Visualizing variation within Global Pneumococcal Sequence Clusters (GPSCs) and country population snapshots to contextualize pneumococcal isolates

MICROBIAL GENOMICS

BIORESOURCE

Gladstone *et al.*, *Microbial Genomics* 2020;6

DOI 10.1099/mgen.0.000357

MICROBIOLOGY SOCIETY

OPEN DATA OPEN MICROBIOLOGY

Visualizing variation within Global Pneumococcal Sequence Clusters (GPSCs) and country population snapshots to contextualize pneumococcal isolates

Rebecca A. Gladstone[1,*], Stephanie W. Lo[1], Richard Goater[2,3], Corin Yeats[2,3], Ben Taylor[2,3], James Hadfield[4], John A. Lees[5], Nicholas J. Croucher[5], Andries J. van Tonder[1,6], Leon J. Bentley[1], Fu Xiang Quah[1], Anne J. Blaschke[7], Nicole L. Pershing[7], Carrie L. Byington[8], Veeraraghavan Balaji[9], Waleria Hryniewicz[10], Betuel Sigauque[11], K.L. Ravikumar[12], Samanta Cristine Grassi Almeida[13], Theresa J. Ochoa[14], Pak Leung Ho[15], Mignon du Plessis[16], Kedibone M. Ndlangisa[16], Jennifer E. Cornick[17], Brenda Kwambana-Adams[18,19], Rachel Benisty[20], Susan A. Nzenze[21,22], Shabir A. Madhi[21,22], Paulina A. Hawkins[23], Andrew J. Pollard[24], Dean B. Everett[25], Martin Antonio[19], Ron Dagan[20], Keith P. Klugman[23], Anne von Gottberg[15], Benjamin J. Metcalf[26], Yuan Li[26], Bernard W. Beall[26], Lesley McGee[26], Robert F. Breiman[23,27], David M. Aanensen[2,3], Stephen D. Bentley[1,*] and The Global Pneumococcal Sequencing Consortium[28]

Abstract

Knowledge of pneumococcal lineages, their geographic distribution and antibiotic resistance patterns, can give insights into global pneumococcal disease. We provide interactive bioinformatic outputs to explore such topics, aiming to increase dissemination of genomic insights to the wider community, without the need for specialist training. We prepared 12 country-specific phylogenetic snapshots, and international phylogenetic snapshots of 73 common Global Pneumococcal Sequence Clusters (GPSCs) previously defined using PopPUNK, and present them in Microreact. Gene presence and absence defined using Roary, and recombination profiles derived from Gubbins are presented in Phandango for each GPSC. Temporal phylogenetic signal was assessed for each GPSC using BactDating. We provide examples of how such resources can be used. In our example use of a country-specific phylogenetic snapshot we determined that serotype 14 was observed in nine unrelated genetic backgrounds in South Africa. The international phylogenetic snapshot of GPSC9, in which most serotype 14 isolates from South Africa were observed, highlights that there were three independent sub-clusters represented by South African serotype 14 isolates. We estimated from the GPSC9-dated tree that the sub-clusters were each established in South Africa during the 1980s. We show how recombination plots allowed the identification of a 20 kb recombination spanning the capsular polysaccharide locus within GPSC97. This was consistent with a switch from serotype 6A to 19A estimated to have occured in the 1990s from the GPSC97-dated tree. Plots of gene presence/absence of resistance genes (*tet*, *erm*, *cat*) across the GPSC23 phylogeny were consistent with acquisition of a composite transposon. We estimated from the GPSC23-dated tree that the acquisition occurred between 1953 and 1975. Finally, we demonstrate the assignment of GPSC31 to 17 externally generated pneumococcal serotype 1 assemblies from Utah via Pathogenwatch. Most of the Utah isolates clustered within GPSC31 in a USA-specific clade with the most recent common ancestor estimated between 1958 and 1981. The resources we have provided can be used to explore to data, test hypothesis and generate new hypotheses. The accessible assignment of GPSCs allows others to contextualize their own collections beyond the data presented here.

DATA SUMMARY

The following resources are accessible via www.pneumogen.net/gps and Figshare, all links are documented in the Supplementary Material (available Fig. S1 in the online version of this article). Input and output files, and scripts used to run BactDating on the GPSC phylogenies are available on Figshare. Roary output files from the 13454 genomes are available on Figshare. Visualizations of gene presence and absence and distribution of recombination blocks across the genome for each GPSC are hosted on Phandango, with the input files available on the Phandango Github [1]. Phylogenetic snapshots are available for each of the 12 countries with >100 isolates, paired with metadata, can be interactively viewed in Microreact (links in the Supplementary Material) [2]. Geographical distribution and other metadata can be interactively viewed in Microreact for each of the 73

000357

common GPSCs (links in the Supplementary Material) [2]. GPSC alignments with recombination masked are available on Figshare. Raw fastq data, assemblies and annotations for 13454 pneumococcal genomes were previously released [3] to the European Nucleotide Archive (ENA) as part of the Global Pneumococcal Sequencing project (GPS) with metadata and accessions in the Supplementary Material (Fig. S1). Raw fastq data and assemblies for pneumococcal isolates from Utah have been deposited in the ENA under the study accession PRJEB34550 and individual accessions are in the Supplementary Material (Fig. S1). The authors confirm all supporting data, code and protocols have been provided within the article or through Supplementary Material files.

Significance as a BioResource to the community

Analysis of the DNA of bacteria, such as the pneumococcus, can provide important insights into how they cause disease. Whole-genome sequencing is increasingly cost effective to comprehensively type bacteria, though the analyses of such datasets often require specialist training. We have taken the results of our analyses and visualize them in an interactive online format to make the results and interpretation more accessible. Any new pneumococcal genome can be easily assigned to Global Pneumococcal Sequence Clusters using PopPUNK with the GPS database, or via Pathogenwatch. Furthermore, we provide international descriptions of GPSC recombination profiles, gene content, estimated age of emergence and population snapshots of geographical regions to provide even greater context. Providing such bioinformatic output in interactive format makes data exploration easier, allowing dissemination of genomic insights into the wider community, and can be used as teaching tools. These resources could also facilitate cross-disciplinary research beyond the original aims of the project, for example mathematical modelling of resistance, serotype switching or geographical spread.

INTRODUCTION

Bacterial typing at the subspecies level to determine the lineage to which a pathogen belongs and distinguish it from others, is an important activity in the study and surveillance of infectious disease. Characteristics such as resistance, geographical spread and association with disease are key features of interest. Understanding their prevalence and distribution across lineages is informative for understanding the role of population structure in pneumococcal disease epidemiology. With the advent of high throughput sequencing, nucleotide variation across the whole genome can be used to cluster genetically related isolates into lineages, which are then a starting point for more detailed analysis. We previously published a genome-derived definition of

