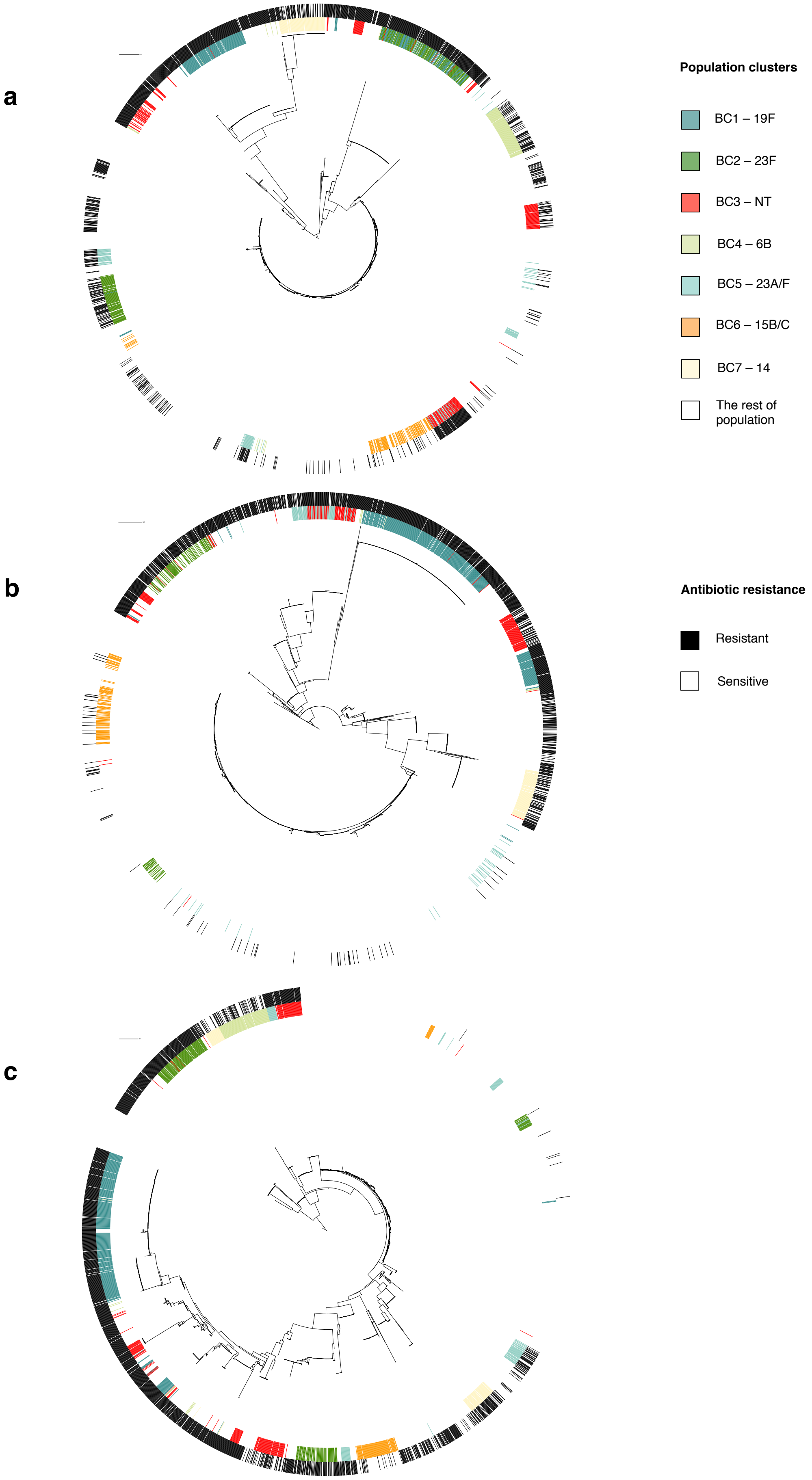
Supplementary information for “Dense genomic sampling identifies highways of pneumococcal recombination”

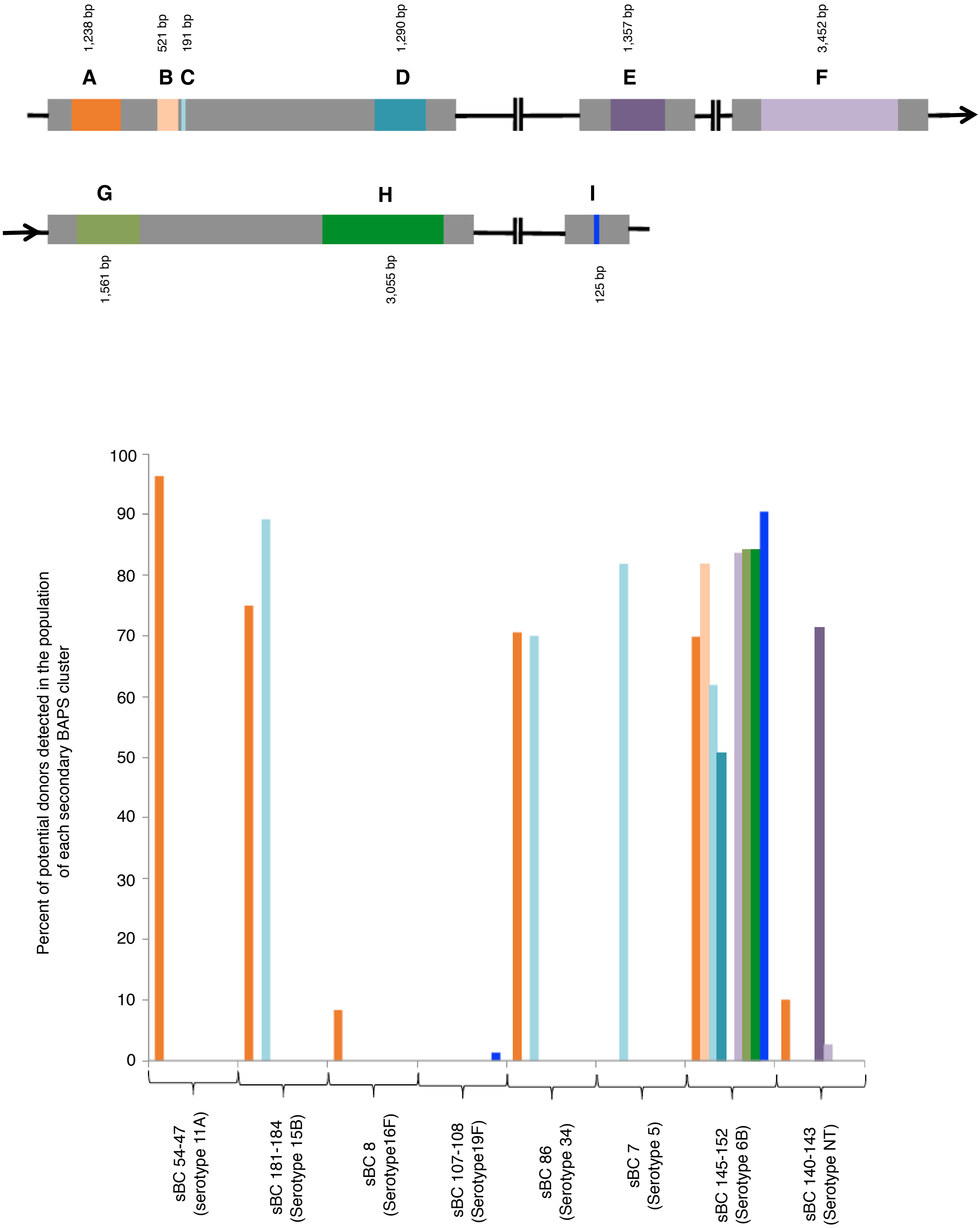
Claire Chewapreecha, Simon R Harris, Nicholas J Croucher, Claudia Turner, Pekka Marttinen, Lu Cheng, Alberto Pessia, David M Aanensen, Alison E Mather, Andrew J Page, Susannah J. Salter, David Harris, Francois Nosten, David Goldblatt, Jukka Corander, Julian Parkhill, Paul Turner and Stephen D Bentley

This file contains
Supplementary Figures 1-9
Supplementary Table 2-9
Supplementary Note
Supplementary References

