Genome Biology and Evolution Advance Access published December 4, 2014 doi:10.1093/gbe/evu263

Global phylogenomic analysis of nonencapsulated *Streptococcus pneumoniae* reveals a deep-branching classic lineage that is distinct from multiple sporadic lineages

Markus Hilty[1,2*], Daniel Wüthrich[3, 4], Susannah J. Salter[5], Hansjürg Engel[1], Samuel Campbell[1], Raquel Sá-Leão[6], Hermínia de Lencastre[6,7], Peter Hermans[8], Ewa Sadowy[9], Paul Turner[10, 11], Claire Chewapreecha [5, 10], Mathew Diggle[12], Gerd Pluschke[13], Lesley McGee[14], Özgen Köseoğlu Eser[15], Donald E. Low[16†], Heidi Smith-Vaughan[17], Andrea Endimiani[1], Marianne Küffer[1], Mélanie Dupasquier[18], Emmanuel Beaudoing[18], Johann Weber[18], Rémy Bruggmann[3, 4], William P. Hanage[19], Julian Parkhill[5], Lucy J. Hathaway[1¶], Kathrin Mühlemann[1,2,†¶] and Stephen D. Bentley[5, 20¶]

1. Institute for Infectious Diseases, University of Bern, Switzerland
2. Department of Infectious Diseases, Inselspital, Bern University Hospital and University of Bern, Switzerland
3. Interfaculty Bioinformatics Unit, University of Bern, Switzerland
4. Swiss Institute of Bioinformatics, Switzerland
5. Wellcome Trust Sanger Institute, Hinxton, UK
6. Instituto de Tecnologia Química e Biológica, University of Lisbon, Portugal
7. Laboratory of Microbiology and Infectious Diseases, The Rockefeller University, New York, USA
8. Laboratory of Pediatric Infectious Diseases, Radboud University Medical Centre, Nijmegen, The Netherlands
9. National Medicines Institute, Warsaw, Poland
10. Shoklo Malaria Research Unit, Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Mae Sot, Thailand
11. Centre for Tropical Medicine, Nuffield Department of Medicine, University of Oxford, Oxford, UK
12. Clinical Microbiology Department, Queens Medical Centre, Nottingham, UK
13. Swiss Tropical and Public Health Institute, University of Basel, Switzerland
14. Respiratory Diseases Branch, Centers for Disease Control and Prevention, Atlanta, USA
15. Department of Microbiology, Medical Faculty, Hacettepe University, Ankara, Turkey
16. Mt Sinai Hospital & Public Health Laboratories, Toronto, Canada
17. Menzies School of Health Research, Charles Darwin University, Darwin, Australia
18. Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland
19. Department of Epidemiology, Center for Communicable Disease Dynamics, Harvard School of Public Health, Boston, Massachusetts, USA
20. Department of Medicine, University of Cambridge, Addenbrookes Hospital, Hills Road, Cambridge, CB2 0SP, UK

† Deceased

¶ Contributed equally

* Corresponding author

Contact details of corresponding author: Markus Hilty, PhD, Institute for Infectious Diseases, University of Bern, Friedbühlstrasse 51, CH-3010 Bern, Switzerland, Phone ++41 31 632 49 83, Fax ++41 31 632 87 66, E-mail: Markus.Hilty@ifik.unibe.ch

Abstract

The surrounding capsule of *Streptococcus pneumoniae* has been identified as a major virulence factor and is targeted by pneumococcal conjugate vaccines (PCV). However, nonencapsulated *Streptococcus pneumoniae* (Non-Ec-*Sp*) have also been isolated globally, mainly in carriage studies. It is unknown if Non-Ec-*Sp* evolve sporadically, if they have high antibiotic non-susceptiblity rates and a unique, specific gene content.

Here, whole genome sequencing of 131 Non-Ec-*Sp* isolates sourced from 17 different locations around the world was performed. Results revealed a deep-branching classic lineage that is distinct from multiple sporadic lineages. The sporadic lineages clustered with a previously sequenced, global collection of encapsulated *S. pneumoniae* (Ec-*Sp)* isolates while the classic lineage is comprised mainly of the frequently identified multi-locus sequences types ST344 (n=39) and ST448 (n=40). All ST344 and nine ST448 isolates had high non-susceptiblity rates to β-lactams and other antimicrobials. Analysis of the accessory genome reveals that the classic Non-Ec-*Sp* contained an increased number of mobile elements, than Ec-*Sp* and sporadic Non-Ec-*Sp*. Performing adherence assays to human epithelial cells for selected classic and sporadic Non-Ec-*Sp* revealed that the presence of a integrative conjugative element (ICE) results in increased adherence to human epithelial cells (*P*=0.005). In contrast, sporadic Non-Ec-*Sp* lacking the ICE had greater growth in vitro possibly resulting in improved fitness.

In conclusion, Non-Ec-*Sp* isolates from the classic lineage have evolved separately. They have spread globally, are well adapted to nasopharyngeal carriage and are able to coexist with Ec-*Sp.* Due to continued use of pneumococcal conjugate vaccines, Non-Ec-*Sp* may become more prevalent.

Keywords: Pneumococcal isolates, whole genome sequencing, comparative genomics, integrative conjugative elements (ICEs), antibiotic non-susceptibility,

Introduction

Streptococcus pneumoniae is an important human pathogen usually surrounded by a polysaccharide capsule which is considered to be a major virulence factor. Some isolates do not possess a capsule and are referred to as nonencapsulated *S. pneumoniae (*Non-Ec-*Sp)*. Although generally considered less virulent than encapsulated strains, Non-Ec-*Sp* are also isolated from sterile sites and make up 10% of those isolated from the nasopharynx (Carvalho, et al. 2003; Finland and Barnes 1977). This may be an underestimate as Non-Ec-*Sp* form small rough colonies on agar plates which might be overlooked. Furthermore, the proportion of pneumococci colonizing the nasopharynx of young children that are Non-Ec-*Sp* may be increasing in the current era of vaccines directed against polysaccharide capsules (Sá-Leão, et al. 2009).

Some Non-Ec-*Sp* clones have been repeatedly isolated worldwide, particularly sequence types (ST) ST344 and ST448 indicating intercontinental spread that may reflect an adaptive advantage in transmission. These types are associated with outbreaks of conjunctivitis (Martin, et al. 2003; Porat, et al. 2006). Non-Ec-*Sp* are also particularly associated with co-colonization with other pneumococci in the nasopharynx, offering frequent opportunities for recombination (Brugger, et al. 2009). Finally, Non-Ec-*Sp* may be repositories of antibiotic resistance genes which could be transferred to encapsulated pneumococci (Chewapreecha, et al. 2014; Hauser, et al. 2004).

Previous studies of encapsulated *S. pneumoniae* (Ec-*Sp*) have attempted to define the pan genome of the species, but the gene content of Non-Ec-*Sp* has not been systematically studied (Obert, et al. 2006). Genes exclusively and consistently present in Non-Ec-*Sp* (i.e. part of the core genome in this group, but not in other pneumococci) may be an adaptation for colonization in the absence of capsule. Sequencing of the capsule region of Non-Ec-*Sp* has revealed the presence of coding sequences (cds) named *aliB-like* open reading frame (ORF)1 and ORF2 due to their homology to *aliB* which encodes the substrate-binding protein of an ABC transporter for branched-chain amino acids (Hathaway, et al. 2004). AliB-like ORF1 and ORF2 have recently been shown to play a role in sensing and responding to

peptides in the environment (Claverys, et al. 2000; Hathaway, et al. 2014). Both AliB and AliB-like ORF2 have been reported to play a role in colonization (Hathaway, et al. 2014; Kerr, et al. 2004). AliB-like ORFs have also been found in the capsule region of the closely related species *S. mitis, S. oralis* and *S. pseudopneumoniae* as well as preceding the capsule genes in pneumococci of serotypes 25F and 38 (Bentley, et al. 2006). Salter *et al*. found another novel gene, a putative surface-anchored protein in place of capsule genes which was designated *nspA* (Salter, et al. 2012). The same gene was called PspK and was revealed to increase adherence to epithelial cells within other studies (Keller, et al. 2013; Park, et al. 2012).

In this work, we sought to employ whole genome sequence (wgs) analysis to define the phylogenetic population structure of Non-Ec-*Sp* and their relationship to Ec-*Sp.* We aimed to reveal whether Non-Ec-*Sp* are members of a separate lineage from Ec-*Sp* and whether there are Non-Ec-*Sp*-specific genes which might shed light on the mechanism of colonization and which we investigated by *in vitro* experiments. We were also interested in the possible role of Non-Ec-*Sp* as reservoirs of antibiotic resistance and so determined the resistance of the strains to a panel of antibiotics and investigated the alleles of relevant genes.