Received 13 December 2019; Accepted 03 March 2020; Published 30 April 2020
Author affiliations: [1]Parasites and microbes, Wellcome Sanger InstituteHinxton, UK; [2]Centre for Genomic Pathogen Surveillance, Wellcome Genome CampusHinxton, UK; [3]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK; [4]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA; [5]Faculty of Medicine, School of Public Health, Imperial College London, UK; [6]Department of Veterinary Medicine, University of Cambridge, Cambridge, UK; [7]Division of Pediatric Infectious Diseases, Department of Pediatrics, School of Medicine, University of Utah, 295 Chipeta Way, Salt Lake City, UT, 84108, USA; [8]University of California Health, Oakland, CA, USA; [9]Christian Medical College, Vellore, India; [10]National Medicines Institute, Division of Clinical Microbiology and Infection Prevention, Warsaw, Poland; [11]Fundação Manhiça / Centro de Investigação em Saúde da Manhiça (CISM), Maputo Mozambique, Instituto Nacional de Saúde, inistério de Saúde, Maputo, Mozambique; [12]Central Research Laboratory, Department of Microbiology, Kempegowda Institute of Medical Sciences Hospital & Research Center, Bangalore, India; [13]Center of Bacteriology, Adolfo Lutz Institute, São Paulo, Brazil; [14]Instituto de Medicina Tropical, Universidad Peruana Cayetano Heredia, Lima, Peru; [15]Department of Microbiology and Carol Yu Centre for Infection, The University of Hong Kong, Queen Mary Hospital, Hong Kong, PR China; [16]Centre for Respiratory Diseases and Meningitis, National Institute for Communicable Diseases, Johannesburg, South Africa; [17]Malawi-Liverpool-Wellcome-Trust, Malawi; [18]NIHR Global Health Research Unit on Mucosal Pathogens, Division of Infection and Immunity, University College London, London, UK; [19]WHO Collaborating Centre for New Vaccines Surveillance, Medical Research Council Unit The Gambia at The London School of Hygiene & Tropical Medicine, Fajara, The Gambia; [20]The Faculty of Health Sciences, Ben-Gurion University of the NegevBeer-Sheva, Israel; [21]Medical Research Council: Respiratory and Meningeal Pathogens Research Unit, University of the Witwatersrand, Johannesburg, South Africa; [22]Department of Science and Technology/National Research Foundation: Vaccine Preventable Diseases, University of the Witwatersrand, Johannesburg, South Africa; [23]Rollins School Public Health, Emory University, GA, USA; [24]Oxford Vaccine Group, Department of Paediatrics, University of Oxford, and the NIHR Oxford Biomedical Research Centre, Oxford, UK; [25]Queens Research Institute, University of Edinburgh, UK; [26]Centers for Disease Control and Prevention, Atlanta, GA, USA; [27]Emory Global Health Institute, Atlanta, GA, USA; [28]See the full list of the The Global Pneumococcal Sequencing Consortium members, in acknowledgements.
*Correspondence: Rebecca A. Gladstone, rg9@sanger.ac.uk; Stephen D. Bentley, sdb@sanger.ac.uk
Keywords: *Streptococcus pneumoniae*; pneumococcal; whole genome sequencing; population structure; recombination; antibiotic resistance; pangenome; phylogenetic dating.
Abbreviations: AMR, antimicrobial resistance; CC, clonal complex; GPS, Global Pneumococcal Sequencing project; GPSC, Global Pneumococcal Sequencing cluster; IPD, invasive pneumococcal disease; MDR, multi drug resistant; MLST, multi locus sequence type; PMEN, Pneumococcal Molecular Epidemiology Network; SNP, single nucleotide polymorphism; ST, sequence type; tMRCA, time to most recent common ancestor.
Repositories: Raw fastq data, assemblies and annotations for 13 454 pneumococcal genomes were previously released to the European Nucleotide Archive (ENA) as part of the Global Pneumococcal Sequencing project (GPS) with metadata and accessions can be found in the Supplementary Material [1]. Raw data and assemblies for pneumococcal isolates from Utah have been deposited in the ENA under the study accession PRJEB34550. Files from downstream bioinformatic analyses can be accessed via www.pneumogen.net/gps.
Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Supplementary Material is available with the online version of this article.

pneumococcal lineages based on an international collection of ~20000 pneumococcal genomes using PopPUNK [3, 4].

The uptake and incorporation of DNA from the environment into the chromosome via transformation and homologous recombination are known to contribute more nucleotide variation than mutation in the pneumococcus [5]. The capacity for recombination varies between lineages and recombination events are unevenly distributed across the genome, with known hotspots in regions under high selection pressure such as the penicillin-binding proteins and the capsular polysaccharide locus [6–8]. Recombination events can be detected and recombination rates quantified, which can be visualized across the genome to further explore recombination dynamics [1, 9]. As well as being biologically interesting, recombination obscures the true phylogenetic signal of vertical descent and is important to account for. Identifying recombination using blocks of dense SNPs lends itself to groups of closely related strains, and therefore relies on robust definitions of lineages.

Recombination can introduce new alleles or new genes into the chromosome. The latter contributes to the pneumococcal pangenome: the complete complement of genes observed in the species. The pangenome is made up of a core set of genes found in all isolates and accessory genes, including some that confer antibiotic resistance, that are variably present across the species. The frequencies of individual accessory genes are suggested to be in an equilibrium that is key in determining population structure [10–12].

The pneumococcal population structure, specifically the lineages observed and the extent to which they are established is known to vary between geographic locations. This is a constraint on local population restructuring after the introduction of pneumococcal conjugate vaccines (PCVs) [3, 11–13]. Comparisons of the presence/absence and prevalence of lineages between countries is also facilitated by the international definition of GPSCs as opposed to dataset-specific designations. Identifying features of the population structure unique to a geographical location and then contextualizing them in the relevant GPSC gives an international perspective. It can give insight into the acquisition and geographical spread of a resistance determinant, the introduction of a lineage into one country from another, and serotype switch or loss of capsule observed in one location.

Dating events in the pneumococcal population structure such as the migration of a lineage or the acquisition of a clinically relevant feature is often useful. After the introduction of PCVs, some lineages were observed to shift from a vaccine serotypes to non-vaccine serotypes. Identifying the recombination events that result in serotype switch and dating it has often revealed that the recombination event and serotype variant was established before vaccine implementation and was subsequently selected [3, 14, 15].

Here we provide population snapshots of 12 countries representing four continents in Microreact [2] annotated with the GPSCs and clinically relevant metadata. We additionally provide the lineage snapshots and recombination and gene presence/absence profiles for 73 common GPSCs interactively in Phandango [1]. We also provide dated GPSC phylogenies to allow the date of genomic events or geographical introductions to be determined. We present examples of how these resources can be used to answer topical pneumococcal questions, with preliminary interpretation, and demonstrate how external data can be assigned to GPSC using Pathogenwatch to aid comparisons between datasets.

METHODS

DNA extraction and sequencing were described previously [3]. Briefly 13454 pneumococcal isolates drawn from pneumococcal disease surveillance programs and/or carriage studies in 30 countries were sequenced on an Illumina HiSeq platform and assembled and annotated as part of the Global Pneumococcal Sequencing project (GPS) [3, 16]. Isolates were clustered into lineages named Global Pneumococcal Sequence Clusters (GPSCs) using assemblies as input for PopPUNK [4]. Serotype, sequence type (ST) and antibiotic susceptibility were previously derived from the genomes [3]. The 13454 assemblies annotated using Prokka were used as input for Roary with default minimum 95% percentage identity for BLASTP, with and without splitting paralogues, with the former presented in Phandango [17, 18].

International GPSC snapshots

Lineage analyses were performed on 73 common GPSCs, representing 782 down to 22 isolates (Supplementary Material). Illumina reads from each isolate were mapped against previously prepared references for each lineage [3, 19]. In brief, where a public closed reference did not exist a high-quality illumina draft assembly was reordered against a complete *S. pneumoniae* genome ATCC 700669 [3], the reference genome accession numbers are provided in the Supplementary Material. The resulting alignment was used as input for Gubbins to identify recombination blocks and create recombination free phylogenies with RAxML [9, 20].

Dating

Taxa dates were recorded in years and the month converted to decimal, in the absence of month data the mid-year value of 0.5 was added. We assessed any temporal signal using the recombination-free phylogenies. BactDating was used to date 73 common GPSCs using the output of Gubbins [21]. We ran the BactDating R package with three replicates and one with randomized tip date through MCMC chains of 100000000 generations sampled every 100000 states with a 10000000 burn-in using the mixed gamma model [21]. The three replicate MCMC chains were deemed to have converged with Gelman diagnostic of approximately 1 for mu, sigma and alpha using the coda R package [22]. We then assessed temporal signal by comparing the first replicate model to a model ran under the same parameters but with randomization of the isolate dates with the modelcompare function of the BactDating package [21]. Finally, we assessed whether the effective sample size (ESS) on the first replicate model

Table 1. Population snapshots

Country	No. of isolates	Percentage IPD	Sampling years
South Africa	4615	63%	1991, 2005–2014
The Gambia	1647	24%	1993,1996–2014
USA	1584	100%	1998–2009
Malawi	1304	43%	1997–2015
Israel	1143	100%	2005–2014
Peru	607	31%	2006–2011
China	504	42%	1995–2001, 2009–2017
Brazil	420	97%	2008–2009, 2012–2013
Nepal	416	16%	2005–2009, 2011–2014
Poland	189	100%	2007–2013
Mozambique	167	100%	2008–2010
India	114	97%	2007–2010, 2013–2016

IPD, Invasive Pneumococcal Disease.

was greater than 200 using the effectiveSize function of the coda R package [22].

As a comparison to BactDating, we additionally ran Bayesian evolutionary analysis software BEAST v2.4.1 on the four sub-clades of GPSC3 [23]. We ran three replicate MCMC chains of 100000000 generations, with a 10000000 burn-in, that were sampled every 1000 states with the discrete gamma model of heterogeneity among sites, the relaxed clock model of nucleotide substitution with the Bayesian skyline tree prior. ESS were greater than 200.

Population snapshots

For the country-based, species-wide analysis, fastq reads from each isolate were mapped using BWA against a complete *S. pneumoniae* genome ATCC 700669 (NCBI accession code FM211187) [19]. The pseudo-genome alignment was then reduced to variant sites using SNP-sites for phylogenetic tree construction using FastTree2 [24, 25]. The SNPs were then reconstructed on the tree [26, 27]. Country-based analysis was only performed on the countries with isolates >100 (n=12, Table 1).

