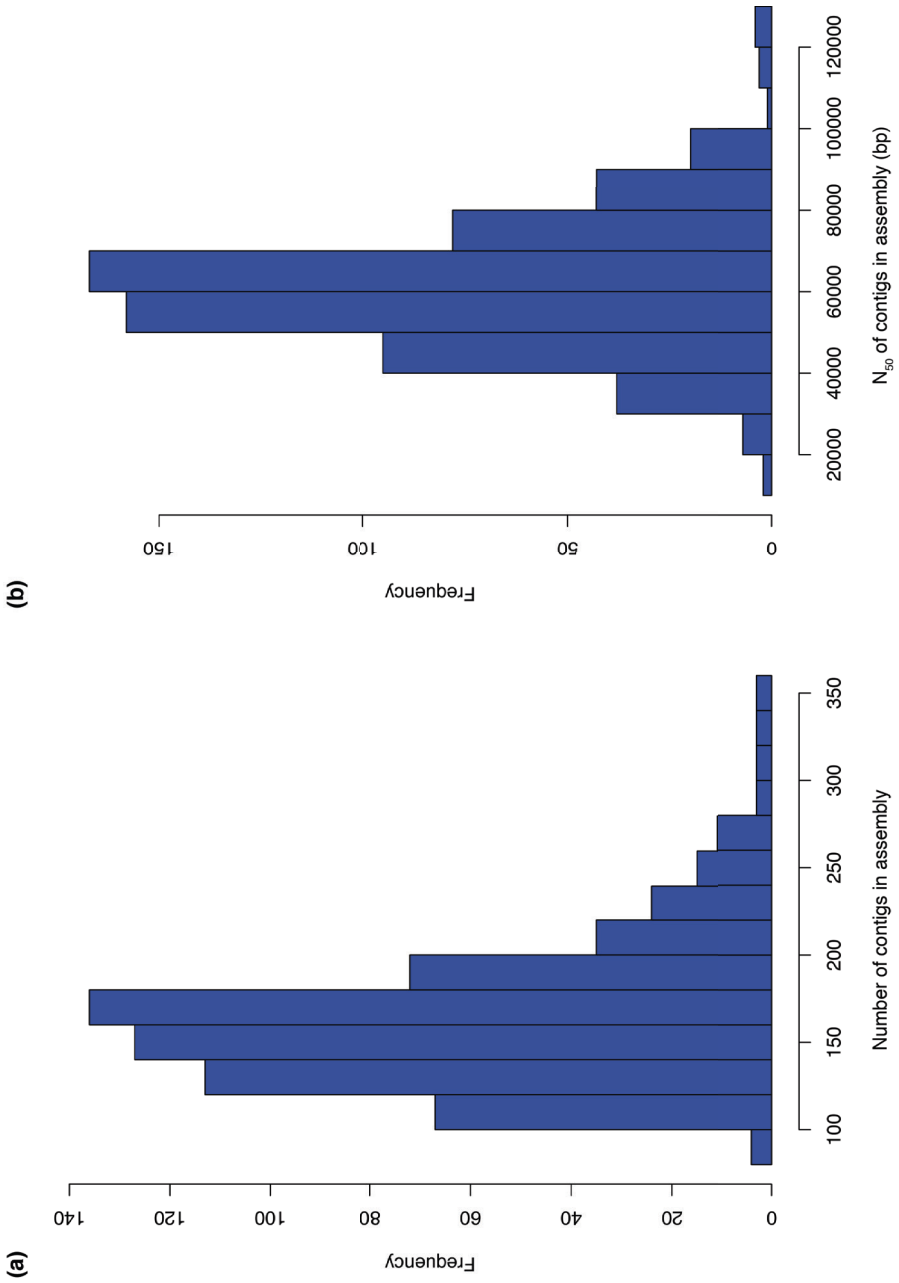
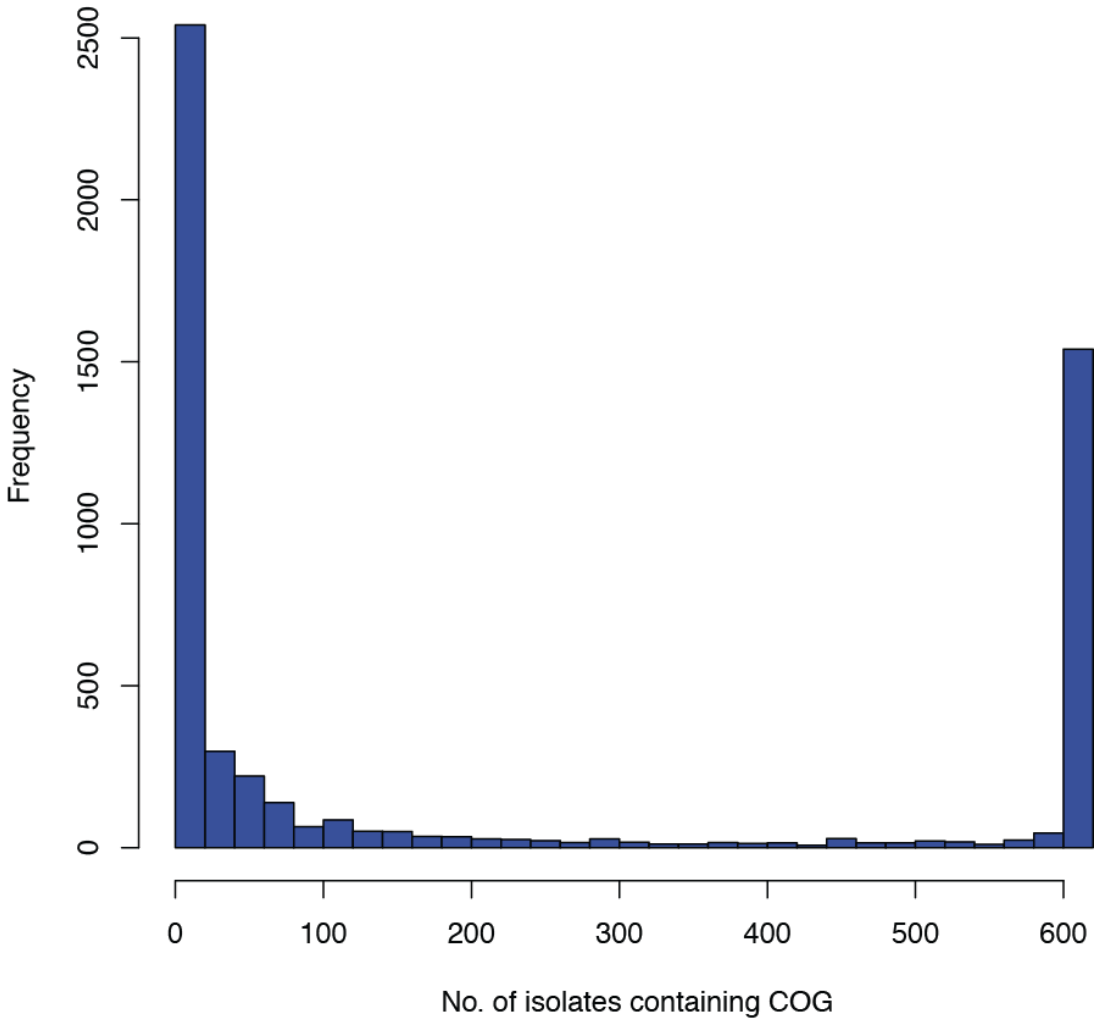
Population genomics of post-vaccine changes in pneumococcal epidemiology

Nicholas J. Croucher, Jonathan A. Finkelstein, Stephen I. Pelton, Patrick K. Mitchell, Grace M. Lee, Julian Parkhill, Stephen D. Bentley, William P. Hanage & Marc Lipsitch

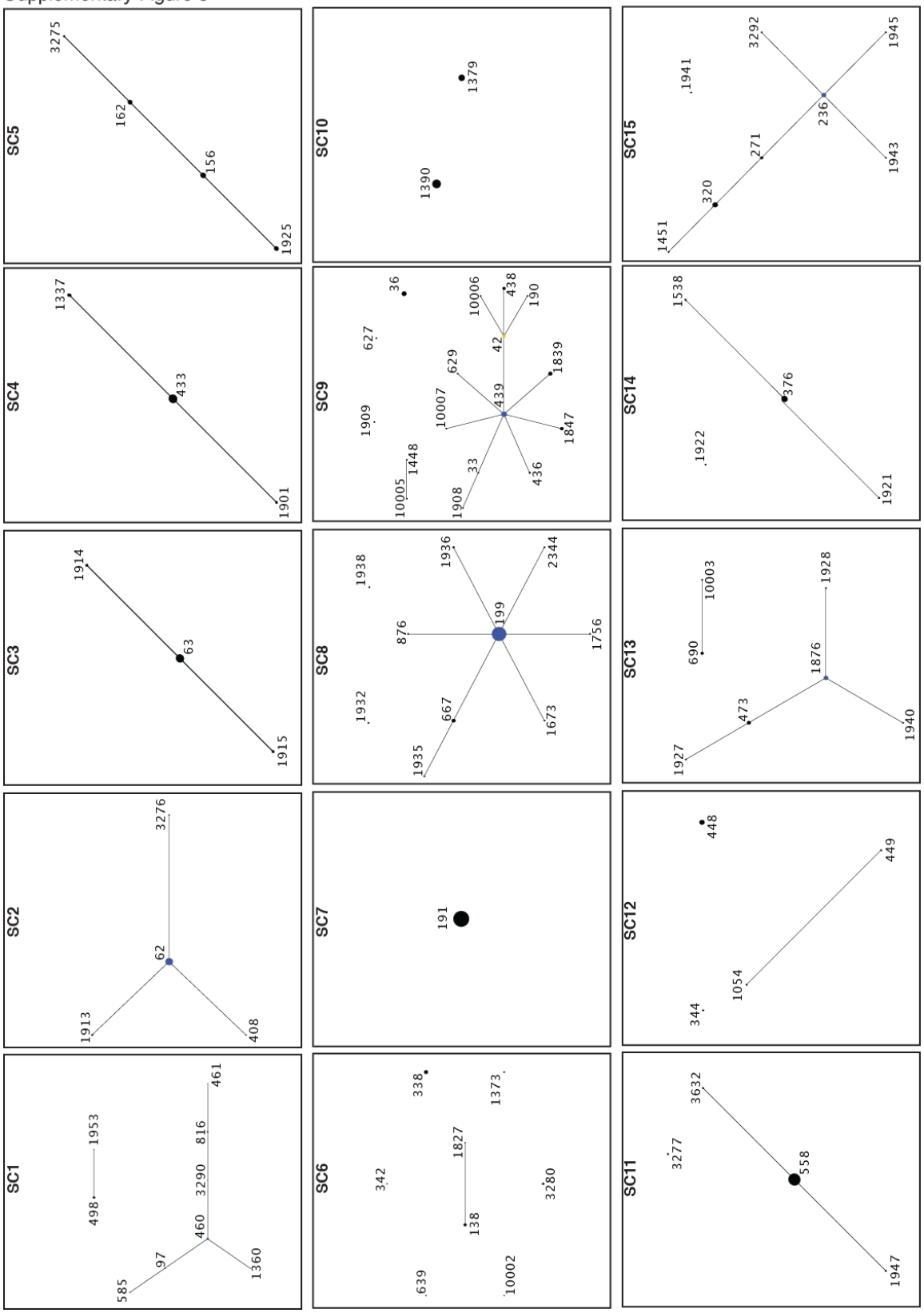


Supplementary Figure 1 Summary statistics of *de novo* assemblies. For each of the 616 assemblies, the distribution of (a) the total number of contigs and (b) the N_{50} of the contigs are shown as histograms.

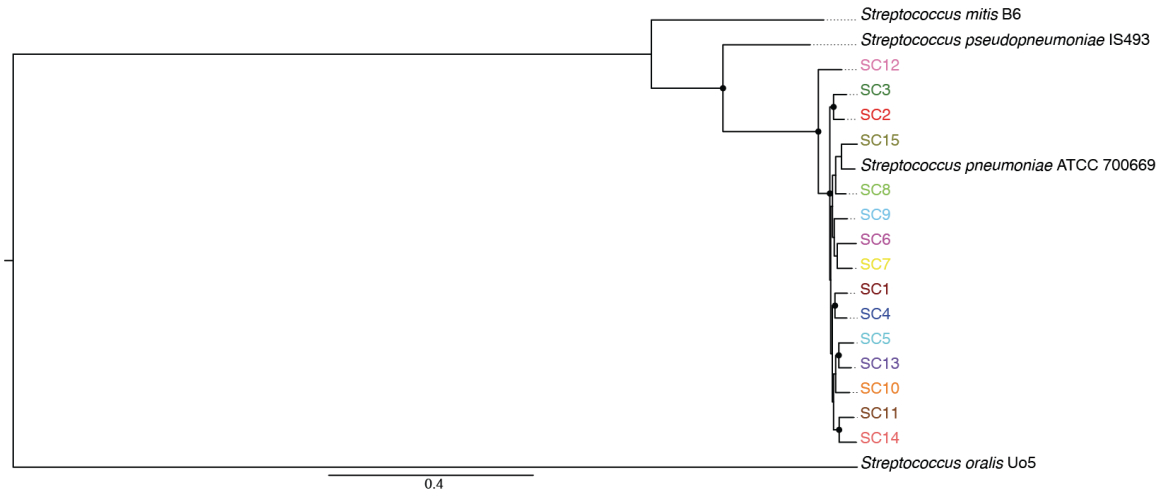


Supplementary Figure 2 Distribution of COGs in the pneumococcal population. This histogram shows the breakdown of COGs by the number of isolates in which they are found. Around 1,500 are present in almost all strains, in accordance with previous estimates of the pneumococcal core genome size. By contrast, the majority of COGs are found in relatively few isolates.