Materials and Methods

Sequencing of a reference isolate

We sequenced the nonencapsulated *S. pneumoniae* (Non-Ec-*Sp)* isolate 110.58 (ST344) on the Pacific Biosciences (PacBio) RS II platform to obtain a fully assembled reference genome. Sequence assembly yielded one major contig of 2.29 Mb and 3 small contigs at low coverage. The major contig was subsequently closed by running a second assembly. This was used as the reference sequence for subsequent analysis and has the accession number CP007593. The reference genome 110.58 consisted of 2 287 774 bp (39.7% GC content). Gene prediction and annotation was performed using Prodigal (version 2.60) and RAST server (http://rast.nmpdr.org/), respectively. There were 2246 predicted coding sequences (cds). Further details are given in the Supplementary methods.

Strain selection and identification of global Non-Ec-*Sp* isolates and antimicrobial susceptibility tests (ASTs)

Further Non-Ec-*Sp* were sourced from 17 different geographical locations including Europe (Portugal, Poland, Scotland, the Netherlands and Switzerland), Africa (Kenya and Ghana), Asia (Turkey and Thailand), Australia, and the Americas (Peru, Massachusetts, Alaska and Canada). The isolates were checked for the presence of contamination by growing on blood agar plates and colony purified. Subsequently, optochin susceptibility was determined for all isolates followed, in case of optochin resistance, bile solubility was tested. In addition, the absence of *cpsA* was verified by real time PCR using the primers as described (Hathaway, et al. 2007). The optochin susceptibility and bile solubility tests are consistent with the species assignment of *Streptococcus pneumoniae*, while the absence of *cpsA* is consistent with the presence of a true, nonencapsulated isolate i.e. an isolate without the first gene of the capsule operon. ASTs were performed for all isolates except the 9 from Australia for which only DNA was available. Disc diffusion and minimal inhibitory concentrations (MIC) for penicillin testing was performed using standardized methods (Clinical and Laboratory Standards 2012). The double-disc diffusion test ('D-Test') for erythromycin and clindamycin was performed as described previously (Asmah, et al. 2009). In total, 216 Non-Ec-*Sp*

isolates were confirmed of which 131 were sequenced. The criteria for selection for sequencing were: i) all ST344 and 448 isolates; ii) all isolates representing a unique geographical location and iii) a variety of antibiotic profiles (Table 1 and Supplementary Table 1).

Whole genome sequencing (wgs) and assembly for Non-Ec-*Sp*

For wgs, multiplexed libraries were created and subsequent sequencing was performed on the Illumina HiSeq platform producing paired-end reads as described (Croucher, et al. 2011). Eleven isolates did first not pass the quality control and were re-isolated and re-sequenced (Supplementary table 2). Finally all the Illumina reads of the 131 samples were then mapped to the reference genome 110.58. In total, a mean of 91.7% (range: 77.7%-99.8%) reads were mapped to the reference genome. This resulted in a 10x genome coverage of 91.8% (Range: 79.5%-100%) (Supplementary Table 2). *De novo* assembly was performed for all isolates using Velvet (Zerbino and Birney 2008).

Calculation of the core Clusters of Orthologous Groups (COGs) and diversity

For calculation of the core COGs we additionally downloaded the assemblies from 44 Ec-*Sp* isolates previously published (Donati, et al. 2010). This included the high quality assembled reference genome *S. pneumoniae* ATCC 700669 (Croucher, et al. 2009). Subsequently, the contigs of the Non-Ec-*Sp* and the Ec-*Sp* isolates were ordered using Mauve (Darling, et al. 2010). For the Ec-*Sp* isolates, ATCC 700669 was used as a reference and for the Non-Ec-*Sp* isolates the PacBioBiosciences assembly of 110.58 was used. Contigs that were shorter than 500 bp were excluded from the assemblies. The ordered contigs were concatenated. Gene prediction was again performed using Prodigal (version 2.60) included in the genome annotation pipeline Prokka (version 1.8) (Hyatt, et al. 2010). All protein sequences identified were compared with each other using BLASTP (version 2.2.27+) (Altschul, et al. 1990). Only protein alignments with a minimal alignment length of ≥70% - with regard to the longer protein - and an alignment similarity of ≥70% were used to construct edges for the graph analysis using NetworkX (http://networkx.github.io, v1.7). Each protein of every isolate was used as node in the graph. Clusters that contain proteins found in all isolates represent the

core genome. Presence and absence of genes compared to the reference genome 110.58 were plotted. ICE proteins were identified using the ICEberg database (Bi, et al. 2012) and complete ICEs assembled using the reference genome 110.58.

As assembly errors may contribute to the variation in gene content we compared the Illumina with the PacBio assembly of 110.58 to estimate the possible divergence. Within Illumina assemblies, we identified 15 genes in 14 COGs not present in the assembly received by PacBiosequencing (i.e 0.72%). We therefore conclude that there was some divergence due to assembly errors but this would not affect the results significantly.

Calculation of the unique COGs within Non-Ec-*Sp*

To define genes that uniquely appear in Non-Ec-*Sp*, all the identified COGs were used. Each COGs was then marked as unique if it was present in ≥ 80 Non-Ec-*Sp* but absent in all 44 Ec-*Sp*. For the newly defined region of diversity 2 (RD_2*Sp*ST344), we additionally calculated the COGs which were present in ≥ 40 Ec-*Sp* but absent in all Non-Ec-*Sp* (see text and Table 2 for details). To enable annotation of the COGs, the sequence of the protein with the most edges to other genes in the clusters was extracted.

Phylogenetic tree construction

Genes present as a single copy in every genome (132 Non-Ec-*Sp* (including 110.58 PacBio), 44 Ec-*Sp*), with the identical nucleotide length were designated as “core”. The ORF of the 363 COGs were aligned separately using Clustal Omega (version 1.1.0) (Sievers, et al. 2011). The resulting alignment files were fused using a Python script. Out of the fused alignment file the phylogenetic tree was constructed using RaxML (version 7.2.8a, raxmlHPC-PTHREADS-SSE3 -m GTRGAMMA -# 50) (Stamatakis 2006). The resulting phylogeny file was visualized using Figtree.

Analysis of ST344 and ST448 within classical Non-Ec-*Sp*

To analyze recent evolution of ST344 and ST448 isolates within the classical Non-Ec-*Sp*, rates of SNP and recombination were determined. To do this, whole-genome alignments were generated for 38 isolates of ST 344 isolates and 40 of ST 448. Non-Ec-*Sp* 110.58 was used as a reference and Illumina reads were mapped using SMALT resulting in two

alignment files. Rates of recombination in ST448 and ST344 were calculated as the ratio of the number of homologous recombination events to the number of mutations (*r*/*m*) (Chewapreecha, et al. 2014; Croucher, et al. 2011). In this study, we used the arithmetic mean of *r*/*m* for a cluster, averaged from the *r*/*m* of each branch within a cluster as previously described (Croucher, et al. 2011). The mean of the distribution of *r*/*m* for the cluster was then obtained. To produce phylogenetic trees from the whole-genome alignments for the isolates of the ST344 and ST448 clones, SNPs due to recombination were excluded.

Adherence to human epithelial cell line Detroit 562 and analysis of planktonic growth

As for measuring planktonic growth, 96-well microtiter polystyrene plates (Thermo Fisher Scientific, Denmark) were used for the isolates Switzerland 4, 11, Thailand 1 and 2 respectively. Wells were filled with a total of 192 µl of liquid chemically defined medium (CDM) supplemented with 5.5 mM glucose. Subsequently 8 µl (dilution = 1:25) of fresh bacterial suspension was added. Plate was sealed with parafilm, and OD450nm was measured on an ELISA plate reader (THERMOmax Microplate Reader, Molecular Devices Corporation, California) every 30 minutes for 30 hours using SOFTmax Pro 3.1.2. This also involved an essential 5 seconds of automatic shaking in the plate reader immediately before every measurement. Each experiment included three technical replicates and each experiment was performed three times.

For adherence assays, Detroit nasopharyngeal epithelial cells (ATCC-CCL-138) were cultured as described (Luer, et al. 2011). Adherence experiments for media including 5.5 mM glucose were subsequently performed nine times (3 times at 3 different time points) as recently described (Schaffner, et al. 2014). Ordinary one way ANOVA was performed to reveal statistical significance among the four isolates chosen for the experiments (Switzerland 4, 11, Thailand 1 and 2).

Results and Discussion

Identification of 'classic' and 'sporadic' Non-Ec-*Sp* by whole genome sequencing (wgs)

Whole genomic sequences were determined for a global collection of 131 Non-Ec-*Sp* from 17 different geographical sites (Table 1 and Supplementary Table 1). Assembled genomes were annotated revealing a total of 1148 clusters of orthologous genes (COGs) which were present in all 131 Non-Ec-*Sp* genomes and therefore represent the Non-Ec-*Sp* core genome. Genomes of recently published, global collection of 44 Ec-*Sp* were added to the dataset reducing the number of core COGs to 858 (Donati, et al. 2010).