External datasets

We included an externally sequenced dataset that was not included in the PopPUNK definition of GPSCs to demonstrate that GPSCs can be assigned to any pneumococcal genome. Seventeen serotype 1 isolates collected from children with invasive disease from Utah, USA between 1996 and 2011 were included. Among them, metadata was available for 14 isolates, of which 13 were associated with complicated pneumonia with empyema. They were whole-genome sequenced on an Illumina HiSeq 2500 platform with 125 bp paired-end reads at the Huntsman Cancer Institute at the University of Utah. The data was imported and assembled as previously described [16]. Assemblies were submitted to Pathogenwatch, which assigns pneumococcal serotypes using SeroBA, and GPSCs using PopPUNK and the GPS reference database [3, 4, 28, 29]. Raw data and assemblies were deposited in the ENA under the study accession PRJEB34550. The isolates from Utah were also mapped to the GPSC31 reference (GenBank accession GCA_901234765) and combined with GPS GPSC31 strains to produce an alignment. Recombiantion was masked with Gubbins to build a phylogenetic representation of the external data and GPSC31 [9].

RESULTS AND INTERPRETATION

Assessing temporal signal in pneumococcal GPSCs

The BactDating models of temporal phylogenetic signal that had converged, were significantly better than the randomized dates model, and had effective population sizes of greater than 200, represented 70% (51/73) of the GPSCs analysed. For the remaining GPSCs, 6/22 did not converge, 11/22 were no better than the randomized dates model, and for 5/22 the effective population sizes were <200 (Supplementary Material Fig. S1). Previously we reported recombination/mutation ratios for the GPSCs and calculated recombination-free pairwise SNP distances for the top 30 GPSCs [3]. We did not observe any extreme values for these metrics in the GPSCs we were unable to date, except smaller sample sizes and a lack of temporal signal.

The estimated dates across the lineages were somewhat consistent, with the average most recent common ancestor (MRCA) for the 51 GPSCs being 1814 [1640–1897] (Fig. 1). Large diverse lineages with strong sub-structure become prohibitively computationally intensive for dating using BEAST, comparing the most recent common ancestor for the four sub-clades CC53, CC62, CC100, CC1012 of GPSC3 from BEAST and BactDating revealed similar estimates (Table 2).

In general the GPSC dates are older but not inconsistent with previous reports of pneumococcal clones, which often represent sub-clades of GPSCs and where the definition of clone, sampling and geographical representation may not have been as extensive [30–32]. PMEN1 a sub-clade of GPSC16, was previously estimated to have shared a common ancestor in 1969 [1958–1977], the BactDating estimate of the MRCA of PMEN1 (ST81) isolates was identical 1969 [1958–1977] [14]. PMEN2 is a sub-clade of GPSC23 and was previously estimated to be of Western European origin between 1962 and 1974, the BactDating estimate of the MRCA of the PMEN2 (ST90) isolates of 1977 [1968–1983] overlaps with that estimate [31]. PMEN14 a sub-clade of GPSC1 was previously estimated to have shared a common ancestor in 1987 [1981–1991], the BactDating estimate of the MRCA of PMEN14 (ST236) isolates was much older and less certain 1885 [1807–1929]. Even when excluding the two most basal isolates of ST236 the estimate of 1949 [1904–1968] does not overlap with the previous estimate [6].

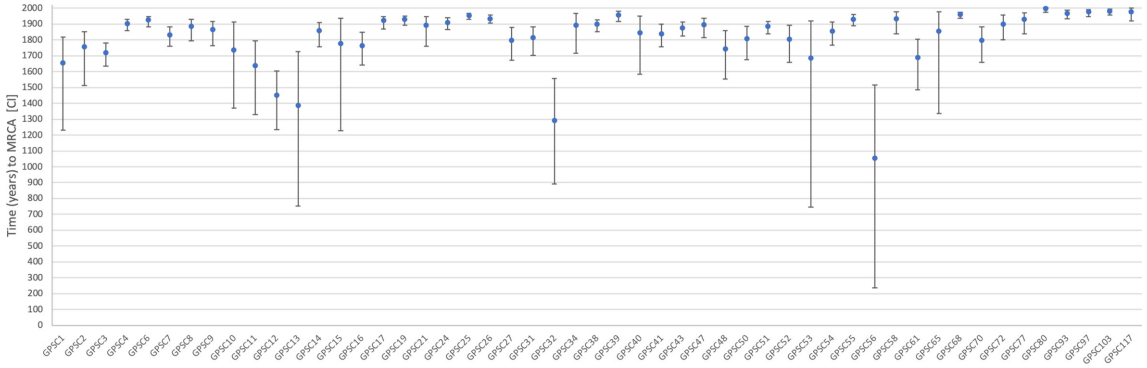

Fig. 1. Point estimates of the year of the most recent common ancestor (MRCA) of each of the 51 GPSCs that could be reliably estimated. The 95% confidence intervals (CI) are plotted.

The conserved range of dates across GPSCs could result from two processes: (1) clones emerge and die at a roughly constant rate, meaning there is an underlying exponential distribution of ages; (2) a clone needs to be old enough to have to be measurably evolving and established enough to be well sampled, preventing younger clones from being dated. Many of these GPSCs represent globally spread clones, the period of human history from 1600 to 1800 is known as the proto-globalisation era, with large scale globalization beginning in the 1820s [33, 34]. This would have included the migration of families with children who may have been colonized with the pneumococcus, providing a new opportunity for global spread of local strains.

Defining the pangenome

In 13454 genomes, only 634 and 868 genes met the core gene definition of present in >=99% or >=95% of the collection respectively when paralogues were spilt. This is likely an underestimate as mis-assemblies, contig breaks and missing annotations can erode the number of genes that meet these criteria. The resultant core gene alignment was 544759 bp long with 147832 variable sites. A further 1957 genes were classified as shell genes (>=15 to <95%) and 24219 as cloud genes (>=1 to <15%). The average number of genes defined as core (>=95%) within lineages (GPSCs) was 1276. The mean core gene number per GPSC was higher than in the total collection in part because genes that are accessory to the species can be common to all isolates of a lineage, and because fewer genes will have been misclassified as non-core due to an accumulation of assembly and annotation errors.

Example 1a: Exploring country phylogenetic snapshots using Microreact

A phylogenetic snapshot of the population diversity in a single country can be clearly visualized interactively in Microreact. The interface is user-friendly and easy to filter on particular features (e.g. serotype), manifestation (carriage or disease), and/or demographic data (e.g. age). The South African population snapshot represents two cross-sectional colonization studies and national IPD surveillance surrounding PCV introductions. Such a dataset allows exploration of the population structure to contextualize numerous scenarios, for example it clearly highlights the diverse genetic background of serotype 14 isolates (Fig. 2a) [35]. The 278 South African isolates expressing serotype 14 in this collection were found within nine unrelated GPSCs – they do not share a common ancestor. Overall, 22 % (61/278) of serotype 14 isolates from South Africa were found in GPSC9 (61/67). They exclusively belonged to clonal complex (CC)63, and 38/61 (63%) were ST63. However, there appeared to be phylogenetic structure within GPSC9/CC63/ST63 (Fig. 2a, b).

Table 2. Comparison of key feasible BEAST runs and BactDating models of phylogenetic dating

Lineage sub-clade	Year range	*N*	BEAST tMRCA	BactDating tMRCA
GPSC3	18	358	Run infeasible	1720.54 [1634.01–1780.92]
GPSC3-CC53	12	152	1879 [1820.49–1921.87]	1926.77 [1907.99–1941.56]
GPSC3-CC62	17	56	1912.82 [1667.10–1967.00]	1929.64 [1908.24–1944.03]
GPSC3-CC100	17	67	1937.13 [1920.19–1950.82]	1919.82 [1899.11–1936.45]
GPSC3-CC1012	17	83	1878 [1808.91–1925.02]	1840.70 [1802.97–1870.36]

CC, Clonal Complex; GPSC, Global Pneumococcal Sequence Cluster; tMRCA, Time to Most Recent Common Ancestor.