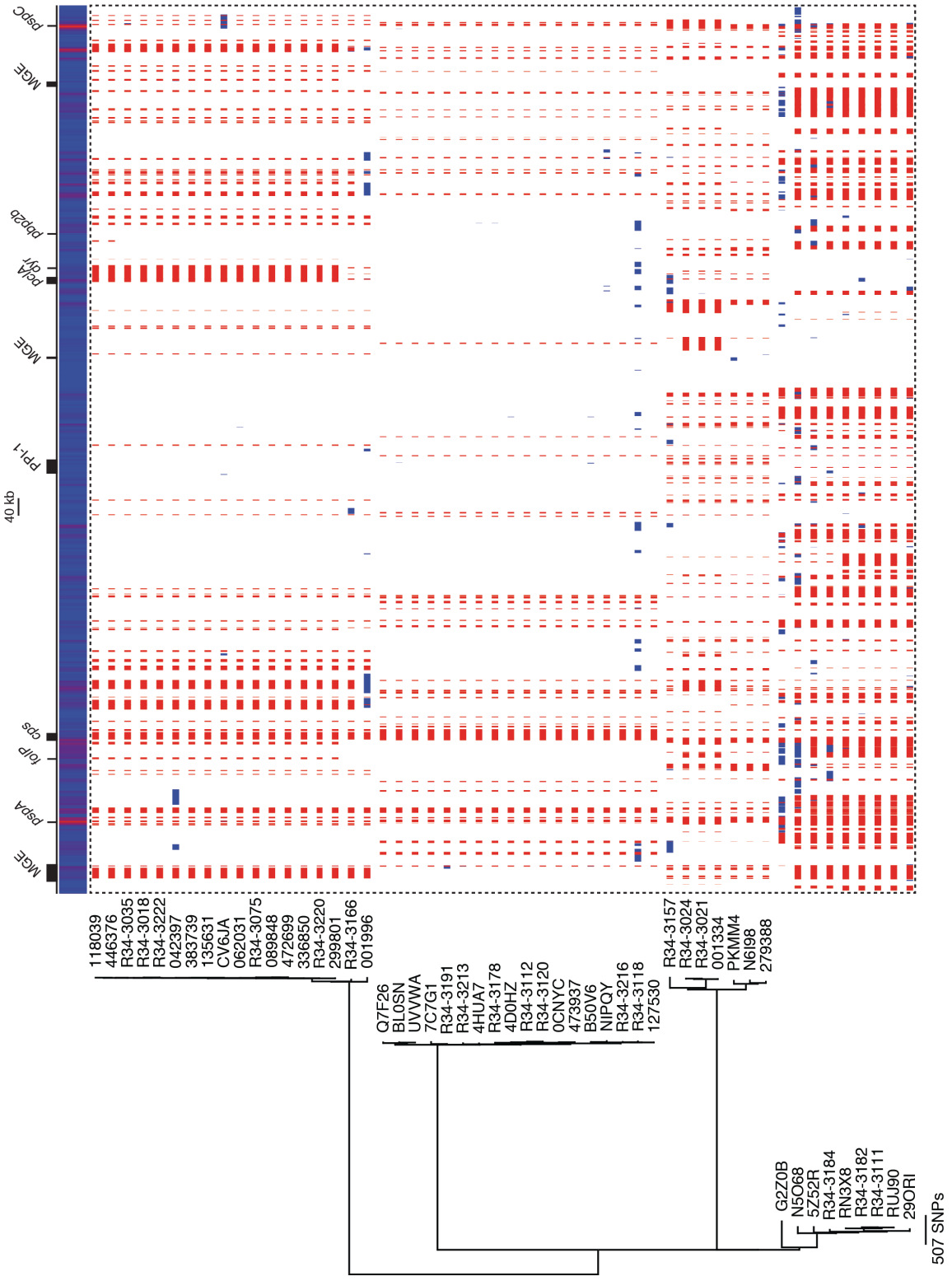
Supplementary Figure 3



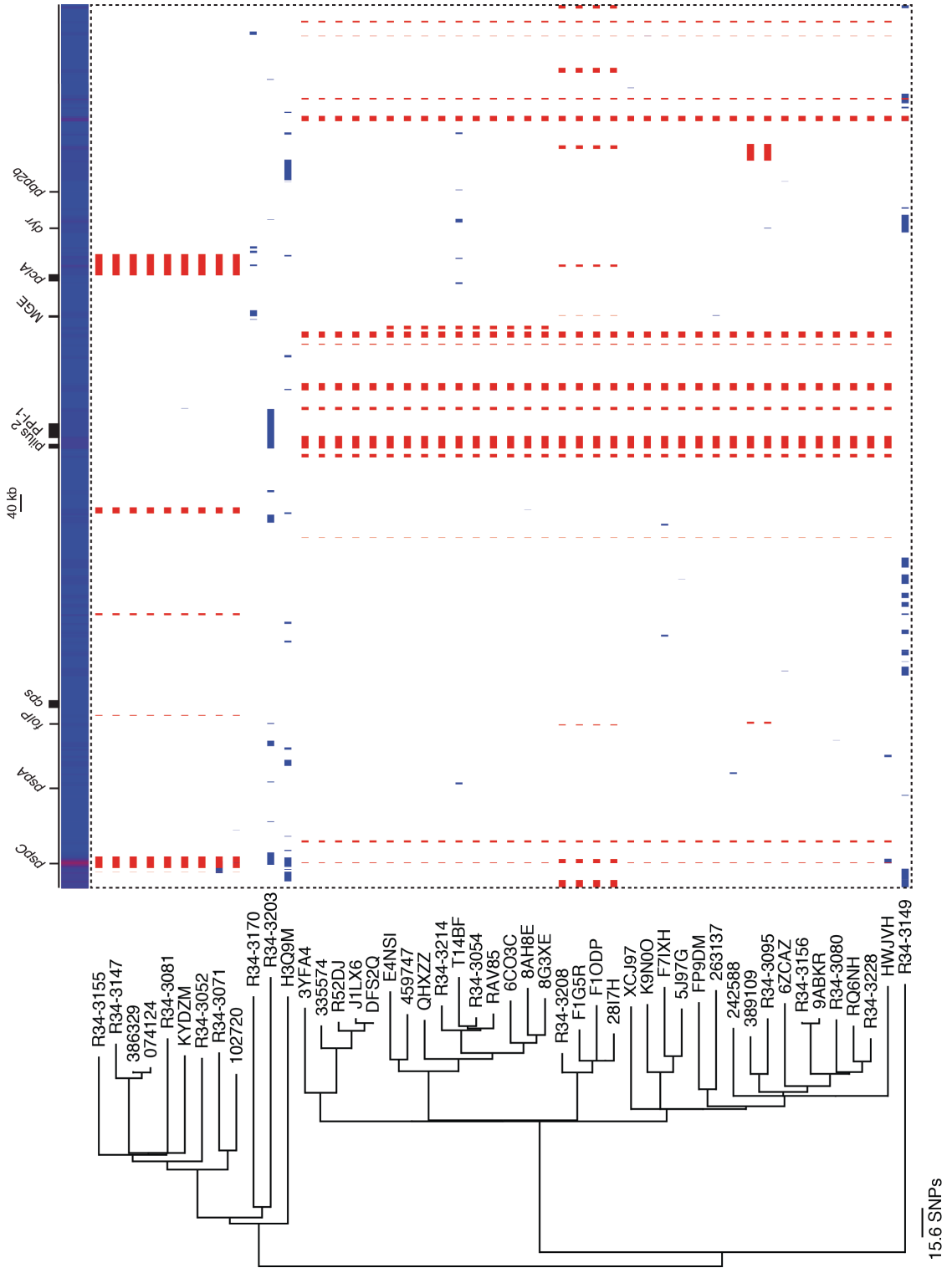
Supplementary Figure 3 Congruence of the whole genome phylogeny with MLST. For each of the fifteen monophyletic sequence clusters, the relationships between the constituent isolates was analyzed by using eBURST to process the sequences of multilocus sequence typing alleles extracted from the Illumina sequence reads. Sequence types represented by numbers over 10000 indicate genotypes not present in the MLST database.



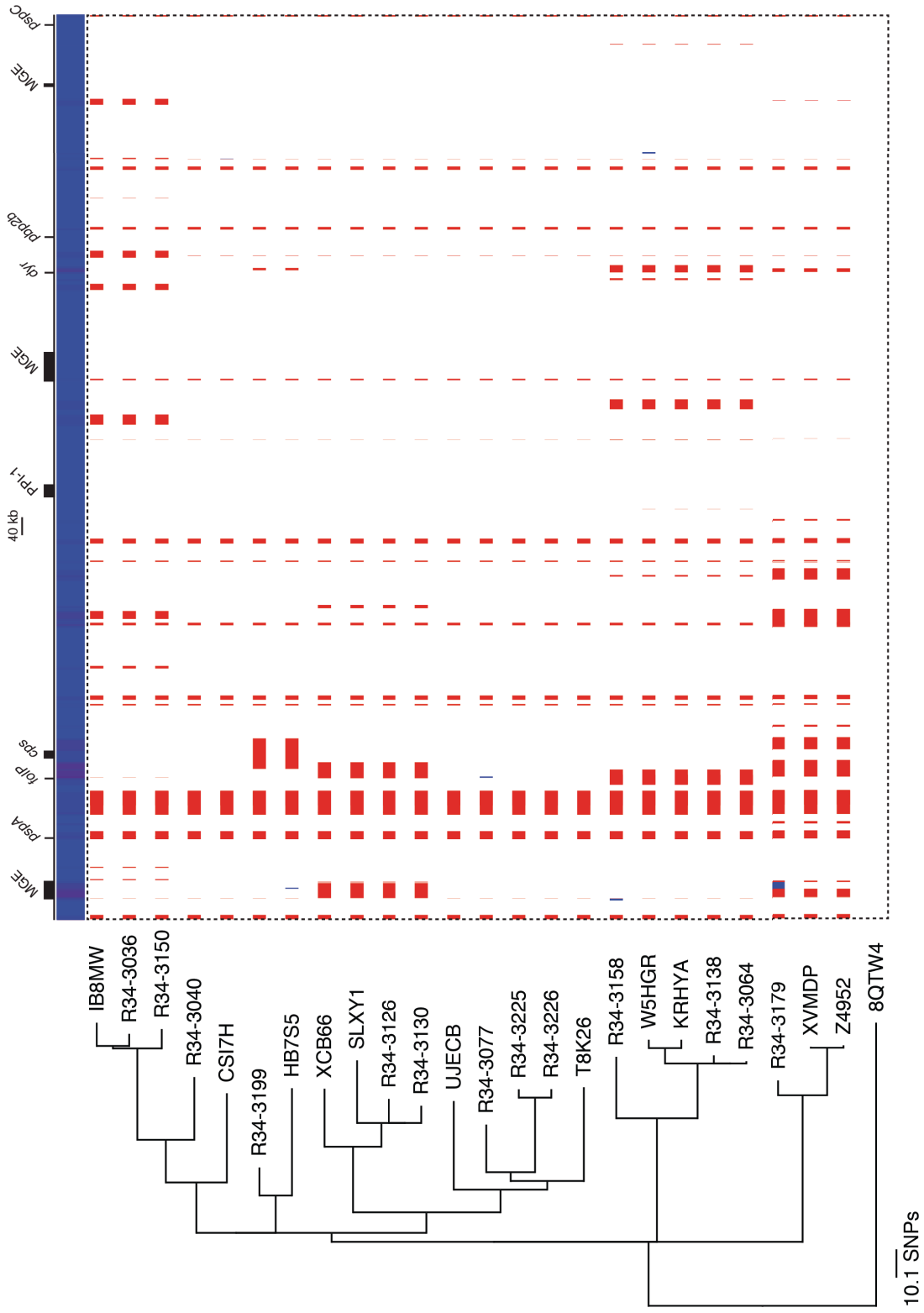
Supplementary Figure 4 Relationships between representatives of mitis group streptococci. This maximum likelihood phylogeny represents the outcome of applying the same analysis used to compare all isolates sequenced in this study to a smaller sample of more diverse genomes. The reference genome for each of the monophyletic sequence clusters, along with a complete pneumococcal genome sequence (*S. pneumoniae* ATCC 700669; EMBL accession code FM211187), were compared to *S. pseudopneumoniae* IS493 (EMBL accession code CP002925), *S. mitis* B6 (EMBL accession code FN568063) and *S. oralis* Uo5 (EMBL accession code FR720602). This lead to the identification of 1,149 core COGs, generating a 1.10 Mb alignment containing 216,225 polymorphic sites. The nodes of the phylogeny supported by a bootstrap value of 95 or greater are indicated by filled circles. This shows the unencapsulated SC12 isolates appear to lie outside the main clade of pneumococci, thereby informing the rooting of the tree in Figure 1.



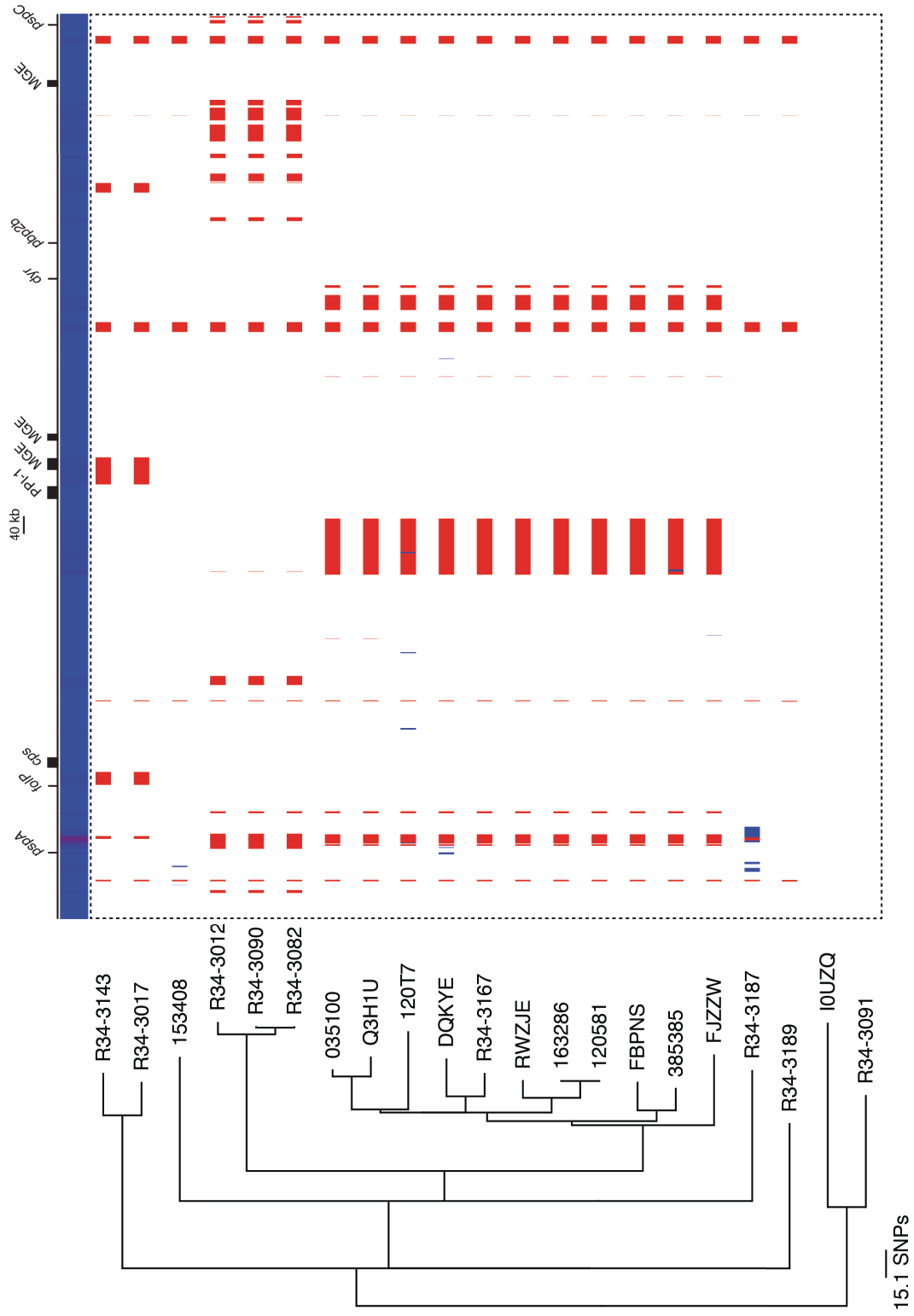
Supplementary Figure 5 Maximum likelihood phylogeny of SC1, displayed as described in Figure 3.