Subsequently only genes present in every isolate as a single copy were selected (764 COGs). In addition, all the genes which did not have the same length in all the isolates were discarded resulting in 363 COGs (Figure 1). By comparing single nucleotide polymorphisms (SNPs) in these 363 COGs, we identified a lineage of exclusively Non-Ec-*Sp* which we have designated as the "classic" lineage (light grey background; Figure 1). Within this single lineage the isolates of ST344 and ST448 form distinct clades.

The remaining Non-Ec-*Sp* were distinct from the classical lineage, clustered with Ec-*Sp* and are part of multiple, sporadic lineages. Therefore, the Non-Ec-*Sp* of these lineages may have been encapsulated strains which lost the capsule 'sporadically' by acquiring the *aliB*-like ORFs, *nspA* (Salter, et al. 2012) or no additional genes (Table 1). Isolates from the recently published BC3-NT lineage, highly prevalent among the isolates from a Thai refugee camp, clustered with one of the sporadic lineages (Chewapreecha, et al. 2014) (Figure 1).

The existence of a distinct lineage of Non-Ec-*Sp* has been previously proposed on the basis of multi locus sequence typing (MLST) data (Croucher, et al. 2013; Hanage, et al. 2006), and is consistent with genome sequencing of a local sample collected in Massachusetts (Croucher, et al. 2013). Here we show that wgs of a global collection of isolates confirms the existence of a divergent Non-Ec-*Sp* group containing sequenc types (STs) 344 and 448 but also other STs (Supplementary table 1).

ST344 and ST448 may be particularly well adapted to colonization of the nasopharynx and conjunctiva. All the Canadian and Scottish isolates of ST448 were isolated from the eye, in concordance with the work of other groups who associated ST448 with outbreaks of conjunctivitis (Porat, et al. 2006). All other ST448 isolates and all ST344 isolates were isolated from the nasopharynx.

Classic and sporadic Non-Ec-*Sp* lineages by accessory genome diversity comparison

In addition to SNPs of core COGs, we investigated whether analysis of distinct gene content in the accessory genome supported the finding of a separate classic Non-Ec-*Sp* lineage. For this, we determined the accessory genome for each Non-Ec-*Sp* and Ec-*Sp* individually and performed pairwise comparisons which confirmed the separation of classic versus sporadic Non-Ec-*Sp* revealed by analysis of the SNPs of the core COGs (Figure 2). Again, the clones ST344 and ST448 were found in the classic lineage (Figure 2).

We also found that the accessory genomes and the genome sizes of classic Non-Ec-*Sp* are at the upper end of the spectrum as compared to the sporadic Non-Ec-*Sp* and the Ec-*Sp* (Figure 3A and 3B). This observation also holds true if only integrative conjugative element (ICE) and prophage proteins within the accessory genomes are analyzed (Figure 3C and D). Therefore, classic Non-Ec-*Sp* may act more likely as a gene reservoir as compared to either sporadic Non-Ec-*Sp* or Ec-*Sp* strains.

Characteristics of COGs within classic Non-Ec-*Sp* which are absent in Ec-*Sp*

In order to investigate whether Non-Ec-*Sp* have their own distinctive gene content, we examined COGs unique to Non-Ec-*Sp* but absent in the published Ec-*Sp* (Donati, et al. 2010) (Supplementary Table 3). We found a total of 116 COGs which were present in ≥ 80% Non-Ec-*Sp* but absent in all 44 Ec-*Sp* (Supplementary Table 4). The location of these and the core COGs as compared to the reference genome 110.58 were plotted in Figure 4.

As expected, 6/116 COGs were identified as unique for ≥ 80 Non-Ec-*Sp* within the capsule region. Non-Ec-*Sp* have no capsule but other genes have been described as being characteristic of this region in nonencapsulated isolates. Furthermore, 13/116 COGs were identified within prophage regions. While three COGs belonged to a prophage remnant, the

remaining COGs belonged to a complete phage which was inserted within exactly the same location of 110.58 as is the φMM1 phage for *S. pneumoniae* ATCC 700669. Production and release of this new prophage in isolate 110.58 when the culture was allowed to progress to lysis was confirmed by electron microscopy and is characteristic for Non-Ec-*Sp* isolates as we found this prophage neither in the sporadic Non-Ec-*Sp* nor in the Ec-*Sp* of this study (Supplementary Figure 1). 17 COGs were dispersed within the genome, while the majority of COGs (n=80) were found within two major regions of differences (RDs) and ICEs (Figure 4 and Supplementary Table 4).

Two complete ICEs are present within classic Non-Ec-*Sp*

Among the unique COGs within ICEs, there were 23 within ICE$_1$*Sp*ST344. Having the reference genome 110.58 allowed us to assemble these ICE proteins to a complete ICE (Figure 5). Though not yet described in *S. pneumoniae*, similar ICEs (~80% nucleotide identity to ICE$_1$*Sp*ST344) were described in *S. suis* (ICE*Ssu*) and *S. dysgalactiae (*ICE*Sde*3396) (Davies, et al. 2009; Holden, et al. 2009) (Figure 5). As for the cargo genes within ICE$_1$*Sp*ST344of Non-Ec-*Sp,* a large surface protein and an epsilon toxin-zeta antitoxin system (*pezAT*) was identified. *PezAT* has been characterized in terms of structure and function and its presence may result in outcompeting other bacteria during colonization (Chan, et al. 2012). However, COGs of ICE$_1$*Sp*ST344 were also found within sporadic Non-Ec-*Sp* (Figure 4). As the two lineages are phylogenetically different, the ICE$_1$*Sp*ST344may have been acquired independently through conjugation or possibly transformation. We therefore hypothesized that COGs of these ICEs could be important for the life cycle of Non-Ec-*Sp.* Indeed, isolates containing ICE$_1$*Sp*ST344 revealed an approximate 2 fold greater adherence to Detroit nasopharyngeal epithelial cells (*P*=0.005; Figure 6A). However, isolates of the BT3-NT lineage were isolated frequently in the refugee camp on the Thailand-Myanmar border despite lacking ICE$_1$*Sp*ST344 (Figure 6A), but showed improved growth (Figure 6B). In contrast, the isolate lacking both, ICE$_1$*Sp*ST344 and RD$_1$ showed poor growth in CDM supplemented with 5.5 M of glucose (Figure 6B). This may mean that having many

mobile elements may result in a fitness cost and that there is a careful balance between fitness, improved adherence and antibiotic resistance.

There were a further 10 unique COGs identified in an additional ICE, called ICE$_2$*Sp*ST344 (Figure 4). In contrast to ICE$_1$*Sp*ST344, ICE$_2$*Sp*ST344 is not unique for Non-Ec-Sp as it has been indentified e.g. in *Streptococcus pneumoniae*Spain23F ST81 (Croucher, et al. 2009). However, although ICE$_2$*Sp*ST344 is somewhat similar to the ICE*Sp*23FST81, an additional putative lantibiotic transporter protein has been found within Non-Ec-*Sp* as previously described for an 'ancestral' *S. pneumoniae* (ICE*Sp*PN1) strain (Wyres, et al. 2013).

Finally, two additional, unique COGs with unknown function were also identified within the Pneumococcal Pathogenicity Island 1 (PPI-1) which has been described as ICE-derived genomic island (Croucher, et al. 2009). As characteristic for ICE$_1$*Sp*ST344 and ICE$_2$*Sp*ST344, the PPi also possesses an epsilon toxin-zeta antitoxin system.

Two major regions of differences (RD$_1$*Sp*ST344 **and** RD$_2$*Sp*ST344**) are uniquely present within Non-Ec-*Sp***

The remaining COGs unique for Non-Ec-*Sp*, were identified within two major regions of differences (RD$_1$*Sp*ST344 and RD$_2$*Sp*ST344). RD$_1$*Sp*ST344 contains 35 COGs for the majority of Non-Ec-*Sp* upstream of SPN23F08840 (translation initiation factor IF-3) which are absent in all Ec-*Sp* isolates (Figure 4). More than half of the COGs (20; 56%) are annotated as hypothetical protein with unknown function. Among those remaining are mobilization and conjugation proteins indicating that the genes of RD$_1$*Sp*ST344 are part of a mobile element. The functional importance of RD$_1$*Sp*ST344 within Non-Ec-Sp has yet to be revealed. However, the presence of RD$_1$*Sp*ST344 had no influence on the adherence potential of *S. pneumoniae* to Detroit nasopharyngeal epithelial cells (Figure 6A).

RD$_2$*Sp*ST344 is situated between the ribosomal protein L33 (SPN23F21670; 110_58_2260) and a zinc ABC transporter (SPN23F2201; 110_58_2284). In contrast to RD$_1$*Sp*ST344, RD$_2$*Sp*ST344 is present in Non-Ec-*Sp* and Ec-*Sp* but there is considerable diversity. Among the genes which are uniquely present within Non-Ec-*Sp* there is a beta galactosidase, an outer membrane lipoprotein and a putative zinc metalloprotease (Table 2). These genes

replace an arginine deaminase system (ADS) which is found within the Ec-*Sp* (Table 2). ADS is thought to influence the arginine metabolism (Kloosterman and Kuipers 2011) and genetic inactivation of the ADS affects colonization and dissemination in mice (Schulz, et al. 2013).