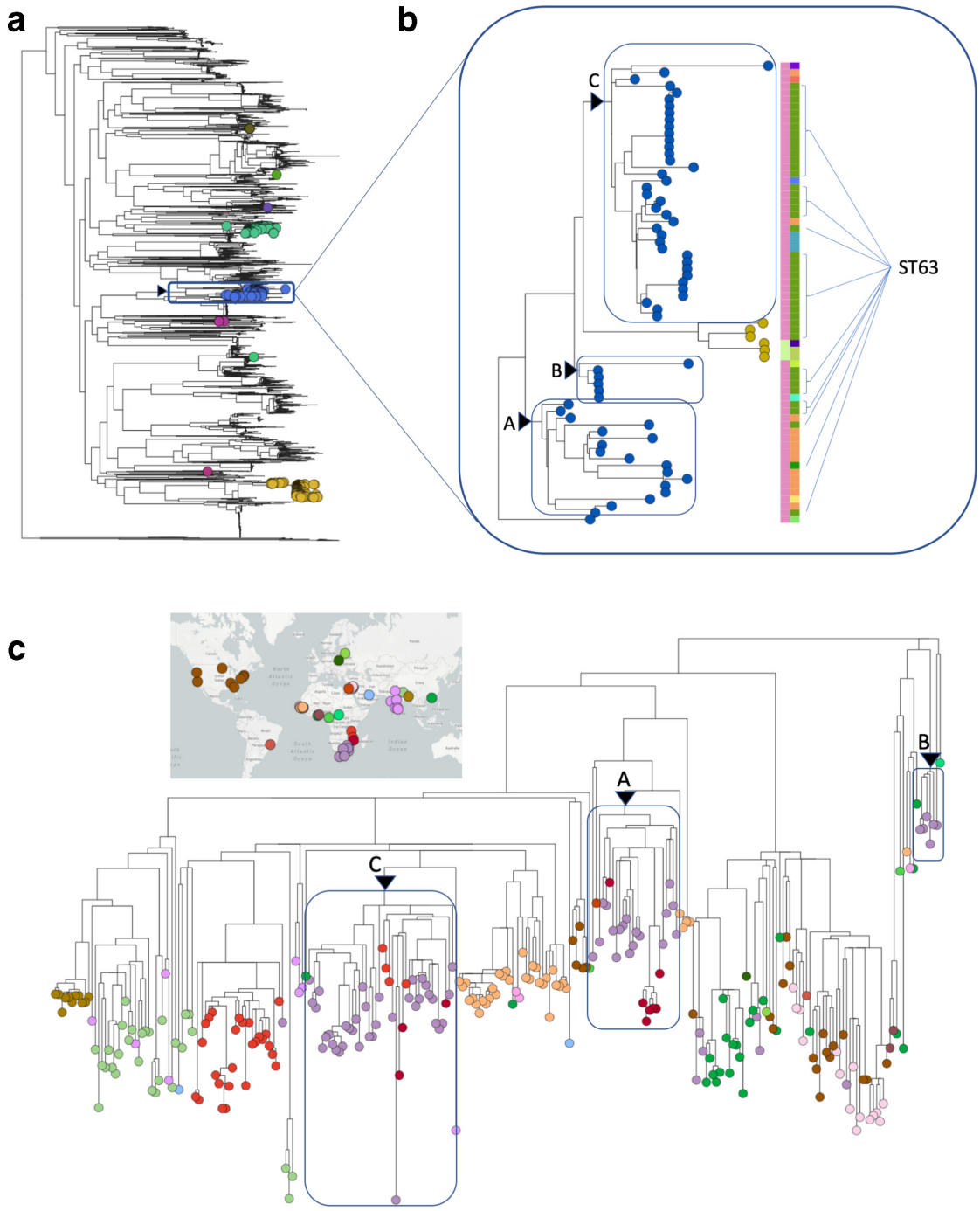

Fig. 2. Contextualizing serotype 14 genotypes in South Africa. (a) Phylogeny of South African pneumococcal population structure, with taxa expressing serotype 14 highlighted by a coloured circle representing their GPSC assignment, GPSC9 (blue) is highlighted with a box. (b) Expanded South African GPSC9 subtree where taxa are coloured by serotype: 14 (blue) and 15A (yellow). The left metadata block is coloured by clonal complex: CC63 (pink), CC12576 (green). The right-hand metadata block is coloured by sequence type: ST63 (green), ST2414 (orange). Three sub-clades (A,B,C) that each contain at least one ST63 isolates are highlighted with a box. (c) The international GPSC9 collection have taxa coloured by country of isolation with a map based key, the three South African (purple) sub-clades (A,B,C) are highlighted in the international GPSC9 collection with a labelled box.

Example 1b. Placing observations in the international context of a GPSC using Microreact

The context of the international GPSC9 phylogeny revealed three sub-clades in South Africa's sample of GPSC9 serotype 14 ST63. Each sub-clade shared a common ancestor with isolates from other geographical regions before they share

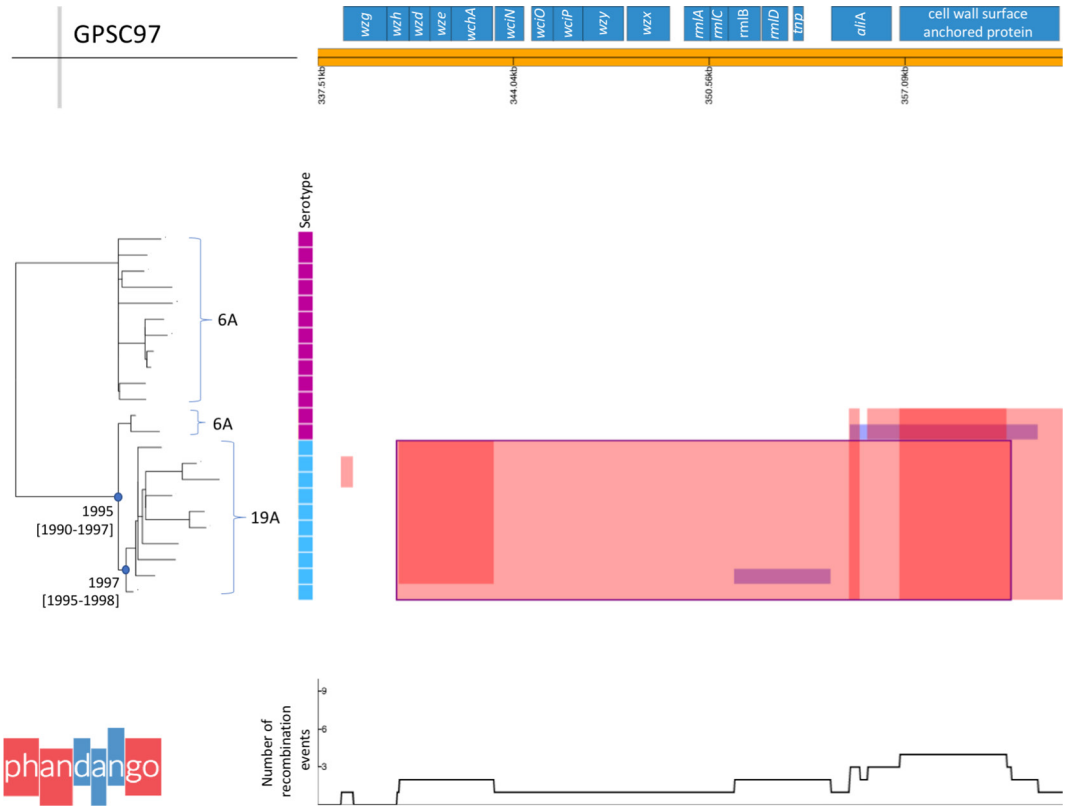

Fig. 3. GPSC97 capsular polysaccharide locus recombination events. Phandango plot of recombination detected with Gubbins, focused on the cps locus, across the GPSC97 phylogeny. Isoaltes are annotated with their serotype in a metablock: 6A (pink) 19A (blue). Recombination blocks span the taxa in which they are detected and the region of genes affected in the reference. Red blocks affect *n*>1, blue blocks affect *n*=1. Overlapping blocks increase the density of the colour. A sliding window of the number of recombination events affecting any one position in the reference is plotted at the underneath. The major recombination block spanning most of the cps locus and common to all 19A isolates is consistent with the serotype switch, and is outlined in blue. The nodes for most recent common ancestors of the 6A and 19A isolates and the 19A isolates are designated with a blue circle, estimated dates and confidence intervals are given.

a common ancestor, suggesting they are not of shared South African descent (Fig. 2c). Using the dated GPSC9 phylogeny we can infer that the South African isolates represent three independently successful sub-clades established around the 1980s (A 1985[1976–1992], B 1995[1987–1999], C 1985[1978–1990]) and shared a common ancestor in the first half of the twentieth century 1933 [1909–1951].