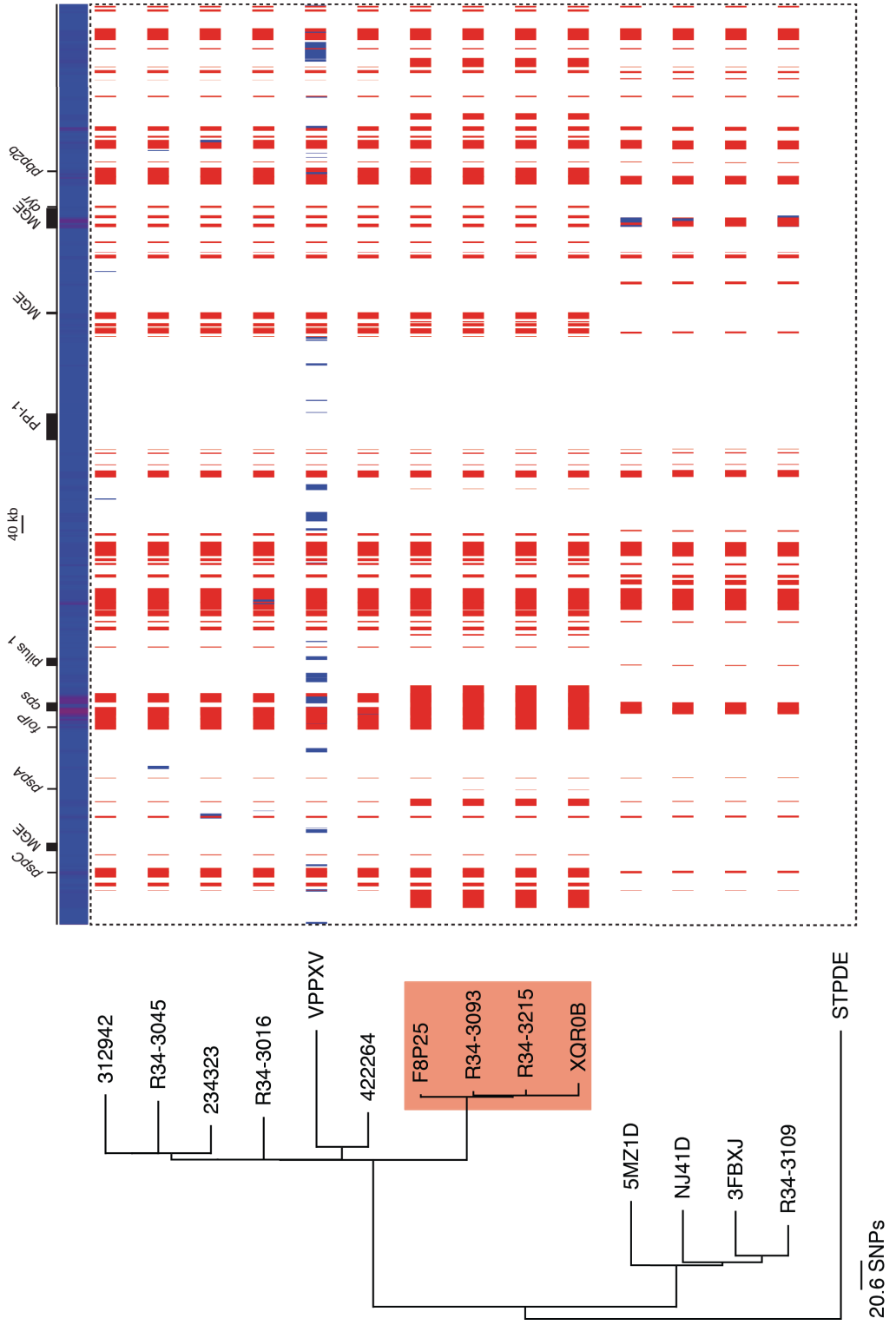
Supplementary Figure 6 Maximum likelihood phylogeny of SC2, displayed as described in Figure 3.



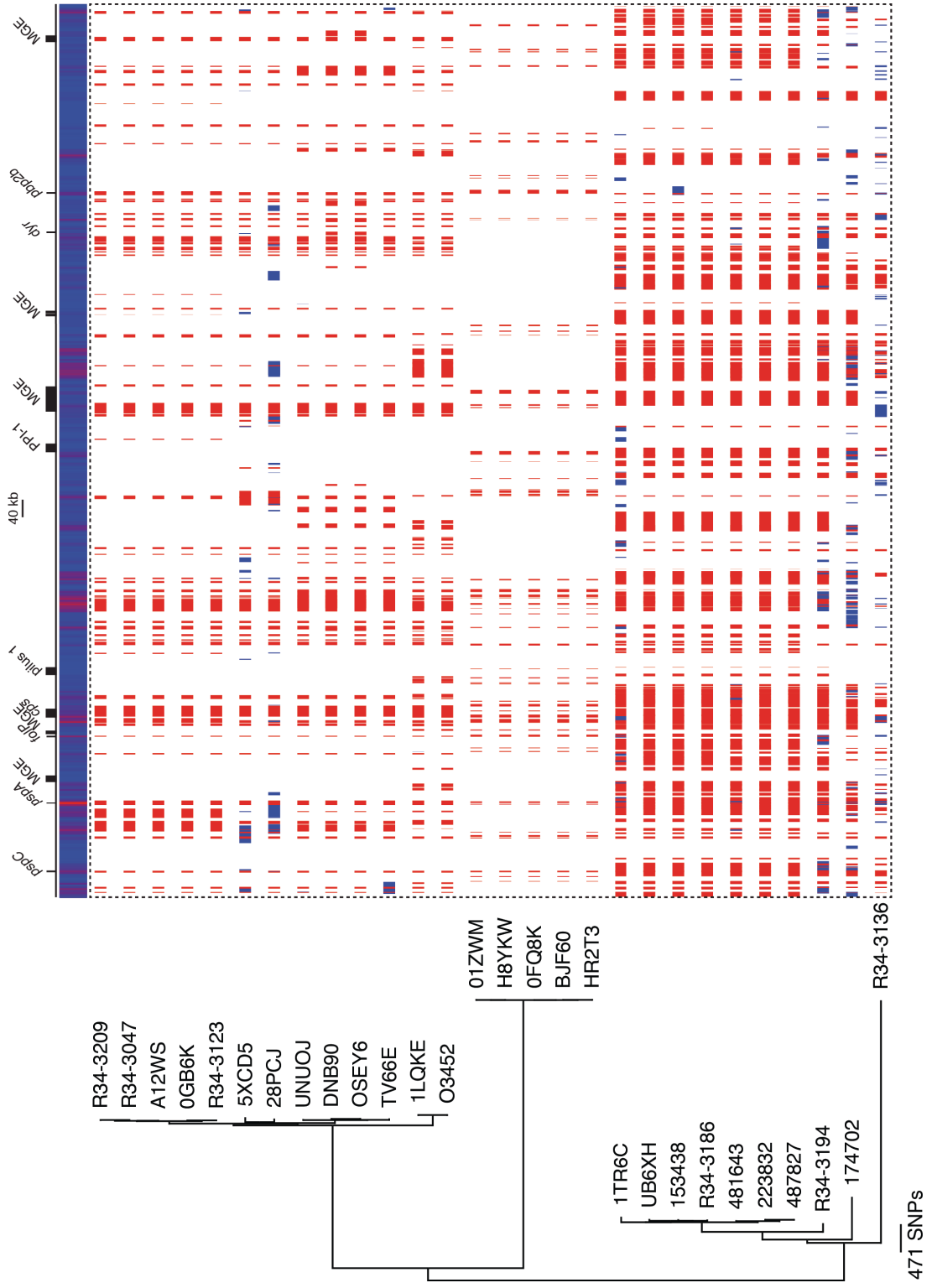
Supplementary Figure 7 Maximum likelihood phylogeny of SC3, displayed as described in Figure 3.



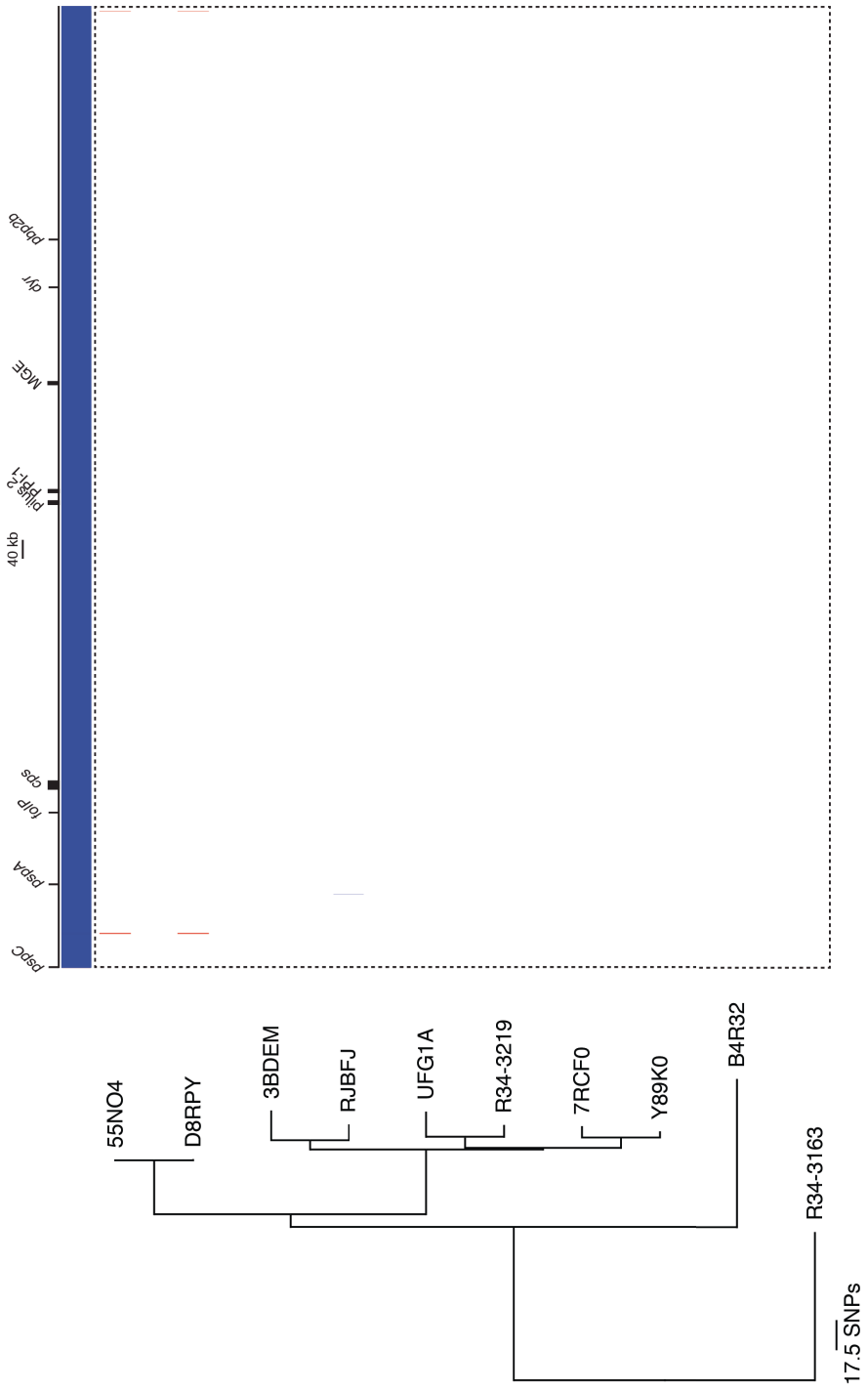
Supplementary Figure 8 Maximum likelihood phylogeny of SC4, displayed as described in Figure 3.



