Supporting Information Appendix

Diverse evolutionary patterns of pneumococcal antigens identified by pangenome-wide immunological screening

Nicholas J. Croucher[1], Joseph J. Campo[2], Timothy Q. Le[2], Xiaowu Liang[2], Stephen D. Bentley[3], William P. Hanage[4], Marc Lipsitch[4]

Affiliations:

[1] Department of Infectious Disease Epidemiology, Imperial College London, W2 1PG, UK

[2] Antigen Discovery Inc., Suite E309, 1 Technology Drive, Irvine, CA 92618, USA

[3] Infection Genomics, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

[4] Center for Communicable Disease Dynamics, Harvard T. H. Chan School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA

Contents

Text S1: Design and construction of the proteome microarray

Proteome microarray design

Clusters of orthologous genes (COGs) were defined and made publicly available previously (1, 2). Initially, 2,089 of the 5,442 COGs were selected to be represented on the microarray, as they were present in 20% or more of the isolates. Each COG's representative was selected as being a sequence of median length; in cases where the length distribution was skewed as a result of misassemblies, more complete representatives were selected manually. Added to this were rare surface structures likely to be immunogenic: three COGs (CLS01943, CLS02796 and CLS02942) were added that represented the three variants of the RrgB type 1 pilus protein (3), five COGs (CLS02869, CLS02870, CLS02871, CLS02872 and CLS02873) were added to represent the components of the type 2 pilus (4), and four COGs (CLS03178, CLS03265, CLS03616 and CLS99466) were added to represent the allelic diversity of PclA (5), alongside the already-included most common variant (CLS01333). Similarly, the three COGs corresponding to Pbp1A, Pbp2X and Pbp2B were represented by three, three and four sequences, respectively, to reflect the allelic diversity associated with different levels of β lactam sensitivity (1). To correct cases where repetitive proteins had not assembled completely, CLS02463 was replaced with PsrP from *S. pneumoniae* ATCC 700669 (6); CLS02794 was replaced with a PblB-like protein from isolate R34-3103; and CLS00097 was replaced with the LytA cellular amidase from *S. pneumoniae* TIGR4 (7) and a prophage amidase from isolate R34-3070 (1).

A more detailed representation of the zinc metalloprotease repertoire was included on the microarray. Therefore the five COGs corresponding to ZmpA and fourteen COGs corresponding to ZmpB were omitted. Instead, the sequences within each set of COGs were trimmed to remove the N terminal region up to, and including, the LPXTG sortase attachment site. A fully-overlapping set of every 15 amino acid segment was extracted from the remaining mature sequence for each representative, and a distribution of pairwise similarities generated based on the number of shared identical 15-mers. Sequences were included in the same cluster if the proportion of shared 15-mers was greater than a threshold (0.575 for ZmpA, 0.475 for ZmpB) derived from the empirical distribution of 15-mer distances (Fig. S1). These were supplemented with the COGs for ZmpC (CLS01991), the most common variant of ZmpD (CLS02608) and a representative of ZmpE from the sequence cluster (SC) 12 isolate R34-3088.

Analysis of PspA and PspC sequences

A related approach was used to study the diversity of PspA and PspC; however, these genes were rarely present as a full-length sequence in the original *de novo* assemblies. Improved draft assemblies were generated from the same sequence read data using the PAGIT pipeline (8). For both genes, the 500 bp immediately upstream and downstream of the corresponding CDSs in *S. pneumoniae* ATCC 700669 (6) were extracted and aligned to each draft assembly using BLASTN (9). All CDSs predicted to occur between these matches, or between these matches and the end of the relevant contig if a break in the assembly occurred between them, were translated and extracted. Sequences were considered surface protein candidates if they contained either a signal peptide (identified by SignalP (10)) or YSIRK motif (identified by Pfam (11)) at their N terminus, or either a choline-binding domain or sortase attachment signal at their C terminus (identified by Pfam (11)). This approach identified 573 PspA candidates and 509 PspC candidates in the 616 isolates. These were filtered to the sets that included both an N and C terminal motif, resulting in a curated set of 207 unique, complete *pspA* DNA sequences and 208 unique, complete *pspC* DNA sequences. After removal of the C termini of these CDSs, which comprised the choline binding domains or sortase attachment site and the subsequent distal sequence, they were used to identify the distribution of allelic diversity across the population at these loci through sequence read mapping.

The approach used was similar to that previously applied to identifying functional variants of the *ivr* locus (12). Illumina reads were mapped against the relevant upstream sequence using BWA (13), and for those read pairs in which only one read matched this region, the unmapped partners were used as a set of queries likely to be matching the 5' region of the CDS of interest. These unmapped reads were then aligned to the database of *pspA* or *pspC*, trimmed as described above, using megablast (9) with a requirement for complete identity over at least 75% of the read's length. This identified candidate variants based on the sequence at the 5' end of the gene; these were concatenated, and used as a reference sequence for a second round of mapping with BWA conducted using all Illumina reads for the isolate. The surface protein sequence with the greatest number of unique hits was then assigned to the isolate. In total, 128 different surface protein sequences were assigned to isolates for each of the *pspA* and *pspC* loci.

An equivalent process to that used for ZmpA and ZmpB was used to cluster these sequences into variants. In addition to the removal of the C terminal motifs for surface attachment, the N terminal motifs for secretion were also trimmed from the PspA and PspC sequences. The pairwise similarities between the remaining central segments were then calculated based on the proportion of 15-mers extracted from each sequence that were shared by both. A threshold for the definition of variants was then selected based on the empirical distribution of pairwise similarities for both proteins: a threshold similarity of 0.525 was used for PspA, giving 39 variants, while a threshold of 0.675 was used for PspC, giving 59 variants.

To validate this approach, phylogenies were constructed to ensure no diversity was lost during the processing of the antigens. The 203 unique *pspA* DNA sequences, following the trimming of both ends to remove secretory and surface attachment motifs as described above, were translated and aligned with those sequences described previously by Hollingshead *et al* (14), as well as representatives from *S. pneumoniae* TIGR4 and ATCC 700669 (6, 7), using MUSCLE (15). The tree shown in Fig. S2 was then constructed with FastTree2 (16); it is coloured according to the six 'clades' defined by Hollingshead *et al*, and the numbers around the edge denote those variants assigned to isolates through mapping. This shows the full diversity of the protein is clearly represented, and assigned to isolates, in the dataset. The same approach was used to compare amino acid translations of the 207 unique *pspC* DNA sequences, trimmed as described above, with those described previously by Brooks-Walter *et al* (17), as well as representatives from *S. pneumoniae* TIGR4 and ATCC 700669. The resulting phylogeny is shown coloured according to the clades defined by Brooks-Walter *et al*, and otherwise annotated as for the PspA tree (Fig. S2). Similarly, the full diversity of the protein is observed across the population, and assigned to isolates. The accuracy of assignation was also checked; for the 128 PspA DNA sequences used to identify variants in this study, mapping assigned 102 back to the isolate from which they were originally extracted, and in the remaining 27 cases, the isolates were assigned to an alternative variant that was >99% identical at the protein level to that originally extracted from their assembly. For the 128 PspC variants, mapping assigned 104 back to the isolate from which they were originally extracted, and in all but five cases, the isolates were assigned to an alternative variant that was >99% identical at the protein level to that originally extracted from their assembly. The anomalies are likely to reflect rare cases of misassemblies.

Proteome microarray construction

A library of partial or complete CDSs cloned into a T7 expression vector pXI has been established at Antigen Discovery, Inc. (ADi, Irvine, CA, USA). This library was created through an *in vivo* recombination cloning process in which PCR-amplified CDSs and a complementary linearized expressed vector were transformed into chemically competent *E. coli*; amplification by PCR and cloning into pXI vector using this high-throughput PCR recombination cloning method is described elsewhere (18). All CDSs in Dataset S1 were cloned for expression as full-length proteins. For those CDSs >3 kb in length, as well attempting to express the full-length CDS, the sequences were also cloned as overlapping fragments, each <3 kb. Of 2,189 CDSs representing the Massachusetts pneumococcal population, 67 were >3 kb and expressed as overlapping fragments.

Proteins were expressed with a 5’ polyhistidine (HIS) epitope and a 3’ hemagglutinin (HA) epitope using an *in vitro* transcription and translation (IVTT) system, the *Escherichia coli* cell-free Rapid Translation System (RTS) kit (5 Prime, Gaithersburg, MD, USA), according to the manufacturer’s instructions. Translated proteins were printed onto nitrocellulose-coated glass AVID slides (Grace Bio-Labs, Inc., Bend, OR, USA) using an Omni Grid Accent robotic microarray printer (Digilabs, Inc., Marlborough, MA, USA). The full microarray was printed on a nitrocellulose ‘pad’, present in triplicate on each slide, allowing three samples to be probed per slide. Microarray chip printing and protein expression were quality checked by probing random slides with anti-HIS and anti-HA monoclonal antibodies with fluorescent labeling.

Proteome microarray sample probing

Microarrays were probed with sera sampled from healthy immunocompetent volunteers from the USA aged between 18 and 40 in a Phase I clinical trial of an *S. pneumoniae* whole cell vaccine (ClinicalTrials.gov identifier: NCT01537185), taken before immunization. Prior pneumococcal infection or vaccination with a different antipneumococcal vaccine was not specified as a criterion for inclusion or exclusion from the trial. Serum samples were diluted 1:100 in a 3 mg mL^{-1} *E. coli* lysate solution in protein arraying buffer (Maine Manufacturing, Sanford, ME, USA) and incubated at room temperature for 30 min. Chips were rehydrated in blocking buffer for 30 min. Blocking buffer was removed, and chips were probed with pre-incubated serum samples using sealed, fitted slide chambers to ensure no cross-contamination of sample between pads. Chips were incubated overnight at 4°C with agitation. Chips were washed five times with TBS-0.05% Tween 20, followed by incubation with biotin-conjugated goat anti-human IgG (Jackson ImmunoResearch, West Grove, PA, USA) diluted 1:200 in blocking buffer at room temperature. Chips were washed three times with TBS-0.05% Tween 20, followed by incubation with streptavidin-conjugated SureLight P-3 (Columbia Biosciences, Frederick, MD, USA) at room temperature protected from light. Chips were washed three times with TBS-0.05% Tween 20, three times with TBS, and once with water. Chips were air dried by centrifugation at 1,000 x g for 4 min and scanned on a GenePix 4300A High-Resolution Microarray Scanner (Molecular Devices, Sunnyvale, CA, USA), and spot and background intensities were measured using an annotated grid file (.GAL). Data were exported in Microsoft Excel.

Proteome microarray data normalisation

Raw spot and local background fluorescence intensities, spot annotations and sample phenotypes were imported and merged in R (19), in which all subsequent procedures were performed. Foreground spot intensities were adjusted by local background by subtraction, and negative values were converted to 1. Next, all foreground values were transformed using the base 2 logarithm. The dataset was normalized to remove systematic effects by subtracting the median signal intensity of the IVTT controls for each sample. Since the IVTT control spots carry the chip, sample and batch-level systematic effects, but also antibody background reactivity to the IVTT system, this procedure normalizes the data and provides a relative measure of the specific antibody binding to the non-specific 'background' antibody binding to the IVTT controls. With the normalized data, a value of 0.0 means that the intensity is no different than the background, and a value of 1.0 indicates a doubling with respect to background.

Classification of proteome microarray probes

The microarray contained 4,504 expressed polypeptides, most of which were full-length proteins, with some corresponding to fragments of proteins encoded by CDSs >3 kb in length. These represented the full proteome of *S. pneumoniae* TIGR4 (7) and 2,190 proteins from the Massachusetts collection of isolates (1). The 2,190 proteins were divided into 2,057 COG, 36 PspA, 57 PspC, 18 ZmpA and 16 ZmpB representatives, along with individual sequences for LytA, a phage amidase, ZmpE, PblB, PsrP and a choline-binding domain. For each person, each protein was represented by the probe recording the highest normalised, $\log_2$ transformed antibody response score, to avoid probes outside of epitopes reducing the recorded immune response to a protein. This meant every protein was represented by one measurement per individual in the study; an all-versus-all Euclidean distance matrix between proteins was constructed using these IgG binding levels, and split into two categories using the R function 'hclust' (19). The 'antibody targets' (ABTs) were the 208 proteins in the group associated with higher levels of IgG binding.

In order to compare this classification with the previous work of Giefing *et al* (20), it was necessary to link the proteome of *S. pneumoniae* TIGR4 (7) to the COGs of the studied pneumococcal population. To avoid the problems of different predicted translation initiation sites, and repetitive regions at the C terminal end of many surface-associated proteins, links were initially made through querying the database of proteins from the study population for an exact match to the middle 50% of the *S. pneumoniae* TIGR4 protein. If this did not find a link, then queries were instead performed with successive 50 aa segments extracted from the N terminal 75% of the protein. Following the manual addition of links to LytA (SP_1937), PsrP (SP_1772), ZmpA (SP_1154), ZmpB (SP_0664), PspA (SP_0117) and PspC (SP_2190), 2,039 of the 2,125 *S. pneumoniae* TIGR4 CDSs were linked to 1,900 COGs; cases of multiple CDSs linking to the same COG represented close paralogues (e.g. IS elements) linking to a single protein on the microarray.

To identify characteristics that were associated with ABTs, a binomial logistic regression was conducted using ABT status as the binary dependent variable in R. The explanatory variables were mean COG coding sequence length (continuous), number of transmembrane helices in the protein predicted by TMHMM version 2 (discrete) (21), presence of signal peptide as predicted by signalP version 4.1 (binary) (22), presence of a lipoprotein processing signal as predicted by Prosite motif PS51257 (binary) (23), and the presence of Pfam domains (binary) (11), limited to those domains found in at least five COGs. The 'safeBinaryRegression' package was used to exclude those explanatory variables that separated the sample points in a manner that would prevent a maximum likelihood estimate. The initial model was then refined using the stepAIC function of the MASS package (24), with model selection based on AIC operating in both forward and backward directions (Table S1). To investigate the relationship between protein characteristics and the strength of the antibody response within the ABTs, the immune response was summarised as the median of the maximally responding probes for each protein across the 35 studied serum samples. A general linear model, assuming a gamma distribution of error values, was then fitted to this continuous dependent variable. The same explanatory variables were used as in the previous logistic regression, except that in this case, the Pfam domains used were those present in at least three ABT protein representatives. The initial fit was again refined using the stepAIC function of the MASS package (Table S2). In both regression analyses, the PspA, PspC, ZmpA and ZmpB proteins were each represented by a single value summarising their immune response, and a single consensus set of protein characteristics (Dataset S1).

Text S2: Detailed functional analysis of ABTs

Of the 208 ABTs, 109 were variants of PspA, PspC, ZmpA or ZmpB. Of the remaining 99, 14 were categorised as degradative enzymes; 16 were categorised as solute binding proteins (SBPs); 24 were categorised as being involved in cell wall metabolism (separately counting four variants of Pbp2B and three variants of Pbp2X), and 15 were categorised as adhesins (including four variants of PclA).