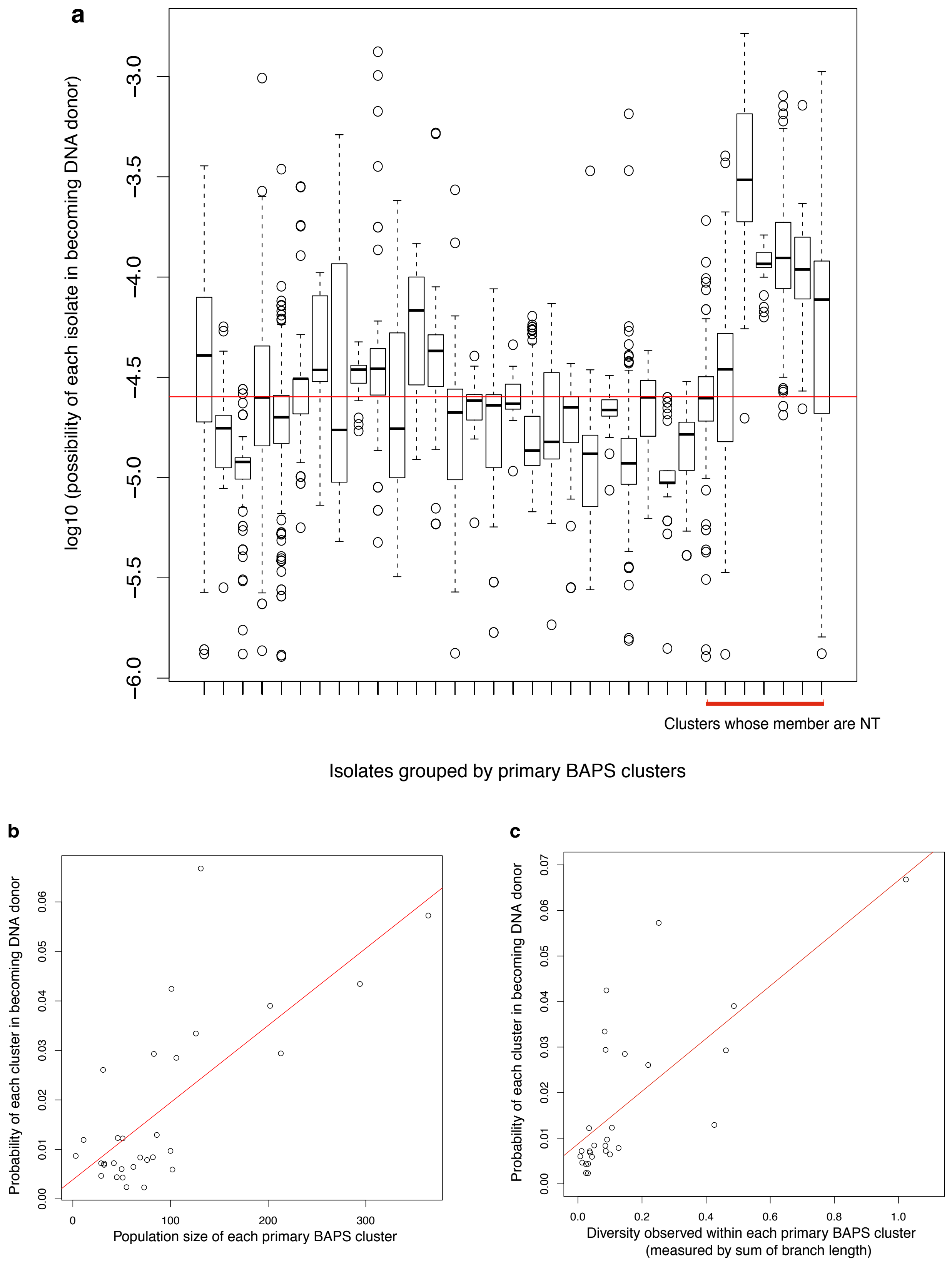
There is an Excel-formatted Supplementary Table 1 detailing epidemiological data associated with strains and accession codes.



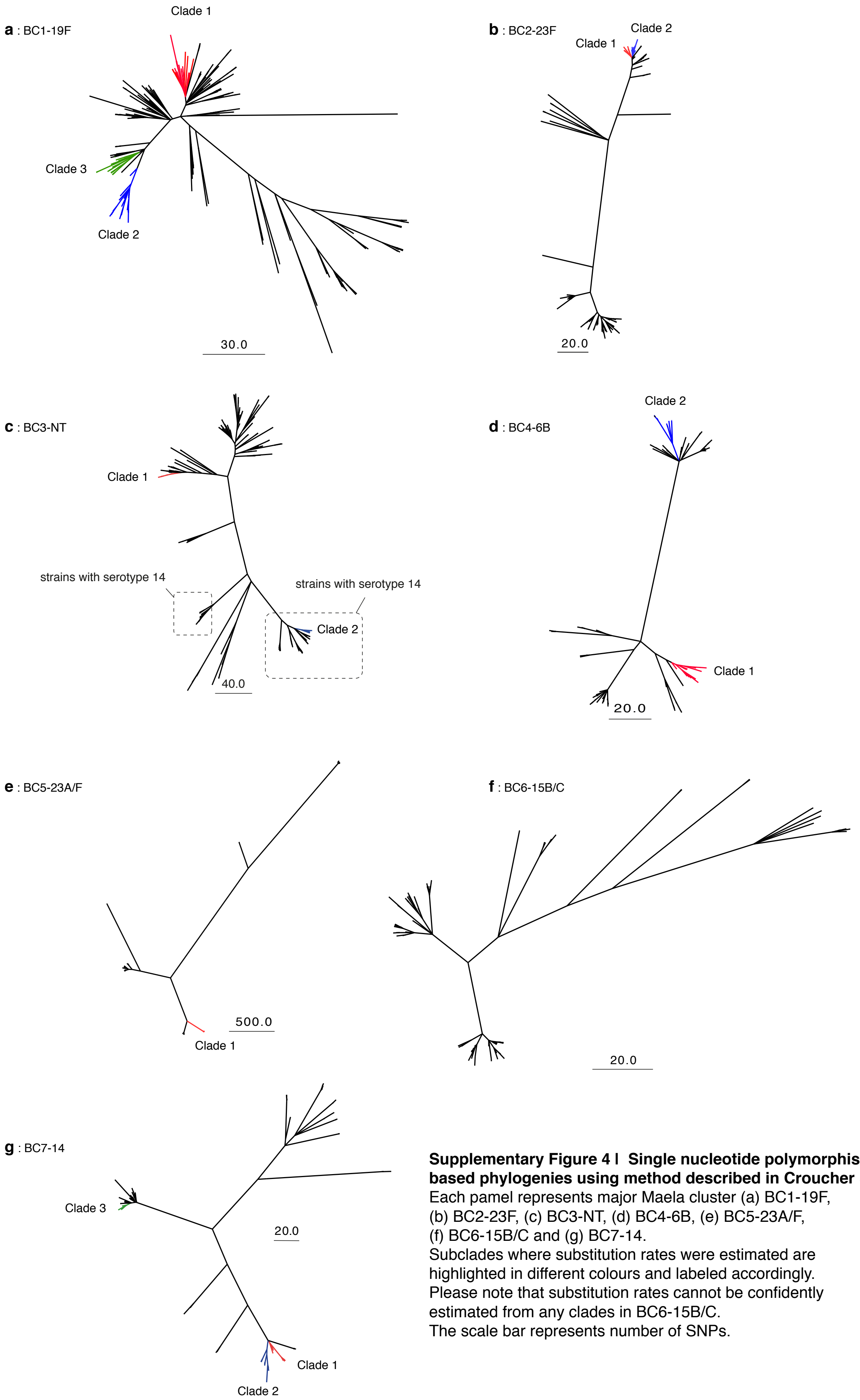
Supplementary Figure 1 | Association between recombining *pbp* genes and resistant phenotypes.
(a) *pbp1a* gene tree. (b) *pbp2b* gene tree. (c) *pbp2x* gene tree. The inner ring is coloured according to dominant population clusters (BC1-7), with the rest of the population appearing in white. The outer ring is coloured according to resistant to penicillin with black and white showing non-susceptibility and susceptibility respectively.