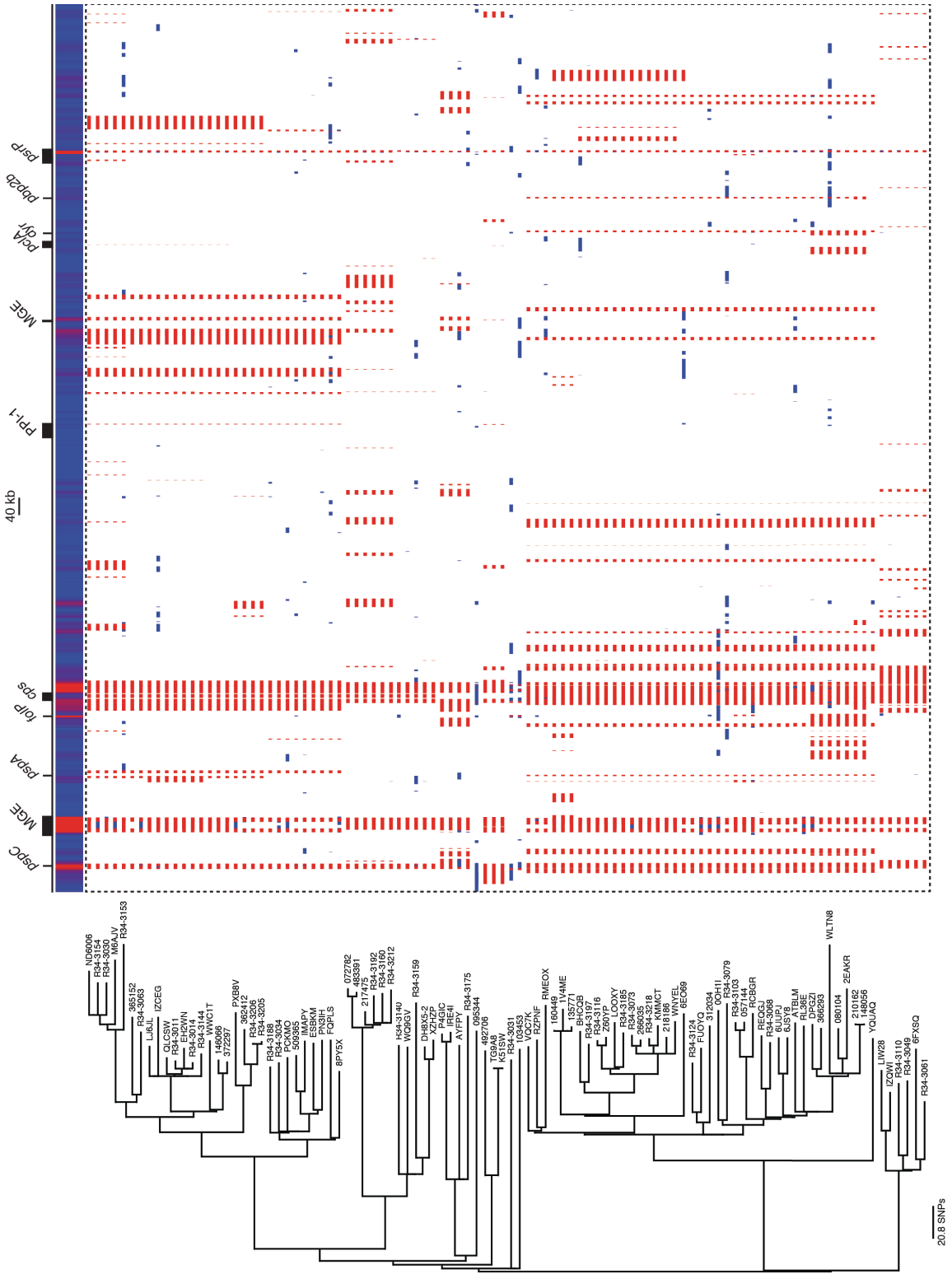
Supplementary Figure 9 Maximum likelihood phylogeny of SC5, displayed as described in Figure 3. Serotype 19A variants of the sequence cluster are indicated by the pink box.



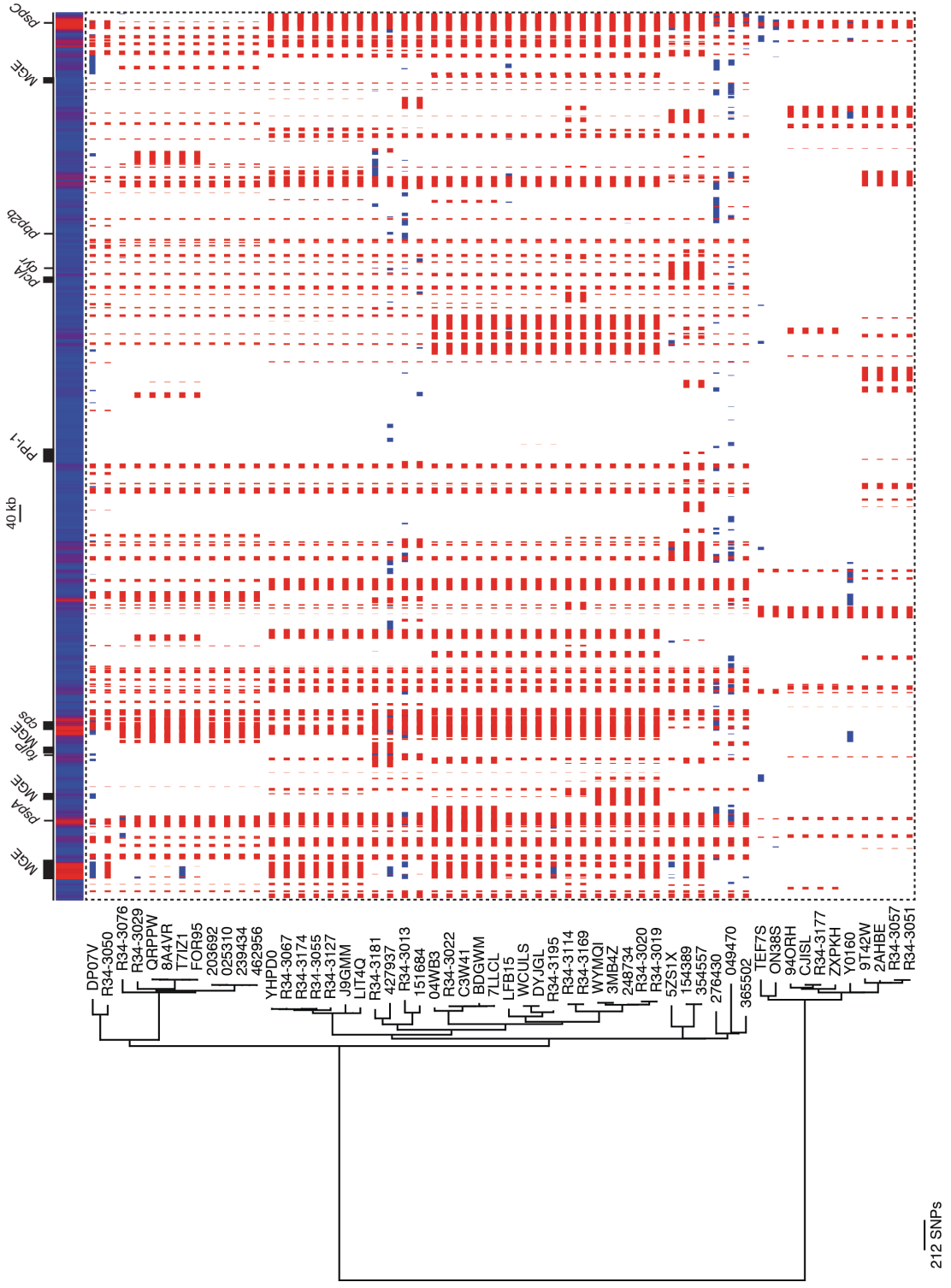
Supplementary Figure 10 Maximum likelihood phylogeny of SC6, displayed as described in Figure 3.



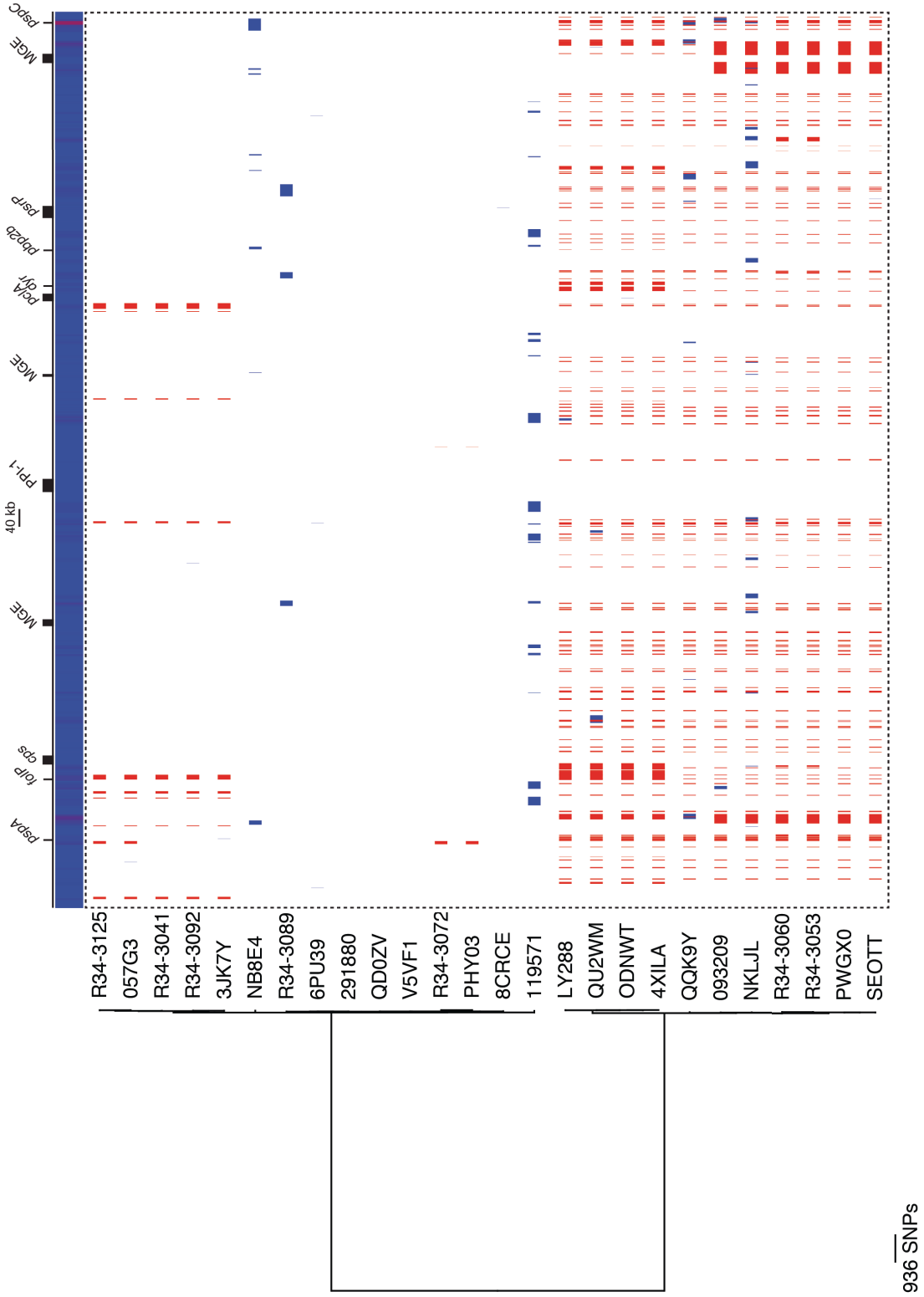
Supplementary Figure 11 Maximum likelihood phylogeny of SC7, displayed as described in Figure 3**.**



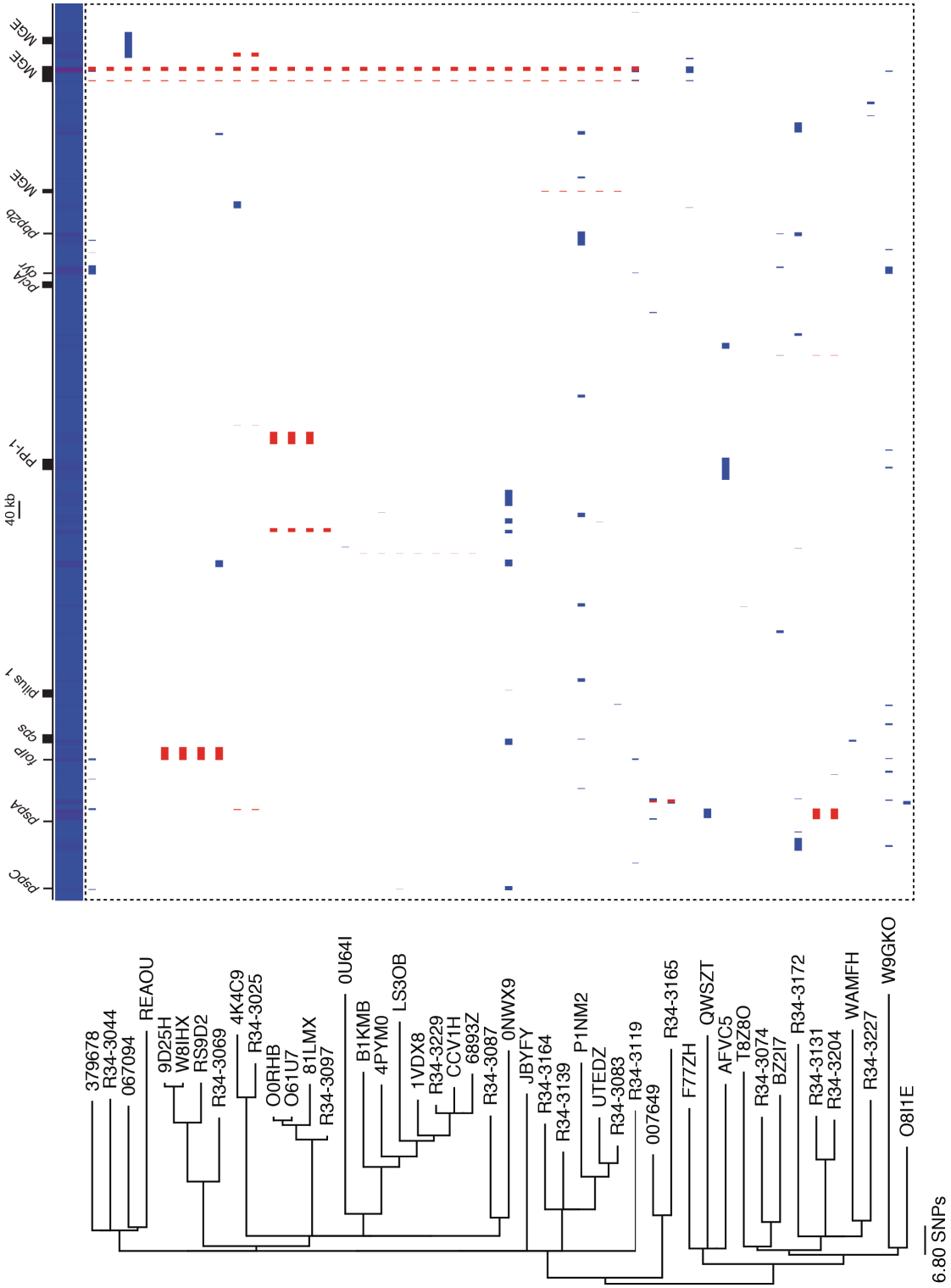
Supplementary Figure 12 Maximum likelihood phylogeny of SC8, displayed as described in Figure 3.