Degradative enzymes

Several of the largest surface-associated proteins were degradative enzymes: in addition to ZmpA, ZmpB and ZmpE, the accessory genome zinc metalloproteases ZmpC and ZmpD were also found to elicit high IgG binding. This was also detected for proteases PrtA (CLS00592) (25), HtrA (CLS00066) (26), and a further uncharacterised peptidase (CLS01541), as well as the polysaccharide and polysaccharide-derivative degrading enzymes HylA (CLS00336) (27), StrH (CLS00136) (28), BgaA (CLS00596) (28), the endo-α-N-acetylgalactosaminidase Eng (CLS00380) (29), an endo-β-N-acetylglucosaminidase (CLS00485), and the neuraminidases NanA (CLS01450), NanB (CLS01445) and NanC (CLS01160) (30).

Solute binding proteins

The transporters that facilitate the import of solutes released by such enzymes were generally not found to be immunogenic, with the exception of the secreted solute-binding protein (SBP) components of ABC transporters, some representatives of which have previously been found to be immunogenic (31, 32). Of the proteins on the array, 44 were associated with a solute-binding protein domain: 18 of these were intracellular proteins involved in gene regulation, of which none were classed as ABTs; 26 had signal peptides and appear to genuinely act as part of transporters, of which 14 were classed as ABTs. Hence these extracellular solute-binding proteins were significantly overrepresented among the ABTs relative to the intracellular proteins sharing the same solute binding domain (Fisher's exact test, OR = 11.8, p = 1.1x10^{-8}). Examples of such ABTs included the oligopeptide or amino acid binding proteins AmiA (CLS01641) (33), AatB (CLS01289) (34), AliA (CLS00379) (33), AliB (CLS01314) (33), GlnPH4 (CLS01088) (35), LivJ (CLS00682) (36), AbpB (CLS01210) (37) and TrpX (CLS00960) (38); the nucleoside binding solute protein PnrA (CLS00766) (39); the sugar-binding proteins MalX (CLS01826) (40) and SatA (CLS01443) (41); two phosphate-binding proteins previously annotated as PstS, which we here propose are distinguished as PstS1 (CLS01803, orthologous with TIGR4 protein SP_2084) (42), and PstS2 (CLS01216, orthologous with TIGR4 protein SP_1400) (34); the siderophore-binding proteins PiaA (CLS00926) and PiuA (CLS01620) (43), and the metal ion binding protein PsaA (CLS01413) (44), which may also have a role in adhesion (45). These proteins account for the majority of ABTs with a lipid attachment site, found to be significantly associated with elevated antibody binding (p = 0.0049; Table S1), which triggers their processing into lipoproteins associated with the external face of the cell membrane.

Cell wall metabolism

Several proteins involved in cell wall remodelling were included amongst ABTs, as indicated by the significant association of the Transpeptidase domain (p = 0.0030) with IgG binding responses. These included subtly differing responses to the multiple variants of penicillin-binding protein 2B (CLS01435) and penicillin-binding protein 2X (CLS00355); the less variable penicillin-binding protein 3 (CLS00792) was also classified as an ABT in this analysis. Three proteins critical to cell division were also found to be immunogenic: the kinase StkP (CLS01487), early cell division protein EzrA (CLS00741), and translocase FtsK (CLS00797) (46). Other examples of cell-wall modifying proteins were the D-Ala-D-Ala peptidase DacB (CLS00584) (47), sortase SrtA

(CLS01065), LysM (CLS01786), Sep (CLS00178) (48), the peptidoglycan deacetylase PgdA (CLS01280), the peptidoglycan transglycoslyase MltG (CLS01305) (49), the two LytR domain proteins LytR (CLS01682) and MsrR (CLS01186) (50), MreC (CLS00046) (51), and the cellular amidases LytA, LytC (CLS01360), LytD/CbpE/Pce (CLS00837), CbpD (CLS00029) and PcsB (CLS00044) (49).

Adhesins

Several of the ABTs have been found to bind host structures and play a role in adhesion. These include the large keratin-binding glycoprotein PsrP (52), the plasmin- and fibronectin-binding sortase-attached adhesins PavB (CLS00114) and PfbA (CLS01587) (53), and the choline-binding protein CbpL (CLS00615) (54), which adheres to collagen, elastin and C reactive protein. For other adhesins, such as PcpA (CLS01852) (55) or PclA (of which four putatively functional variants were identified: CLS01333, CLS03178, CLS03616 and CLS99466) (5), the exact ligand has not yet been identified. Some of the proteins associated with the highest levels of IgG binding share the histidine triad motif: these are PhtA (CLS02425), PhtB/D (CLS00902), PhtE (CLS00903) and a truncated version of PhtE (CLS00904 and CLS02083). Evidence has been found for these proteins having a role in adhesion (56), inhibiting complement deposition (57) and acquisition of divalent cations (58). Finally, the two pneumococcal pili have also been found to mediate attachment to host surfaces (4, 59). The PitB backbone protein of the type 2 pilus elicited high IgG binding in this study (4) (Fig. S6). However, the three variants of the RrgB pilus 1 backbone were not classified as ABTs, despite eliciting an above-average level of IgG binding (Fig. S6).

Other ABTs

Of the 30 uncategorised ABTs, there were multiple membrane-associated proteins of unknown function. Additionally, the function of the choline binding protein CbpC/J

(CLS00389) remains unclear at present (60). However, some shared related sets of functions. One group maintained proteins in their correctly folded state, particularly on the cell surface. The extracellular Eftx1 and MsrAB2 system (the ABTs CLS00607 and CLS00608) reduces methionine sulphoxide back to methionine (61); the lipoprotein PrsA (CLS00885) is an extracellular foldase (62); and the lipoprotein SlrA (CLS00702) is an extracellular peptidyl prolyl *cis*-*trans* isomerase (63). The GroEL chaperonin (CLS01654) was also identified as an ABT, although it is generally cytosolic. Similarly, the specificity subunit of the SpnIV restriction-modification system (CLS00804) was also found to bind IgG at an elevated level (12, 64). Three cytosolic metabolic enzymes were identified as ABTs: the alcohol dehydrogenase AdhP (CLS00321), the dihydrolipoamide dehydrogenase AcoL (CLS01023), and the Glf capsule polysaccharide synthesis gene (CLS02157). The Wzg membrane protein (CLS00362) that regulates capsule synthesis was also identified as an ABT, as was a two component system histidine kinase (CLS00156) of unknown function. Other integral membrane proteins classed as ABTs were the paralogous transporters ComB (CLS00123) and BlpB (CLS00507), which mediate the efflux of the competence stimulating peptide and *blp* bacteriocin-like peptides, respectively (65). A further uncharacterised protein involved in efflux (CLS00718) also bound IgG at a high level. Further functional characterisation will determine how many of these ABTs it is possible to assign to one of the four main categories; else they may form one or more new functional groupings of antigens.

Figure S1

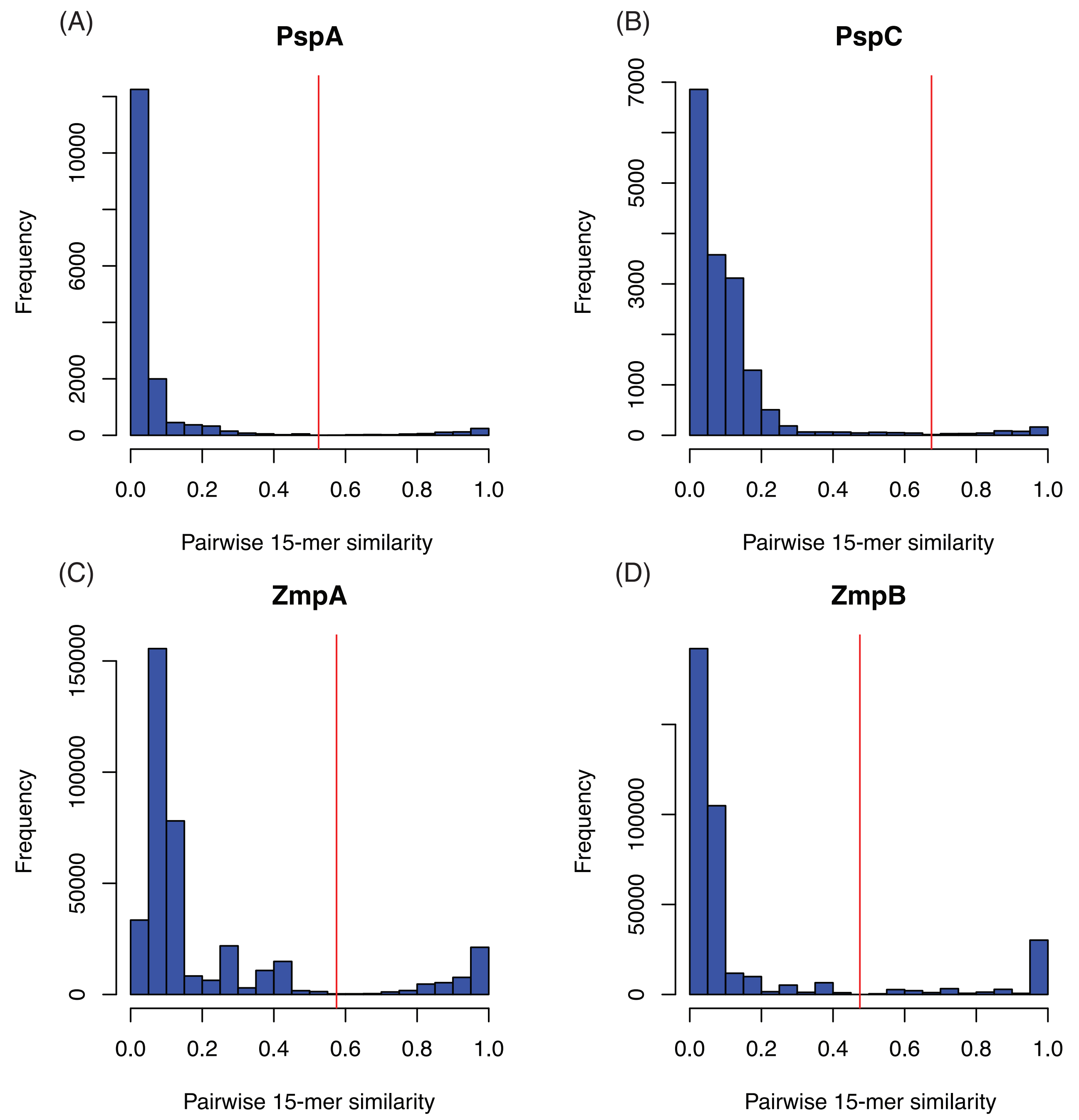

(A)
PspA
Frequency
0
2000
6000
10000
0.0
0.2
0.4
0.6
0.8
1.0
Pairwise 15-mer similarity
(B)
PspC
Frequency
0
1000
3000
5000
7000
0.0
0.2
0.4
0.6
0.8
1.0
Pairwise 15-mer similarity
(C)
ZmpA
Frequency
0
50000
100000
150000
0.0
0.2
0.4
0.6
0.8
1.0
Pairwise 15-mer similarity
(D)
ZmpB
Frequency
0
50000
100000
0.0
0.2
0.4
0.6
0.8
1.0
Pairwise 15-mer similarity

Figure S1 Defining variants for the four core variable antigens. Each histogram shows the distribution of pairwise similarities between the central segments of representatives of (A) PspA, (B) PspC, (C) ZmpA, and (D) ZmpB, following the removal of N and C terminal motifs for secretion and surface attachment (see Text S1). For each pair of proteins, the similarity metric was calculated as the proportion of all 15 amino acid fragments found in the central segment of either protein that was exactly matched in both sequences. The vertical red lines on each plot represent the thresholds that were used as a cut-off to define distinct variants: 0.525 for PspA; 0.675 for PspC; 0.575 for ZmpA, and 0.475 for ZmpB.