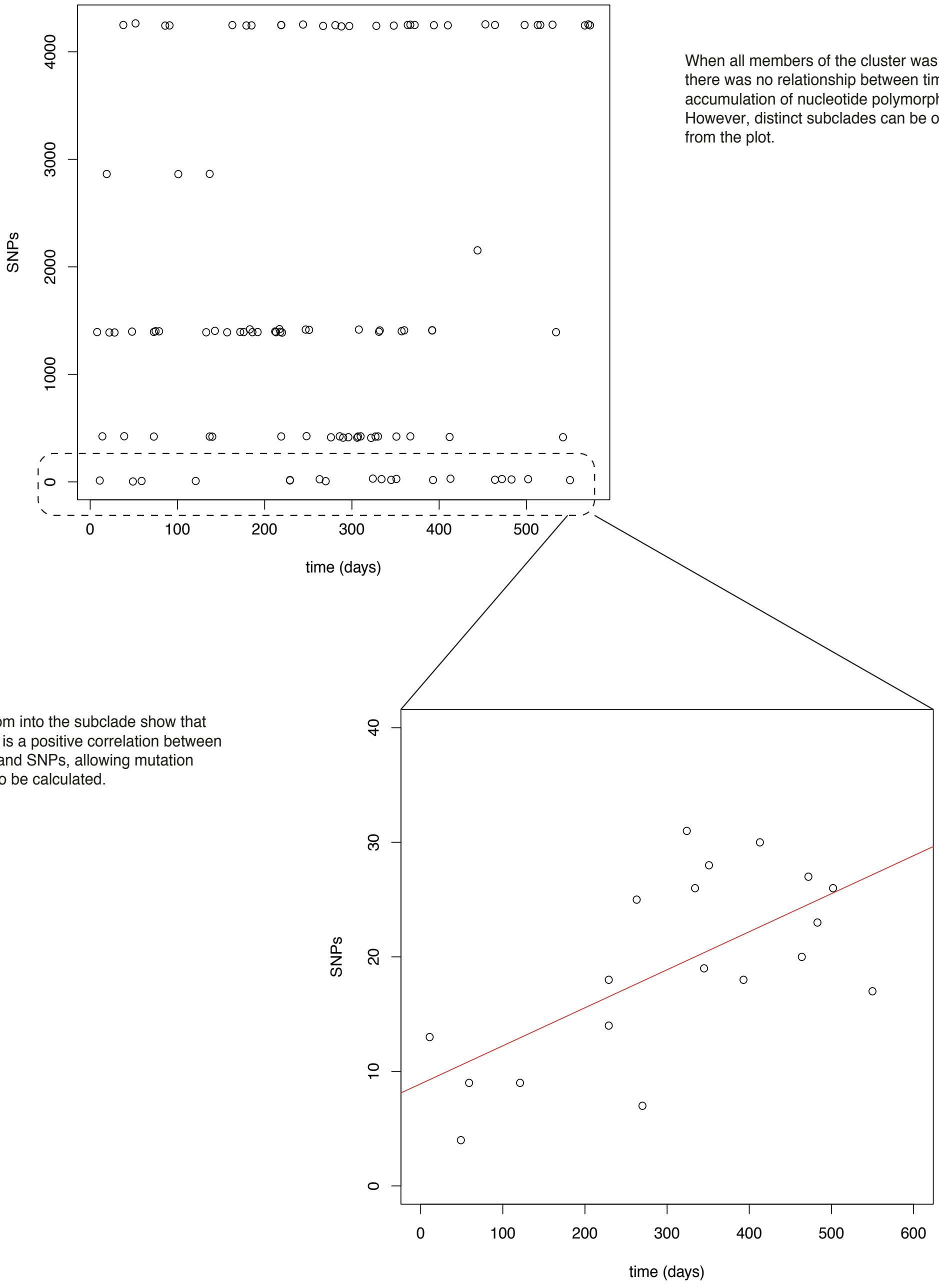
Supplementary Figure 2 | All possible recombination donors for recombinant fragments detected in one isolate.
Top panel: Nine predicted recombination fragments in SMRU1452 (fragment A - I) are highlighted in different colours and are ordered according to their locations on the genome with their size labeled. Bottom panel: The bar chart presents the possible sources of each recombinant fragment based on the above colour scheme. The y-axis gives the proportion of hits detected per population of particular lineage. For example, recombination fragment A of 1,236 bp length found their identical match in 96.29%, 75%, 70.59%, 69.84%, 10% and 8.33 % in the population of secondary BAPS clusters of serotype 11A, 15B, 34, 6B, NT and 16F respectively.



Supplementary figure 3 I Potential donors characterised by isolates and by BAPS primary clusters.
(a) Boxplots represent distribution of donation probability of isolates within each cluster. A red bar represents a mean frequency of donation event of any isolates (2.53 x 10-5) i.e. each isolate has a probability of 1/39,537 to donate DNA in recombination event. (b) and (c) respectively show positive correlations between potential donor clusters (based on primary BAPS clusters) and cluster population size, and separately cluster diversity



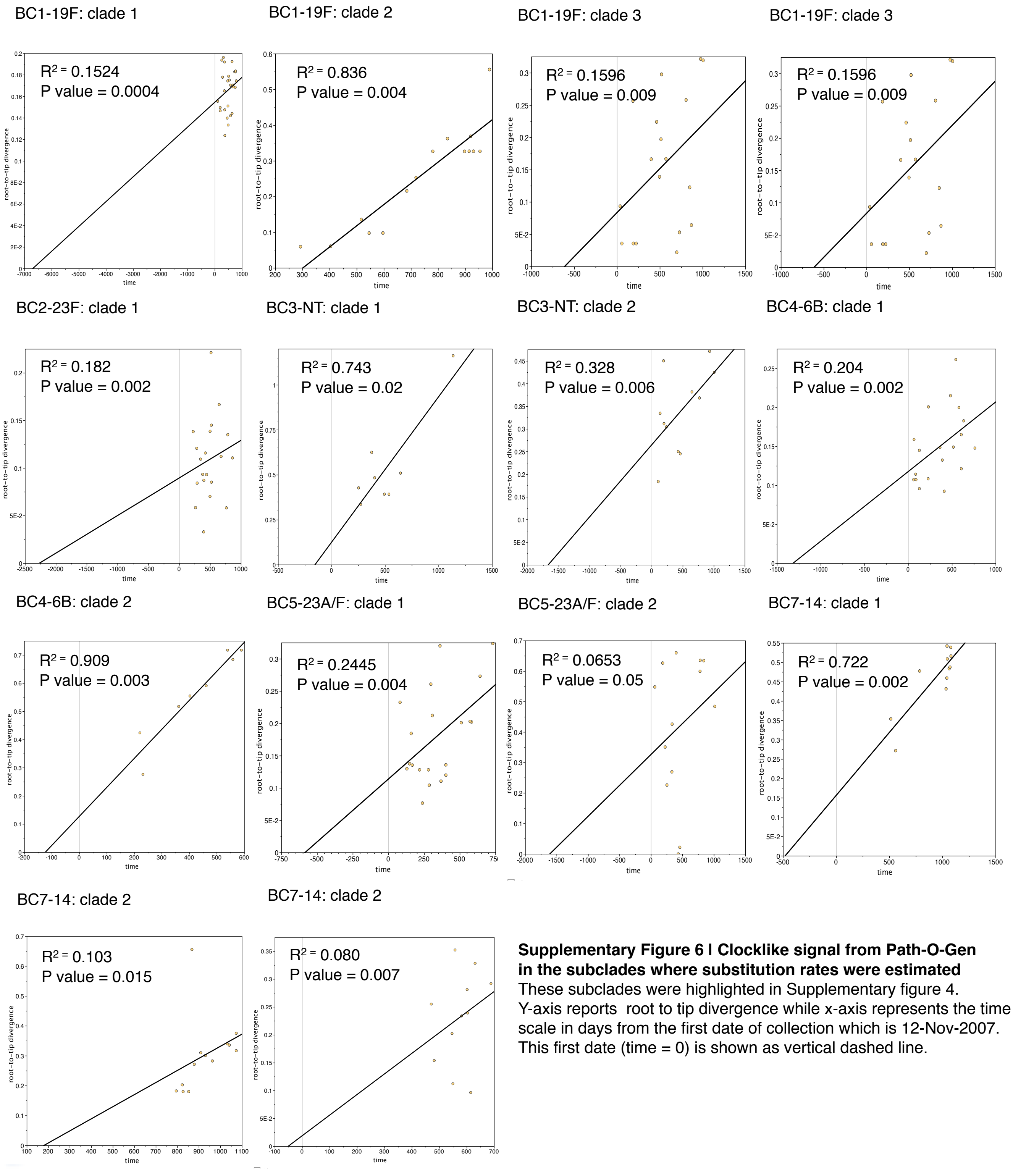
Supplementary Figure 4 l Single nucleotide polymorphism (SNPs) based phylogenies using method described in Croucher *et al.*
Each pamel represents major Maela cluster (a) BC1-19F, (b) BC2-23F, (c) BC3-NT, (d) BC4-6B, (e) BC5-23A/F, (f) BC6-15B/C and (g) BC7-14.
Subclades where substitution rates were estimated are highlighted in different colours and labeled accordingly.
Please note that substitution rates cannot be confidently estimated from any clades in BC6-15B/C.
The scale bar represents number of SNPs.