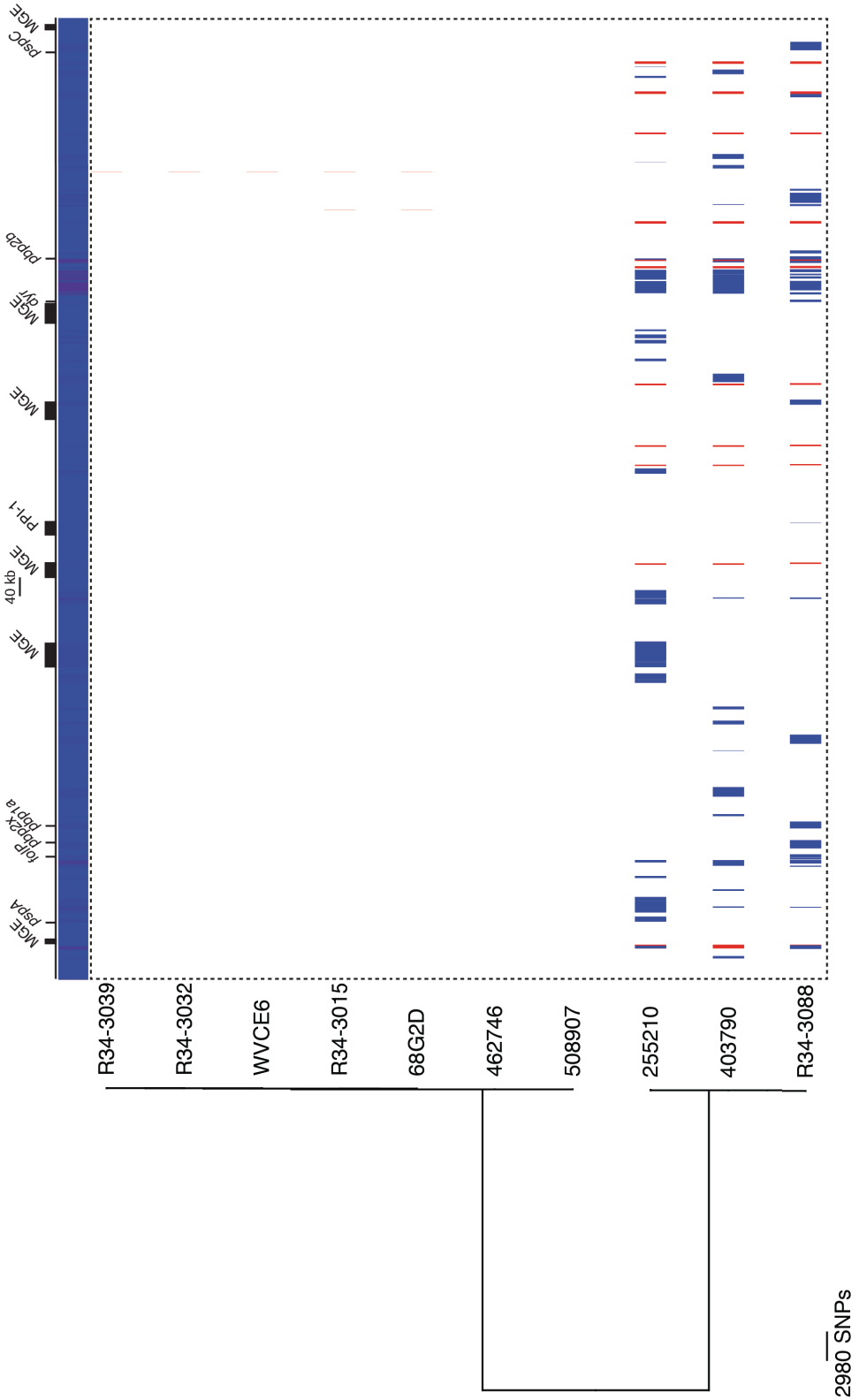
Supplementary Figure 13 Maximum likelihood phylogeny of SC9, displayed as described in Figure 3**.**



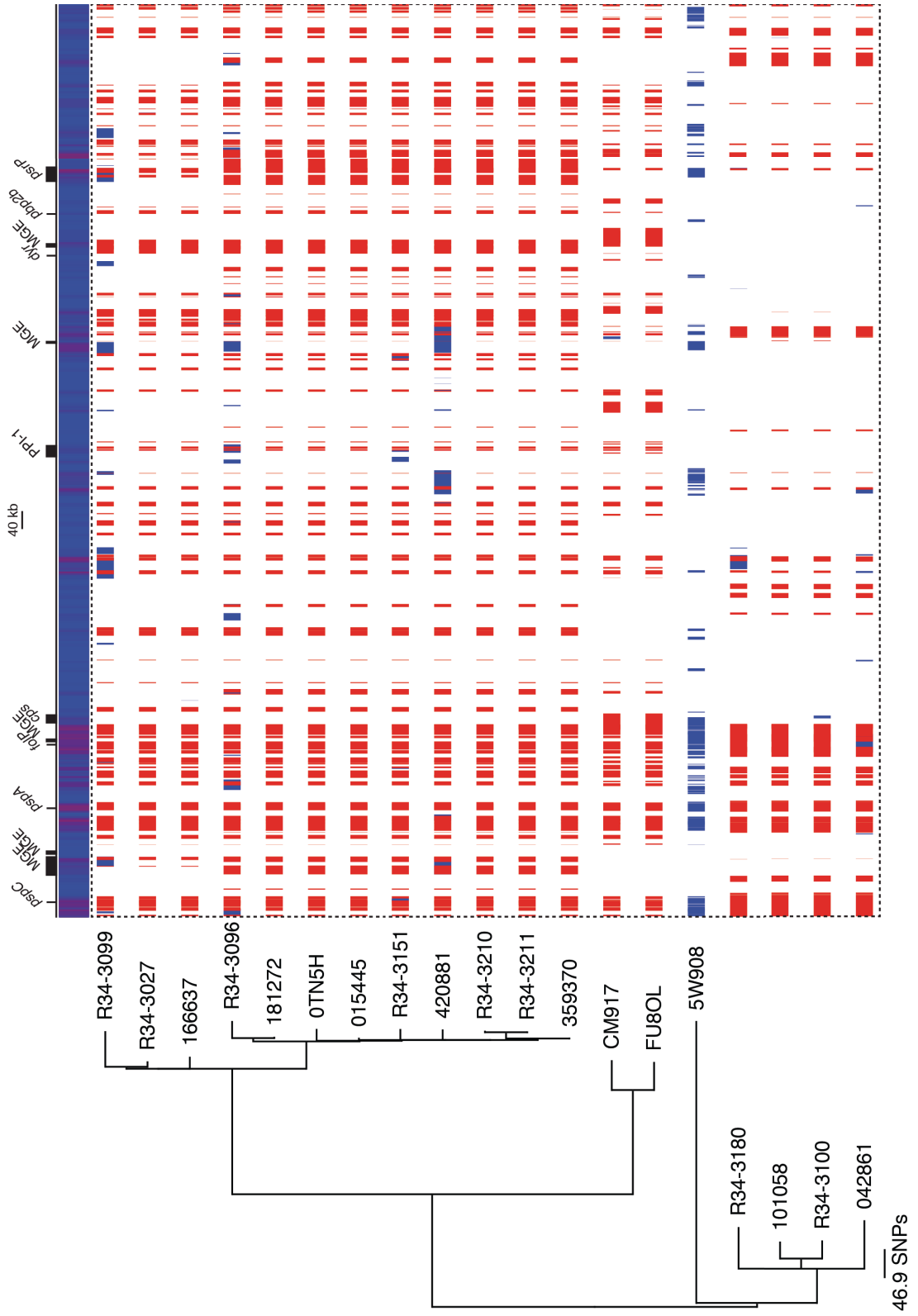
Supplementary Figure 14 Maximum likelihood phylogeny of SC10, displayed as described in Figure 3.



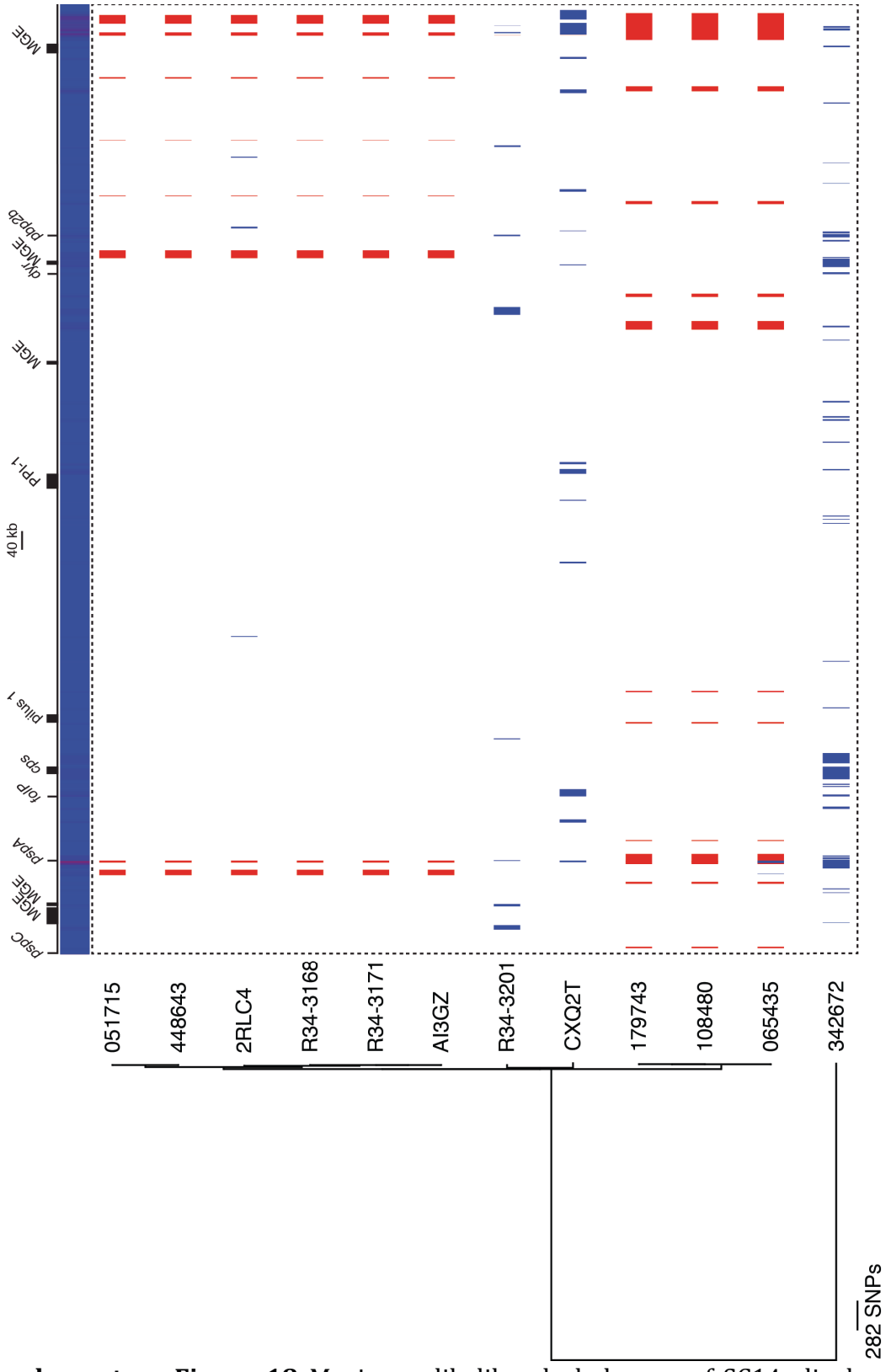
Supplementary Figure 15 Maximum likelihood phylogeny of SC11, displayed as described in Figure 3.



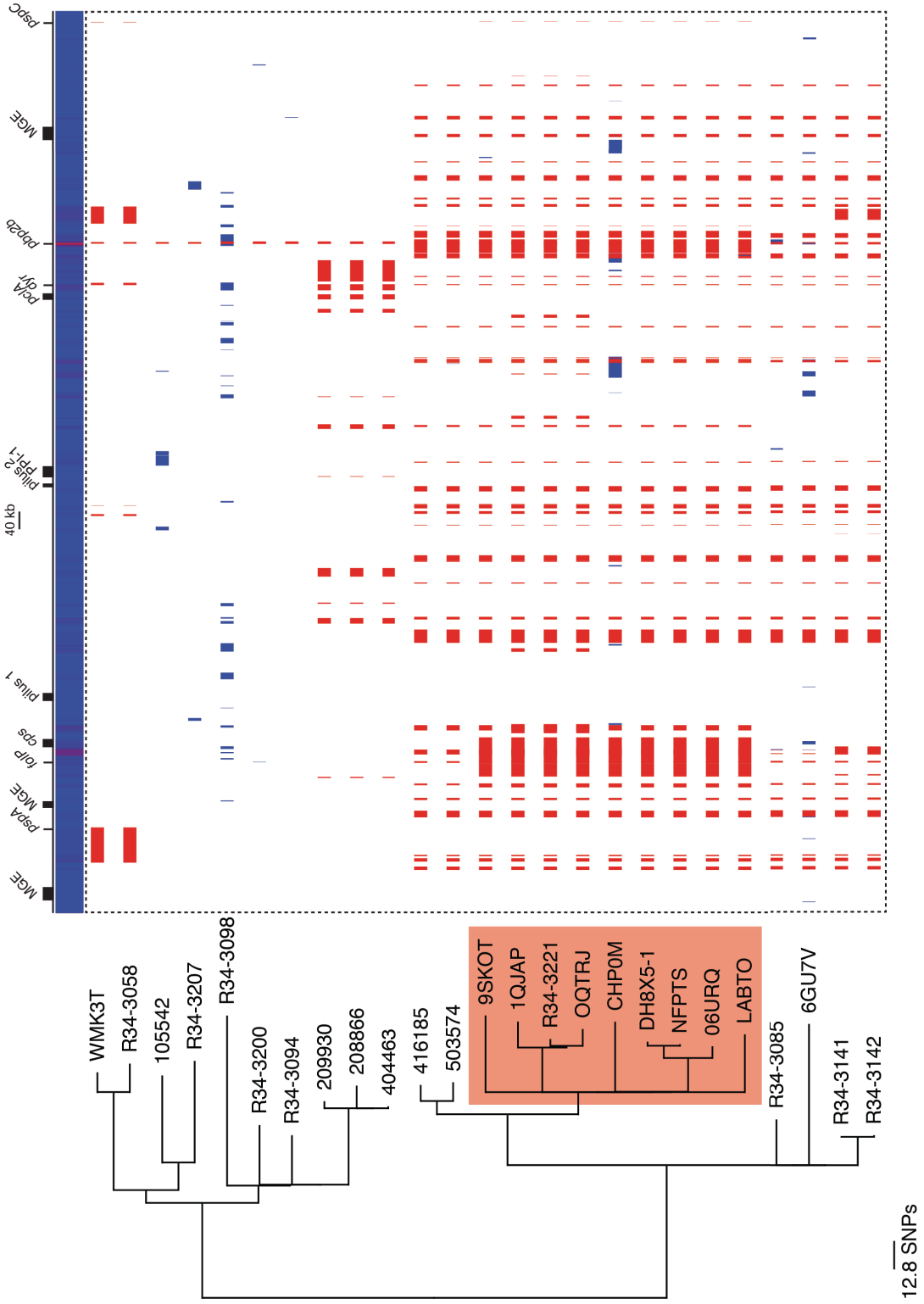
Supplementary Figure 16 Maximum likelihood phylogeny of SC12, displayed as described in Figure 3.