Figure S2

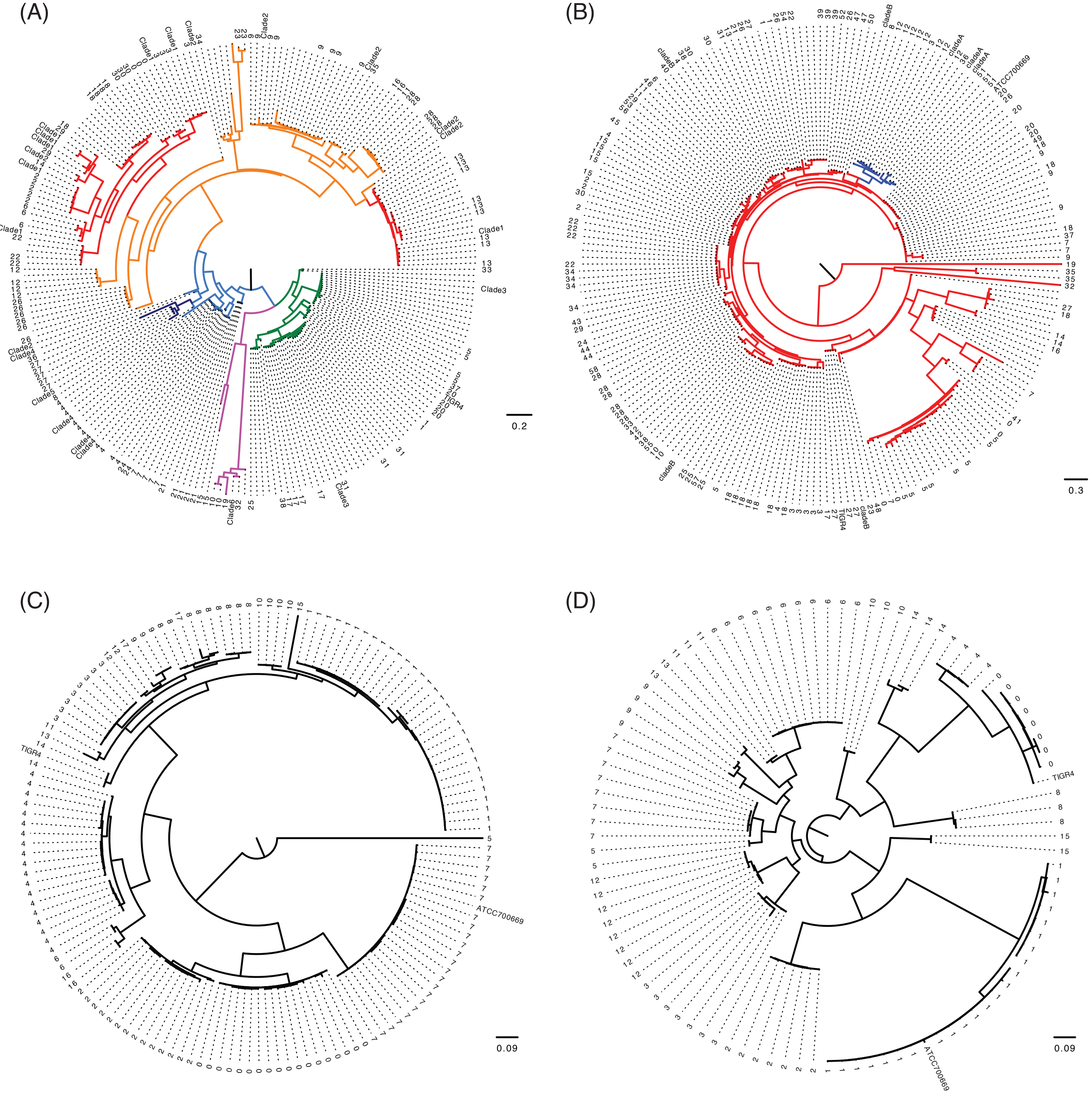


Figure S2 Phylogenies of variable core antigen proteins. (A) Phylogeny of the PspA proteins. This FastTree2 phylogeny was constructed from the unique PspA protein sequences from this isolate collection, following removal of the N-terminal signal peptide and C-terminal surface attachment motifs, alongside representatives from Hollingshead *et al.* (28) that were used to define six 'clades'. The parts of the tree corresponding to each set of PspA sequences are coloured red (clade 1), orange (clade 2), green (clade 3), light blue (clade 4), dark blue (clade 5), and purple (clade 6). Those protein sequences that were assigned to isolates in the collection by the mapping-based approach described in the Methods section are annotated with the number of the variant to which they were assigned by the analysis shown in Fig. S1. (B) Phylogeny of the PspC proteins. This FastTree2 phylogeny is displayed as in panel (A), except that the two colours of the tree denote parts that correspond to clade A (blue) and clade B (red), as defined by Brooks-Walter *et al*. (29) (C) Phylogeny of ZmpA proteins. This FastTree2 phylogeny includes all unique mature protein sequences from the population, annotated with the number of the variant to which they were assigned by the analysis shown in Fig. S1, as well as the representatives from *S. pneumoniae* ATCC 700669 and TIGR4. (D) Phylogeny of ZmpB proteins. This FastTree2 phylogeny includes all unique mature protein sequences from the population, annotated with the number of the variant to which they were assigned by the analysis shown in Fig. S1, as well as the representatives from *S. pneumoniae* ATCC 700669 and TIGR4.

Figure S3

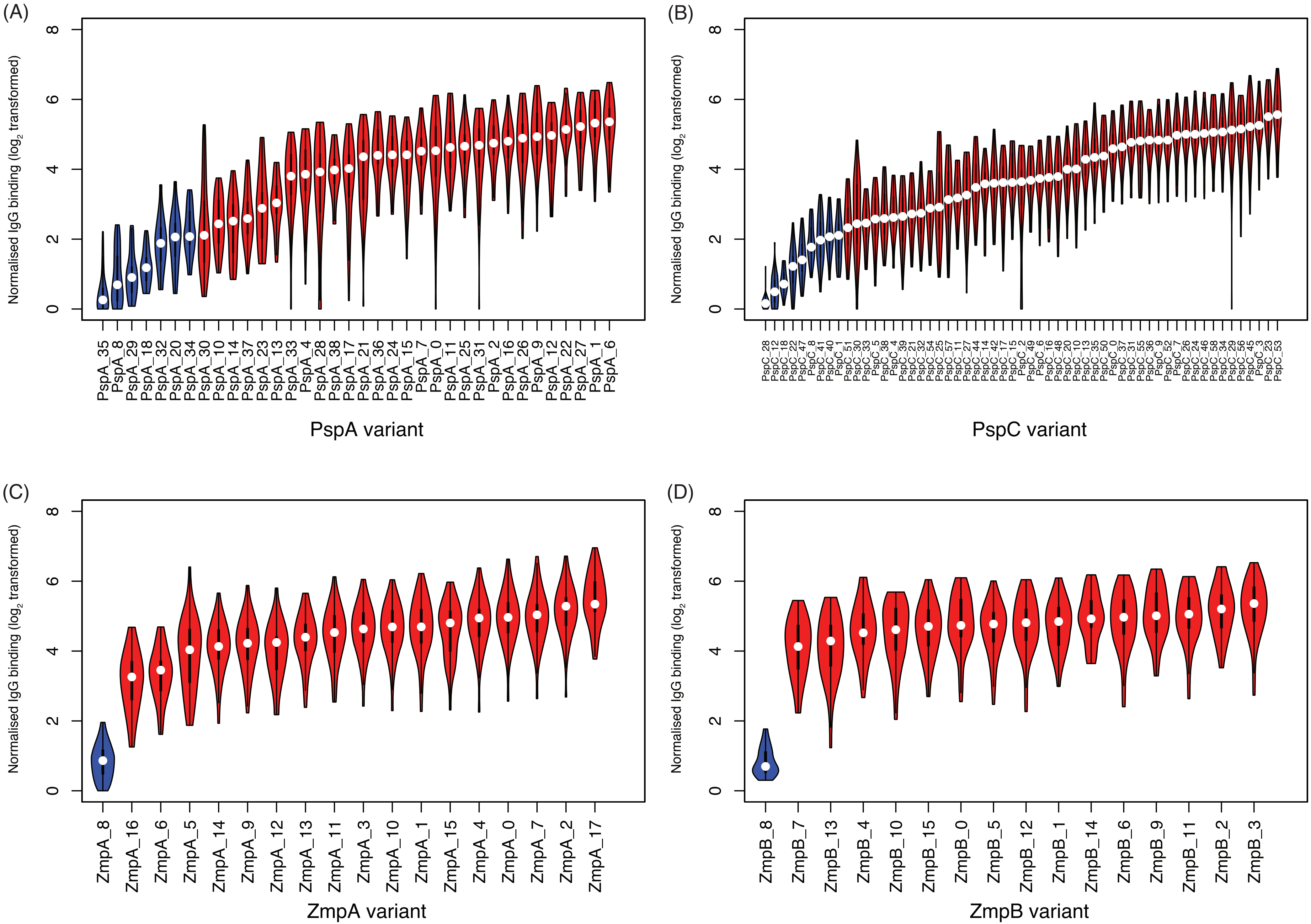
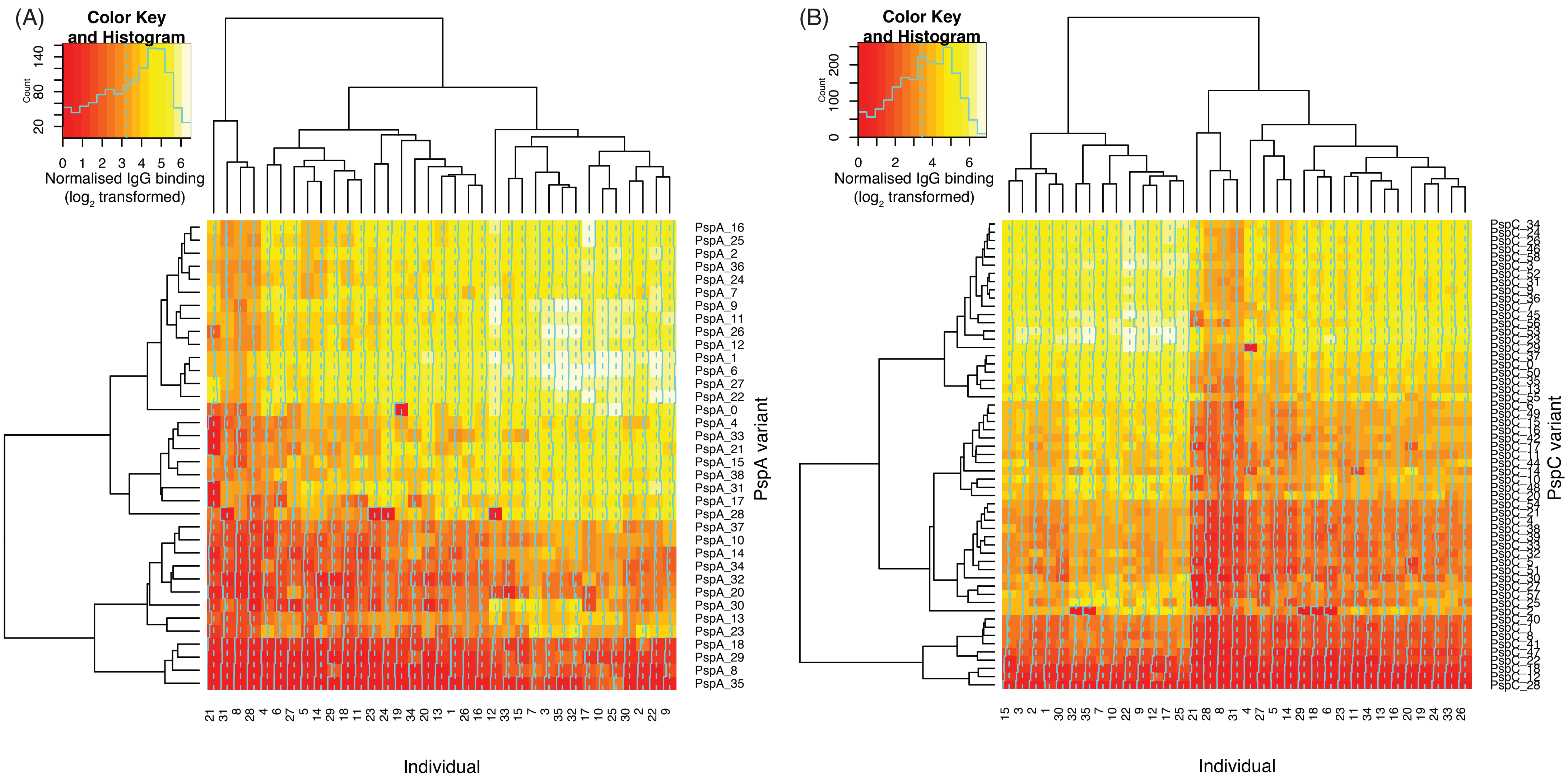

(A)
Normalised IgG binding (log_2 transformed)
0
2
4
6
8
PspA_35
PspA_8
PspA_29
PspA_18
PspA_32
PspA_20
PspA_34
PspA_30
PspA_10
PspA_14
PspA_37
PspA_23
PspA_13
PspA_33
PspA_4
PspA_28
PspA_38
PspA_17
PspA_21
PspA_36
PspA_24
PspA_15
PspA_7
PspA_0
PspA_11
PspA_25
PspA_31
PspA_2
PspA_16
PspA_26
PspA_9
PspA_12
PspA_22
PspA_27
PspA_1
PspA_6
PspA variant
(B)
Normalised IgG binding (log_2 transformed)
0
2
4
6
8
PspC_28
PspC_12
PspC_18
PspC_22
PspC_47
PspC_8
PspC_41
PspC_40
PspC_1
PspC_51
PspC_30
PspC_33
PspC_5
PspC_38
PspC_4
PspC_39
PspC_21
PspC_32
PspC_54
PspC_25
PspC_57
PspC_11
PspC_27
PspC_44
PspC_14
PspC_42
PspC_17
PspC_15
PspC_2
PspC_49
PspC_6
PspC_16
PspC_48
PspC_20
PspC_10
PspC_13
PspC_35
PspC_50
PspC_0
PspC_37
PspC_31
PspC_55
PspC_36
PspC_9
PspC_52
PspC_7
PspC_26
PspC_24
PspC_46
PspC_58
PspC_34
PspC_29
PspC_56
PspC_45
PspC_3
PspC_23
PspC_53
PspC variant
(C)
Normalised IgG binding (log_2 transformed)
0
2
4
6
8
ZmpA_8
ZmpA_16
ZmpA_6
ZmpA_5
ZmpA_14
ZmpA_9
ZmpA_12
ZmpA_13
ZmpA_11
ZmpA_3
ZmpA_10
ZmpA_1
ZmpA_15
ZmpA_4
ZmpA_0
ZmpA_7
ZmpA_2
ZmpA_17
ZmpA variant
(D)
Normalised IgG binding (log_2 transformed)
0
2
4
6
8
ZmpB_8
ZmpB_7
ZmpB_13
ZmpB_4
ZmpB_10
ZmpB_15
ZmpB_0
ZmpB_5
ZmpB_12
ZmpB_1
ZmpB_14
ZmpB_6
ZmpB_9
ZmpB_11
ZmpB_2
ZmpB_3
ZmpB variant

Figure S3 IgG binding elicited by variants of the four core variable antigens: (A) PspA, (B) PspC, (C) ZmpA, and (D) ZmpB. Each violin plot corresponds to the normalised $\log_2$-transformed IgG binding response to a single representative of each protein across the 35 serum samples. The variants are ordered by the median binding response across the samples, with red plots indicating those variants that were categorised as ABTs by the overall classification displayed in Fig 1A, and blue plots indicating those that were not.