Example 2: Relating recombination events to changes in phenotype using Phandango

Recombination is known to facilitate clinically relevant changes in phenotype such as serotype switch. GPSC97 (CC1339/CC376), only observed in the USA in this collection, expressed either serotype 6A (13/23, 57%) or 19A (10/23, 43%). ST1339 was first reported expressing 6A and 19A in pubMLST in 2002 and 2006, respectively, in the USA. The earliest isolation date of 6A and 19A within GPSC97 in this collection was 1998 and 1999, respectively. There was phylogenetic structure to the serotypes expressed, and ancestral reconstruction of serotype suggests the clone originally expressed 6A. Visual inspection of recombination blocks in GPSC97 across the genome allowed the identification of a 20495 bp recombination event spanning capsular polysaccharide locus (*cps*) genes *wzh* to *aliA*, which coincided with the change in serotype (Fig. 3). There was evidence of four other shorter recombination events affecting 1–12 isolates in the *cps* locus, which did not result in a change in serotype (Fig. 3). The date for the most recent common ancestor (MRCA) of the 19A strains was estimated to be 1997 [1995–1998], the MRCA for the 19A strains and the 6A strains was estimated to be 1995 [1990–1997] therefore the recombination event likely occurred between 1990 and 1998. This recombination event did not extend out to *pbp2X* and therefore had no affect on penicillin susceptibility. It was previously determined that all of GPSC97 had an identical *pbp2X* allele, and a predicted MIC of >=2 $\mu g\ l^{-1}$ [3].

Example 3: Detecting antibiotic gene acquisition associated with mobile genetic elements using Phandango

Recombination and integration of mobile genetic elements (MGEs) can result in the acquisition of multiple accessory

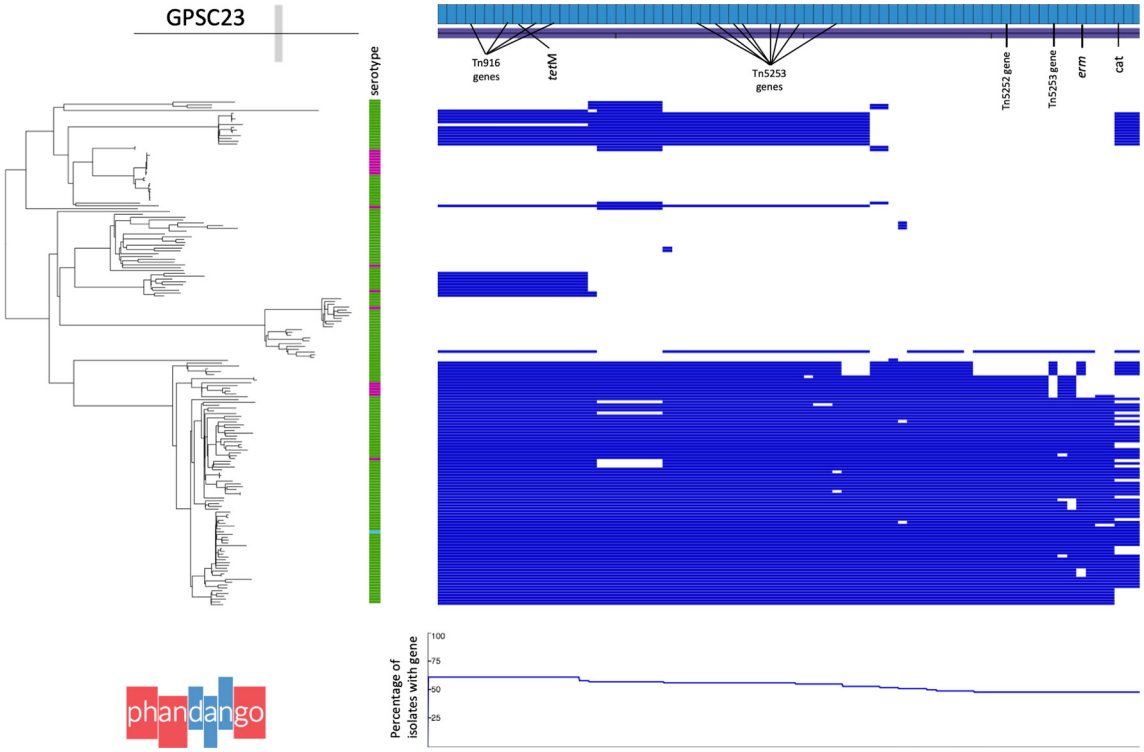

Fig. 4. GPSC23 acquired resistance gene presence and absence. Phandango plot of Roary gene presence and absence, focused on the genes with prevalence and phylogenetic patterns similar to acquired resistance genes: *tetM* 61%(110/180) *ermB* 43%(77/180) and *cat* 40%(72/180), across the GPSC23 phylogeny. Isolates are annotated with their serotype in a metablock: 6B (green), 6A (pink) and 19A (blue). Genes are shown as light blue bricks along the top and are sorted left to right by the proportion of isolates they are observed in. Presence (blue) and absence (white) of genes is plotted with respect to each isolates phylogenetic placement. A graph of the proportion of isolates the gene is observed in is plotted at the underneath.

Table 3. Pathogenwatch output

	Summary
No. assemblies processed	17
No. analyses performed	102
Time taken	3 min
No. contigs	38–58
GC content	39.5–39.6%
Assembly length	2.04–2.14 Mb
Species	*Streptococcus pneumoniae* (100%, 17/17)
Serotype	1 (100%, 17/17)
ST	ST227 (59%, 10/17) ST306 (24%, 4/17) ST304 (12%, 2/17) ST4288 (6%, 1/17)
Strain	GPSC31 (100%, 17/17)
AMR determinants	None identified

AMR, Antimicrobial resistance; Mb, Megabases.

genes in a single event, for example multiple acquired resistance genes can be carried by a single MGE. Members of GPSC23 (CC273) were commonly (106/180, 59%) predicted to be multidrug resistant (>=3 classes). This lineage is globally disseminated; observed in 13 countries representing Africa, Asia, Europe, North and South America. GPSC23 encompasses multiple Pneumococcal Molecular Epidemiology Network (PMEN) multidrug-resistant (MDR) clones: PMEN2 Spain^{6B} (ST90), PMEN17 Maryland^{6B} (ST384) and PMEN22 Greece^{6B} (ST273). This provides further confirmation on the relatedness of PMEN2 and PMEN22 reported previously [31]. Altogether, 33 % (59/180) of GPSC23 have the acquired resistance genes *cat*, *ermB*/*mefA* and *tet*M conferring resistance to chloramphenicol, erythromycin and tetracycline, respectively. The phylogenetic structure to the presence and absence of these genes in GPSC23 is variable (Fig. 4). The 58/59 isolates that carry *cat*, *erm*B and *tet*M fall in a sub-clade of 87 isolates encompassing PMEN2 and PMEN22. There appeared to be stable maintenance of *tet*M *n*=87/87, two independent losses of *erm*B *n*=76/87 and multiple independent losses of the *cat* gene *n*=58/87. Genes that are present in a similar proportion of isolates and similar phylogenetic pattern to these acquired resistance genes include those annotated as Tn*916*, Tn*5253* and Tn*5252* indicative of a composite transposon with gene loss since its integration (Fig. 4). Such variation