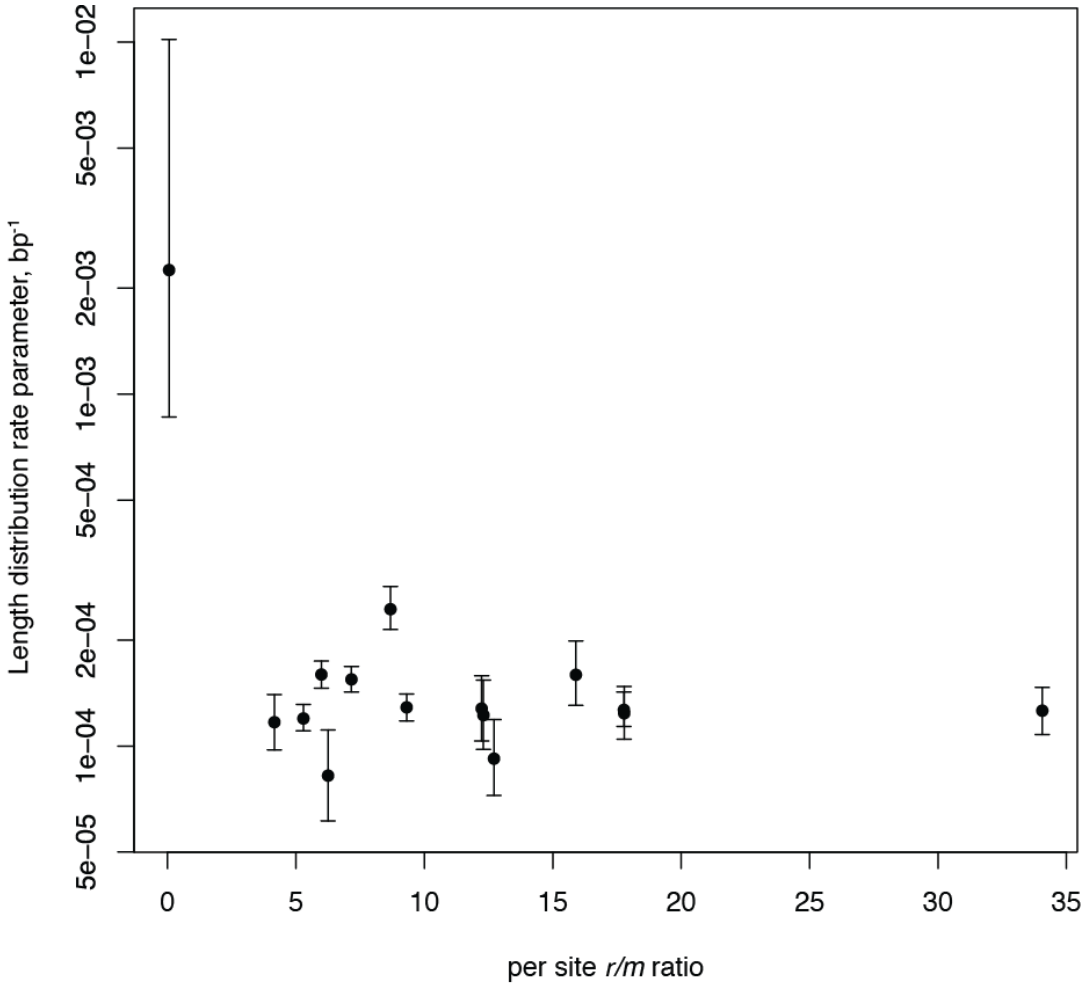
Supplementary Figure 17 Maximum likelihood phylogeny of SC13, displayed as described in Figure 3.



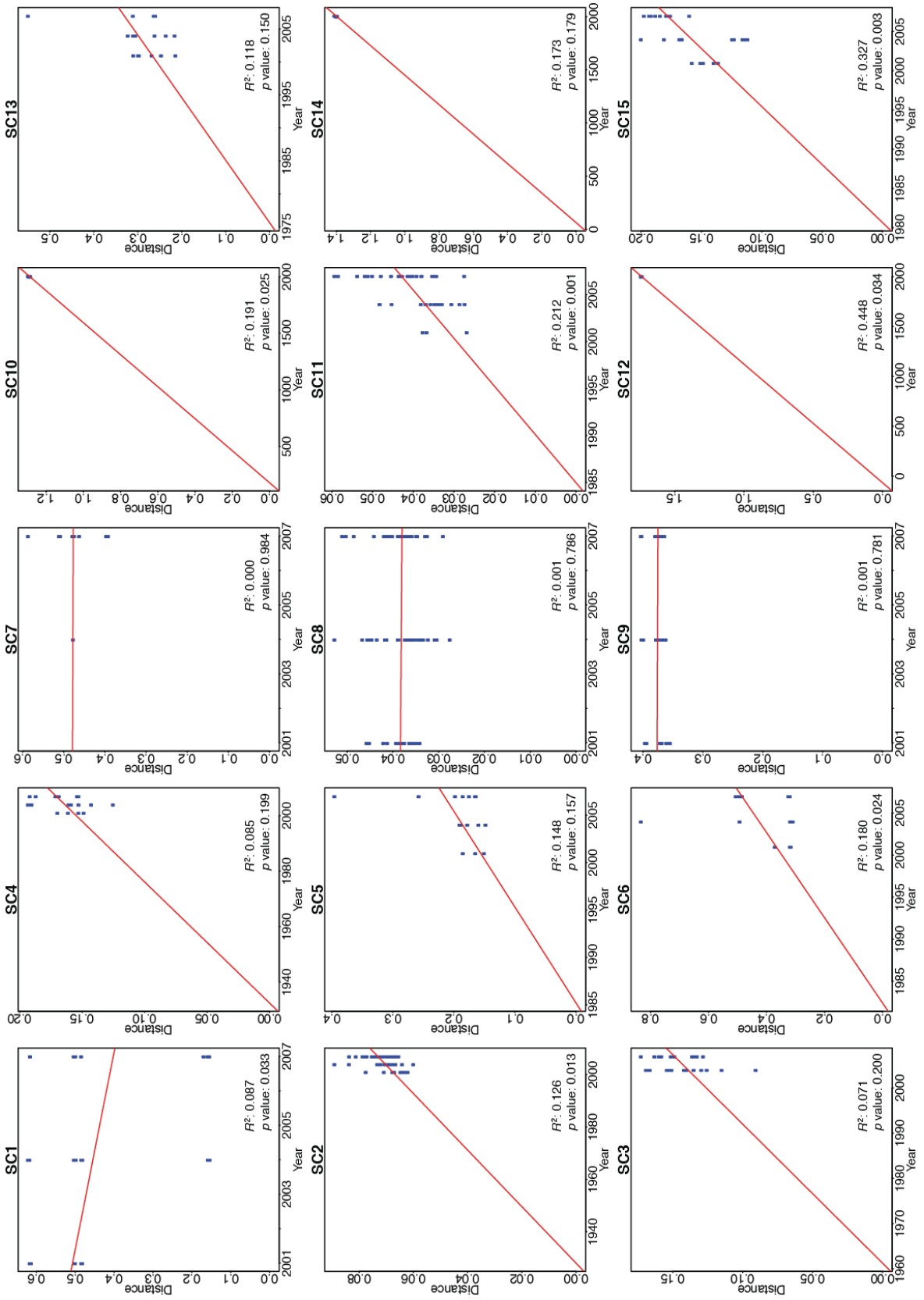
Supplementary Figure 18 Maximum likelihood phylogeny of SC14, displayed as described in Figure 3.



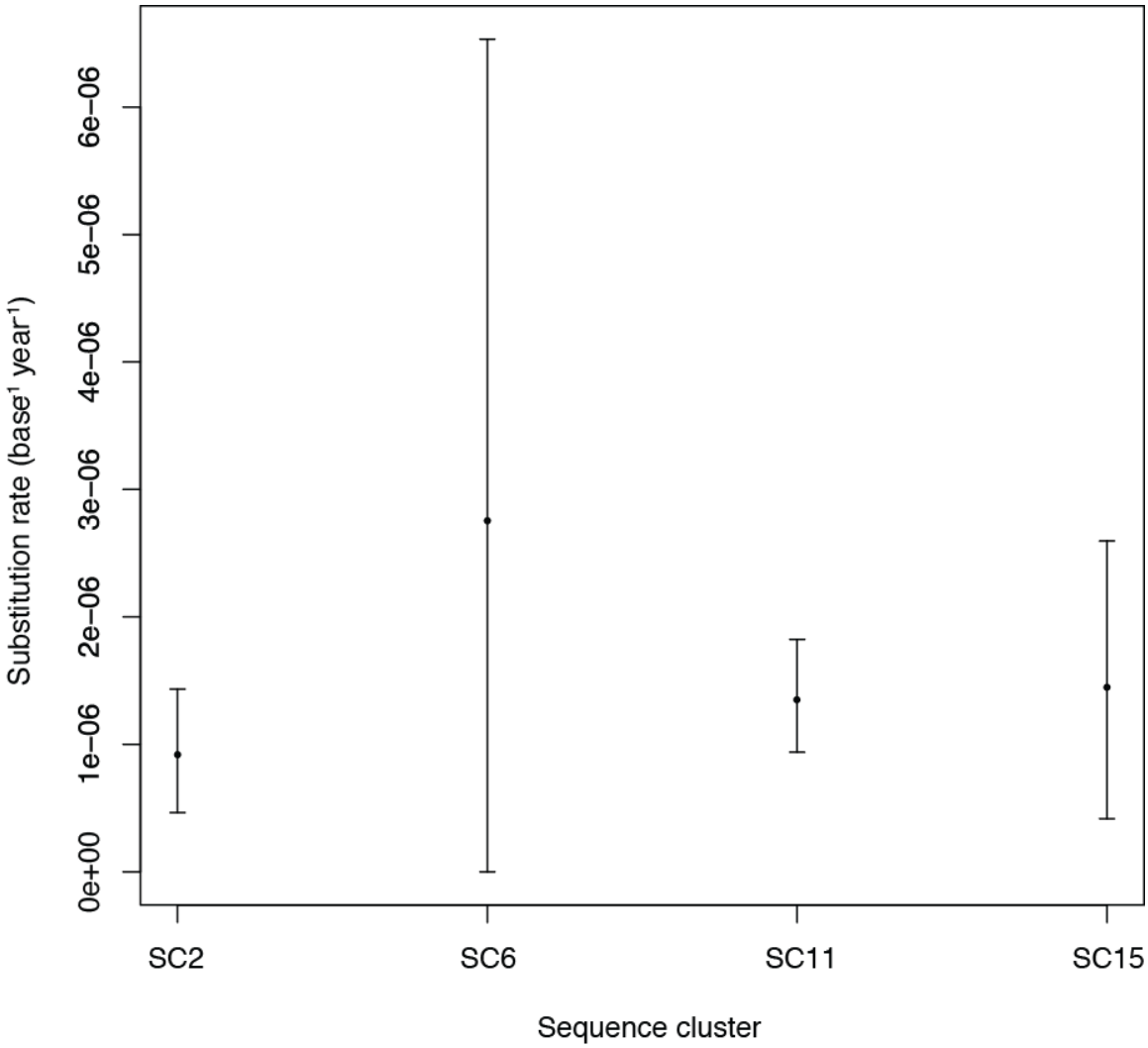
Supplementary Figure 19 Maximum likelihood phylogeny of SC15, displayed as described in Figure 3. Serotype 19A variants of the sequence cluster are indicated by the pink box.



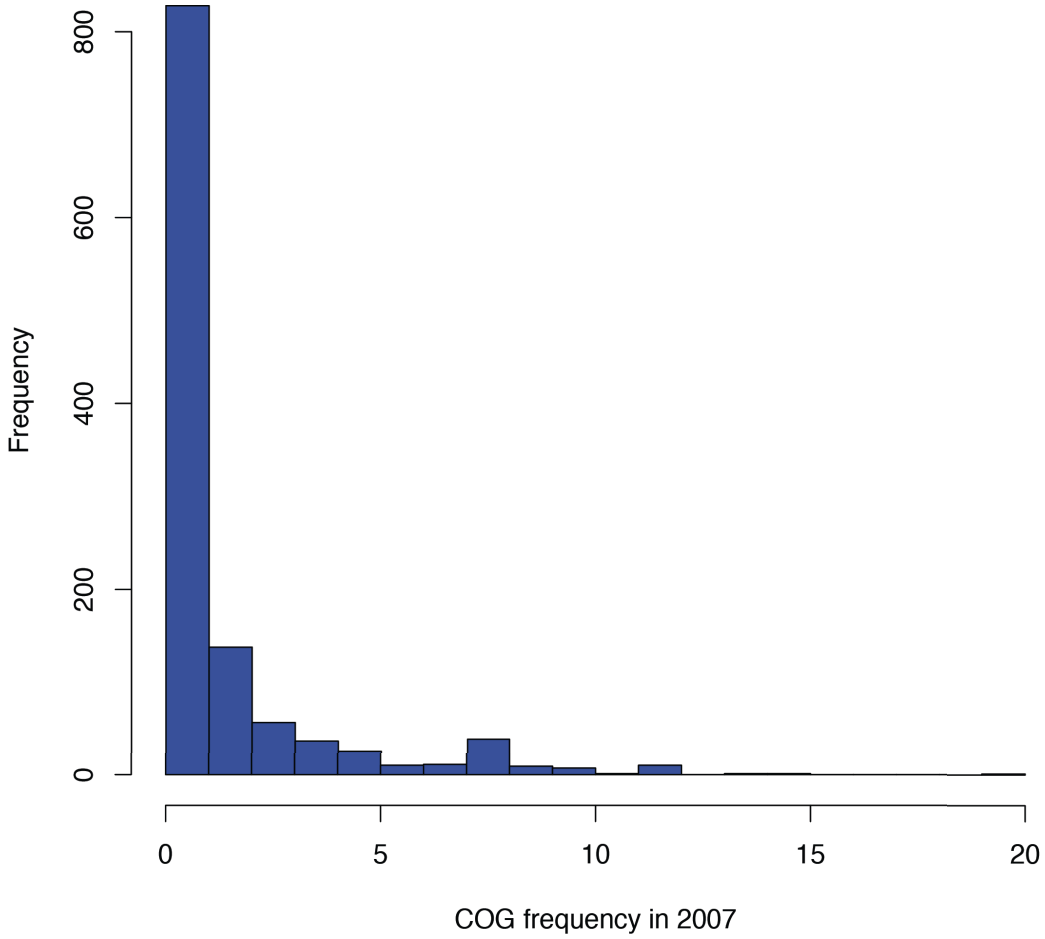
Supplementary Figure 20 Distribution of detected transformation event sizes. An exponential size distribution was parameterised using the lengths of detected transformation events in each sequence cluster individually. This gave an estimate of the rate parameter with 95% confidence intervals, plotted on the y axis. On the x axis is plotted the per site r/m statistic for the SC. The highest rate parameter estimate corresponds to SC7, in which only three homologous recombinations were identified, providing little data to use in calculating the parameter. The other estimates are consistent and similar to that estimated from PMEN1 (1.58×10^{-4} bp^{-1}), with no evidence that higher r/m values are the consequence of longer transformation events.