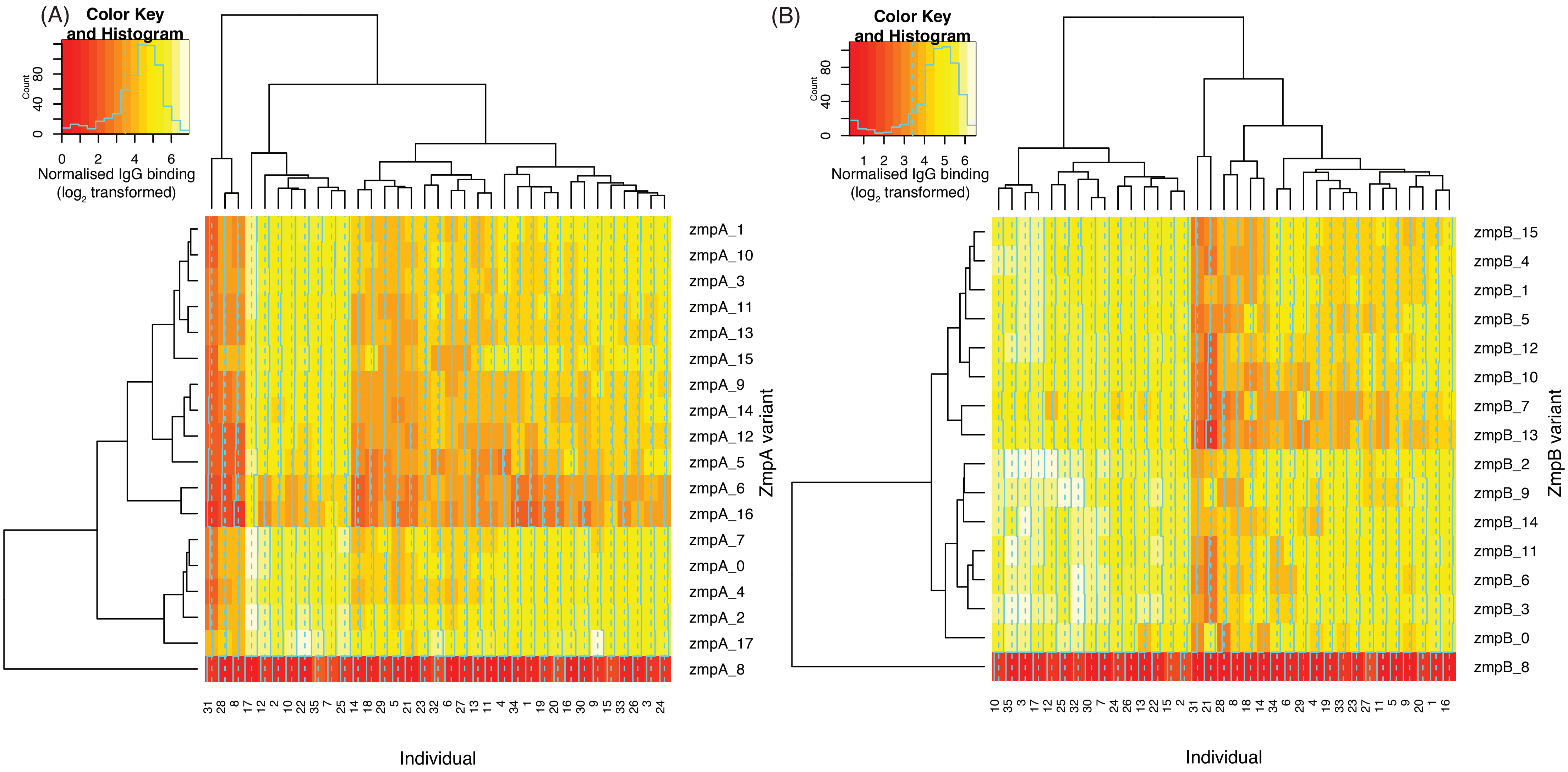
Figure S4



(A)
Color Key and Histogram
Count
Normalised IgG binding (log$_2$ transformed)
PspA variant
Individual
(B)
Color Key and Histogram
Count
Normalised IgG binding (log$_2$ transformed)
PspC variant
Individual

Figure S4 Details of IgG binding to PspA and PspC variants across the serum samples. (A) Heatmap showing the normalised $\log_2$-transformed IgG binding to each of the PspA variants across the 35 serum samples. Each row corresponds to a variant of the antigen, and each column to a different serum sample. Both the serum samples and PspA variants are hierarchically clustered based on the similarity of their IgG binding profiles. (B) Heatmap showing the IgG binding to each of the PspC variants, displayed as described in panel (A).

Figure S5



(A)
Color Key
and Histogram
Count
0 40 80
0 2 4 6
Normalised IgG binding
(log_2 transformed)
zmpA_1
zmpA_10
zmpA_3
zmpA_11
zmpA_13
zmpA_15
zmpA_9
zmpA_14
zmpA_12
zmpA_5
zmpA_6
zmpA_16
zmpA_7
zmpA_0
zmpA_4
zmpA_2
zmpA_17
zmpA_8
ZmpA variant
31 28 8 17 12 2 10 22 35 7 25 14 18 29 5 21 23 32 6 27 13 11 4 34 1 19 20 16 30 9 15 33 26 3 24
Individual
(B)
Color Key
and Histogram
Count
0 40 80
1 2 3 4 5 6
Normalised IgG binding
(log_2 transformed)
zmpB_15
zmpB_4
zmpB_1
zmpB_5
zmpB_12
zmpB_10
zmpB_7
zmpB_13
zmpB_2
zmpB_9
zmpB_14
zmpB_11
zmpB_6
zmpB_3
zmpB_0
zmpB_8
ZmpB variant
10 35 3 17 12 25 32 30 7 24 26 13 22 15 2 31 21 28 8 18 14 34 6 29 4 19 33 23 27 11 5 9 20 1 16
Individual

Figure S5 Details of IgG binding to ZmpA and ZmpB variants across the serum samples. (A) Heatmap showing the normalised log_2-transformed IgG binding to each of the ZmpA variants across the 35 serum samples, displayed as described in Fig. S4. (B) Heatmap showing the normalised log_2-transformed IgG binding to each of the ZmpB variants, displayed as described in Fig. S4.

Figure S6

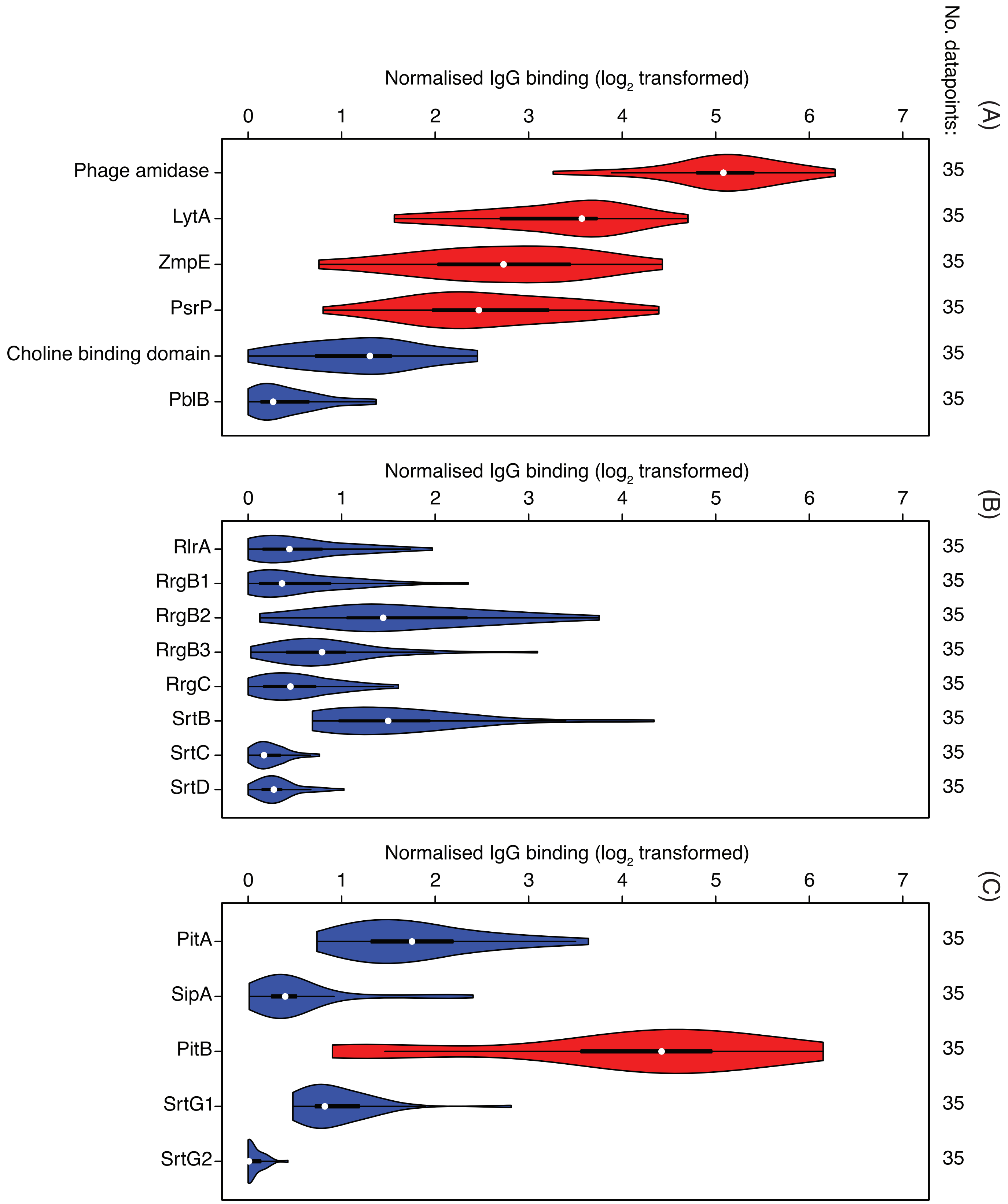
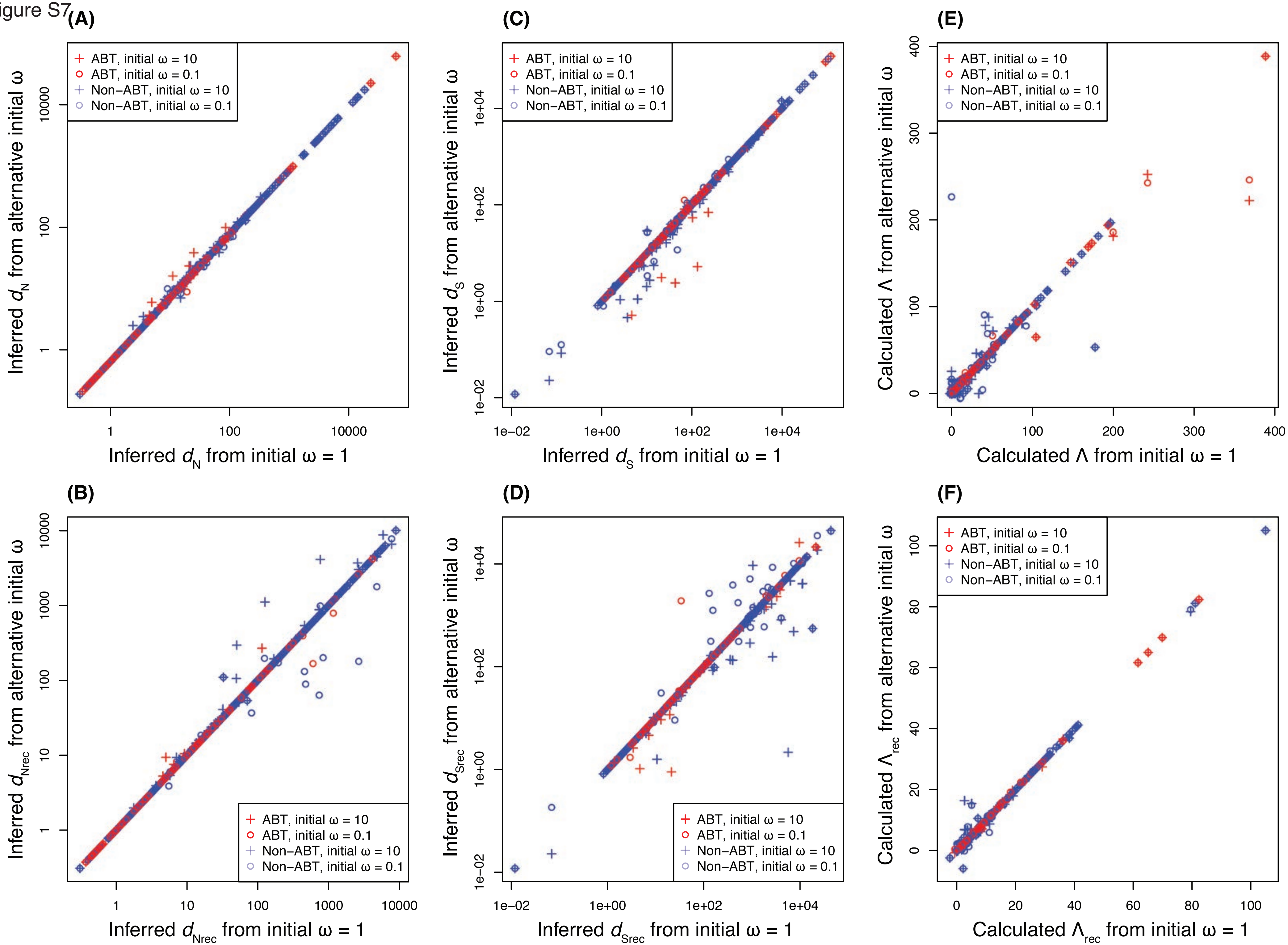

No. datapoints:
(A)
Normalised IgG binding (log$_2$ transformed)
0
1
2
3
4
5
6
7
Phage amidase
LytA
ZmpE
PsrP
Choline binding domain
PblB
35
35
35
35
35
35
(B)
Normalised IgG binding (log$_2$ transformed)
RlrA
RrgB1
RrgB2
RrgB3
RrgC
SrtB
SrtC
SrtD
35
35
35
35
35
35
35
35
(C)
Normalised IgG binding (log$_2$ transformed)
PitA
SipA
PitB
SrtG1
SrtG2
35
35
35
35
35

Figure S6 Violin plots showing the IgG binding to selected pneumococcal proteins. (A) Normalised log_2-transformed IgG binding to the six 'additional' probes that did not correspond to COGs or core variable antigens. The violin plots for those proteins that were classified as ABTs in the overall analysis shown in Fig 1(A) are coloured red; those for proteins not classified as ABTs are coloured blue. (B) IgG binding to type I pilus proteins. The responses to the three RrgB 'clades' or variants are shown separately. No RrgA proteins were included on the microarray because these proteins were divided between two COGs, each of which was too rare to be included as part of the studied proteome. (C) IgG binding to type II pilus proteins. The violin plot for the only protein classified as a ABT, PitB, is coloured red.

