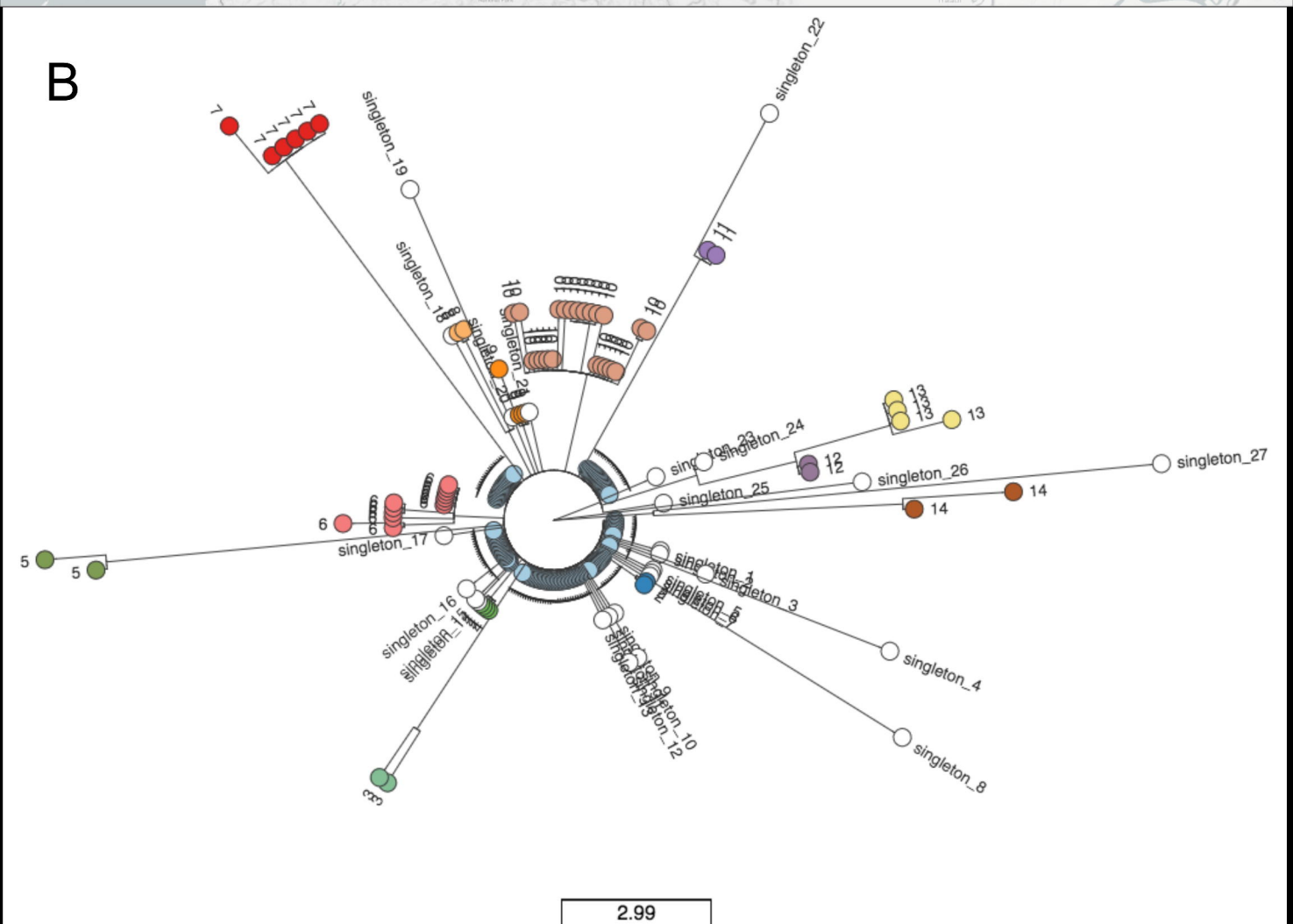
6: rhierBAPS level 4

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9

7: rhierBAPS level 5

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9


A
Samraong
Poipet
Sisophon
Siem Reap
Battambang
Pailin
Pursat
Preah Vihear
Stueng Saen
Kampong Chhnang
Kampong Cham


B
singleton_22
singleton_19
singleton_18
singleton_24
singleton_25
singleton_26
singleton_27
singleton_17
singleton_16
singleton_4
singleton_8
2.99