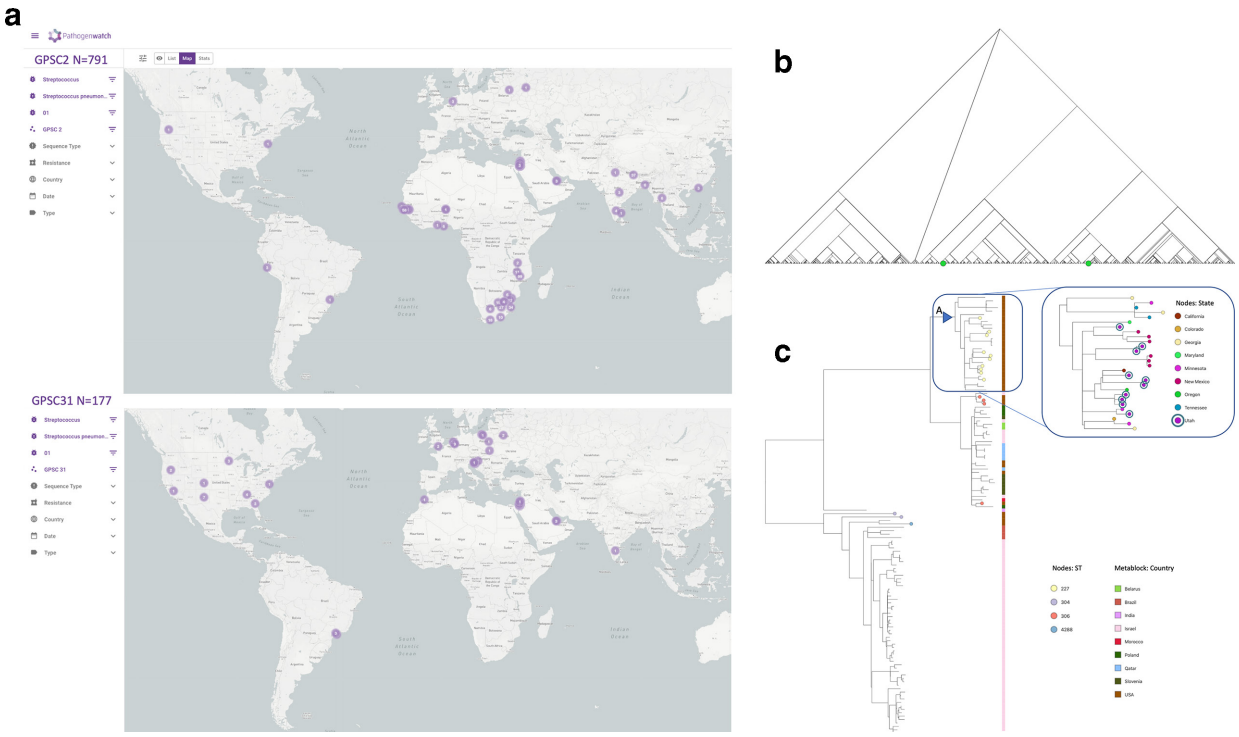

Fig. 5. Giving context to serotype 1 isolates from Utah. (a) Different geographical distribution of serotype 1 genomes belonging to GPSC2 and GPSC31 on Pathogenwatch. (b) Serotype 1 isolates in the GPS project (*n*=893/13,454) fall exclusively into two lineages (green): GPSC2 (*n*=782) and GPSC31 (*n*=111), which are in different parts of the species-wide tree and do not share a recent common ancestor. (c) Utah isolates are highlighted on a and coloured by their ST (see key), the metablock shows the country of isolation across the GPC31 tree structure, and USA state on the expanded subtree. Triangle A denotes the most recent common ancestor 1973 [1958-1981] of the USA sub-clade in which the majority (10/17) of the Utah serotype 1 isolates were found.

is often difficult to assemble with short read data, the gene presence absence plot suggests a composite Tn*916*-Tn*5253* element accounts for most of the acquired resistance genes in this lineage. Using the dated tree for GPSC23 we estimate the date of acquisition to be around 1966 [1953–1975]. This fits all with previous work reporting inactivation or loss of resistance genes from a composite Tn5253-type integrative and conjugative element in PMEN2 in a similar timeframe [31].

Example 4a: Contextualizing other datasets using the GPSCs and Pathogenwatch

Raw reads or assemblies can be dragged and dropped onto Pathogenwatch, which will QC the assemblies and check if they belong to the pneumococcus. Then it will assign their multi-locus ST, GPSC, serotype and report the detection of key genetic determinants of antimicrobial resistance (AMR). The assemblies from 17 serotype 1 isolates from Utah, which are external to the dataset used to define the GPSCs, were assigned by Pathogenwatch to GPSC31 (Table 3).

The assignment of GPSCs to these isolates allows them to be put in a global context. All 20027 pneumococcal genomes previously used to define the GPSCs are deposited in the Pathogenwatch [3]. Serotype 1 isolates in Pathogenwatch are only found in two lineages: GPSC2 and GPSC31, the database can be filtered on metadata, for example by strain (GPSC) and serotype, to view the geographical distribution, which differs between the two GPSCs (Fig. 5a). Differential geographical distributions of serotype 1 clones has been previously reported in the literature [36, 37]. Indeed within the GPS collection GPSC2 alone accounted for all the serotype 1 isolates from South Africa, Malawi and The Gambia, whilst GPSC31 was more common than GPSC2 in Israel and the USA, consistent with finding the Utah isolates only in GPSC31 [3].

These two lineages only express serotype 1 but do not share a recent ancestor in the species-wide tree (Fig. 5b). GPSC2 is composed of clonal complexes CC217, CC3581 and CC615, whilst GPSC31 includes CC5784, CC306 and ST227. ST227 was the predominant ST seen in children with pneumococcal empyema in Utah before conjugate vaccine introduction [38]. The relationship between these clones using genome-wide variation is consistent with previous designations of lineages A, B and C inferred from MLST where A represents GPSC31 and B and C are sub-clades of GPSC2 [36]. The tMRCA of the GPSC2 was estimated as 1655 [1230–1818] and for GPSC31 1814 [1702–1882], suggesting the serotype 1 CPS was transferred into a second lineage before the twentieth century.

The clones of serotype 1 have also been associated with different manifestations of disease, with lineage A (GPSC31)

associated with pneumonia with or without empyema in Europe and North America and lineage B/C (GPSC2) with bacteremia and meningitis in Africa [37, 39, 40]. In Utah serotype 1 disease has been associated with pleural empyema and the STs of GPSC31 [38]. Serotype 1, 100% ST 227, was identified in 50% of children with empyema in Utah in the decade before the licensing of pneumococcal conjugate vaccine [41]. Serotype 1 was identified in 33% of children with empyema after PCV7 introduction [38] but was relatively rare in other USA populations. Finally the absence of AMR determinants (Table 3) in the Utah isolates is typical of serotype 1, 72% of GPSC2 and 99% of GPSC31 were previously predicted to be pan-susceptible to the five antibiotics tested [3].

Example 4b: Contextualizing additional data within a GPSC

After mapping the Utah isolates to the GPSC31 reference (GenBank accession GCA_901234765) and masking recombination using Gubbins we observed that 10/17 of the Utah isolates fell into a GPSC31 sub-clade of 28 isolates that were only isolated in the USA (Fig. 5c). These ten Utah isolates were spread across this sub-clade but GPS USA isolates were basal, this sub-clade in the dated GPSC31 tree shared a MRCA in 1962 [1941–1975]. The remaining seven isolates clustered in two other regions of the tree, which had broader geographical representation (Fig. 5c).

DISCUSSION

Large-scale bacterial genomic data generation is increasingly common in the next-generation sequencing era. The availability of raw data, assemblies and annotation currently deposited in the European nucleotide archive (ENA) are an invaluable open data resource with huge potential for research beyond the original research scope, especially if paired with metadata. However, visualizing results in an interactive tool can make the data vastly more accessible than the availability of primary data files [1, 2]. Constructing a bioinformatic workflow involving genotyping, identifying recombination and robustly building and dating phylogenies, is a specialized activity requiring resources and training. Automated genotyping from raw reads or assemblies combined with databases of genomes vastly increases accessibility, and the utility of private sequencing data in the context of public data processed in a standardized manner. Furthermore, interactive phylogenies of published data combined with metadata allows those who have yet to harness such bioinformatic skills to explore the output of genomic analyses and gain experience of its uses and interpretation.

Nonetheless, for those with sufficient bioinformatic experience, the interactive element is far superior to static figures and is invaluable for exploring the data to test and generate hypotheses. The interactive data that we present here, does still require knowledge of how to interpret phylogenetic trees and the limitations of bioinformatic methods. The tool Roary allows the rapid determination of pan genomes in large datasets from annotated assemblies. It relies on user-defined percentage identity thresholds, unsupervised algorithms for defining a gene, and assembly errors can lead to underestimation of core genes and a likely overestimation of singleton genes [17]. Gubbins readily identifies differences in SNP densities across the genome between isolates of the same lineage as recombination [9]. It relies on robust definitions of a lineage such as GPSCs, but is limited in its detection of recombination that does not result in a sufficient difference in SNP density, such as recombination within highly similar strains [9]. It does allow the identification of SNP dense recombination events that alter important phenotypes as we demonstrate here for serotype. Phandango allows such recombination data to be displayed interactively across the genome making recombination hotspots readily visible. Phylogenetic reconstruction relies on the vertical accumulation of variation, and therefore the identification of recombination and masking from lineage alignments is important. Many models currently exist that infer phylogenies with different advantages and disadvantages but all rely on the quality of the alignment [42]. Phylogenetic dating also depends on robust masking of recombination, the subsequent phylogeny, as well as sufficient temporal sampling and the presence of temporal signal. Dated trees allow us to add a temporal context to other observations such as capsular switch events, geographical dissemination and gene acquisition.