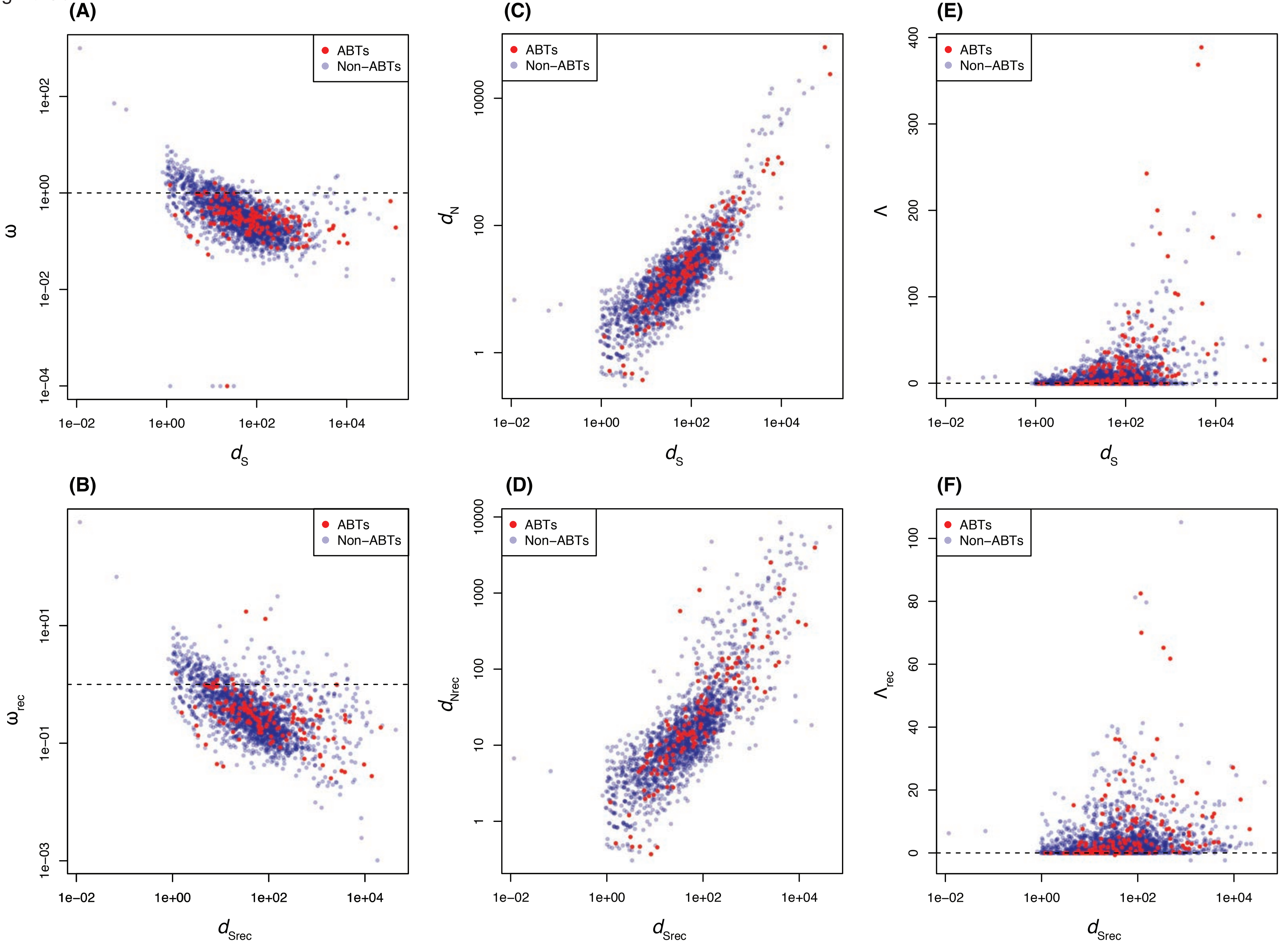
Figure S7



(A)
ABT, initial ω = 10
ABT, initial ω = 0.1
Non−ABT, initial ω = 10
Non−ABT, initial ω = 0.1
Inferred d_N from alternative initial ω
Inferred d_N from initial ω = 1
(B)
Inferred d_{Nrec} from alternative initial ω
Inferred d_{Nrec} from initial ω = 1
(C)
Inferred d_S from alternative initial ω
Inferred d_S from initial ω = 1
(D)
Inferred d_{Srec} from alternative initial ω
Inferred d_{Srec} from initial ω = 1
(E)
Calculated Λ from alternative initial ω
Calculated Λ from initial ω = 1
(F)
Calculated Λ_{rec} from alternative initial ω
Calculated Λ_{rec} from initial ω = 1

Figure S7 Evidence for convergence of PAML analyses of codon alignments. (A) Maximum likelihood estimates of non-synonymous (d_N) distances for codon alignments with different starting ω values. The estimates displayed in Fig. 2 were initiated with ω = 1; these estimates are compared with those from the same data initiated with ω = 0.1 (circles) or ω = 10 (crosses). Points are red for ABTs and blue for non-ABTs. That the majority of point lie on the diagonal indicates independent PAML analyses are converging on the same result despite different starting value parameters. (B) Maximum likelihood estimates of d_{Nrec} distances for codon alignments segmented at putative recombination breakpoints. Data are displayed as described in panel (A). (C) Maximum likelihood estimates of synonymous (d_S) distances for codon alignments. Data are displayed as described in panel (A). (D) Maximum likelihood estimates of d_{Srec} distances for codon alignments segmented at putative recombination breakpoints. Data are displayed as described in panel (A). (E) Maximum likelihood estimates of the log likelihood difference between the fits of models 2a and 1a (Λ) for codon alignments. Data are displayed as described in panel (A). (F) Maximum likelihood estimates of Λ_{rec} for codon alignments segmented at putative recombination breakpoints. Data are displayed as described in panel (A).

Figure S8



(A)
ABTs
Non–ABTs
ω
1e+02
1e+00
1e−02
1e−04
1e−02
1e+00
1e+02
1e+04
d
S
(C)
ABTs
Non–ABTs
d
N
10000
100
1
1e−02
1e+00
1e+02
1e+04
d
S
(E)
ABTs
Non–ABTs
Λ
400
300
200
100
0
1e−02
1e+00
1e+02
1e+04
d
S
(B)
ABTs
Non–ABTs
ω
rec
1e+01
1e−01
1e−03
1e−02
1e+00
1e+02
1e+04
d
Srec
(D)
ABTs
Non–ABTs
d
Nrec
10000
1000
100
10
1
1e−02
1e+00
1e+02
1e+04
d
Srec
(F)
ABTs
Non–ABTs
Λ
rec
100
80
60
40
20
0
1e−02
1e+00
1e+02
1e+04
d
Srec

Figure S8 Power to detect selection acting on codon alignments at different levels of sequence divergence. In each plot, the synonymous divergence (d_S), or the synonymous divergence based on codon alignments segmented at putative recombination breakpoints (d_{Srec}), is used as a measure of genetic distance. (A) Distribution of ω relative to d_S based on codon alignments. (B) Distribution of ω_{rec} relative to d_{Srec}. (C) Distribution of non-synonymous divergence (d_N) relative to d_S. (D) Distribution of d_{Nrec} relative to d_{Srec}. (E) Distribution of log likelihood difference between the fits of models 2a and 1a (Λ) relative to d_S. (F) Distribution of Λ_{rec} relative to d_{Srec}.

Figure S9

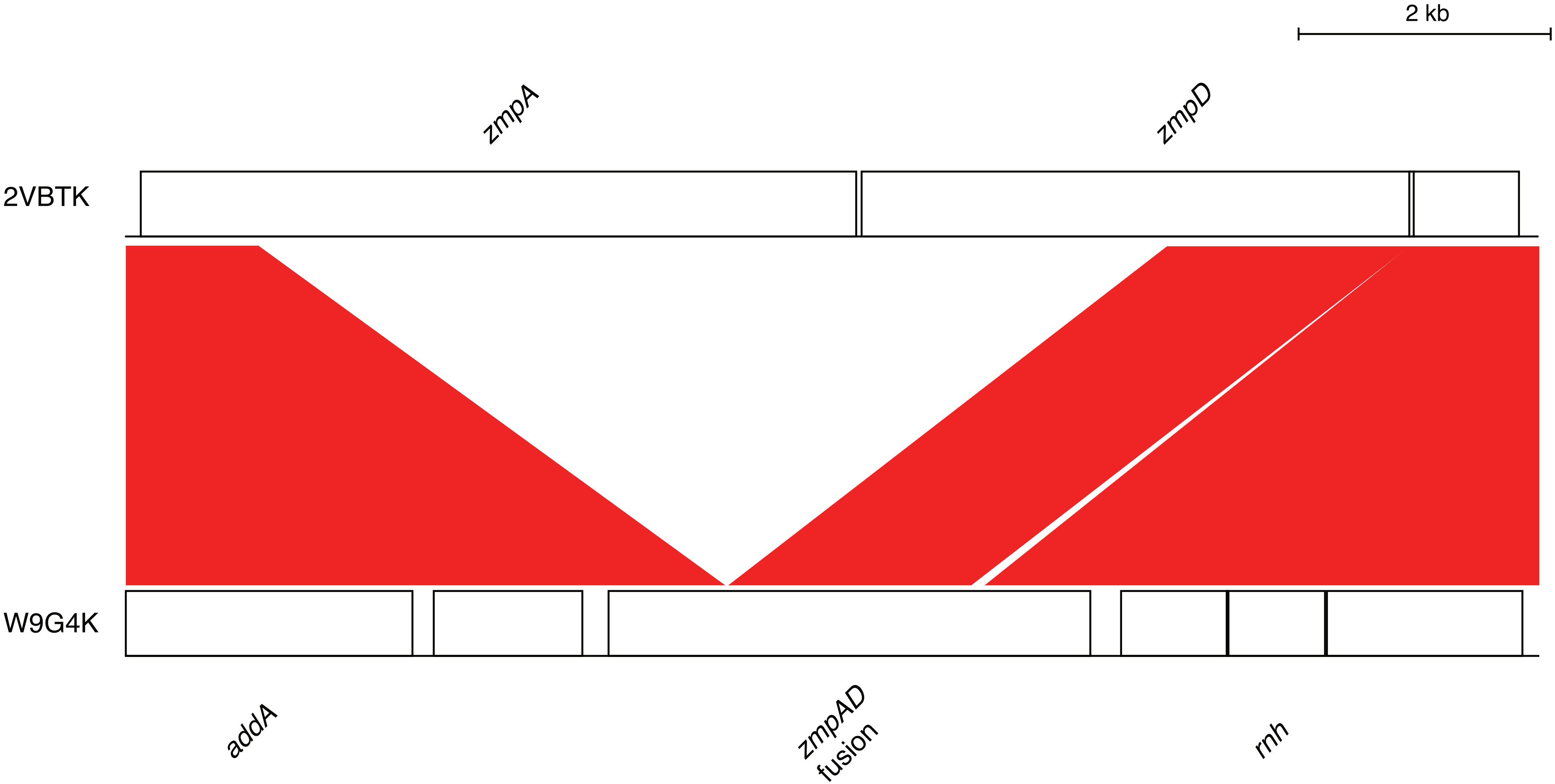
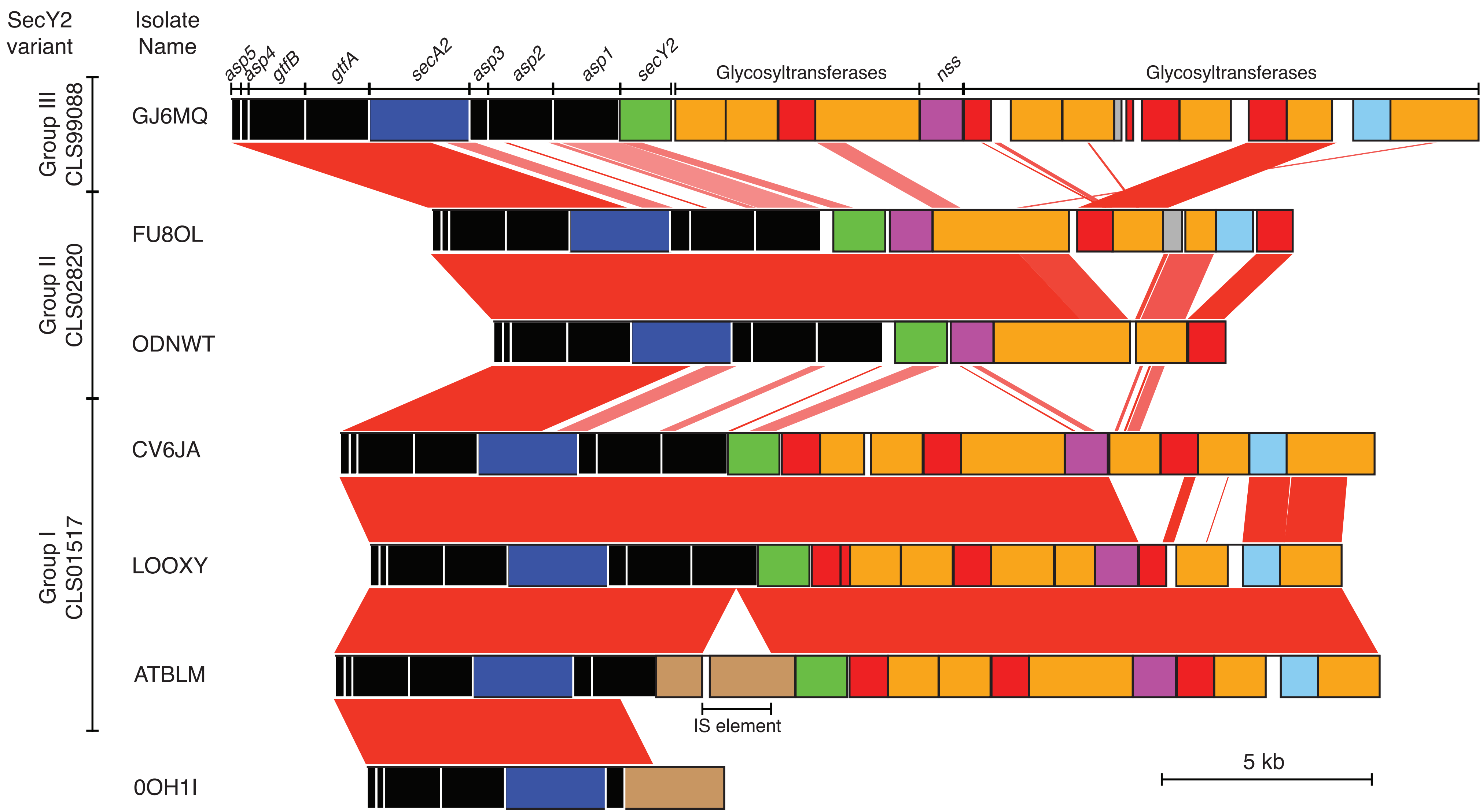


Figure S9 Formation of a novel fusion ZmpAD protein. The figure shows orthologous loci from two closely-related serotype 7C isolates, 2VBTK and W9G4K. Red bands connect regions of sequence similarity, with the band's shade indicating the strength of the match identified by BLAT. This alignment suggests a ~7.2 kb deletion has fused the orthologues of the *zmpA* and *zmpD* genes of 2VBTK into a single coding sequence in W9G4K.

Figure S10



SecY2 variant
Isolate Name
Group III CLS99088
Group II CLS02820
Group I CLS01517
GJ6MQ
FU8OL
ODNWT
CV6JA
LOOXY
ATBLM
0OH1I
asp5
asp4
gtfB
gtfA
secA2
asp3
asp2
asp1
secY2
Glycosyltransferases
nss
Glycosyltransferases
IS element
5 kb

Figure S10 Alignment of the non-repetitive parts of PsrP-encoding genomic islands. Each sequence represents a different variant of the PsrP-encoding genomic islands, excluding only the large repetitive CDS encoding the serine-rich protein itself. Individual CDSs are coloured according to their putative function, following the scheme shown in Fig. 5: black CDSs encode the accessory secretion proteins and GtfAB proteins; green CDSs encode SecY2 proteins; blue CDSs encode the SecA2 proteins; purple CDSs encode the glycosyltransferase Nss; orange and red CDSs encode other likely glycosyltransferases; brown CDSs indicate pseudogenes, and grey CDSs encode proteins of unknown function. Red bands connect regions of sequence similarity; the intensity of each band's shading indicates the strength of the similarity, as calculated by BLAT. The sequences are grouped by the variant of the SecY2 protein they encode: group I (CLS01517), group II (CLS02820), and group III (CLS99088).