High antibiotic resistance within Non-Ec-*Sp*An alternative hypothesis for the global success of ST344 and ST448 is a high antibiotic resistance in these clones. Therefore, 123 of the 131 isolates were tested for susceptibility to the main antibiotic classes (Table 1). The three major genes which determine β-lactam susceptibility or resistance are the penicillin binding proteins (PBPs) found on either side of the 'capsule' locus (*pbp2x* and *pbp1a*) and elsewhere in the genome (*pbp2b*). Therefore, we analyzed the sequences of all three *pbp* genes to determine allelic combinations associated with increased MIC towards penicillin (Supplementary figures 2-4).

Overall, the majority of ST448 and other classic Non-Ec-*Sp* isolates were fully susceptible to penicillin (Table 1). In contrast, the classic Non-Ec-*Sp* isolates of ST344 had a MIC of at least 0.06 µg/ml but not more than 2.0 µg/ml for penicillin (Table 1). This may indicate a well-adapted balance between a basic tolerance towards penicillin without having the cost for fitness loss as shown recently for isolates with very high penicillin resistance (Trzcinski, et al. 2006).

With regard to non-β-lactam antibiotic resistance, 95% of ST344 isolates (n=36) were resistant towards both tetracycline and erythromycin (Table 1). The presence of *tet(M)* was confirmed in all the tetracycline non-susceptible isolates. However, three additional isolates carrying the *tet(M)* gene were found to be susceptible due to disruption of the gene leading to a frame shift mutation.

As for macrolide resistance, 36 isolates of the classic (ST344) and 7 of the sporadic lineages were fully resistant due to the presence of the *ermB* rRNA methylase. Four isolates initially presented erythromycin resistance but clindamycin susceptibility despite *ermB* being present. Subsequent erythromycin-clindamycin 'D' test revealed inducible expression of the resistance phenotype for these isolates. Additionally, five isolates presented intermediate erythromycin resistance due to the presence of the *mel-mef* efflux pump. The presence of

the *mel-mef* efflux pump is normally not as strongly associated with sequence clades, an observation consistent with our data (Croucher, et al. 2013) (Table 1). The *ermB* gene has been shown to be carried by the *Tn*917 or Omega resistance cassettes (Croucher, et al. 2013). The same was true for the Non-Ec-*Sp* isolates for which the Tn917 and/or Omega resistance cassettes were identified within ICE$_2$*Sp*ST344 (Figure 4).

As for sulfa drugs, ST448 isolates were generally susceptible to trimethoprim-sulfamethoxazole with the exception of those from Thailand (n=9) and Kenya (n=2). Within these, resistance-conferring mutations were detected in the genes *dyr* (encoding dihydrofolate reductase) and *folP* (dihydropteroate synthase). For ST344 all but the oldest isolate of the collection (Alaska 1 from 1998) were resistant to sulfa drugs (Table 1) indicating a worrying evolution of ST344 towards high antibiotic resistance.

Evolutionary rates within ST344 and 448

Previous studies revealed that *S. pneumoniae* recombines at different rates across different lineages. In particular, it has been found that a specific NT-lineage (BC3-NT) identified in a refugee camp on the Thailand-Myanmar border had very high frequencies of uptake and donation of DNA fragments compared to encapsulated lineages (Chewapreecha, et al. 2014). However, as BC3-NT strains clustered with the sporadic Non-Ec-*Sp* isolates within our global collection (Figure 1), we wondered whether the classic ST344 and 448 had similar recombination rates to BC3-NT. Furthermore, we were interested in analyzing whether the hot spot for recombination differs from those previously published within different lineages. In order to determine recombination rates, whole genome alignments were generated for ST344 and ST448 isolates using the PacBio sequenced genome 110.58 as a reference. We then calculated the rates of recombination in ST344 and ST448 as the ratio of the number of homologous recombination events to the number of mutations (*r*/*m*) according to previously described methods (Chewapreecha, et al. 2014; Croucher, et al. 2011). The estimates for the r/m by recombination event per mutation were 0.18 (95% CI 0.12-0.26) and 0.15 (95% CI 0.02-0.28) for the ST344 and ST448 isolates, respectively. These values are lower than for the recently described, sporadic Non-Ec-*Sp*, BC3-NT for which r/m was 0.3 (Chewapreecha,

et al. 2014) and approximately similar as for multidrug resistant (MDR) Ec-*Sp* PMEN14 (0.17), PMEN1 (0.1) and PMEN2 (0.17) (Croucher, et al. 2013; Croucher, et al. 2014). The low r/m values within ST344 and ST448 are surprising as it is generally assumed that the absence of the capsule as a physical barrier facilitates DNA exchange (Chewapreecha, et al. 2014). However, the low values suggest that the classic Non-Ec-*Sp* clones ST344 and ST448 are stable, fit and well adapted even though they lack the major virulence factor of the capsule.

Recombination hotspots within ST448 and ST344

Despite a general absence of recombination in ST448, frequent recombination sites were detected for the isolates from Australia (n=1) and the refugee camp on the Thailand-Myanmar border (n=8) (Supplementary Figure 5). A driving force for the latter is probably the high antibiotic consumption within the camp as the recombination hotspots are *pbp2x*, *pbp2b* and *pbp1a*, genes involved in penicillin resistance. Recombination within these genes was also detected within the worldwide ST344 collection, though the evolution went the 'opposite' way i.e. towards gaining antibiotic susceptibility (Supplementary Figure 6). The 4 outgroup isolates of this clone (Portugal 21, Switzerland 10, Portugal 9, Alaska 1) had a mean MIC for penicillin of 0.6 µg/ml (±0.2 µg/ml) (Supplementary table 1). Of the remaining ST344 isolates, all but two had lower penicillin resistance. This could be the result of opposing selection pressures: antibiotic-mediated selection versus the cost of acquiring resistance on strain fitness. However, as the number of sequenced isolates was limited, this inference may change with further sequencing. Penicillin resistant strains have been shown previously to be less fit than susceptible strains in the absence of penicillin (Trzcinski, et al. 2006). For the ST448 isolates, all but nine were fully susceptible to penicillin (8 isolates from Thailand and one from Australia; Table 1). It is possible that there is a higher antibiotic pressure present in these two settings and therefore an expansion of a resistant sub-clone from within the susceptible lineage. Further hotspots of recombination affect choline binding proteins like *lytB* (Supplementary Figures 5 and 6).

In conclusion, due to continued use of pneumococcal conjugate vaccines which target the pneumococcal capsule, Non-Ec-*Sp* may become more prevalent in the future. This is why a thorough understanding of the types of Non-Ec-*Sp* and their importance is urgently needed. Performing wgs of a global collection of Non-Ec-*Sp* we clearly showed a distinct lineage of classical Non-Ec-*Sp*. This lineage is dominated by, but not restricted to, ST344 and 448. The other Non-Ec-*Sp* named here as “sporadic” cluster with encapsulated strains. Compared to Ec-*Sp*, classic Non-Ec-*Sp* have a very high number of mobile elements (e.g. ICEs) resulting in an increased genome size and adherence to epithelial cells but a lowered potential of planktonic growth as shown in *in vitro* experiments.

Supplementary Material

Supplementary methods, figures S1–S6 and tables S1-S4 are available at Genome Biology and Evolution online.

Acknowledgments

We thank Suzanne Aebi for excellent technical assistance. This work was partly supported by a grant from the Swiss National Science Foundation [31003A_133157/1] awarded to K. M. and currently led by L.J.H. Work at the Wellcome Trust Sanger Institute was supported by core grant no. 098051. P.T. was supported by a Wellcome Trust Clinical Research Training Fellowship, grant no 083735/Z/07. S.D.B. is also supported by the Cambridge BRC. W.P.H. was supported by the National Institutes of Health under Award Number R01 AI106786-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The purchase of the Pacific Biosciences (PacBio) RSII instrument was financed in part by the *Loterie Romande* through the *Fondation pour la Recherche en Médecine Génétique.*

We wish to thank the CDC Arctic Investigations Program for providing the Alaska isolates for this analysis. Isolates contributed by L.M. were from the Pneumococcal Global Strain Bank, a project supported by PATH.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403-410.

Asmah N, Eberspacher B, Regnath T, Arvand M. 2009. Prevalence of erythromycin and clindamycin resistance among clinical isolates of the *Streptococcus anginosus* group in Germany. Journal of Medical Microbiology. 58:222-227.

Bentley S, et al. 2006. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet. 2:e31.

Bi D, et al. 2012. ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria. Nucleic Acids Res. 40:D621-626.

Brugger S, Hathaway L, Mühlemann K. 2009. Detection of *Streptococcus pneumoniae* strain cocolonization in the nasopharynx. J Clin Micro. 47:1750-1756.