Resource maintenance

Here we present phylogenetic analyses for 73 lineages and 12 countries, those that had reasonable sample numbers for such analyses. As the the GPS collection grows it will be possible to perform the analyses on further lineages and country population snapshots. Any resulting Microreact or phandango instances will be created and made accessible via the resources tab of the www.pneumogen.net/gps/. Adding further samples into existing phylogenies with ease and without repeating significant analyses is an area in which further bioinformatic development is needed. The definition of GPSCs will be updated, to include novel clusters as more sequence data is generated including those from external data if provided via www.pneumogen.net/gps/. Subsequent versions will be made available at www.pneumogen.net/gps/assigningGPSCs.html and incorporated into Pathogenwatch.

Microreact and Pathogenwatch are developed and maintained by the Centre for Genomic Surveillance (CGPS) at the Big Data Institute, University of Oxford and Wellcome Genome Campus, Hinxton, UK (http://pathogensurveillance.net). The underlying methods and software used to genotype rely on robust established schemes and databases, or bespoke pipelines and emerging software, with the latter two under continuous review and development to improve accuracy and reproducibly. CGPS have made comprehensive documentation available here: https://cgps.gitbook.io/pathogenwatch/. Those who wish to make use of Pathogenwatch to genotype their data (e.g. example 4a), or to visualize phylogenies representing their own data and metadata, with or without combining with GPS data in Microreact (e.g. example 4b) can do so privately by logging in. Any uploaded sequence data,

phylogeny or associated metadata is not made available to other users or collaborators, or used in the internal analyses of the centre for Genomic Surveillence, without explicit consent. The Phandango website processes data 'dragged and dropped' in the users web browser with no data leaving the user's computer.

We show that the GPS collection of interactive results, dated phylogenies and GPSC definitions give useful insights for common research questions in the study of pneumococcus. This applies to exploring the GPS collection or external data. The data however also have use beyond the pneumococcal community, as a teaching resource, or for mathematical modelling of features relevant to other pathogens, to understand the evolution of antimicrobial resistance, geo-temporal spread of clones and the phylogenetic histories of populations.

Funding information
This study was co-funded by the Bill and Melinda Gates Foundation (grant code OPP1034556), the Wellcome Sanger Institute (core Wellcome grants 098051 and 206194) and the U.S. Centers for Disease Control and Prevention. The funding sources had no role in isolate selection, analysis, or data interpretation. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. AJB was supported by the Primary Children's Hospital Foundation Edward B. Clark II Endowed Chair in Pediatrics at the University of Utah. NLP was supported by the NIH/NIAID Microbial Pathogenesis T32-AI055434. CLB was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR002538 and the HA and Edna Benning Presidential Endowment. The corresponding authors had full access to the data and are responsible for the final decision to submit for publication.

Acknowledgements
We would like to thank all members of the GPS consortium for their collaborative spirit and determination during the monumental task of sampling, extracting and sequencing this dataset, and all contributions to experimental design and input into this manuscript. We also would like to thank members of teams 284 and 81 at the Wellcome Sanger Institute (WSI) for their advice and critique and the pathogen informatics team at the WSI for the pipelines and expertise that made genomic analysis at this scale possible. GPS consortium members: Patrick E. Akpaka, Department of Paraclinical Sciences, The University of the West Indies, St. Augustine, Trinidad and Tobago; Krow Ampofo, Division of Pediatric Infectious Diseases, Department of Pediatrics, School of Medicine, University of Utah, 295 Chipeta Way, PO BOX 581289, Salt Lake City, UT, 84108, USA; Houria Belabbès, Ibn Rochd university-hospital center-Casablanca; Godfrey Bigogo, Centre for Global Health Research, Kenya Medical Research Institute, Kisumu, Kenya; Abdullah W Brooks, International Centre for Diarrheal Diseases Research, Dhaka, Bangladesh; Philip E. Carter, Institute of Environmental Science and Research Limited, Kenepuru Science Centre, Porirua, Zealand; Stuart C. Clarke, Faculty of Medicine and Institute of Life Sciences, University of Southampton, UK; Alejandra Corso, Instituto Nacional de Enfermedades Infecciosas, Argentina; Maria Cristina de Cunto Brandileone, Center of Bacteriology, Adolfo Lutz Institute, São Paulo, Brazil; Alexander Davydov, Belarusian State Medical University, Minsk, Belarus; The Republican Research and Practical Center for Epidemiology and Microbiology, Minsk, Belarus; Idrissa Diawara, Faculty of Sciences and Techniques of Health, Mohammed VI University of Health Sciences (UM6SS); Sanjay Doiphode, Hamad Medical Corporation, Doha, Qatar; Ekaterina Egorova, Gabrichevsky Epidemiology and Microbiology Research Institute, Moscow, Russia; Naima Elmdaghri, Laboratoire of Microbiology, Faculty of Medicine and Pharmacy and Ibn Rochd University Hospital Center, Casablanca, Morocco; Özgen Köseoglu Eser, Hacettepe University Faculty of Medicine, Department of Medical Microbiology, 06100, Ankara, Turkey; Diego Faccone, Instituto Nacional de Enfermedades Infecciosas, Argentina; Rebecca Ford, Papua New Guinea Institute of Medical Research, PO Box 60, Goroka, 441, Eastern Highlands Province, Papua New Guinea; Paula Gagetti, Instituto Nacional de Enfermedades Infecciosas, Argentina; Noga Givon-Lavi, The Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel; Md Hasanuzzaman, Child Health Research Foundation, Department of Microbiology, Dhaka Shishu Hospital, Dhaka 1207, Bangladesh; Kristina G. Hulten, Department of Pediatrics, Baylor College of Medicine, Houston TX; Margaret Ip, Department of Microbiology, Chinese University of Hong Kong; Aurelie Kapusta, Department of Human Genetics, University of Utah, 15 North 2030 East, Salt Lake City, UT 84112; Rama Kandasamy, Oxford Vaccine Group, Department of Paediatrics, University of Oxford, and the NIHR Oxford Biomedical Research Centre, Oxford, UK; Tamara Kastrin, Department of Medical Microbiology, Institute of Public Health of the Republic of Slovenia, Grabloviceva 44, 1000 Ljubljana, Slovenia; Jeremy Keenan, Francis I. Proctor Foundation, University of California, San Francisco, San Francisco, California, United States of America; Pierra Y. Law, Department of Microbiology and Carol Yu Centre for Infection, The University of Hong Kong, Queen Mary Hospital, Hong Kong, China; Deborah Lehmann, Telethon Kids Institute, the University of Western Australia, Perth, WA; Jennifer Moïsi, Agence de Médecine Préventive, Paris, France; Helio Mucavele, Fundação Manhiça, Centro de Investigação em Saúde da Manhiça (CISM), Maputo, Moçambique; Michele Nurse-Lucas, Department of Paraclinical Sciences, The University of the West Indies, St. Augustine, Trinidad and Tobago; Stephen K Obaro, University of Nebraska Medical Center, Omaha, USA; Metka Paragi, National Laboratory of Health, Environment and Food, Centre for Medical Microbiology, Department for Public Health Microbiology, Grablovičeva 44, 1000, Ljubljana, Slovenia; Ewa Sadowy, Department of Molecular Microbiology, National Medicines Institute, 00–725 Warsaw, Poland; Samir K. Saha, Child Health Research Foundation, Dhaka, Bangladesh; Eric Sampane-Donkor, Department of Medical Microbiology, School of Biomedical and Allied Health Sciences University of Ghana, Accra, Ghana; Shamala Devi Sekaran, MAHSA University, Selangor, Malaysia; Sadia Shakoor, Department of Pathology and Laboratory Medicine and Department of Paediatrics and Child Health, The Aga Khan University, Karachi 74800, Pakistan; Shrijana Shrestha, Patan Academy of Health Sciences, Kathmandu, Nepal; Anna Skoczynska, National Reference Centre for Bacterial Meningitis, Department of Epidemiology and Clinical Microbiology, National Medicines Institute, Warsaw, Poland Kwan; Soo Ko, Department of Molecular Cell Biology, Samsung Biomedical Research Institute, Sungkyunkwan University School of Medicine, Suwon, South Korea; Somporn Srifuengfung, Department of Microbiology, Faculty of Medicine Siriraj Hopital, Mahidol University, Bangkok, Thailand; Peggy-Estelle Tientcheu, Vaccines and Immunity Theme, MRC Unit, The Gambia; Leonid Titov, The Republican Research and Practical Center for Epidemiology and Microbiology, Minsk, Belarus; Paul Turner, Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK; Yulia Urban, Gabrichevsky Epidemiology and Microbiology Research Institute, Moscow, Russia; Jennifer Verani, Respiratory Diseases Branch, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, United States of America; Elena Voropaeva, Gabrichevsky Epidemiology and Microbiology Research Institute, Moscow, Russia; Nicole Wolter, Centre for Respiratory Diseases and Meningitis, National Institute for Communicable Diseases, Johannesburg, South Africa.