Figure S11

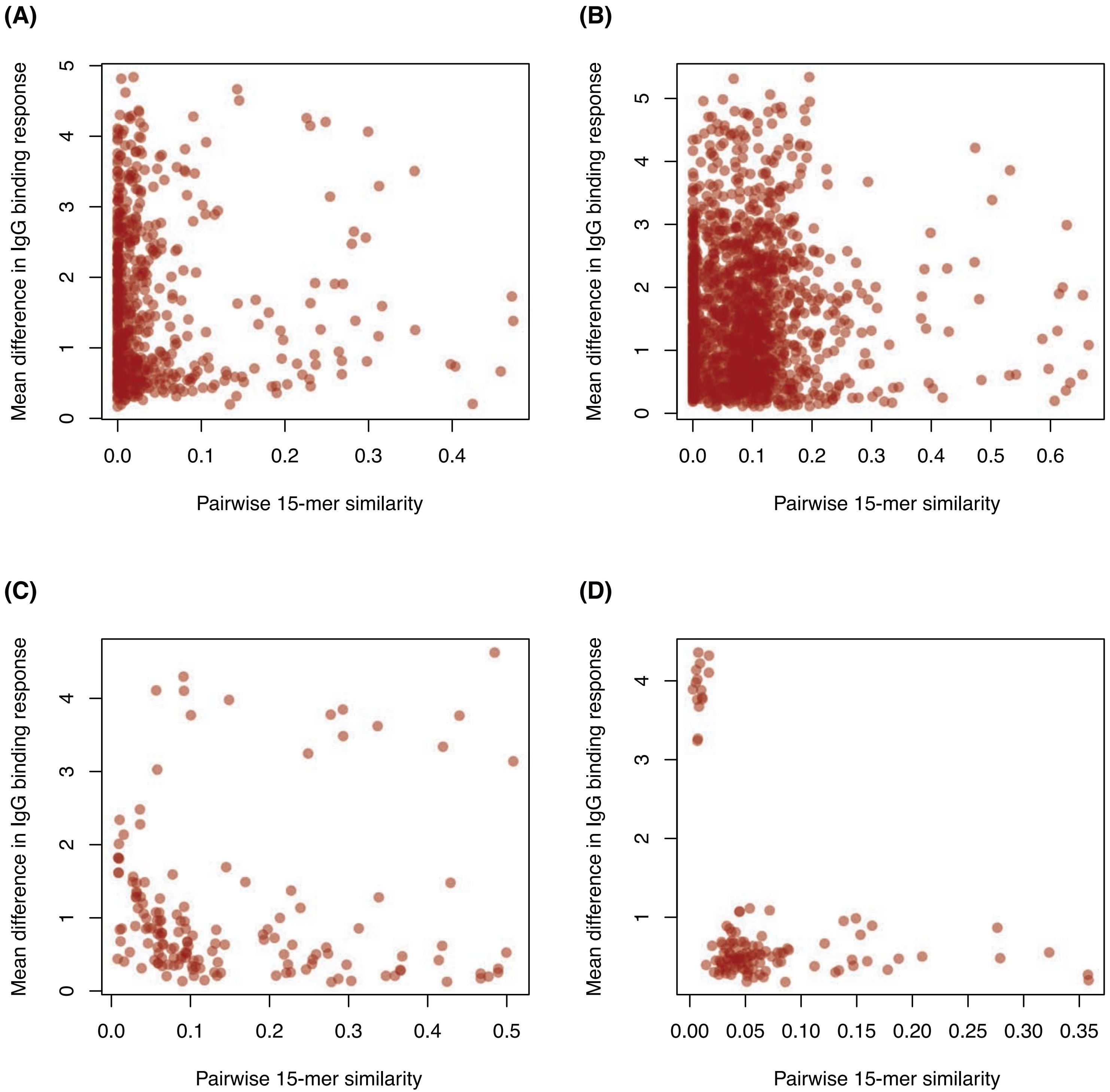

(A)
Mean difference in IgG binding response
Pairwise 15-mer similarity
(B)
Mean difference in IgG binding response
Pairwise 15-mer similarity
(C)
Mean difference in IgG binding response
Pairwise 15-mer similarity
(D)
Mean difference in IgG binding response
Pairwise 15-mer similarity

Figure S11 Scatterplot showing the relationship between genetic divergence of variants and IgG binding for (A) PspA, (B) PspC, (C) ZmpA and (D) ZmpB. Each plot shows the sequence similarity between each pairwise comparison of variants, as calculated by the 15-mer based analysis shown in Fig. S1, against the mean difference in IgG binding across the 35 study participants (Fig S4 & S5). If individuals' immune reactions were similar to clusters of related variants, a negative relationship would be expected in each scatterplot. The only distinct cluster of points corresponds to a single divergent variant of ZmpB, associated with the atypical unencapsulated SC12 strains, which elucidates a consistently lower measured IgG binding response (Fig. S3). The absence of a clear relationship indicates there is no signal of greater immune cross-reactivity between similar variants.

Figure S12

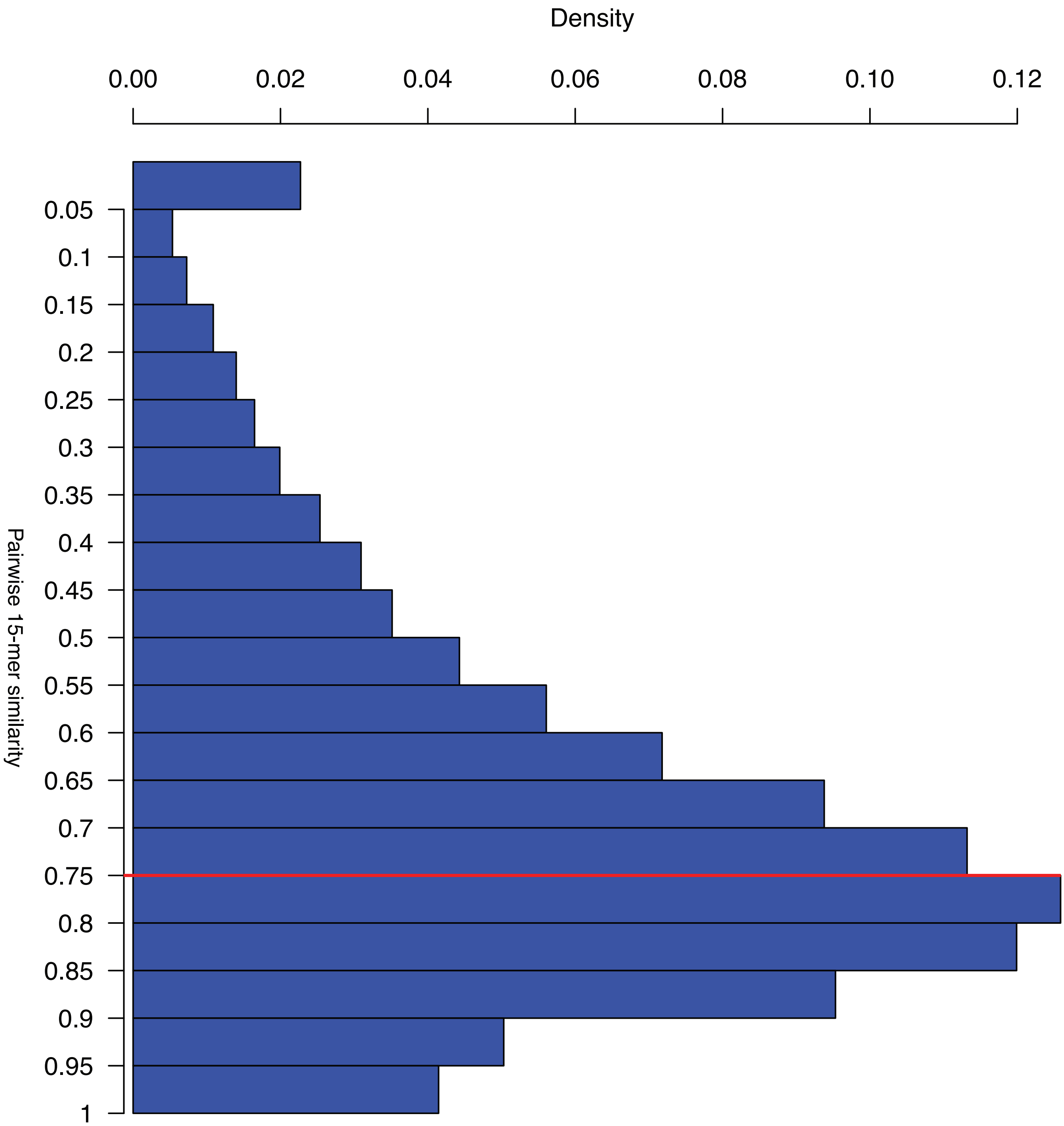
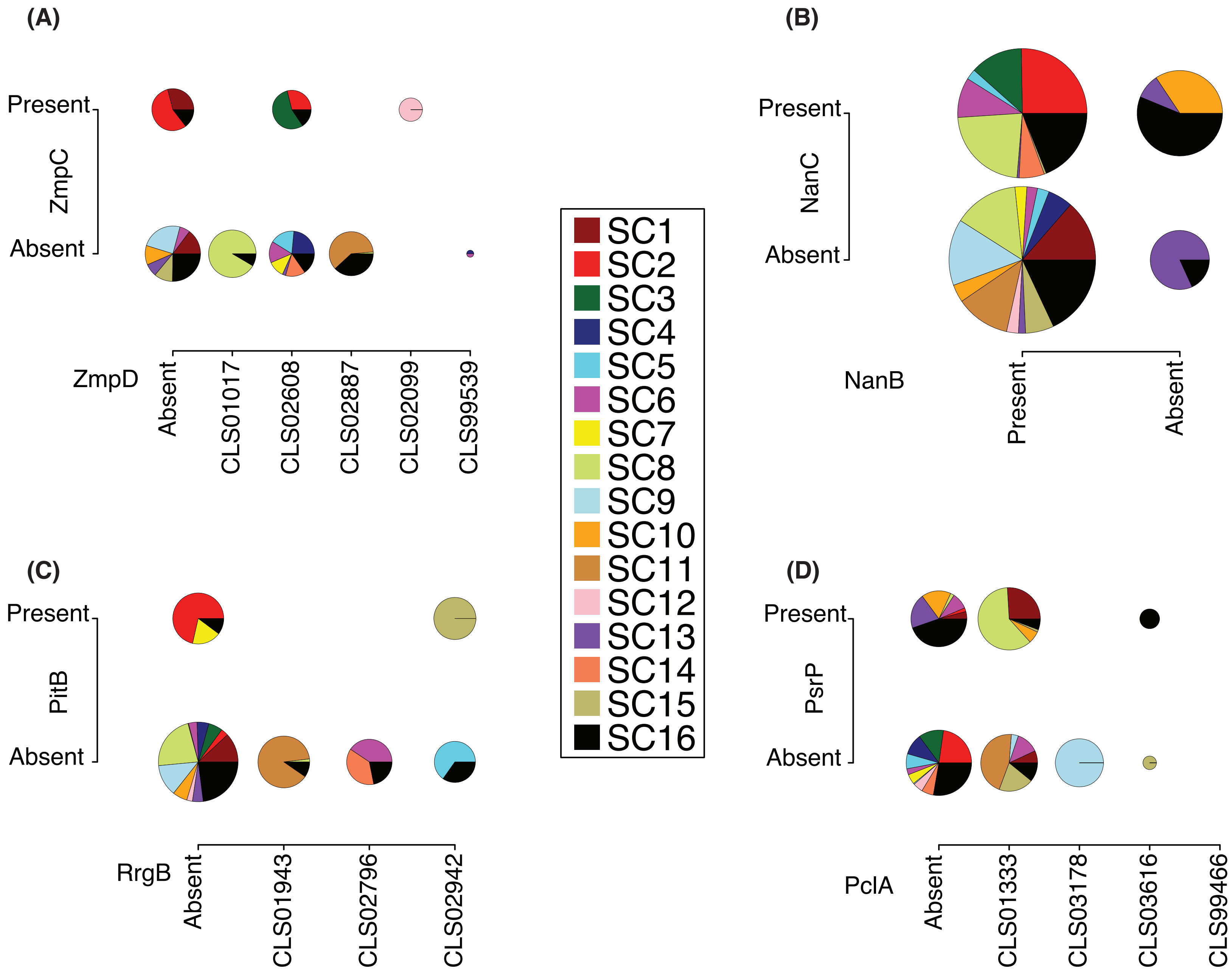

Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Pairwise 15-mer similarity
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1

Figure S12 Distribution of pairwise similarities between sequences within the same COG, calculated using those COGs included on the microarray. The codon alignments used were those analysed in Fig. 2. This analysis excluded the PspA, PspC, ZmpA and ZmpB sequences, and analysed the distinct groups of *pbp1a*, *pbp2x* and *pbp2b* sequences independently. Sequence similarity was calculated as the proportion of the full set of overlapping 15 amino acid fragments found in either protein that was exactly matched in both sequences. The red line represents the cutoff (0.75) that was used to define distinct variants based on these comparisons.

Figure S13



(A)
Present
ZmpC
Absent
ZmpD
Absent
CLS01017
CLS02608
CLS02887
CLS02099
CLS99539
(B)
Present
NanC
Absent
NanB
Present
Absent
(C)
Present
PitB
Absent
RrgB
Absent
CLS01943
CLS02796
CLS02942
(D)
Present
PsrP
Absent
PclA
Absent
CLS01333
CLS03178
CLS03616
CLS99466
SC1
SC2
SC3
SC4
SC5
SC6
SC7
SC8
SC9
SC10
SC11
SC12
SC13
SC14
SC15
SC16

Figure S13 Co-occurrence of functionally-related accessory ABTs. (A) Co-occurrence of *zmpC* and *zmpD*, both surface-associated zinc metalloproteases. All ZmpC proteins belong to a single COG (CLS01991), whereas ZmpD representatives were distributed across the five displayed COGs. Pie charts mark combinations of these ABTs observed in the population. The diameter of the pie chart is proportional to the logarithm of the number of times the combination was observed, and the chart itself shows the genetic background of the relevant isolates in terms of sequence clusters, annotated according to the core genome phylogeny shown in Fig 4, as indicated by the key. (B) Co-occurrence of the neuraminidase-encoding genes *nanB* and *nanC*. Both the NanB (CLS01445) and NanC (CLS01160) proteins each corresponded to single COGs. Data on the observed antigenic combinations are displayed as described in panel (A). (C) Co-occurrence of *rrgB* (a gene on the type I pilus islet) and *pitB* (a gene on the type II pilus islet). The PitB proteins all belong to the COG CLS02871, whereas the RrgB proteins are split between three COGs, each corresponding to a previously defined 'clade'. Data on the observed antigenic combinations are displayed as described in panel (A). (D) Co-occurrence of the *pclA* and *psrP* genes, both encoding large surface adhesins. All PsrP-encoding islands were considered equivalent, and their distribution inferred using that of the SecA2 protein (CLS01513). Only the four full-length PclA protein COGs were considered to represent the antigen being present in an isolate; truncated CDSs were included within the 'absent' category. Data on the observed antigenic combinations are displayed as described in panel (A).