When all members of the cluster was considered, there was no relationship between time and accumulation of nucleotide polymorphisms. However, distinct subclades can be observed from the plot.

A zoom into the subclade show that there is a positive correlation between time and SNPs, allowing mutation rate to be calculated.

Supplementary Figure 5 I Demonstration that clocklise signals can be detected from the subclades but not from the whole population
Each dominant cluster is comprised of more than a single subclade that co-evolve together. This plot used BC5-23A/F as an example.
The clock signal cannot be detected in the whole cluster as they are confounded by signals of the subclades.



BC1-19F: clade 1

BC1-19F: clade 2

BC1-19F: clade 3

BC1-19F: clade 3

BC2-23F: clade 1

BC3-NT: clade 1

BC3-NT: clade 2

BC4-6B: clade 1

BC4-6B: clade 2

BC5-23A/F: clade 1

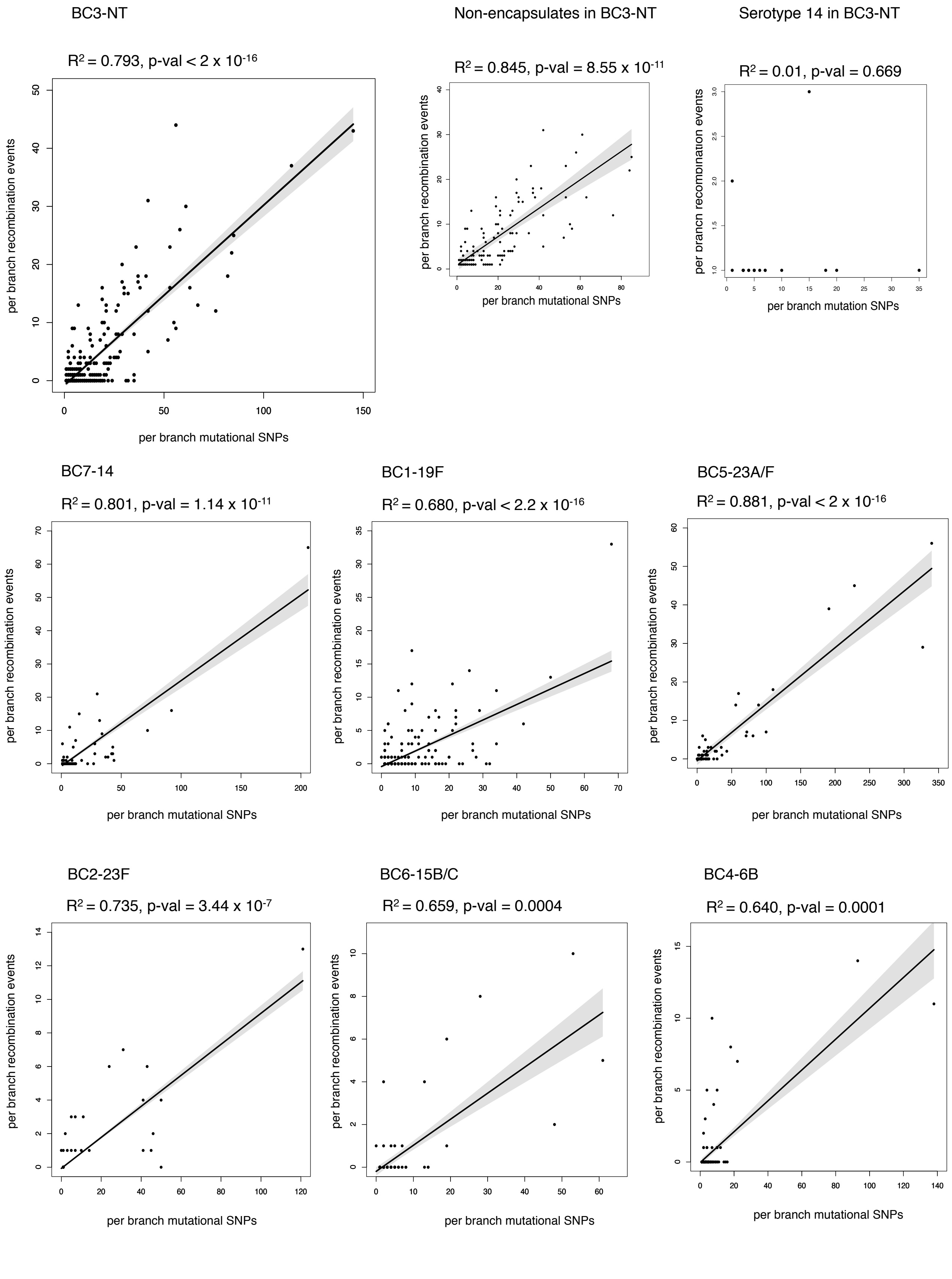
BC5-23A/F: clade 2

BC7-14: clade 1

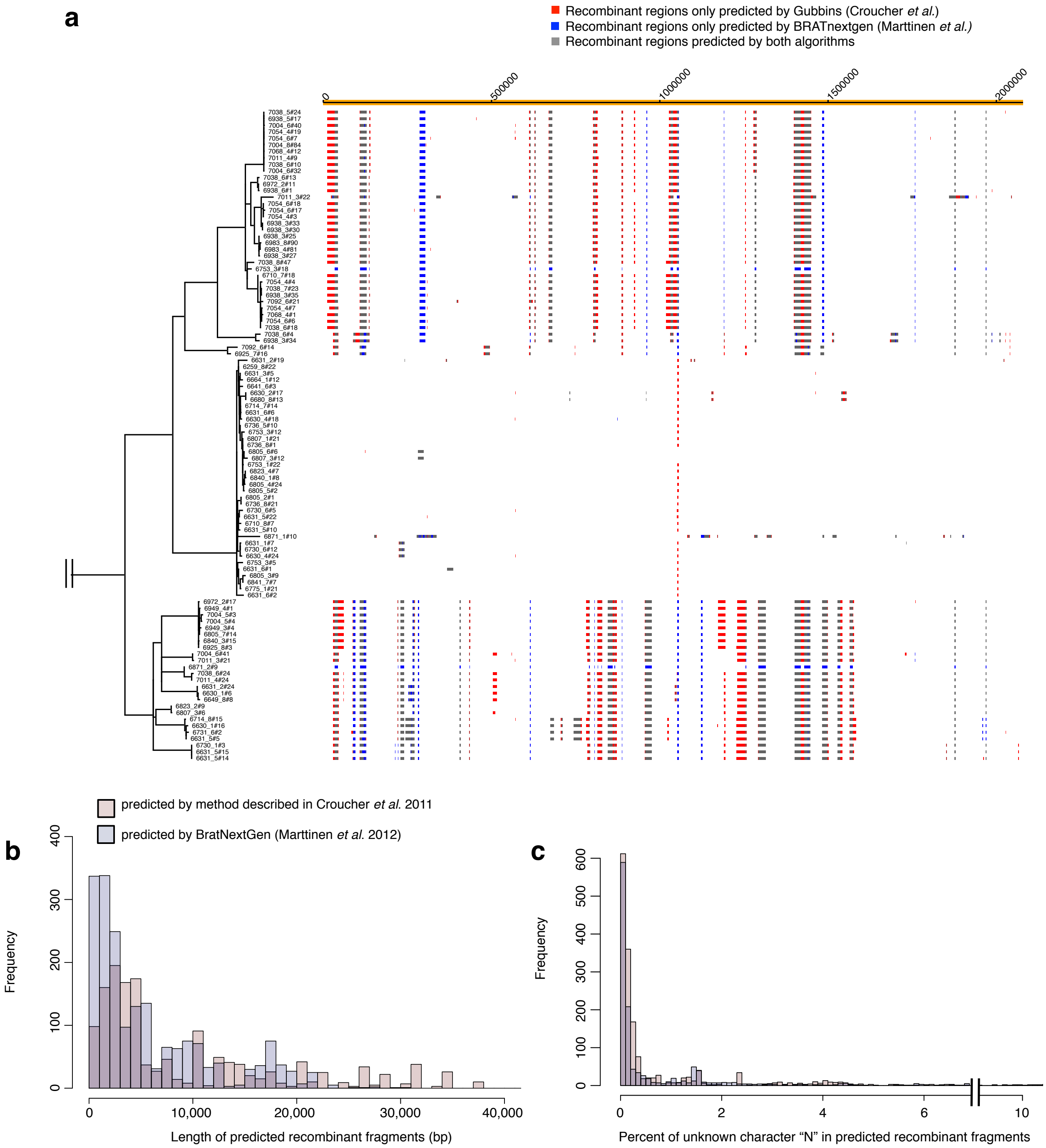
BC7-14: clade 2

BC7-14: clade 2

Supplementary Figure 6 I Clocklike signal from Path-O-Gen in the subclades where substitution rates were estimated
These subclades were highlighted in Supplementary figure 4. Y-axis reports root to tip divergence while x-axis represents the time scale in days from the first date of collection which is 12-Nov-2007. This first date (time = 0) is shown as vertical dashed line.