Carvalho M, Steigerwalt A, Thompson T, Jackson D, Facklam R. 2003. Confirmation of nontypeable *Streptococcus pneumoniae*-like organisms isolated from outbreaks of epidemic conjunctivitis as *Streptococcus pneumoniae*. J Clin Micro. 41:4415-4417.

Chan WT, Moreno-Cordoba I, Yeo CC, Espinosa M. 2012. Toxin-antitoxin genes of the Gram-positive pathogen *Streptococcus pneumoniae*: so few and yet so many. Microbiol Mol Biol Rev. 76:773-791.

Chewapreecha C, et al. 2014. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 46:305-309.

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

Clinical and Laboratory Standards I. 2012. Performance standards for antimicrobial susceptibility testing, twenty-second informational supplement. Document M100-S22. CLSI, Wayne, Pa. 32:104-107.

Croucher NJ, et al. 2013. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet. 45:656-663.

Croucher NJ, et al. 2014. Variable recombination dynamics during the emergence, transmission and 'disarming' of a multidrug-resistant pneumococcal clone. BMC Biol. 12:49.

Croucher NJ, et al. 2011. Rapid pneumococcal evolution in response to clinical interventions. Science. 331:430-434.

Croucher NJ, et al. 2009. Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone *Streptococcus pneumoniae* Spain23F ST81. J Bacteriol. 191:1480-1489.

Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 5:e11147.

Davies MR, Shera J, Van Domselaar GH, Sriprakash KS, McMillan DJ. 2009. A novel integrative conjugative element mediates genetic transfer from group G Streptococcus to other {beta}-hemolytic Streptococci. J Bacteriol. 191:2257-2265.

Donati C, et al. 2010. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol. 11:R107.

Finland M, Barnes M. 1977. Changes in occurrence of capsular serotypes of *Streptococcus pneumoniae* at Boston City Hospital during selected years between 1935 and 1974. J Clin Micro. 5:154-166.

Hanage WP, Kaijalainen T, Saukkoriipi A, Rickcord JL, Spratt BG. 2006. A successful, diverse disease-associated lineage of nontypeable pneumococci that has lost the capsular biosynthesis locus. Journal of Clinical Microbiology. 44:743-749.

Hathaway LJ, Battig P, Muhlemann K. 2007. In vitro expression of the first capsule gene of *Streptococcus pneumoniae*, cpsA, is associated with serotype-specific colonization prevalence and invasiveness. Microbiology. 153:2465-2471.

Hathaway LJ, et al. 2014. *Streptococcus pneumoniae* detects and responds to foreign bacterial peptide fragments in its environment. Open Biol. 4:130224.

Hathaway LJ, Stutzmann Meier P, Battig P, Aebi S, Muhlemann K. 2004. A homologue of aliB is found in the capsule region of nonencapsulated *Streptococcus pneumoniae*. J Bacteriol. 186:3721-3729.

Hauser C, Aebi S, Muhlemann K. 2004. An internationally spread clone of *Streptococcus pneumoniae* evolves from low-level to higher-level penicillin resistance by uptake of penicillin-binding protein gene fragments from nonencapsulated pneumococci. Antimicrob Agents Chemother. 48:3563-3566.

Holden MT, et al. 2009. Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen *Streptococcus suis*. PLoS One. 4:e6072.

Hyatt D, et al. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 11:119.

Institute. CLS. 2010. M02-A10. Performance standards for antimicrobial disk susceptibility tests; approved standard tenth edition. Wayne, PA: Clinical Laboratory Standards Institute.

Keller LE, et al. 2013. PspK of *Streptococcus pneumoniae* increases adherence to epithelial cells and enhances nasopharyngeal colonization. Infect Immun. 81:173-181.

Kerr A, et al. 2004. The Ami-AliA/B permease of *Streptococcus pneumoniae* is involved in nasopharyngeal colonization but not in invasive disease. Infect Immun. 72:3902-3906.

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. 286:44594-44605.

Luer S, Troller R, Jetter M, Spaniol V, Aebi C. 2011. Topical curcumin can inhibit deleterious effects of upper respiratory tract bacteria on human oropharyngeal cells *in vitro*: potential role for patients with cancer therapy induced mucositis? Supportive Care in Cancer: Official Journal of the Multinational Association of Supportive Care in Cancer. 19:799-806.

Martin M, et al. 2003. An outbreak of conjunctivitis due to atypical *Streptococcus pneumoniae*. N Engl J Med. 348:1112-1121.

Obert C, et al. 2006. Identification of a candidate *Streptococcus pneumoniae* core genome and regions of diversity correlated with invasive pneumococcal disease. Infect Immun. 74:4766-4777.

Park IH, et al. 2012. Nontypeable pneumococci can be divided into multiple cps types, including one type expressing the novel gene pspK. MBio. 3.

Porat N, et al. 2006. The important role of nontypable *Streptococcus pneumoniae* international clones in acute conjunctivitis. J Infect Dis. 194:689-696.

Sá-Leão R, et al. 2009. Changes in pneumococcal serotypes and antibiotypes carried by vaccinated and unvaccinated day-care centre attendees in Portugal, a country with widespread use of the seven-valent pneumococcal conjugate vaccine. Clin Microbiol Infect. 15:1002-1007.

Salter SJ, et al. 2012. Variation at the capsule locus, cps, of mistyped and non-typable *Streptococcus pneumoniae* isolates. Microbiology. 158:1560-1569.

Schaffner TO, et al. 2014. A point mutation in *cpsE* renders *Streptococcus pneumoniae* nonencapsulated and enhances its growth, adherence and competence. BMC Microbiol. 14:210.

Schulz C, Petruschka L, Hammerschmidt S. 2013. Strain specific regulation of the arginine deiminase system (ADS) in *Streptococcus pneumoniae* and impact of the ADS on virulence. XI European Meeting on the Molecular Biology of the Pneumococcus (EuroPneumo 2013).110-111.

Sievers F, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 7:539.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688-2690.

Trzcinski K, Thompson CM, Gilbey AM, Dowson CG, Lipsitch M. 2006. Incremental increase in fitness cost with increased beta -lactam resistance in pneumococci evaluated by competition in an infant rat nasal colonization model. J Infect Dis. 193:1296-1303.

Wyres KL, et al. 2013. Evidence of antimicrobial resistance-conferring genetic elements among pneumococci isolated prior to 1974. BMC Genomics. 14:500.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821-829.

Table 1: Characteristics of global nonencapsulated *Streptococcus pneumoniae* (Non-Ec-*Sp*) isolates

	Classic nonencapsulated			Sporadic nonencapsulated (%)
	ST344 (%)	ST448 (%)	Other 'classic' (%)	
Total	38 (100)	40 (100)	11 (100)	42 (100)
'capsule' (cps) genes				
aliB-like ORF1	38 (100)	40 (100)	11 (100)	17 (40)
aliB-like ORF2	38 (100)	40 (100)	11 (100)	18 (43)
nspA	0	0	0	11 (26)
'Real' *cps* genes	0	0	0	7 (17)
No genes	0	0	0	6 (14)
Antibiogram*				
Tetracycline, r	36 (95)	3 (8)	2 (18)	15 (36)
Levofloxacin, r	0	0	0	0
Vancomycin, r	0	0	0	0
Chloramphenicol, r	0	0	0	1 (2)
Trimethoprim-Sulfamethoxazole, r	37 (97)	11 (28)	7 (39)	29 (69)
Erythromycin, r	36 (95)	0	0	7 (17)
Erythromycin, i	1 (3)	3 (8)	0	1 (2)
Clindamycin, $cMLS_{B,}$	32 (84)	0	0	7 (17)
Clindamycin, $iMLS_B$	4 (11)	0	0	0
Penicillin (MIC)				
<0.06 µg/ml	0	31 (78)	9 (82)	6 (14)
0.06-0.25 µg/ml	32 (84)	0	2 (18)	9 (21)
0.38-2 µg/ml	6 (16)	9 (23)	0	12 (28)
>2 µg/ml	0	0	0	2 (5)
Presence of resistance genes				
Presence of *tetM*	38 (100)	3 (8)	2 (18)	16 (38)
Presence of *ermB*	36 (95)	0	0	7 (17)
Presence of *mefE*	1 (3)	3 (8)	0	1 (2)
Geographical origin				
Portugal	20 (53)	2 (5)	3 (27)	5 (12)
Thailand		9 (23)		17 (40)
Switzerland	11 (29)	2 (5)		
Canada		9 (23)		2 (5)
Australia		1 (3)		8 (19)
Massachusetts	1 (3)	6 (15)	1 (9)	
Poland	3 (8)			3 (7)
Scotland	1 (3)	4 (10)		
Alaska	1 (3)	1 (3)	1 (9)	1 (2)
Netherlands		3 (8)	1 (9)	
Peru		1 (3)		2 (5)
Kenya		2 (5)	1 (9)	
India				3 (7)
Turkey	1 (3)		1 (9)	
Ghana			2 (18)	
Nepal			1 (9)	
Mongolia				1 (2)