Figure S14

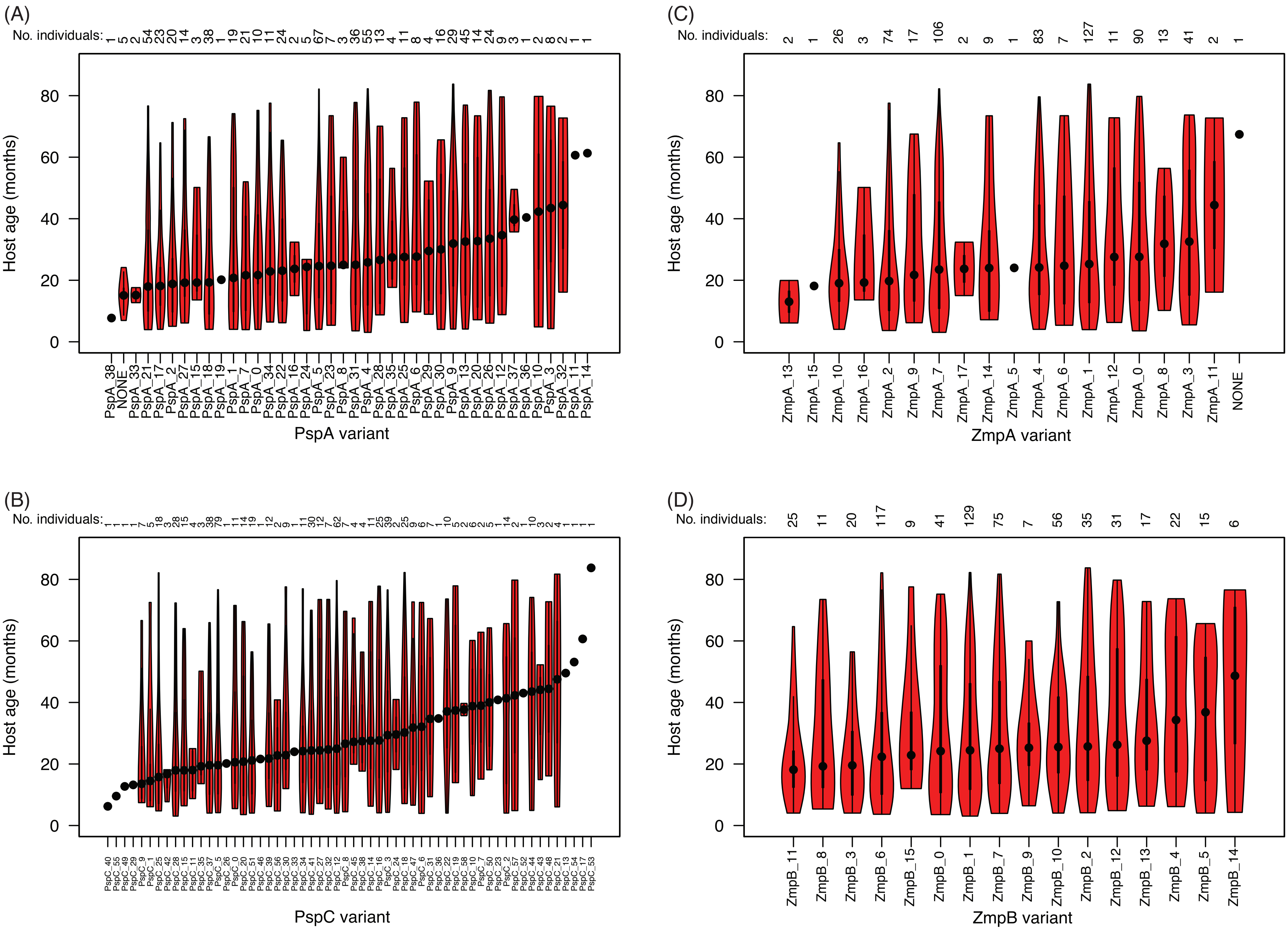
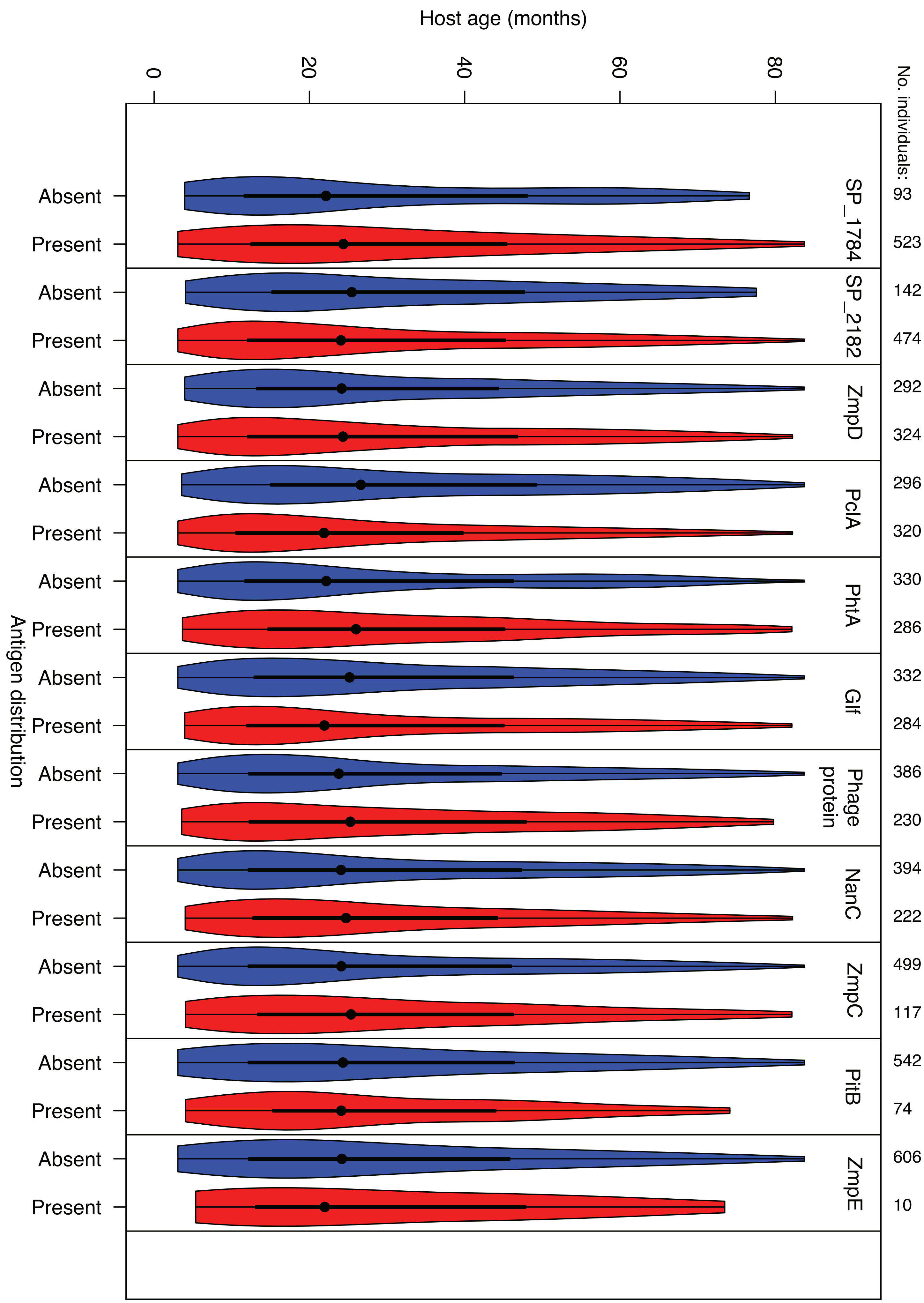

(A)
No. individuals:
Host age (months)
PspA variant
(C)
No. individuals:
Host age (months)
ZmpA variant
(B)
No. individuals:
Host age (months)
PspC variant
(D)
No. individuals:
Host age (months)
ZmpB variant

Figure S14 Violin plots showing the range of host ages associated with the variants of the four variable core antigenic loci. Each violin plot is annotated with the number of host age measurements it represents at the top of the plot. (A) Host ages associated with the different variants of PspA. These data do not provide any significant evidence of age stratification of variants (Kruskal-Wallis test, p = 0.425). (B) Host ages associated with the different variants of PspC. These data do not provide any significant evidence of age stratification of variants (Kruskal-Wallis test, p = 0.799). (C) Host ages associated with the different variants of ZmpA. These data do not provide any significant evidence of age stratification of variants (Kruskal-Wallis test, p = 0.492). (D) Host ages associated with the different variants of ZmpB. These data do not provide any significant evidence of age stratification of variants (Kruskal-Wallis test, p = 0.183).

Figure S15



Host age (months)
0
20
40
60
80
No. individuals:
SP_1784
Absent 93
Present 523
SP_2182
Absent 142
Present 474
ZmpD
Absent 292
Present 324
PclA
Absent 296
Present 320
PhtA
Absent 330
Present 286
Glf
Absent 332
Present 284
Phage protein
Absent 386
Present 230
NanC
Absent 394
Present 222
ZmpC
Absent 499
Present 117
PitB
Absent 542
Present 74
ZmpE
Absent 606
Present 10
Antigen distribution

Figure S15 Violin plots showing the ages of hosts associated with the presence of accessory genome ABTs. Each violin plot is annotated with the number of host age measurements it represents at the top of the plot. For each indicated ABT, the red plot displays the range of host ages associated with bacteria carrying the relevant ABT, whereas the blue plot displays the range of host ages associated with bacteria lacking the same ABT.

Figure S16

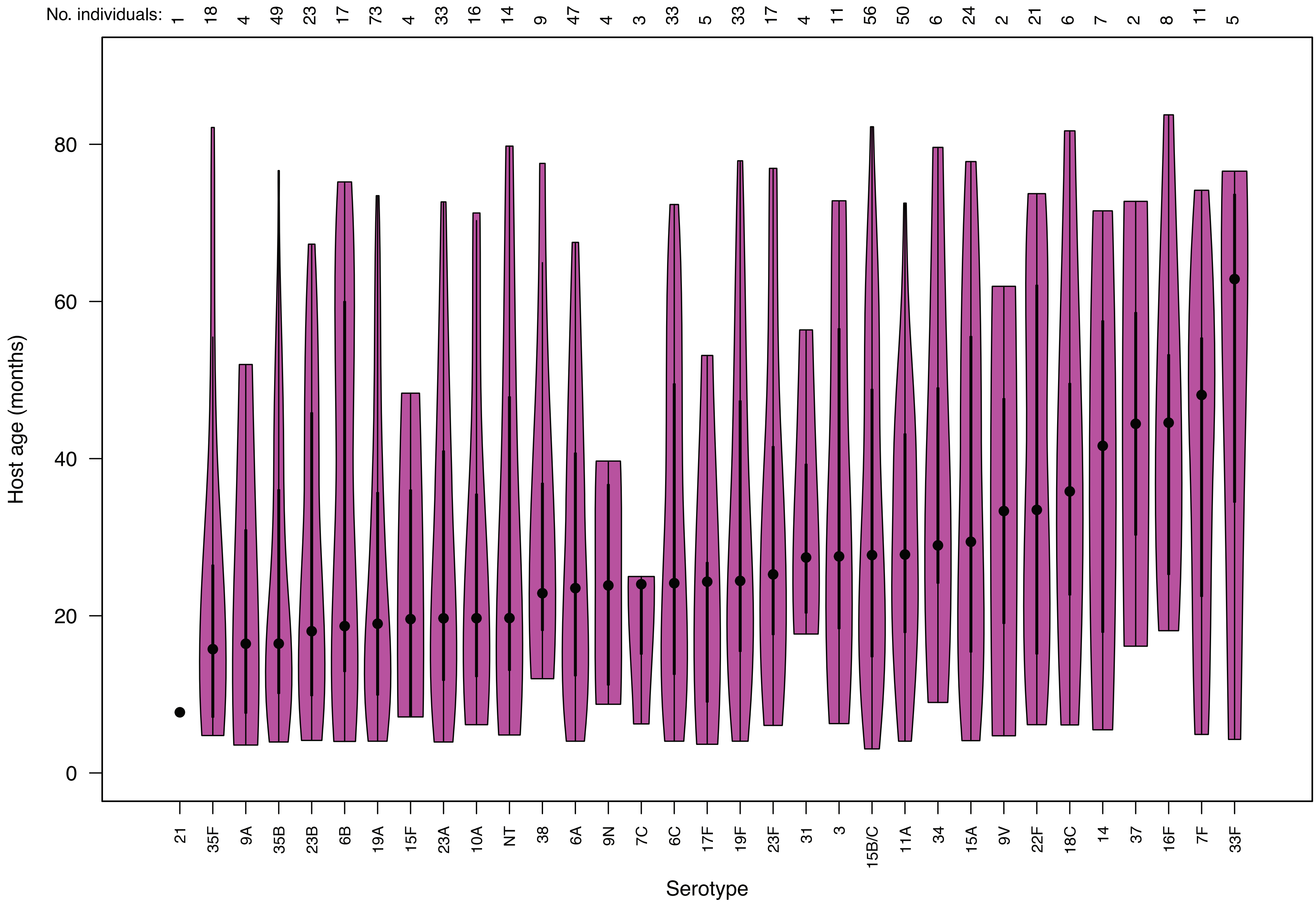

No. individuals:
1 18 4 49 23 17 73 4 33 16 14 9 47 4 3 33 5 33 17 4 11 56 50 6 24 2 21 6 7 2 8 11 5
Host age (months)
0
20
40
60
80
21 35F 9A 35B 23B 6B 19A 15F 23A 10A NT 38 6A 9N 7C 6C 17F 19F 23F 31 3 15B/C 11A 34 15A 9V 22F 18C 14 37 16F 7F 33F
Serotype

Figure S16 Violin plots showing the range of host ages associated with each observed serotype. Each violin plot is annotated with the number of host age measurements it represents at the top of the plot. These data do not provide any significant evidence of age stratification of serotypes (Kruskal-Wallis test, p = 0.149).

Table S1 Protein features associated with antibody binding targets. This multivariable logistic binary regression analysis fitted a model combining the explanatory variables of different protein characteristics to the binary dependent variable of whether or not a protein was an ABT. The analysis removed variables preventing a maximum likelihood estimate, followed by stepwise model selection based on AIC values. The table lists the features found to significantly associate with being identified as an ABT: the protein's length, having a signal peptide for secretion, and possessing the listed functional motifs. The 'significance level' column marks $p < 0.1$ with '.', $p < 0.05$ with '*', $p < 0.01$ with '**', and $p < 0.001$ with '***'.