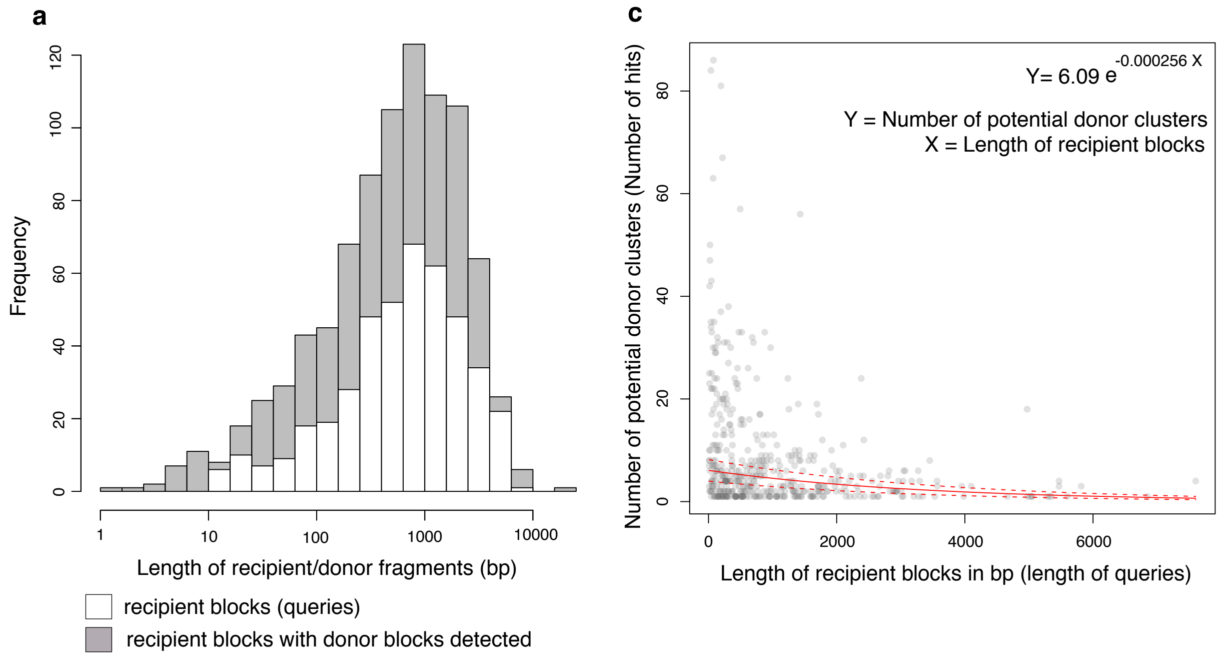
Supplementary Figure 7 l Recombination per mutation (r/m) of each cluster calculated by linear regression
Due to large sample size available in our studies, we alternatively calculated the ratio of recombination events (y - axis) over point mutations (x - axis) observed on each branch from the slope (r/m) of the linear regression. The number is tabulated in Supplementary Table 4. For comparison of r/m by linear regression between clusters, all the data are ranked to accommodate the non-parametric ANCOVA analysis.



d

	methods	1st Qu	Median	Mean	3rd Qu
Length of recombination fragments in B	Croucher *et al*	2403	4558	10150	14350
	Marttinen *et al.*	1189	3427	5777	8500
	Co-predicted	1177	3057	5418	21263
Percent of "N" in recombinant fragments in C	Croucher *et al*	0.084	0.200	6.05	2.36
	Marttinen *et al.*	0.066	0.153	5.59	2.00

Supplementary Figure 8 l Comparisons of two recombination detecting methods
(a) presents a genome view of recombination fragments predicted by both algorithms. Recombination regions are aligned with taxa on phylogenetic tree (left). Genome coordinate is labeled on top. Recombination regions exclusively predicted by methods described in Croucher *et al.* and Marttinen *et al.* are highlighted in red and blue respectively. Overlapping regions predicted by both algorithms are highlighted in dark grey.
(b) A histogram showing length of recombination fragments (bp) predicted by two algorithms. (c) Sequence quality of recombination fragments predicted by two algorithms reported as percent "N". For (b) and (c), fragments predicted by tools described in Croucher *et al.* and Marttinen *et al.* are shaded in red and blue respectively. The values given by (b) and (c) are summarised in (d).



b

Length of fragments (bp)	min	1st Qu	Median	Mean	3rd Qu	Max
All recipient blocks (queries)	3	197	604	1072	1412	22970
Recipient blocks with donor blocks detected (identical hits)	10	275	741	1162	1513	6846

Supplementary Figure 9 I Identifying donor blocks from recipient block identity
Donor blocks are sequences that show identical match to the recipient blocks. (a) A histogram showing distribution of length of recipient blocks from recombination events detected at the tip of the phylogenies. Shaded in grey are all recipient blocks used as quries for blast searches. In white are recipient blocks where identical hits were detected from the rest of the population. Table (b) summarises the values from distribution in (a). (c) A plot showing association between the length of sequence queries (recipient blocks) and the diversity of detected hits (potential donor blocks classified by secondary BAPS clusters). The data was modeled as exponential decay with the line of best fit (red line) and the 95% confidence interval (dashed red lines)

Supplementary table 2

Distribution of non-typable serotype (NT) within Maela.

Categories		**The whole NT in Maela population**	**NT within BC3-NT cluster**
Non-pneumococcal streptococci		11	0
Intact capsule locus but not expressing capsule (19A – 1 isolates, and serogroup 6-like cluster – 3 isolates)		4	0
Group NT1: cps deletion/partial deletion		42	16
Group NT2: putative surface protein NspA		258	111
Group NT3: *aliB* genes	**Group NT3.1**: contain two *aliB* genes, *glf* pseudogene, *ntaAB* toxin/antitoxin	107	1
	Group NT3.2: contain *ISSpn10, aliB-1, aliB-2* pseudogene, *glf* pseudogene, *ntaAB*	51	0
	Group NT3.3: two *aliB* genes, *glf*	13	0
	Group NT3.4: *aliB-2, glf*	26	0
total		512	128

Supplementary table 3

Nucleotide substitution rates estimated by BEAST. Mean nucleotide substitution rate, the lower bound of the 95% highest posterior density (HPD) and the upper bound of the 95% HPD were tabulated respectively.

	mutation rate (substitutions per site per year)		
	mean	**lower bound**	**upper bound**
BC1-19F clade_1	2.60 x 10^{-6}	1.58 x 10^{-6}	3.71 x 10^{-6}
BC1-19F clade_2	2.86 x 10^{-6}	1.69 x 10^{-6}	4.07 x 10^{-6}
BC1-19F clade_3	2.35 x 10^{-6}	1.15 x 10^{-6}	3.68 x 10^{-6}
BC2-23F clade_1	1.83 x 10^{-6}	9.96 x 10^{-7}	2.70 x 10^{-6}
BC2-23F clade_2	1.45 x 10^{-6}	2.89 x 10^{-7}	2.66 x 10^{-6}
BC3-NT clade_1	4.49 x 10^{-6}	1.61 x 10^{-6}	9.78 x 10^{-6}
BC3-NT clade_2	2.39 x 10^{-6}	8.49 x 10^{-7}	4.38 x 10^{-6}
BC4-6B clade_1	3.08 x 10^{-6}	1.29 x 10^{-6}	5.13 x 10^{-6}
BC4-6B clade_2	3.07 x 10^{-6}	4.36 x 10^{-7}	9.83 x 10^{-6}
BC5-23A/F	3.26 x 10^{-6}	1.02 x 10^{-6}	6.60 x 10^{-6}
BC7-14 clade_1	2.79 x 10^{-6}	1.16 x 10^{-6}	4.67 x 10^{-6}
BC7-14 clade_2	4.81 x 10^{-6}	5.31 x 10^{-7}	9.77 x 10^{-6}
BC7-14 clade_3	4.39 x 10^{-6}	2.65 x 10^{-6}	8.14 x 10^{-6}

Supplementary table 4 Recombination per mutation (r/m) calculated from linear regression and arithmetic mean.