* 1 Sequence type (ST)448) and 8 'other' isolates from Australia had no antibiogram; r (resistant) and i (intermediate) according to disc diffusion test and CLSI guidelines (Institute. 2010). Constitutively ($cMLS_B$) Macrolide-Lincosamide-Streptogramin B or inducible ($iMLS_B$) phenotype; MIC (Minimal Inhibitory Concentration) in µg/ml

Table 2: Characteristics of cluster of orthologous genes (COGs) of the region of diversity (RD_2*Sp*ST344)

SPN23F ID*	110_58 ID	No of Non-Ec-*Sp* with the gene (%)**	Number of Ec-*Sp* with the gene (%)***	Annotation COGs
NA	02261	95 (72)	0	hypothetical protein
NA	02262	93 (70)	0	putative protein encoded in hyper variable junctions of pilus gene clusters
NA	02263	95 (72)	0	hypothetical protein
NA	02264	105 (80)	0	site-specific tyrosine recombinase XerD
NA	02274	93 (70)	0	hypothetical protein
NA	02277	92 (70)	0	hypothetical protein
NA	02279	93 (70)	0	Beta-galactosidase
NA	02280	90 (68)	0	Immunoglobulin A1 protease precursor****
NA	02283	94 (71)	0	Outer membrane lipoprotein P4
21800	NA	0	42 (95)	arginine deiminase *arcA*
21890	NA	0	44 (100)	alcohol dehydrogenase, iron-containing; *medH*
21900	NA	0	44 (100)	L-fucose isomerase *fucI*
21990	NA	0	40 (91)	fuculokinase *fucK*
22000	NA	0	41 (93)	putative fucose phosphotransferase system

* Numbering according to (Croucher, et al. 2009).

** COGs present ≥80 nonencapusulated *S. pneumoniae* (Non-Ec-*Sp)* but completely absent in 44 published Ec-*Sp* (Donati, et al. 2010).

*** According to the collection of 44 published *Ec-Sp* (Donati, et al. 2010). COGs present ≥40 Ec-*Sp* but absent in all Non-Ec-*Sp* are included.

****Gene is annotated as a putative zinc metalloprotease (ZmpC) by a recent study (Keller, et al. 2013).

Figure legends

Fig. 1: Core genome tree of 131 nonencapsulated and 44 encapsulated *Streptococcus pneumoniae*. 131 sequences of nonencapsulated *S. pneumoniae* (Non-Ec-*Sp*) isolates from 17 different countries or states of America were included (shown in red, green and blue). The 'classic' lineage of exclusively Non-Ec-*Sp* (light grey) contains the isolates of ST448 (green) and ST344 (red). Additionally published encapsulated pneumococci (indicated in pink) (Donati, et al. 2010) clustered within other Non-Ec-*Sp* (blue). COGs that are present in every isolate as a single copy and, in addition, having the same sequence length were selected for building the tree (363 COGs). The maximum-likelihood phylogeny was generated using 24342 polymorphic sites within a 278778 bp codon alignment. Three isolates of the recently published BC3-NT clade are also indicated (yellow) (Chewapreecha, et al. 2014).

Fig. 2: Accessory genome diversity of 131 nonencapsulated and 44 encapsulated *Streptococcus pneumoniae*. The classic lineage of exclusively Non-Ec-*Sp* is indicated (light grey) and contains the isolates of ST448 (green) and ST344 (red). The remaining isolates are sporadic Non-Ec-*Sp* and Ec-*Sp*. To construct the input array data all isolates were compared pairwise. Of each pair the COGs (70% coverage, 70% similarity) which are not part of the core-genome and are present in both isolates were counted. The isolates in the x-axis are ordered according to the Manhattan distance constructed dendrogram and in the y-axis by the phylogeny constructed on the core genome.

Fig. 3: Size and composition of accessory genome. The total number of accessory genes (**A**) of each isolate was determined by number of homologous groups (70 % coverage, 70 % similarity) which are present in the specific strain and not part of the core genome. The total genome sizes of all isolates are indicated **(B**). To detect genes of possible integrative and conjugative elements (ICEs) (**C**) in the different isolates, all coding sequences of each strain were aligned to the ICEberg database [1] (March 2014) using BLASTP (version 2.2.28+) [2]. Every coding sequence that showed homology (70 % coverage, 70 % similarity) to at least one ICE-protein was assumed to be ICE related. To determine the number of proteins that

originate by phages **(D**), the genomes were scanned for possible prophage region using PhiSpy (version 2.2) [3]. All the cds that lie in the region of a predicted prophage region were counted for each isolate separately.

Fig. 4: Presence and absence of coding sequences (cds) of nonencapsulated and encapsulated *S. pneumoniae* compared to a ST344 reference genome. Each row illustrates a specific isolate genome with the presence or absence of cds as defined in the reference genome (110.58). The rows are ordered by the phylogeny of the isolates (indicated on the left). The X-axis represents the genomic position of the cds in the genome of 110.58. The figure was constructed using R (version 3.1.0 alpha) with the "ggplot2" package. Two large surface proteins (*Sp*) were not correctly assembled in the illumina sequenced isolates and are therefore lacking. The reference genome (110.58) contains the genes *aliB*-like ORF1 and ORF2 at the cps (capsule) operon (Hathaway, et al. 2004) which were absent and replaced by cps genes in Ec-*Sp*. ICE; Integrative and conjugative element: RD; region of diversity: cps; capsule operon: ϕ; prophage

Fig. 5: Genetic organization and nucleic acid alignment of Integrative and conjugative elements (ICEs). Artemis comparison tool (ACT) was used for the nucleic acid alignment of the ICE*Ssu*$_{BM4407}$1 (*S. suis)*, ICE$_1$*Sp*ST344 and ICE*Sde*$_{3396}$ (*S. dysagalagticae*). Individual cds are indicated in light blue. Conserved cds beween the three ICEs are indicated by red shading. Cds which were unique for ICE$_1$*Sp*ST344 are indicated in dark blue.

Tn; transposon: *pezAT*; Toxin-Antitoxin Genes: *lrtA*; reverse transcriptase A: *prtP*; PII-type proteinase precursor

Fig. 6: Adherence of *S. pneumoniae* to human epithelial cells (A) and in vitro growth (B). Detroit nasopharyngeal epithelial cell lines were exposed to Non-Ec-*Sp* containing or lacking ICE$_1$*Sp*ST344 and RD$_1$*Sp*ST344, respectively **(6A)**. Adherence was determined at 30 minutes and calculated as the proportion of recovered bacteria to the inoculum. Experiments were repeated three times (3 times on 3 different time days). Swiss and Thai strains were classic and sporadic Non-Ec-*Sp*, respectively. For the nine adherence experiments, ordinary one-way ANOVA resulted in *P*=0.0052. See text for details. MLST; Multi locus sequence

type. Isolates of the BC3-NT lineage have been recently defined (Chewapreecha, et al. 2014).

Measurment of planktonic growth was done in 96-well microtiter polystyrene plates and OD_{450nm} was measured on an ELISA plate reader every 30 minutes for 30 hours (**6B**). Each experiment included three technical replicates and each experiment was performed three times.

Figure 1:

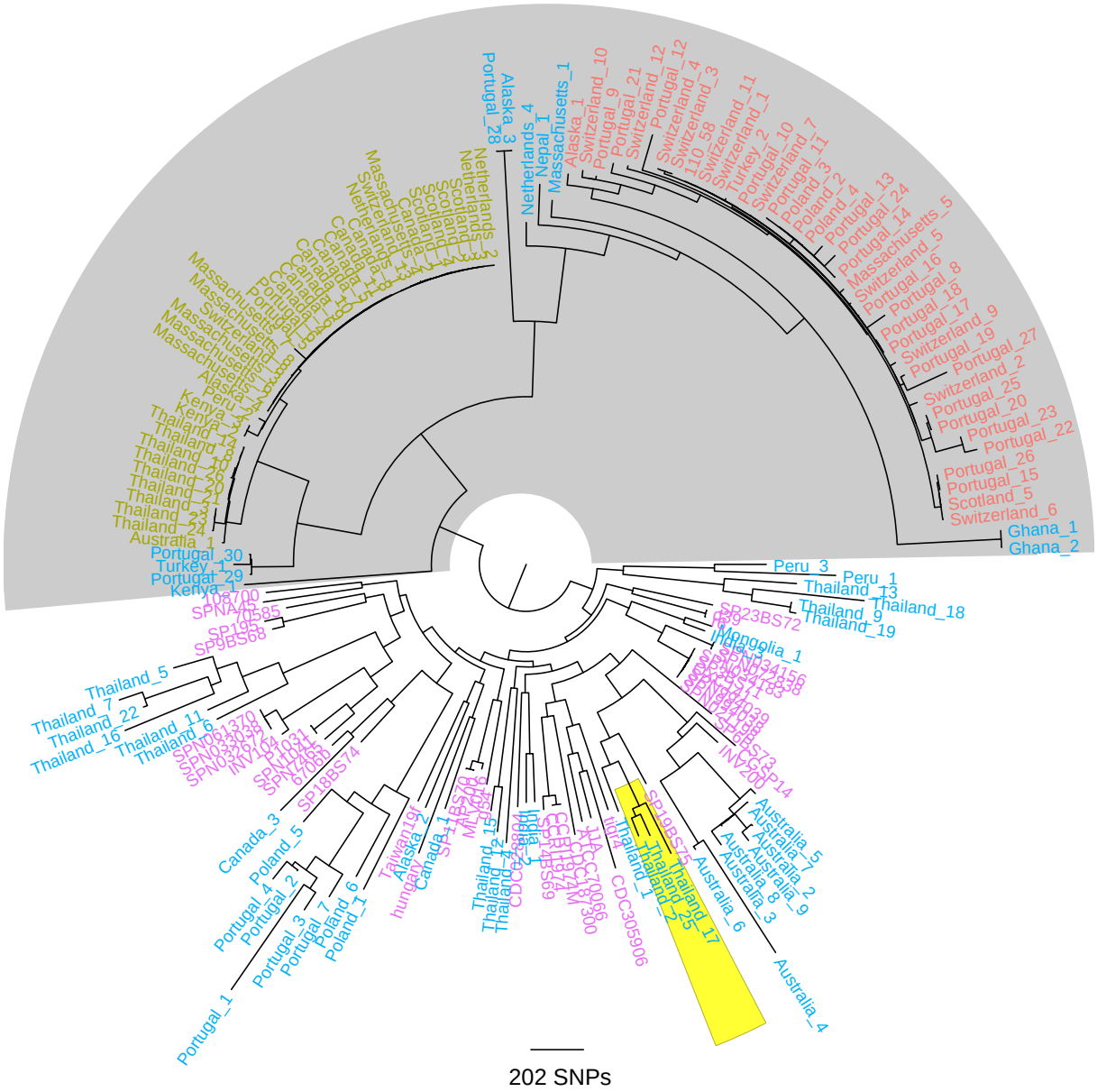

Figure 2:

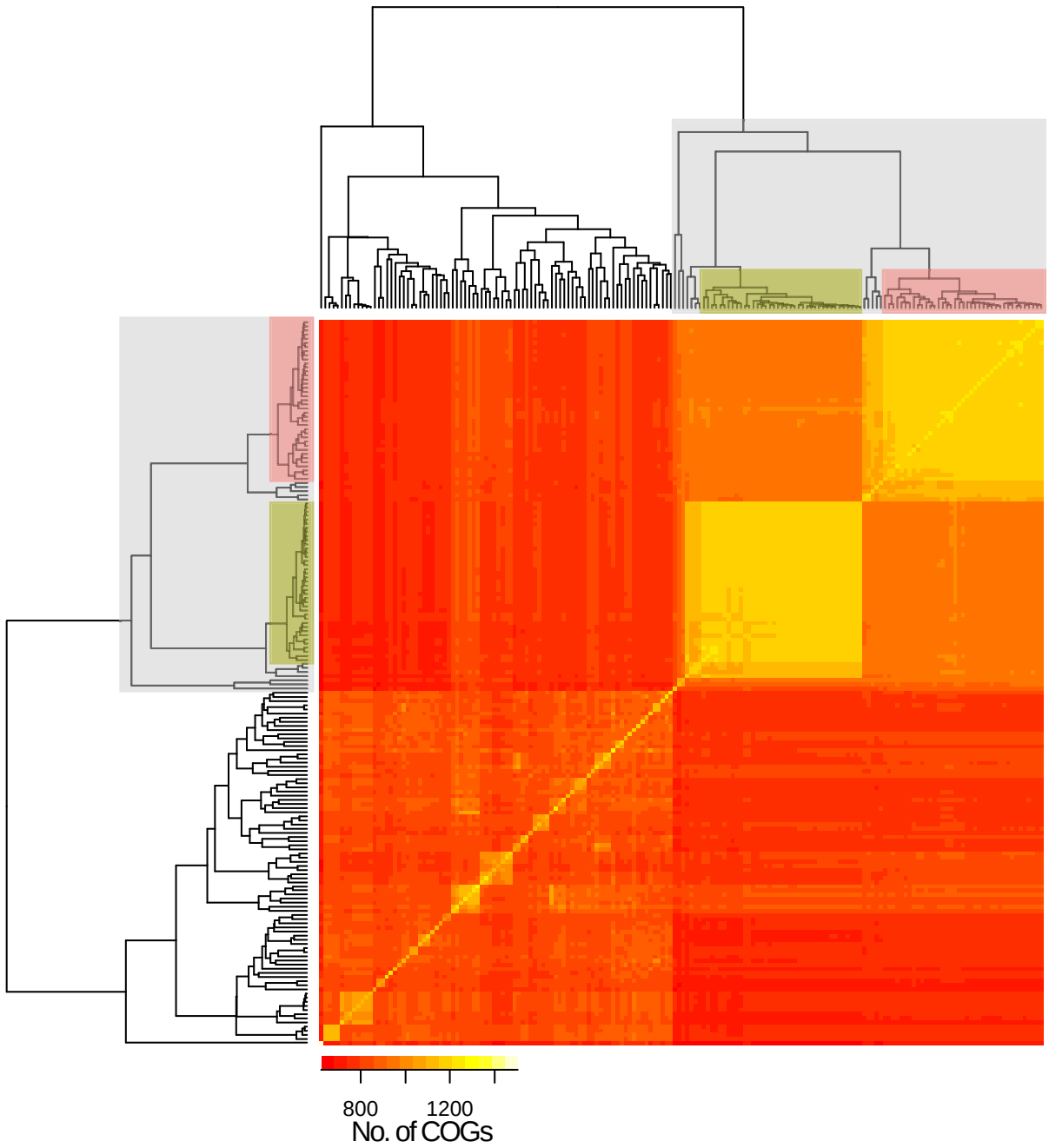

Figure 3:

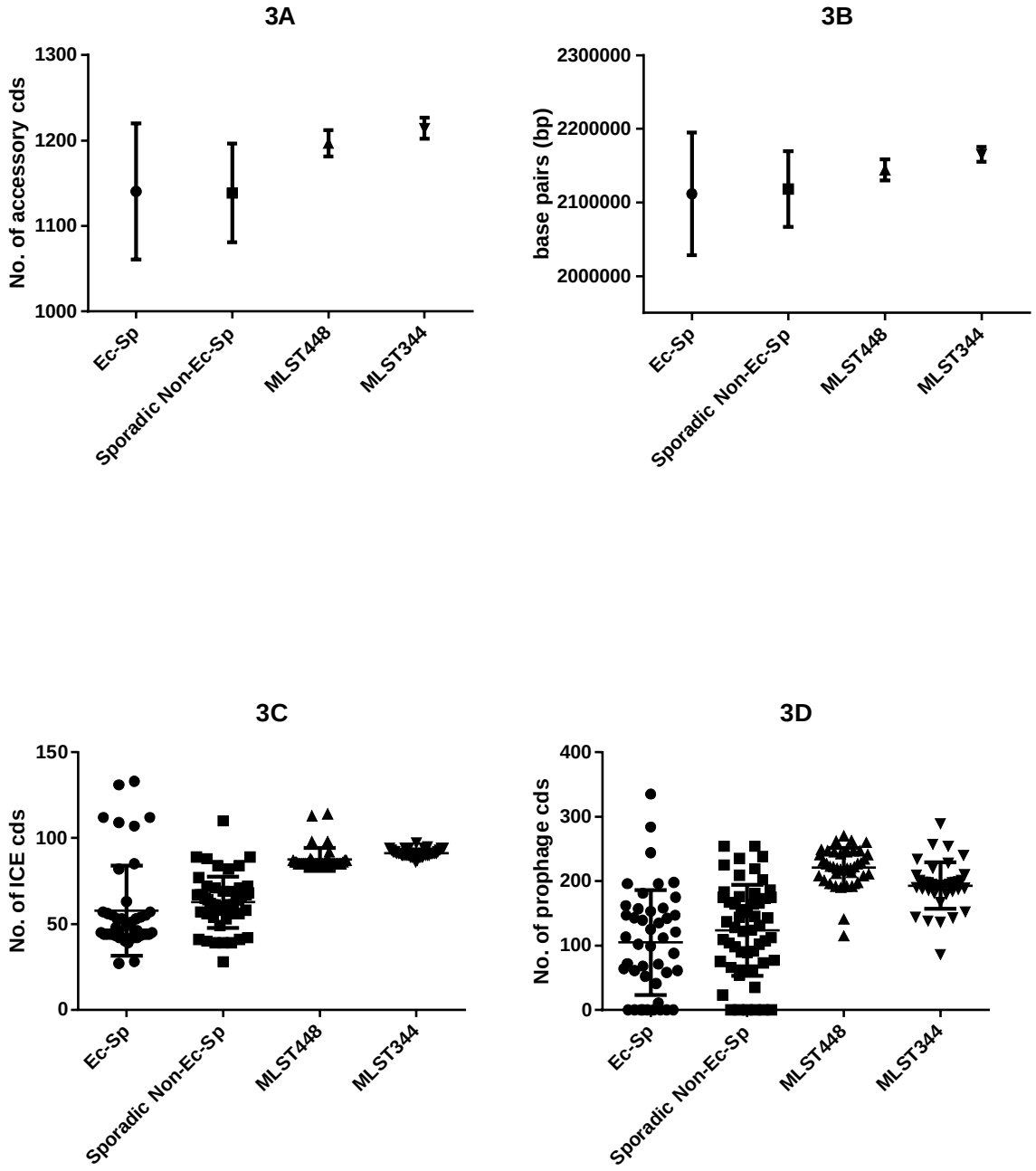

Figure 4:

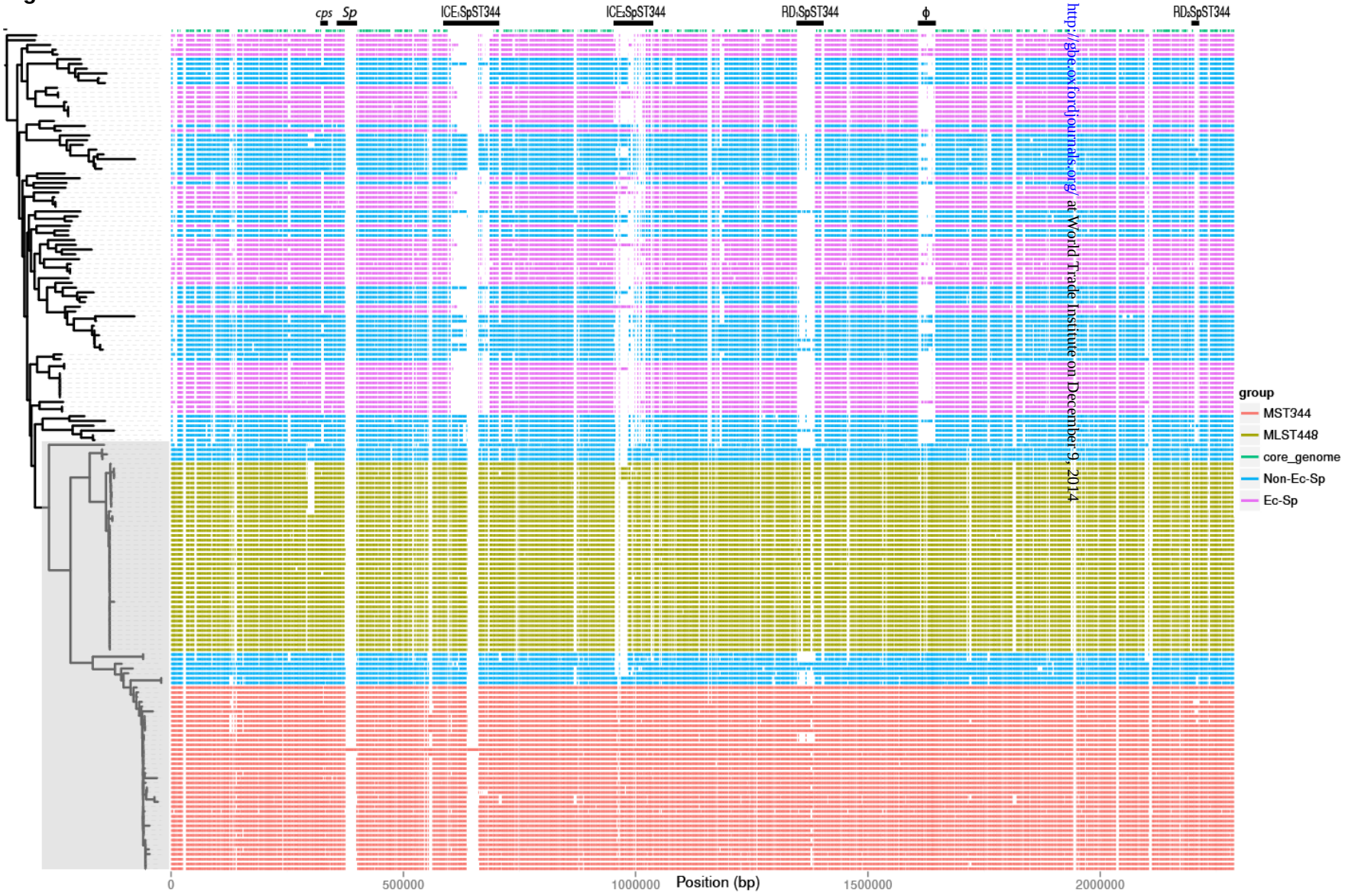

Figure 5:

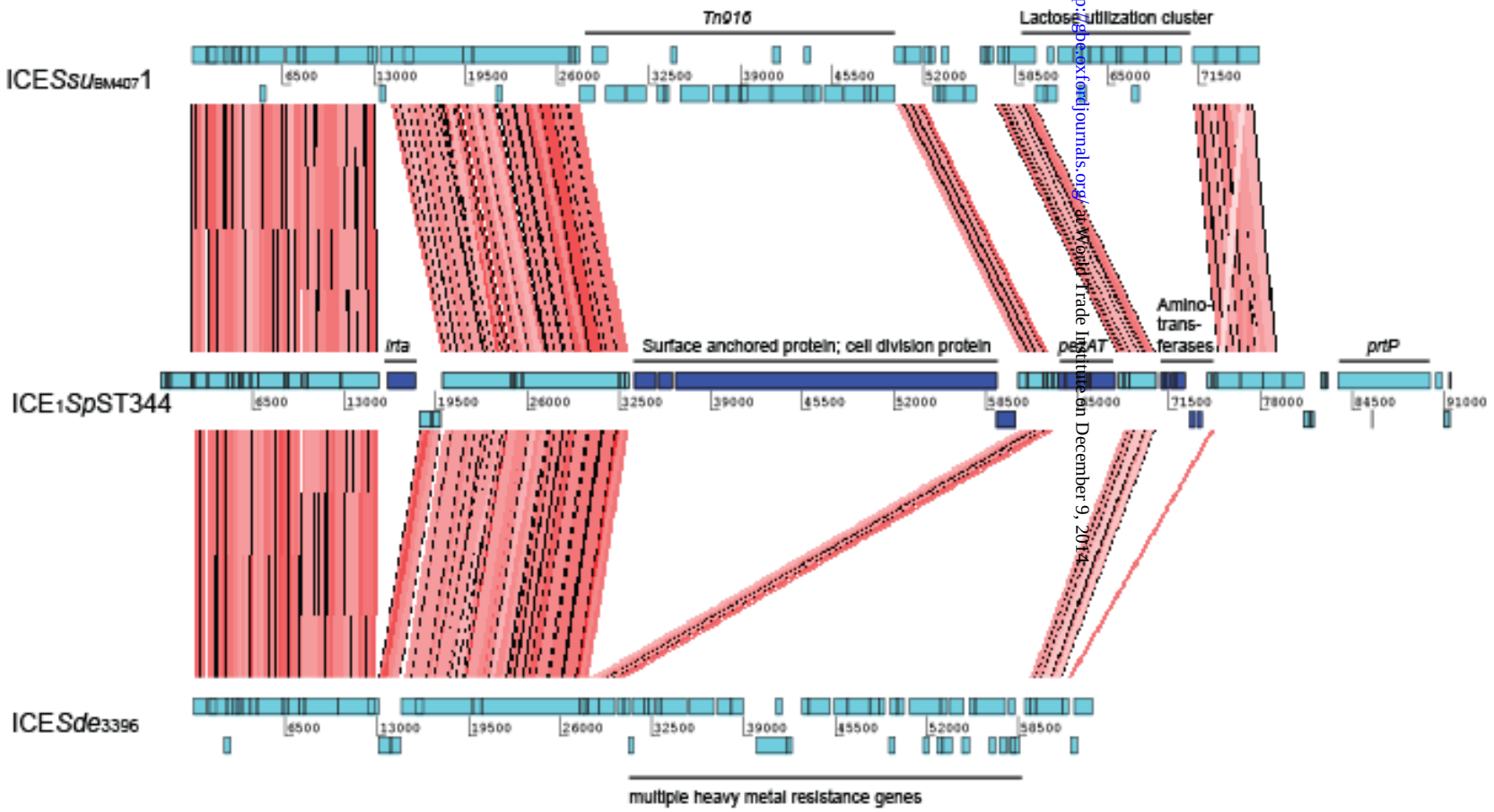

Figure 6