Feature	Estimate	Standard Error	z Value	Pr(>|z|)	Significance Level
Length	0.0007331	0.0001786	4.106	4.03E-05	***
Signal peptide	3.0113121	0.3622169	8.314	<2.00E-16	***
Lipoprotein motif	1.3103604	0.4658557	2.813	0.00491	**
ADH_zinc_N domain	2.6913921	1.1333986	2.375	0.01757	*
CBD	3.4696616	0.8163899	4.25	2.14E-05	***
HisKA domain	2.3065463	1.1112429	2.076	0.03793	*
Peptidase_M26_C domain	3.1660345	1.8254579	1.734	0.08285	.
SBP_bac_8 domain	-1.591431	1.1989284	-1.327	0.18438	
Histidine triad motif	5.309944	1.1971175	4.436	9.18E-06	***
Transpeptidase domain	2.8268601	0.9515744	2.971	0.00297	**
YSIRK motif	4.1100924	1.4311293	2.872	0.00408	**

Table S2 Domains associated with higher antibody binding within the set of antibody binding targets. The multivariable gamma family general linear regression used the median antibody binding value across the serum samples as the dependent variable, and protein characteristics as the explanatory variables. This model was optimised by stepwise model selection based on AIC values. The table lists the features found to significantly associate with higher levels of IgG binding: signal peptides, Gram positive anchors for sortase-mediated attachment to the cell surface, LysM domains involved in peptidoglycan metabolism, and histidine triad domains. The same significance level annotation is used as in Table S1.

Feature	Estimate	Standard Error	t Value	Pr(>|t|)	Significance Level
Signal peptide	-0.04831	0.01354	-3.566	0.000577	***
CBD	-0.04262	0.01995	-2.137	0.035259	*
Sortase attachment motif	-0.06293	0.01865	-3.374	0.001085	**
LysM domain	-0.07563	0.03004	-2.518	0.013535	*
Histidine triad motif	-0.11037	0.02489	-4.434	2.55E-05	***

Dataset S1 Summary of the proteins analysed in this study. This dataset includes the identifiers, immunogenicity data, physiochemical properties, and evolutionary analyses results for all proteins included in this study. Some rows correspond to archetypes of diverse loci, which are associated with physiochemical properties and information regarding the encoding locus in whole genomes, but lack the results of evolutionary analyses because they were too diverse to align. In a complementary manner, the rows that correspond to individual variants of proteins correspondingly lack information on physiochemical properties, but are each associated with the output of evolutionary analyses, where a sufficient number of representatives were available.

(see Excel spreadsheet for dataset)

Supporting Information References

1. Croucher NJ, et al. (2013) Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet* 45(6):656–663.
2. Croucher NJ, et al. (2015) Population genomic datasets describing the post-vaccine evolutionary epidemiology of *Streptococcus pneumoniae*. *Sci Data* 2:150058.
3. Moschioni M, et al. (2008) *Streptococcus pneumoniae* contains 3 *rlrA* pilus variants that are clonally related. *J Infect Dis* 197(6):888–896.
4. Bagnoli F, et al. (2008) A second pilus type in *Streptococcus pneumoniae* is prevalent in emerging serotypes and mediates adhesion to host cells. *J Bacteriol* 190(15):5480–5492.
5. Paterson GK, Nieminen L, Jefferies JM, Mitchell TJ (2008) PclA, a pneumococcal collagen-like protein with selected strain distribution, contributes to adherence and invasion of host cells. *FEMS Microbiol Lett* 285(2):170–176.
6. Croucher NJ, et al. (2009) Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone *Streptococcus pneumoniae*$^{\mathrm{Spain23F}}$ ST81. *J Bacteriol* 191(5):1480–1489.
7. Tettelin H, et al. (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* 293(5529):498–506.
8. Swain MT, et al. (2012) A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nat Protoc* 7(7):1260–1284.
9. Camacho C, et al. (2009) BLAST+: architecture and applications. *BMC*

Bioinformatics 10:421.

10. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340(4):783–795.
11. Punta M, et al. (2012) The Pfam protein families database. *Nucleic Acids Res* 40(Database issue):D290–301.
12. Croucher NJ, et al. (2014) Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat Commun* 5:5471.
13. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14):1754–1760.
14. Hollingshead SK, Becker R, Briles DE (2000) Diversity of PspA: mosaic genes and evidence for past recombination in *Streptococcus pneumoniae*. *Infect Immun* 68(10):5889–5900.
15. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
16. Price MN, Dehal PS, Arkin AP (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3):e9490.
17. Brooks-Walter A, Briles DE, Hollingshead SK (1999) The *pspC* gene of *Streptococcus pneumoniae* encodes a polymorphic protein, PspC, which elicits cross-reactive antibodies to PspA and provides immunity to pneumococcal bacteremia. *Infect Immun* 67(12):6533–6542.
18. Davies DH, et al. (2005) Profiling the humoral immune response to infection by using proteome microarrays: high-throughput vaccine and diagnostic antigen discovery. *Proc Natl Acad Sci U S A* 102(3):547–552.
19. R Development Core Team (2011) *R: A language and environment for*

statistical computing (R Foundation for Statistical Computing, Vienna).

20. Giefing C, et al. (2008) Discovery of a novel class of highly conserved vaccine antigens using genomic scale antigenic fingerprinting of pneumococcus with human antibodies. *J Exp Med* 205(1):117–131.
21. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 305:567–580.
22. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* 8(10):785–786.
23. Sigrist CJ, et al. (2013) New and continuing developments at PROSITE. *Nucleic Acids Res* 41:D344–7.
24. Venables WN, Ripley BD (2002) Modern Applied Statistics with S. *Issues of Accuracy and Scale* (March):868.
25. Bethe G, et al. (2001) The cell wall-associated serine protease PrtA: A highly conserved virulence factor of *Streptococcus pneumoniae*. *FEMS Microbiol Lett* 205(1):99–104.
26. Sebert ME, Palmer LM, Rosenberg M, Weiser JN (2002) Microarray-based identification of *htrA*, a *Streptococcus pneumoniae* gene that is regulated by the CiaRH two-component system and contributes to nasopharyngeal colonization. *Infect Immun* 70(8):4059–67.
27. Berry A, et al. (1994) Cloning and nucleotide sequence of the *Streptococcus pneumoniae* hyaluronidase gene and purification of the enzyme from recombinant *Escherichia coli*. *Infect Immun* 62(3):1101–1108.
28. Zähner D, Hakenbeck R (2000) The *Streptococcus pneumoniae* beta-

galactosidase is a surface protein. *J Bacteriol* 182(20):5919–21.

29. Marion C, et al. (2009) Identification of a pneumococcal glycosidase that modifies O-linked glycans. *Infect Immun* 77(4):1389–96.

30. Pettigrew MM, Fennie KP, York MP, Daniels J, Ghaffar F (2006) Variation in the presence of neuraminidase genes among *Streptococcus pneumoniae* isolates with identical sequence types. *Infect Immun* 74(6):3360–3365.

31. Briles DE, et al. (2000) Intranasal immunization of mice with a mixture of the pneumococcal proteins PsaA and PspA is highly protective against nasopharyngeal carriage of *Streptococcus pneumoniae*. *Infect Immun* 68(2):796–800.

32. Jomaa M, et al. (2006) Immunization with the iron uptake ABC transporter proteins PiaA and PiuA prevents respiratory infection with *Streptococcus pneumoniae*. *Vaccine* 24(24):5133–5139.

33. Claverys JP, Grossiord B, Alloing G (2000) Is the Ami-AliA/B oligopeptide permease of *Streptococcus pneumoniae* involved in sensing environmental conditions? *Res Microbiol* 151(6):457–463.

34. Khandavilli S, et al. (2008) Maturation of *Streptococcus pneumoniae* lipoproteins by a type II signal peptidase is required for ABC transporter function and full virulence. *Mol Microbiol* 67(December 2007):541–557.

35. Härtel T, et al. (2011) Impact of glutamine transporters on pneumococcal fitness under infection-related conditions. *Infect Immun* 79(1):44–58.

36. Basavanna S, et al. (2009) Screening of *Streptococcus pneumoniae* ABC transporter mutants demonstrates that *livJHMGF*, a branched-chain amino acid ABC transporter, is necessary for disease pathogenesis. *Infect Immun* 77(8):3412–3423.

37. Kloosterman TG, Kuipers OP (2011) Regulation of arginine acquisition and virulence gene expression in the human pathogen *Streptococcus pneumoniae* by transcription regulators ArgR1 and AhrC. *J Biol Chem* 286(52):44594–44605.

38. Panina EM, Vitreschak AG, Mironov AA, Gelfand MS (2003) Regulation of biosynthesis and transport of aromatic amino acids in low-GC Gram-positive bacteria. *FEMS Microbiol Lett* 222(2):211–220.

39. Saxena S, Khan N, Dehinwal R, Kumar A, Sehgal D (2015) Conserved Surface Accessible Nucleoside ABC Transporter Component SP0845 Is Essential for Pneumococcal Virulence and Confers Protection *In Vivo*. *PLoS One* 10(2):e0118154.

40. Nieto C, Espinosa M, Puyet A (1997) The maltose/maltodextrin regulon of *Streptococcus pneumoniae*. Differential promoter regulation by the transcriptional repressor MalR. *J Biol Chem* 272(49):30860–5.

41. Marion C, Burnaugh AM, Woodiga SA, King SJ (2011) Sialic acid transport contributes to pneumococcal colonization. *Infect Immun* 79(3):1262–9.

42. Soualhine H, et al. (2005) A proteomic analysis of penicillin resistance in *Streptococcus pneumoniae* reveals a novel role for PstS, a subunit of the phosphate ABC transporter. *Mol Microbiol* 58:1430–1440.

43. Jomaa M, et al. (2005) Antibodies to the iron uptake ABC transporter lipoproteins PiaA and PiuA promote opsonophagocytosis of *Streptococcus pneumoniae*. *Infect Immun* 73(10):6852–9.

44. Lawrence MC, et al. (1998) The crystal structure of pneumococcal surface antigen PsaA reveals a metal-binding site and a novel structure for a putative ABC-type binding protein. *Structure* 6(12):1553–1561.

45. Anderton JM, et al. (2007) E-cadherin is a receptor for the common protein pneumococcal surface adhesin A (PsaA) of *Streptococcus pneumoniae*. *Microb Pathog* 42(5-6):225–236.

46. Pinho MG, Kjos M, Veening J-W (2013) How to get (a)round: mechanisms controlling growth and division of coccoid bacteria. *Nat Rev Microbiol* 11(9):601–14.

47. Barendt SM, Sham LT, Winkler ME (2011) Characterization of mutants deficient in the L,D-carboxypeptidase (DacB) and WalRK (VicRK) Regulon, involved in peptidoglycan maturation of *Streptococcus pneumoniae* serotype 2 strain D39. *J Bacteriol* 193:2290–2300.

48. Turner MS, Hafner LM, Walsh T, Giffard PM (2004) Identification and Characterization of the Novel LysM Domain-Containing Surface Protein Sep from *Lactobacillus fermentum* BR11 and Its Use as a Peptide Fusion Partner in Lactobacillus and Lactococcus. *Appl Environ Microbiol* 70(6):3673–3680.

49. Tsui H-CT, et al. (2016) Suppression of a Deletion Mutation in the Gene Encoding Essential PBP2b Reveals a New Lytic Transglycosylase Involved in Peripheral Peptidoglycan Synthesis in *Streptococcus pneumoniae* D39. *Mol Microbiol*. doi:10.1111/mmi.13366.

50. Johnsborg O, Håvarstein LS (2009) Pneumococcal LytR, a protein from the LytR-CpsA-Psr family, is essential for normal septum formation in *Streptococcus pneumoniae*. *J Bacteriol* 191(18):5859–64.

51. Land AD, Winkler ME (2011) The requirement for pneumococcal MreC and MreD is relieved by inactivation of the gene encoding PBP1a. *J Bacteriol* 193(16):4166–4179.

52. Shivshankar P, Sanchez C, Rose LF, Orihuela CJ (2009) The *Streptococcus pneumoniae* adhesin PsrP binds to Keratin 10 on lung cells. *Mol Microbiol* 73(4):663–679.

53. Yamaguchi M, Terao Y, Mori Y, Hamada S, Kawabata S (2008) PfbA, a novel plasmin- and fibronectin-binding protein of *Streptococcus pneumoniae*, contributes to fibronectin-dependent adhesion and antiphagocytosis. *J Biol Chem* 283(52):36272–36279.

54. Frolet C, et al. (2010) New adhesin functions of surface-exposed pneumococcal proteins. *BMC Microbiol* 10:190.

55. Khan MN, Sharma SK, Filkins LM, Pichichero ME (2012) PcpA of *Streptococcus pneumoniae* mediates adherence to nasopharyngeal and lung epithelial cells and elicits functional antibodies in humans. *Microbes Infect* 14(12):1102–10.

56. Khan MN, Pichichero ME (2012) Vaccine candidates PhtD and PhtE of *Streptococcus pneumoniae* are adhesins that elicit functional antibodies in humans. *Vaccine* 30(18):2900–2907.

57. Ogunniyi AD, et al. (2009) Pneumococcal histidine triad proteins are regulated by the Zn^{2+}-dependent repressor AdcR and inhibit complement deposition through the recruitment of complement factor H. *FASEB J* 23(3):731–738.

58. Loisel E, et al. (2011) Biochemical characterization of the histidine triad protein PhtD as a cell surface zinc-binding protein of pneumococcus. *Biochemistry* 50(17):3551–8.

59. Barocchi MA, et al. (2006) A pneumococcal pilus influences virulence and host inflammatory responses. *Proc Natl Acad Sci U S A* 103(8):2857–2862.

60. Pérez-Dorado I, Galan-Bartual S, Hermoso JA (2012) Pneumococcal surface proteins: When the whole is greater than the sum of its parts. *Mol Oral Microbiol* 27(4):221–245.

61. Saleh M, et al. (2013) Molecular architecture of *Streptococcus pneumoniae* surface thioredoxin-fold lipoproteins crucial for extracellular oxidative stress resistance and maintenance of virulence. *EMBO Mol Med* 5(12):1852–1870.

62. Jakob RP, et al. (2015) Dimeric structure of the bacterial extracellular foldase PrsA. *J Biol Chem* 290(6):3278–3292.

63. Hermans PWM, et al. (2006) The streptococcal lipoprotein rotamase A (SlrA) is a functional peptidyl-prolyl isomerase involved in pneumococcal colonization. *J Biol Chem* 281(2):968–976.

64. Manso AS, et al. (2014) A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. *Nat Commun* 5:5055.

65. de Saizieu A, et al. (2000) Microarray-based identification of a novel *Streptococcus pneumoniae* regulon controlled by an autoinduced peptide. *J Bacteriol* 182(17):4696–4703.