SC1
R²: 0.087
p value: 0.033
SC2
R²: 0.126
p value: 0.013
SC3
R²: 0.071
p value: 0.200
SC4
R²: 0.085
p value: 0.199
SC5
R²: 0.148
p value: 0.157
SC6
R²: 0.180
p value: 0.024
SC7
R²: 0.000
p value: 0.984
SC8
R²: 0.001
p value: 0.786
SC9
R²: 0.001
p value: 0.781
SC10
R²: 0.191
p value: 0.025
SC11
R²: 0.212
p value: 0.001
SC12
R²: 0.448
p value: 0.034
SC13
R²: 0.118
p value: 0.150
SC14
R²: 0.173
p value: 0.179
SC15
R²: 0.327
p value: 0.003
Distance
Year

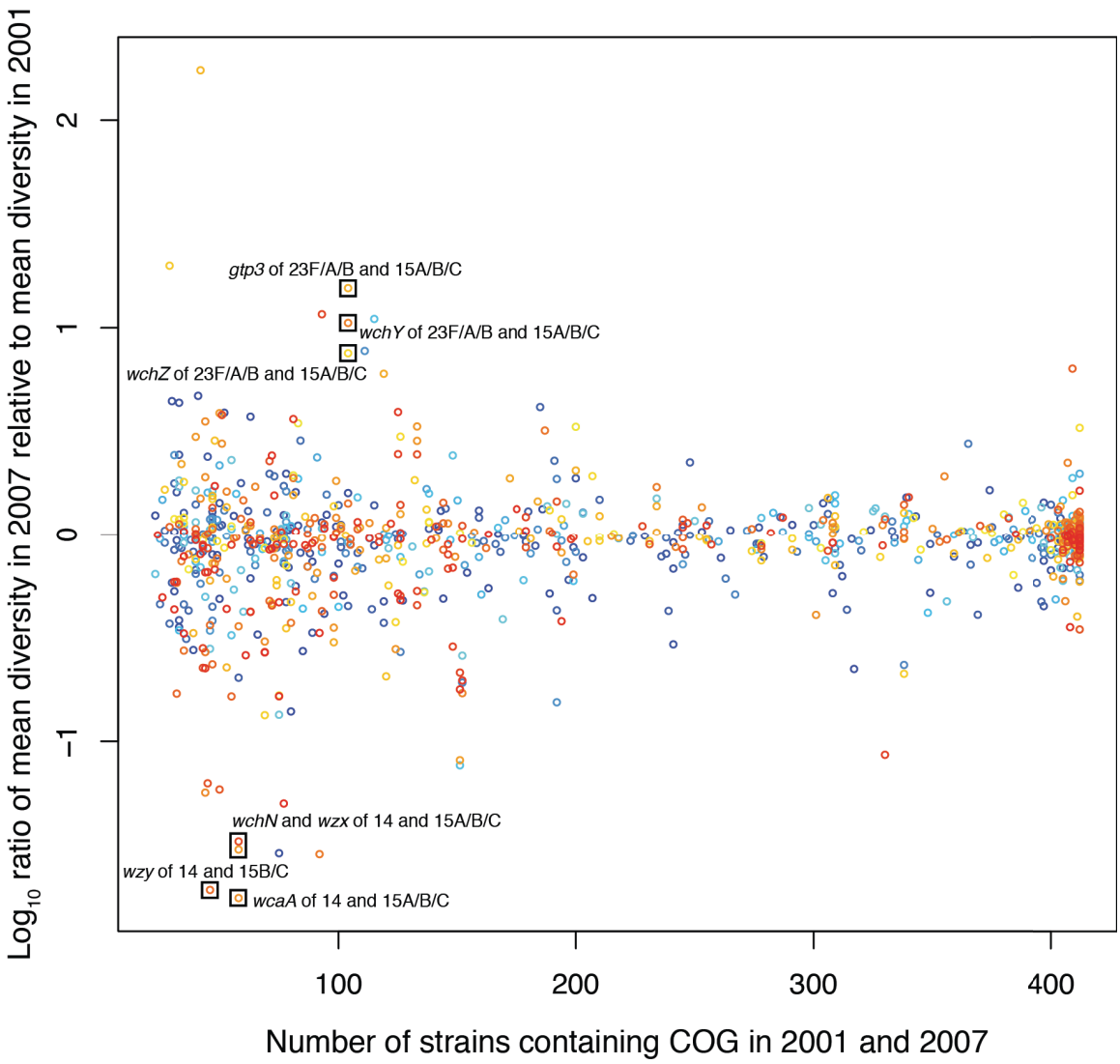
Supplementary Figure 21 Divergence of taxa from each monophyletic sequence cluster's most recent common ancestor over time. Using the phylogeny of each sequence cluster (Supplementary Figure 5-Supplementary Figure 19), Path-O-Gen was used to select the root node to maximize the positive correlation of year of a strain's isolation with its genetic distance from the root. Such a relationship indicates a signal of a 'molecular clock', with strains measurably diversifying from their last common ancestor over time. Four sequence clusters (SC2, SC6, SC11 and SC15) appear to have originated within the last century and exhibit a significant positive correlation at the $p = 0.05$ threshold.



Supplementary Figure 22 Estimates of point mutation rate. For the four sequence clusters that showed evidence of a molecular clock signal, the tree and alignment of point mutation sites was analysed using BEAST. A relaxed molecular clock model, with a lognormal distribution of rates, was used. The displayed values represent the median estimate of the Euclidean mean rate of mutation, with the accompanying 95% credibility intervals, for each sequence cluster. The estimates are consistent with one another, and with a previous estimate from the analysis of a collection of isolates from the PMEN1 lineage (1.57×10^{-6} base^{-1} year^{-1}).

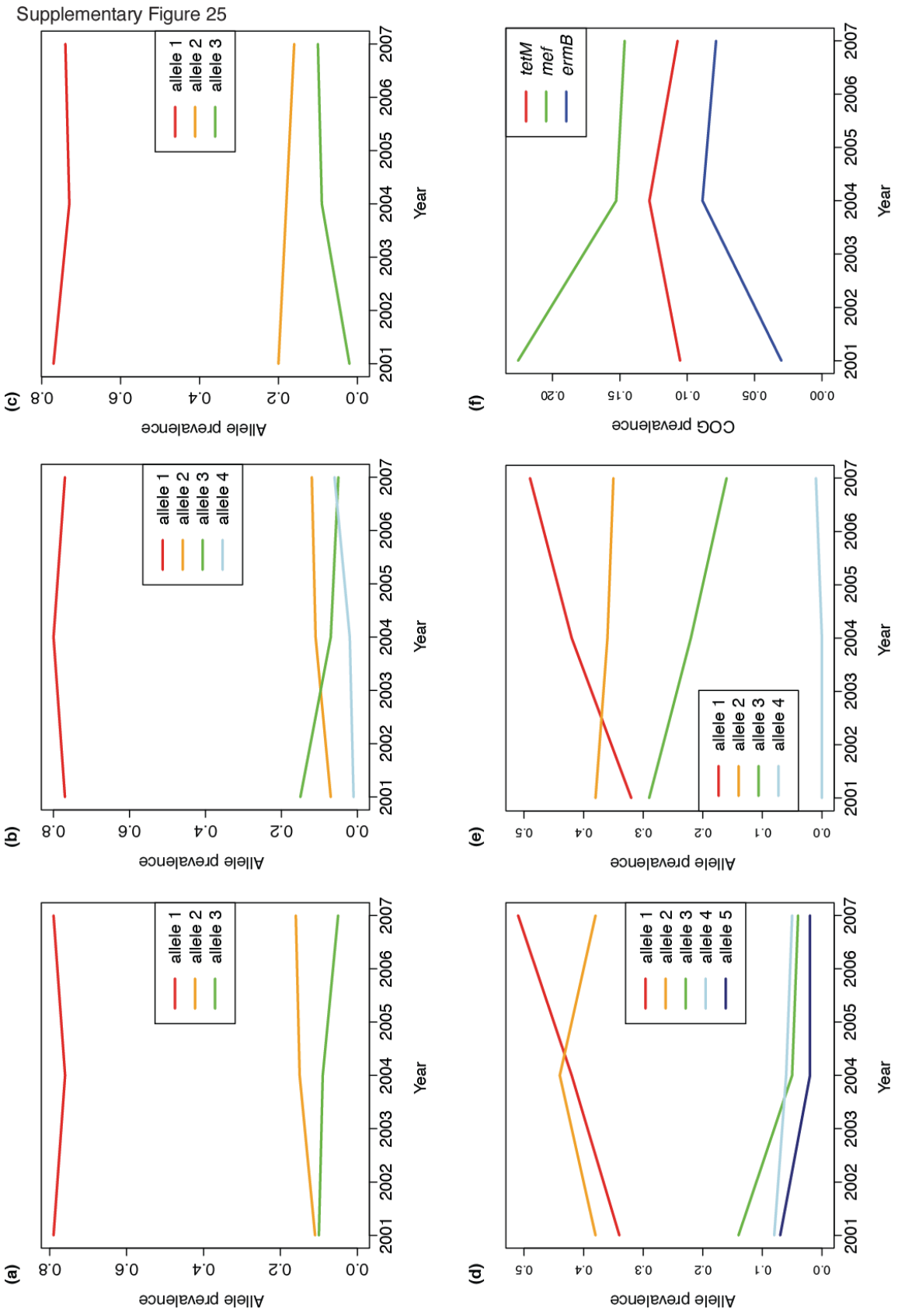


Supplementary Figure 23 Frequency distribution of COGs in the 2007 sample (n = 1176) absent from the 2001 sample. These COGs are not displayed in Figure 4, as their odds ratios are infinite. Many COGs not found in 2001 are also absent from the 2007 sample; these are present at low frequencies in the collection, and are only found in the 2004 sample. The majority of the remainder are found only at low frequency in 2007; the most common is present in 20 strains. This represents the complementary misassembly to the false positive COG observed to decrease in frequency in Figure 4; this variation appears to be an artifact resulting from the differences in read lengths used in the sequencing of isolates from 2001 and 2007.

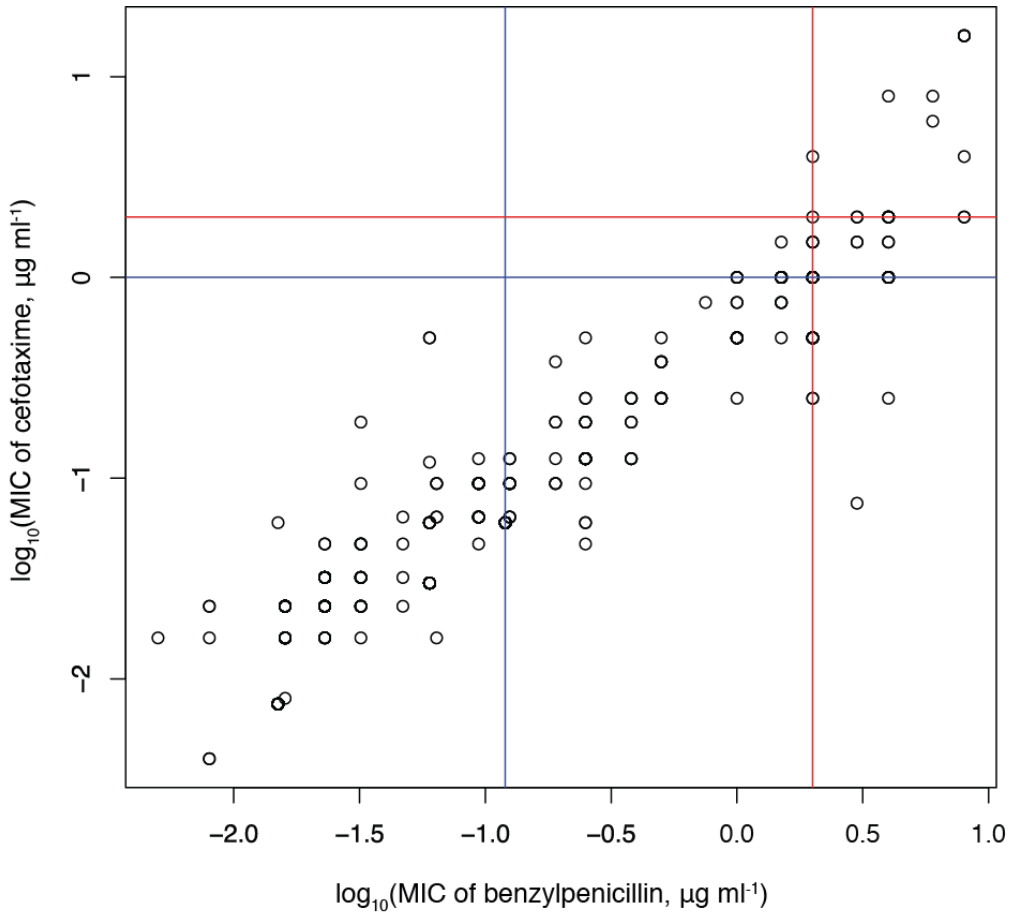


Supplementary Figure 24 Changes in diversity within COGs. For each COG found in more than ten strains in both 2001 and 2007, the mean pairwise Kimura distance between all members of the pair present in 2001 and in 2007 were separately calculated. This plot shows the base ten logarithm of the ratio of the mean Kimura distances in the two years, with values greater than zero indicating higher diversity in 2007. Points are ordered by their total frequency in the two years and colored according to their length as in Figure 4. COGs involved in capsule biosynthesis that show a large change in diversity between 2001 and 2007 are labeled.