Testing clusters	**Population size**	**r/m**		**Hypothesis**	**Test and p-value**	
		Estimated by linear regression **(95% confident interval)**	**Estimated by arithmetic mean** **(95% confident interval)**		**ANCOVA (difference in slope calculated by linear regression)**	**Kruskal-Wallis (difference in arithmetic mean)**
BC1-19F	365	0.233 (0.210 – 0.256)	0.229 (0.195 – 0.263)	r/m of BC3-NT > other clusters	1.10 x 10^{-3}	1.76 x 10^{-5}
BC2-23F	213	0.092 (0.068 – 0.117	0.140 (0.068 – 0.212)			
BC3-NT	202	0.310 (0.284 – 0.336)	0.320 (0.289 – 0.351)			
BC4-6B	126	0.107 (0.005 – 0.209	0.132 (0.037 – 0.227)			
BC5-23A/F	106	0.147 (0.132 – 0.162)	0.146 (0.100 – 0.192)			
BC6-15B/C	102	0.122 (0.070 – 0.174)	0.200 (0.115 – 0.285)			
BC7-14	102	0.257 (0.211 – 0.304)	0.148 (0.086 – 0.210)			
NT within BC3-NT	128	0.341 (0.288 – 0.395)	0.343 (0.307 – 0.379)	within BC3-NT, r/m of NT > serotype 14	NA	2.44x 10^{-3}
Serotype 14 within BC3-NT	74	Assumptions of the linear regression models were not met	0.203 (0.164 – 0.242)			

Supplementary table 5

Trend in antibiotic consumption based on the Burmese border guidelines (1994 – 2010)

Year	Co-trimoxazole consumption	B-lactam consumption
1994	Co-trimoxazole was the primary treatment for non-severe pneumonia, otitis media, urinary tract infection, and dysentery.	Ampicillin was used for severe pneumoniae, meningitis and for infections in pregnant women where co-trimoxazole was contra-indicated.
1999	As in 1994.	As in 1994, but with amoxicillin being used as second line treatment for non-severe pneumoniae (if no improvement with co-trimoxazole).
2002/2003	As in 1999 but ciprofloxacin replaced co-trimoxazole for dysentery.	As in 1999 with ceftriaxone appearing as an alternative drug for meningitis and typhoid.
2007	Amoxicillin now replaced co-trimoxazole for non-severe pneumonia.	Amoxicillin was recommended as primary treatment for non-severe pneumonia while ceftriaxone was first line for meningitis.

Supplementary table 6

Potential donors for each recombinant fragment detected in isolate SMRU1452. Nine recombinant fragments (recipient blocks A - I) were detected in this isolate. Potential donor clusters were identified for each recipient block. “1” and “0” indicate donor, and non-donor respectively.

		Recipient blocks								
		A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
donor clusters	**sBC 54-57**	1	0	0	0	0	0	0	0	0
	sBC181-184	1	0	1	0	0	0	0	0	0
	sBC8	1	0	0	0	0	0	0	0	0
	sBC107-108	0	0	0	0	0	0	0	0	1
	sBC 86	1	0	1	0	0	0	0	0	0
	sBC 7	0	0	1	0	0	0	0	0	0
	sBC 145-152	1	1	1	1	0	1	1	1	1
	sBC 140-143	1	0	0	0	1	1	0	0	0

Supplementary table 7

References used for mapping and mapping coverage generated for each dominant cluster.

strain	reference	accession number	serotype	ST	use in the study	genome size (bp)	% mapping
Spanish23F (ATCC 700669)	Public database	FM211187	23F	81	map against the whole population to determine coarse structure	2221315	82.33
Taiwan19F-14	Public database	CP000921	19F	236	unique referenece for BC1-19F	2112148	96.79
INV200	Public database	FQ312029	14	9	unique reference for BC3-NT	2093317	91.42
G54	Public database	CP001015	19F	63	unique reference for BC7-14	2078953	96.22
SMRU1949	new reference from draft assembled genome	NA	23F	802	unique reference for BC2-23F	1935768	96.53
SMRU2513	new reference from draft assembled genome	NA	6B	315	unique reference for BC4-6B	1991123	95.92
SMRU1861	new reference from draft assembled genome	NA	23F	2218	unique reference for BC5-23A/F	1896242	94.91
SMRU1478	new reference from draft assembled genome	NA	15C	4209	unique reference for BC6-15B/C	1933435	96.73

Supplementary table 8

Numbers of combined recombination events or recombinant blocks detected in this study.

	Number of combined recombination events or recombinant blocks		
	Total events or blocks	Events or blocks associated with mobile genetic elements	Events or blocks associated with recombination hotspots
A) All recombination events predicted by method described in Croucher et al.	2,209	132	450
B) Recombination detected at the external node of method in Croucher et al., which is likely to represent recent recombination	620	34	94
C) Recombinant blocks predicted by method described in Martinen et al. that show overlap with regions in B)	928	Excluded from the analysis	71
D) Recipient blocks in C) where potential donors can be found	443	Excluded from the analysis	50

Supplementary table 9

Associations between recombining strains, antibiotic resistant phenotypes and temporal changes in recombining trend in details.

		Recombination vs no recombination		**Recent vs older recombination**	
		no. of strains undergoing recombination at loci of interests	**no. of strains with no recombination at loci of interests**	**more recent recombination (external branches)**	**Older recombination (internal nodes)**
Beta-lactam	resistant	795	120	25	770
	sensitive	150	146	6	144
co-trimoxazole	resistant	873	210	10	863
	intermediate	55	20	5	50
	sensitive	45	8	4	41

Supplementary note

Detailed sample collection

Nasopharyngeal swabs used in this study was part of a carriage cohort conducted by Turner *et al* [1,2]. Written consent was gained from all participants prior to enrollment in the study. Ethical approval was granted by the Ethics Committee of the Faculty of Tropical Medicine, Mahidol University, Thailand (MUTN 2009-306) and the Oxford Tropical Research Ethics Committee, Oxford University UK (031-06). The longitudinal study was performed between 2007-2010 from 528 infants as well as from 242 mothers, yielding 6,747 pneumococci from 11,829 cultured nasopharyngeal swabs. Serotyping was performed using latex-agglutination and Quellung reaction. All potentially non-typeable pneumococci were confirmed by bile solubility and absence of capsular swelling with Omniserum (SSI Diagnostica, Denmark). Penicillin and co-trimoxazole susceptibilities were determined by disk diffusion following current CLSI guidelines. Penicillin MIC was determined by E-test if disk diffusion test indicated potential penicillin non-susceptibility (1 µg oxacillin disk zone diameter of < 20 mm)

Control for sample mix up through determination of serotype and sequence type

Serotype and Multi Locus Sequence Typing (MLST) were derived from Illumina read data as described in [3] and comparing to serotype detected from latex-agglutination and Quellung reaction for quality control purposes. For non-typeable (NT) serotypes, methods as explained in Salter *et al.* [4] were used to confirm the absence of a capsule locus or the presence of alternative genes detected between *dexB* and *aliA*. Diversity of the NT category detected in this population is summarized in **Supplementary table 2.**

Sequence assembly

3,085 strains were *de novo* assembled multiple times using Velvet [5], where the kmer size was varied between 60% and 90% of the read length. The assembly with the best N50 was chosen. Contigs shorter than the insert size length were filtered out because they are most likely misassemblies. The sequencing data were then used to improve further the assembly. The contigs were iteratively scaffolded and extended 16 times using SSPACE [6] beginning with the contigs where the greatest number of reads overlap. Gaps, denoted by 1 or more N's were targeted for closure by running 120 iterations of GapFiller [7], cycling between BWA [8] and Bowtie [9], beginning where the greatest number of reads overlapped. A final QC step was performed on each assembly, with the reads mapped back to the assembly using SMALT 0.5.7.