Supplementary Figure 25

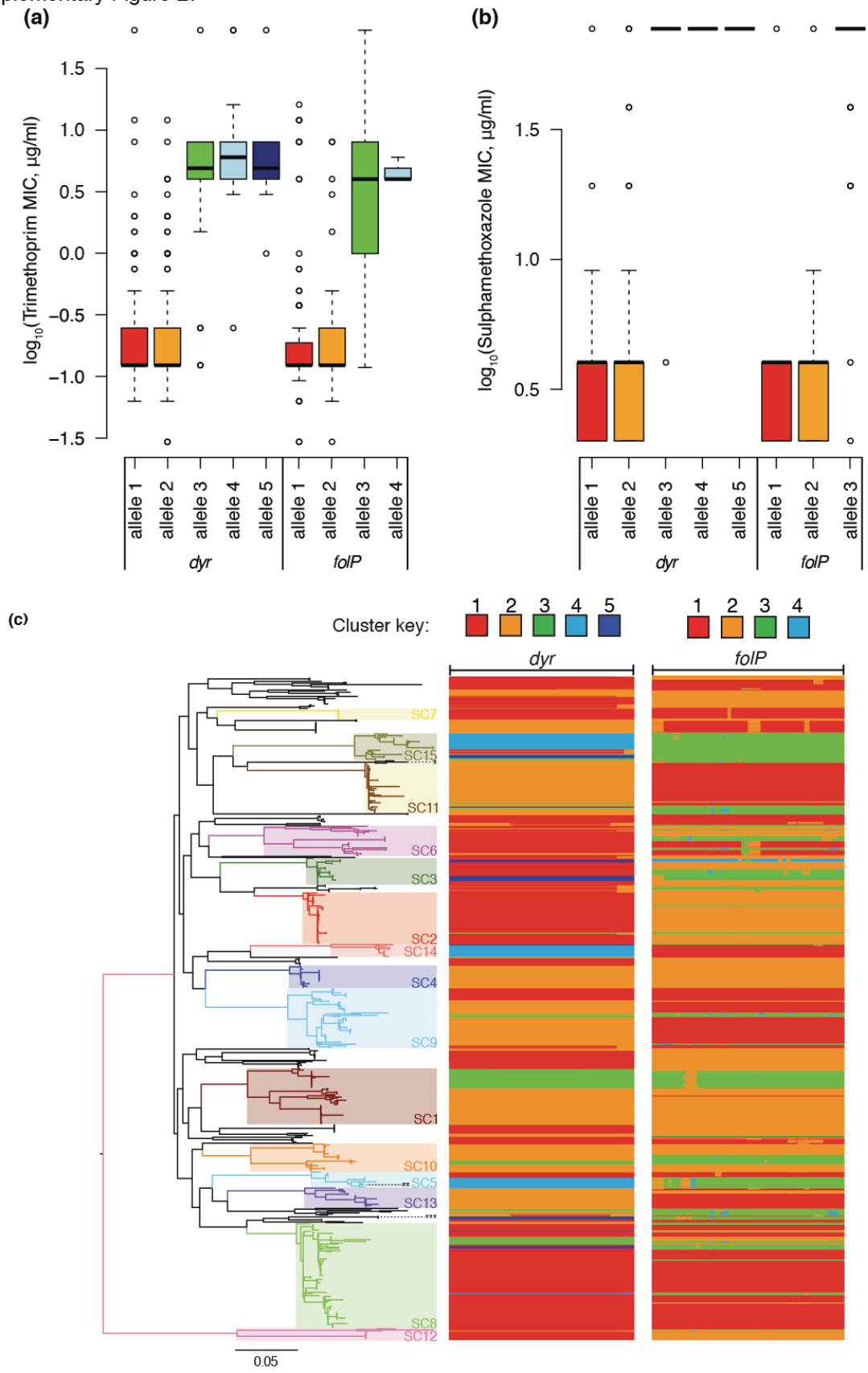


Supplementary Figure 25 Changes in frequency of sequences associated with resistance. For the penicillin binding proteins (a) *pbp1a*, (b) *pbp2b* and (c) *pbp2x* shown in Figure 6, the changes in allele frequency over the three sampled timepoints are displayed. The equivalent changes for the alleles of (d) *dyr* (encoding dihydrofolate redutase) and (e) *folP* (encoding dihydropteroate synthase), as presented in the analysis in Supplementary Figure 27 are also shown. The changes in the frequency of COGs associated with resistance to tetracycline and macrolides, the distributions of which are displayed in Figure 5, are shown in panel (f).

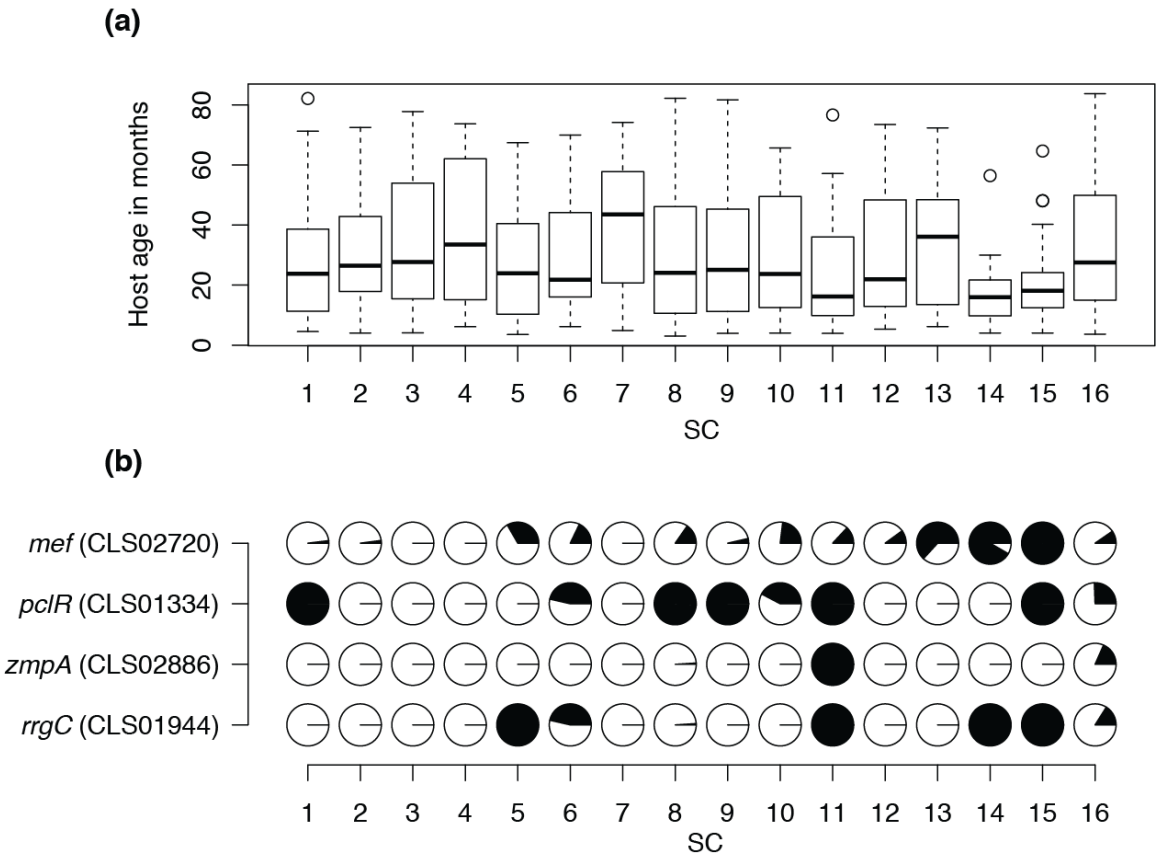


Supplementary Figure 26 Distribution of beta lactam minimum inhibitory concentrations (MICs) in the sampled population. Each isolate's MIC of benzylpenicillin is plotted against the equivalent value for ceftriaxone on logarithmic scales. The blue lines indicate the pre-2008 breakpoint for each drug at which strains were regarded as having intermediate resistance, rather than being fully susceptible; the red lines indicate the thresholds above which strains were regarded as fully resistant. This plot shows that strains have highly correlated levels of resistance to the two drugs (n = 610, R^2 = 0.57, p < 2.2e-16), with no subpopulation maintaining resistance to only one agent or the other. It also finds ceftriaxone to be generally more effective against *S. pneumoniae* than benzylpenicillin per unit concentration across all detected levels of resistance.

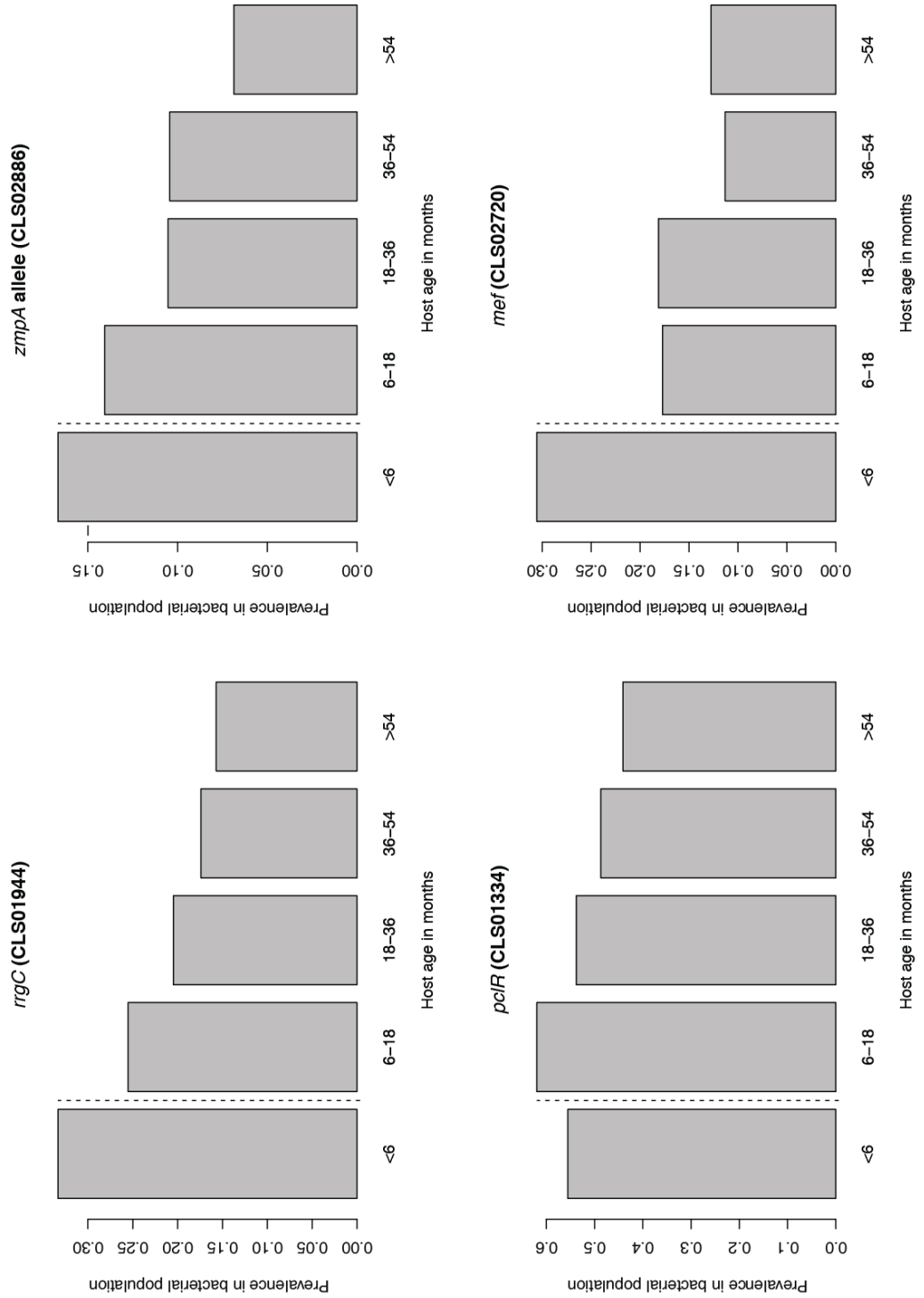
Supplementary Figure 27



Supplementary Figure 27 Changes in sulfa drug resistance. (a) Resistance to trimethoprim. Both *dyr*, encoding dihydrofolate redutase, and *folP*, encoding dihydropteroate synthase, were independently clustered on the basis of sequence similarity using nextgenBRAT; this produced five clusters for *dyr* and four clusters for *folP*. Box and whisker plots display the distribution of trimethoprim MICs associated with strains possessing each of these alleles. Alleles 3, 4 and 5 of *dyr* are associated with resistance to this antibiotic; each of these contains the I100L substitution thought to cause trimethoprim resistance. (b) Box and whisker plot equivalent to panel (a), but showing MICs to sulfamethoxazole. Allele 3 of *folP* is associated with resistance to this antibiotic; these sequences have various small insertions around amino acid S61 that are associated with sulfamethoxazole resistance. No data are available on allele 4, found in SC3, as isolates were only tested for resistance to sulfamethoxazole in 2001 when this genotype was absent from the collection. All of these alleles contain an arginine insertion adjacent to amino acid S91, and would therefore be expected to be resistant to sulfamethoxazole. (c) Distribution of *dyr* and *folP* alleles throughout the pneumococcal population. The phylogeny displayed in Figure 1 is shown in a linear form on the left, with the multidrug-resistant lineages PMEN1 (***), PMEN3 (**) and PMEN15 (*) labeled with asterisks. The two columns on the right represent the independent analyses for the *dyr* and *folP* genes involved in sulfa drug resistance. These are comprised of one row for each taxon in the tree. Blocks are colored according to the group to which the sequence belongs, as indicated by the key at the top of the column.



Supplementary Figure 28 Distribution of age-associated COGs across the bacterial population. (a) Distribution of host ages colonized by each sequence cluster. The boxplot for each sequence cluster shows the range of host ages found to be colonized by the bacteria of that genotype. (b) Distribution of age-associated COGs. A pie chart shows the prevalence of each of the surface structures associated with young children in each SC, with the black segments indicating presence and the white segments absence.



Supplementary Figure 29 Distribution of surface structures across host age ranges. For each of the four gene clusters highlighted as being negatively associated with host age in Figure 8, the proportion of sampled hosts carrying the COG in each age range is displayed.

Supplementary Tables

Supplementary Table 1 Epidemiological data associated with isolates and accession codes associated with data deposited in the European Nucleotide Archive (separate spreadsheet file).

Supplementary Table 2 Non-typeable pneumococci outside of SC12 detected in the dataset.

Isolate	Sequence type	Sequence cluster	Capsule locus
R34-3053	1379	10	Capsule type 6C
CH2006	3288	16	NspA-type
CL3012	3288	16	NspA-type
R34-3108	2011	16	Deleted