Coarse mapping using single reference

To estimate the whole population structure, reads from all 3,085 samples were mapped onto a single reference genome, *S. pneumoniae* ATCC700669 [EMBL accession code FM211187] [10] using SMALT 0.5.7 to generate a coarse but sufficient alignment for determining the population structure. The 2,221,315 bp reference gave on average 82.33 percent mapping coverage. Bases were called using the method described in [11].

Fine mapping with unique reference for each studied cluster

Further analyses within focused clusters required genome alignment with higher resolution. To improve the resolution given by coarse mapping against a single reference, closely related references were employed for finer mapping in 7 dominant clusters. We first screened streptococcal whole genome archive which are publicly available for reference that have identical sequence type or differ from members of each cluster within double loci variants (*ddl* was omitted as this gene is potentially linked with penicillin binding resistant gene and thus appear to be diverse in penicillin-resistant isolates). When more than one potential reference was available, a reference that gave the highest mapping coverage was then selected for each cluster. Mapping was done as described previously. The final alignment included indels using the method described in [11].

When public references were not available, our own references were made from selected samples within the cluster. References were created by assembly as described earlier and ordered relative to its closest references using ABACAS v2.5.1 [12] and ACT [13]. Annotations were directly transferred from *S. pneumoniae* ATCC 700669 followed by manual curation. References used for mapping within the focused cluster and the mapping scores are listed in **Supplementary table 7**.

Estimating population structure

Based on the coarse mapping against the core genome of *S. pneumoniae* ATCC 700669, the BAPS software v6.0 [14-17] was used to estimate the population structure. The estimation algorithms in BAPS are based on non-reversible stochastic optimisation, which overcomes the challenge of doing inference from large data sets with complex structure, as shown recently by applications to bacterial populations of several different species [18-21]. As described in [22,23], we used BAPS in a hierarchical manner to resolve the population structure at a fine level of detail. First, the module for clustering individual strains was applied to obtain the posterior mode partition into primary clusters based on 5 runs of the estimation algorithm. Data from each of these primary clusters were then analysed again with BAPS in an identical manner to obtain secondary clustering within each primary cluster. As noted in [22,23], this approach captures efficiently the population structure at a finer level when the genetic differences between the major lineages in data are so large that they mask more subtle signals of divergences present within a lineage. In this dataset, 33 primary and 183 secondary clusters were determined (**Supplementary table 1**).

Detection of serotype switches

States of changes in serotypes were counted based on parsimony reconstruction of serotypes onto the phylogenetic tree represented in **Figure 1A**.

Selection of dominant BAPS clusters and construction of individual cluster phylogeny

Out of 11 large primary clusters that comprised more than 100 isolates, 7 primary clusters appear to have members that either share the same serotype/serogroup or differ by a few MLST locus variants, suggesting that isolates within these clusters are not too distant to exceed the limitations of recombination detection tools. As a result, BC1-19F (1st BAPS cluster 16, n = 365 isolates), BC2-23F (1st BAPS cluster 4, n = 213 isolates), BC3-NT (1st BAPS cluster 10, n = 202 isolates), BC4-6B (1st BAPS cluster 25, n = 126 isolates), BC5-23A/F (1st BAPS cluster 22, n = 106 isolates), BC6-15B/C (1st BAPS cluster 32, n = 102 isolates) and BC7-14 (1st BAPS cluster 21, n = 102 isolates) which collectively represent 39.4% of the population, were selected. The phylogeny based on single nucleotide substitution was constructed for each selected cluster using the method described in [3], which separate signals of single nucleotide substitution from recombination and use only substitution sites for building the phylogenetic tree. The phylogeny for each cluster is shown in **Supplementary figure 4a – 4g**.

Estimation and comparison of mutation rates of individual clusters

Following separation of recombination SNPs from mutation SNPs, rooted, time – measured phylogenies can be inferred from point mutation data. There was difficulty in correlating the accumulation of SNPs through time from the whole cluster due to narrow time frame. **Supplementary figure 5** shows the case when combined signals were considered using BC5-23A/F as an example. A linear regression cannot be determined from the plot as a whole when signals from all subclades are combined. However, there are distinct clusters representing small subclades where there are good correlations between SNP accumulation and time. When considering each subclade individually, the temporal signal and clocklikeness of each subclade phylogenies were captured by Path-O-Gen (http://tree.bio.ed.ac.uk/software/pathogen) (**Supplementary figure 6**). Also, the subclades where substitution rates were derived are highlighted in **Supplementary figure 4a – 4g**. The mutation rate from each subclade was calculated with BEAST [24] using the skyline population size prior and a relaxed lognormal clock model (tabulated in **Supplementary table 3**). Comparison of nucleotide substitution rate between different clusters was conducted using the Kruskal-Wallis test.

Estimation of level of homologous recombination within individual clusters

Recombination levels of 7 dominant clusters were calculated given numbers of recombination events and number of single polymorphic sites produced by the algorithm described in [3]. Recombination signals localised in the regions of mobile genetic elements were removed so that only homologous recombination were considered.

The number of recombination events per number of point mutations (r/m) observed in each branch was calculated. For the numerator (r), we used the number of recombination events instead of numbers of polymorphic sites introduced by recombination as originally given in [25]. As numbers of introduced polymorphic sites depend on the genetic distance between donor and recipient, our method help reduce this bias for our cross-lineage comparison. Using the number of recombination events, the r/m calculated here is expected to be lower than reported earlier [3].

The ratio r/m was calculated and compared by two different approaches.

By modeling the relationship between recombination events and mutations as a linear regression (**Supplementary figure 7**), using the ranked recombination events as the outcome, and ranked number of SNPs as the predictor variable. The slope, representing r/m, was calculated and is reported in **Supplementary table 4**. Where the assumptions of linear regression were met, an ANCOVA test was used to test whether or not there was a significant difference in r/m between each cluster.

And by using the arithmetic mean of r/m of a cluster, averaged from the r/m of each branch within a cluster. For each major cluster, the r/m was calculated separately for each branch, and the mean of the distribution of the r/m for the cluster reported in **Supplementary table 4**. The Kruskal-Wallis test was used to test for differences in r/m between clusters calculated by arithmetic mean.

Associations between *pbp1a, pbp2b, pbp2x, folA, folP* recombining strains, antibiotic resistant phenotypes and temporal changes in antibiotic consumptions

The DNA sequences of the above genes, whose allelic forms are known to confer resistant to β-lactam (*pbp1a, pbp2b, pbp2x*) and co-trimoxazole (*folA, folP*), and appear to be recombination hotspots, were extracted from the draft genome assembly. Phylogenies of individual gene trees: *pbp1a*, *pbp2b*, *pbp2x* (**Supplementary figure**

1a to 1c); *folA* and *folP* (**Figure 4b**), as well as concatenated phylogenetic tree for *pbp genes* (**Figure 4a**) were estimated with RAxML v7.0.4 [26] using a GTR model with a gamma correction for site rate variation using 100 bootstraps.

The trend of recombination was estimated through the detected phenotypes observed in the presence and absence of recombination in the sub-population including 7 most prevalent clusters. Please note that 5 isolates with missing phenotypes (**Supplementary table 1**) were not included in this analysis. Based on the prediction of recombination from 7 dominant clusters, strains undergoing recombination at *pbp1a*, *pbp2b, pbp2x, folA* or *folP* and their phenotypic resistance to β-lactam and co-trimoxazole were compared against the strains with no recombination events observed at these sites (**Supplementary table 9).** Statistical difference between the recombining group and the non-recombining group was estimated with two – tailed Fisher's exact test. Alternative *murM* and *murN* genes associated with high β-lactam resistant [27] were also considered. However, only two candidates with partial matches were observed and thus less likely to explain trends in β-lactam resistant.

Temporal trends in recombination were determined by comparing the phenotype difference between strains undergoing recent recombination (recombination events predicted at the external branches) to strains whose ancestors had undergone recombination (recombination events predicted at the internal nodes). Using two – tailed Fisher's exact test, statistical difference between these groups was estimated.

Trends in antibiotic consumption obtained from recommended treatments are tabulated in **Supplementary table 5**.

Searching for potential of recombination donors given recipient blocks

Based on the sequence identity of recombination fragments detected in recipient strains, potential donors from the rest of population were assessed. Number of recipient blocks used for this analysis is summarized in **Supplementary table 8**. To maximise detection specificity and reduce false positives, several criteria were applied as follows:

i) Using two independent algorithms for predicting recombinant fragments in the recipients

The recombination detection method described in [3] and BRATnextgen [28] were used to reduce false positive and refine recombination boundaries. Comparisons of the length of predicted recombinant fragments and percent of unknown characters "N" found in the fragments from both algorithms were summarized in **Supplementary figure 8b and 8c**respectively. The overlap regions predicted by both algorithms are highlighted in **Supplementary figure 8a,** presenting the phylogeny and recombination detected in BC7-14, one of the smallest among other prevalent clusters.

ii) Only recent recombination events were considered

Only recent recipient blocks detected at the external branch (identified using the algorithm described in [3]) and coinciding with recombination sites predicted by BRATnextgen were used in the analysis. A focus on recent recombination occurring at the external branch alone reduces the chances of the donor detection being confounded by subsequent recombination events.

As the studied exchanges are constrained to be relatively recent by occurring on external branches of the phylogeny, no limitations were placed on the pairings of sequence donors and recipients based on date of isolation. This is because the short sampling time frame makes it likely that all sequenced isolates represent genotypes extant in the camp over the period in which the studied recombinations are likely to have happened.

iii) No unknown mapping character 'N' is allowed

As ‘N’s, unknown nucleic acid residues generate no specific match, recipient and donor blocks were checked for sequence quality such that no sequences with ‘N’s were allowed.

iv) Checking length of recipient queries

We checked the length of recipient blocks (blast queries) and the number of potential donors detected (blast hits) to examine the specificity of the search. Our queries ranged between 10 – 6,846 bp, with a mean length of 1,162 bp. Distributions of the length of the queries (recipient blocks) used for blast and queries where identical hits were detected are similar (**Supplementary figure 9)**, suggesting that there is no length bias for the blast search. As expected, the number of hits decreases as the length of the query increases and this can be modeled with a negative exponential function (**Supplementary figure 9**).

v) Hits must be identical matches

Recombination recipient blocks were used as query sequences for nucleotide blast searches for potential donor blocks from 3,085 draft assembled genomes. Using blastall v 2.2.15 [29], a search against the genome of the recipient itself was used as a positive control. Any hits that have an exact match to the score given by the positive control are likely to be potential donors. Although recombination may lead to insertions or deletions over the recombining region, indels were not considered here as the searches were performed in closely related species, which require high specificity for the recipient-donor relationship to be drawn.

Calculating the probability of a single isolate acting as a donor for a recipient

For each recipient isolate, “n” potential donor isolates were identified, and each donor isolate was assigned a probability of “1/n” of having been the donor. Isolates showing no particular hit for a particular search were given a probability of 0. The total frequency of each isolate in being the donor was represented by the sum of the above probabilities from all potential donation events. Isolates were grouped into lineages based on their population cluster from BAPS. The boxplots (**Supplementary figure**

3a) show the distribution of probabilities of isolates within that cluster acting as a donor. However, identical matches could come from shared recipients of recombinant fragments as well as true donors. As we cannot distinguish the two events apart based on the sequenced isolate alone, the results should be interpreted with caution.

Filtrations of potential BAPS donor clusters

To reduce the random hits from non-related clusters, we only allowed i) clusters most commonly detected as sources for each recipient isolate and ii) clusters detected as the sole source of recombinant DNA in each recipient isolate. To illustrate with an example, a case where multiple potential donor clusters were identified from nine recipient blocks of a single recipient isolate SMRU1452 (**Supplementary figure 2**) were tabulated in **Supplementary table 6**. A hit was scored as "1" while no hit was scored "0". The most common sources came from sBC145-152 and thus was selected based on the first criterion. Also, sole-source donors were detected in sBC145-152 (sole sources for recipient blocks B, D, G and H) and sBC140-143 (recipient block E). Thus, the clusters were selected according to the second criterion.

Evaluating the relationship between cluster size, cluster diversity, and probability of being a donor

Following cluster filters, a similar procedure as described above for calculating the probability of a single isolate acting as a donor was performed at the cluster level to calculate the probabilities of particular clusters being recombination event donors for each recipient isolate. For a search matching "n" potential donor clusters for each recipient block, each identified cluster was assigned a probability of "1/n" of being the donor. Clusters that did not contain any potential donor isolates were given a probability of 0. The total probability of each cluster having acted as donors was calculated as the sum of above probabilities.

We then tested whether or not there are correlations between the probability of being donors and other two features, the cluster population size and the diversity within the cluster. Population size presenting the number of isolates detected in a particular cluster was plotted against the cluster probability of being donors **(Supplementary figure 3b)**. The diversity within a cluster was calculated as the sum of total branch lengths, which is proportional to the number of polymorphic sites (including both mutation and recombination) observed in that particular cluster. The diversity observed within each cluster was plotted against the overall probability of each cluster acting as donors (**Supplementary figure 3c**). Please note that 4 primary BAPS clusters with members on polyphyletic branches were excluded from this analysis. Based on the Spearman ranking correlation, the association between both features and the probability of becoming donors were estimated.

References

1. Turner, P. *et al.* A longitudinal study of Streptococcus pneumoniae carriage in a cohort of infants and their mothers on the Thailand-Myanmar border. *PloS one* **7**, e38271 (2012).
2. Turner, C. *et al.* High rates of pneumonia in children under two years of age in a South East Asian refugee population. *PloS one* **8**, e54026 (2013).
3. Croucher, N.J. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430-4 (2011).
4. Salter, S.J. *et al.* Variation at the capsule locus, cps, of mistyped and non-typable Streptococcus pneumoniae isolates. *Microbiology* **158**, 1560-9 (2012).
5. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821-9 (2008).
6. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-9 (2011).
7. Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome biology* **13**, R56 (2012).
8. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
9. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25 (2009).
10. Croucher, N.J. *et al.* Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone Streptococcus pneumoniaeSpain23F ST81. *Journal of bacteriology* **191**, 1480-9 (2009).
11. Harris, S.R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469-74 (2010).
12. Assefa, S., Keane, T.M., Otto, T.D., Newbold, C. & Berriman, M. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25**, 1968-9 (2009).
13. Carver, T.J. *et al.* ACT: the Artemis Comparison Tool. *Bioinformatics* **21**, 3422-3 (2005).

14. Corander, J., Waldmann, P. & Sillanpaa, M.J. Bayesian analysis of genetic differentiation between populations. *Genetics* **163**, 367-74 (2003).
15. Corander, J. & Tang, J. Bayesian analysis of population structure based on linked molecular information. *Mathematical biosciences* **205**, 19-31 (2007).
16. Corander, J., Marttinen, P., Siren, J. & Tang, J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC bioinformatics* **9**, 539 (2008).
17. Tang, J., Hanage, W.P., Fraser, C. & Corander, J. Identifying currents in the gene pool for bacterial populations using an integrative approach. *PLoS computational biology* **5**, e1000455 (2009).
18. Hanage, W.P., Fraser, C., Tang, J., Connor, T.R. & Corander, J. Hyper-recombination, diversity, and antibiotic resistance in pneumococcus. *Science* **324**, 1454-7 (2009).
19. Corander, J., Connor, T.R., O'Dwyer, C.A., Kroll, J.S. & Hanage, W.P. Population structure in the Neisseria, and the biological significance of fuzzy species. *Journal of the Royal Society, Interface / the Royal Society* (2011).
20. Mutreja, A. *et al.* Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**, 462-5 (2011).
21. Cheng, L., Connor, T.R., Siren, J., Aanensen, D.M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Molecular biology and evolution* (2013).
22. Willems, R.J. *et al.* Restricted gene flow among hospital subpopulations of Enterococcus faecium. *mBio* **3**, e00151-12 (2012).
23. Cheng, L., Connor, T.R., Siren, J., Aanensen, D.M. & Corander, J. Hierarchical and Spatially Explicit Clustering of DNA Sequences with BAPS Software. *Molecular biology and evolution* **30**, 1224-8 (2013).
24. Drummond, A.J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology* **7**, 214 (2007).
25. Feil, E.J., Maiden, M.C., Achtman, M. & Spratt, B.G. The relative contributions of recombination and mutation to the divergence of clones of Neisseria meningitidis. *Molecular biology and evolution* **16**, 1496-502 (1999).
26. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-90 (2006).
27. Smith, A.M. & Klugman, K.P. Alterations in MurM, a cell wall muropeptide branching enzyme, increase high-level penicillin and cephalosporin resistance in Streptococcus pneumoniae. *Antimicrobial agents and chemotherapy* **45**, 2393-6 (2001).
28. Marttinen, P. *et al.* Detection of recombination events in bacterial genomes from large population samples. *Nucleic acids research* **40**, e6 (2012).
29. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-10 (1990).