Author contributions
Rebecca A. Gladstone: conceptualization, methodology, project administration, data curation, investigation, formal analysis, visualization, writing – original draft preparation. Stephanie W. Lo: methodology, project administration, data curation, formal analysis, visualization, writing – review and editing. Richard Goater software: visualization, writing – review and editing. Corin Yeats software: visualization. Ben Taylor software: visualization. James Hadfield software: visualization. John A Lees: formal analysis. Nicholas J. Croucher: writing – review and editing. Andries J. van Tonder methodology: formal analysis, writing – review and editing. Leon J. Bentley: formal analysis, visualization. Anne J. Blaschke: resources, conceptualization, writing – review and editing. Nicole L Pershing: resources, conceptualization. Carrie L. Byington: conceptualization, writing – review and editing. Veeraraghavan Balaji: resources. Waleria Hryniewicz: resources. Betuel Sigauque: resources. K.L. Ravikumar: resources. Maria Cristina de Cunto Brandileone:

resources. Theresa J. Ochoa: resources. Pak Leung Ho: resources. Mignon du Plessis: resources. Kedibone M. Ndlangisa: resources. Jennifer E. Cornick: resources. Brenda Kwambana-Adams: resources. Rachel Benisty: resources. Susan A. Nzenze: resources. Shabir A. Madhi: resources. Paulina A. Hawkins: resources, project administration, data curation. Andrew J Pollard: resources, writing – review and editing. Dean B. Everett: conceptualization, resources. Martin Antonio: conceptualization, resources. Ron Dagan: resources. Keith P. Klugman: conceptualization, funding. Anne von Gottberg: conceptualization, resources. Lesley McGee: conceptualization, resources, funding, project administration. Robert F. Breiman: conceptualization. David M. Aanensen: conceptualization, software, visualization. Stephen D. Bentley: conceptualization, funding, project administration, writing – review and editing. The Global Pneumococcal Sequencing Consortium: resources, review and editing.

Conflicts of interest

Dr Gladstone reports PhD studentship from Pfizer, outside the submitted work; Dr Lees reports grants from Pfizer, outside the submitted work; Dr Madhi reports grants from BMGF, during the conduct of the study; grants and personal fees from BMGF, grants from Pfizer, grants from GSK, grants from Sanofi, grants from BIOVAC, outside the submitted work; Dr Dagan reports grants and personal fees from Pfizer, during the conduct of the study; grants and personal fees from MSD, personal fees from MeMed, outside the submitted work; Dr von Gottberg reports grants and other from Pfizer, during the conduct of the study; grants and other from Sanofi, outside the submitted work; Dr Bentley reports personal fees from Pfizer, personal fees from Merck, outside the submitted work.

Ethical statement

Isolates for this study were selected from retrospective bacterial collections in each country participating in GPS. Appropriate approvals for use of isolates was obtained from each institution contributing isolates. No tissue material or other biological material was obtained from humans. All information regarding these isolates was anonymized.

Data Bibliography

1. Gladstone RA, Lo SW, Lees JA, Croucher NJ, van Tonder AJ, *et al*. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine* 2019;43:338–346.

References

1. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM *et al*. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics* 2018;34:292–293.
2. Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J *et al*. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom* 2016;2:e000093.
3. Gladstone RA, Lo SW, Lees JA, Croucher NJ, van Tonder AJ *et al*. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine* 2019;43:338–346.
4. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW *et al*. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res* 2019;29:304–316.
5. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM *et al*. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet* 2013;45:656–663.
6. Croucher NJ, Chewapreecha C, Hanage WP, Harris SR, McGee L *et al*. Evidence for soft selective sweeps in the evolution of pneumococcal multidrug resistance and vaccine escape. *Genome Biol Evol* 2014;6:1589–1602.
7. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P *et al*. Dense genomic sampling identifies highways of pneumococcal recombination. *Nat Genet* 2014;46:305–309.
8. Mostowy RJ, Croucher NJ, De Maio N, Chewapreecha C, Salter SJ *et al*. Pneumococcal capsule synthesis locus CPS as evolutionary hotspot with potential to generate novel serotypes by recombination. *Mol Biol Evol* 2017;34:2537–2554.
9. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA *et al*. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015;43:e15.
10. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ *et al*. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol* 2010;11:R107.
11. Corander J, Fraser C, Gutmann MU, Arnold B, Hanage WP *et al*. Frequency-Dependent selection in vaccine-associated pneumococcal population dynamics. *Nat Ecol Evol* 2017;1:1950–1960.
12. Azarian T, Martinez PPP, Arnold BJ, Grant LR, Corander J *et al*. Prediction of post-vaccine population structure of Streptococcus pneumoniae using accessory gene frequencies. *bioRxiv* 2018;420315.
13. Gladstone RA, Devine V, Jones J, Cleary D, Jefferies JM *et al*. Pre-vaccine serotype composition within a lineage signposts its serotype replacement - a carriage study over 7 years following pneumococcal conjugate vaccine use in the UK. *Microb Genom* 2017;3:e000119.
14. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J *et al*. Rapid pneumococcal evolution in response to clinical interventions. *Science* 2011;331:430–434.
15. Golubchik T, Brueggemann AB, Street T, Spencer CCA, Spencer CCA *et al*. Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event. *Nat Genet* 2012;44:352–355.
16. Page AJ, De Silva N, Hunt M, Quail MA, Parkhill J *et al*. Robust high-throughput prokaryote *de novo* assembly and improvement pipeline for Illumina data. *Microb Genom* 2016;2:e000083.
17. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S *et al*. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3693.
18. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
19. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.
20. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 2006;22:2688–2690.
21. Didelot X, Croucher NJ, Bentley SD, Harris SR, Wilson DJ. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res* 2018;46:e134.
22. Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. *R News* 2006;6:7–11.
23. Drummond AJ, Rambaut A. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 2007;7:214.
24. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;5:e9490.
25. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T *et al*. *SNP-sites*: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* 2016;2:e000056.
26. Sankoff D. Minimal mutation trees of sequences. *SIAM J Appl Math* 1975;28:35–42.
27. Pupko T, Pe'er I, Shamir R, Graur D. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol Biol Evol* 2000;17:890–896.
28. Epping L, van Tonder AJ, Gladstone RA, Bentley SD *et al*, The Global Pneumococcal Sequencing Consortium. SeroBA: rapid high-throughput serotyping of Streptococcus pneumoniae from whole genome sequence data. *Microb Genom*;4.
29. Centre for Genomic Pathogen Surveillance. https://cgps.gitbook.io/pathogenwatch/ September 2019.
30. Wyres KL, Lambertsen LM, Croucher NJ, McGee L, von Gottberg A *et al*. The multidrug-resistant PMEN1 pneumococcus is a paradigm for genetic success. *Genome Biol* 2012;13:R103.

31. Croucher NJ, Hanage WP, Harris SR, McGee L, van der Linden M *et al*. Variable recombination dynamics during the emergence, transmission and 'disarming' of a multidrug-resistant pneumococcal clone. *BMC Biol* 2014;12:49.
32. Azarian T, Mitchell PK, Georgieva M, Thompson CM, Ghouila A *et al*. Global emergence and population dynamics of divergent serotype 3 CC180 pneumococci. *PLoS Pathog* 2018;14:e1007438.
33. O'Rourke KH, Williamson JG. When did globalisation begin? *Eur Rev Econ Hist* 2002;6:23–50.
34. Van Der Bly MCE. 'Proto-Globalization', in: Encyclopaedia of Global Studies, Sage. https://www.academia.edu/14444466/_Proto-Globalization_in_Encyclopaedia_of_Global_Studies_Sage_2012 2012.
35. Ndlangisa KM, du Plessis M, Wolter N, de Gouveia L, Klugman KP *et al*. Population snapshot of *Streptococcus pneumoniae* causing invasive disease in South Africa prior to introduction of pneumococcal conjugate vaccines. *PLoS One* 2014;9:e107666.
36. Brueggemann AB, Spratt BG. Geographic distribution and clonal diversity of *Streptococcus pneumoniae* serotype 1 isolates. *J Clin Microbiol* 2003;41:4966–4970.
37. Blumental S, Granger-Farbos A, Moïsi JC, Soullié B, Leroy P *et al*. Virulence factors of *Streptococcus pneumoniae*. Comparison between African and French invasive isolates and implication for future vaccines. *PLoS One* 2015;10:e0133885.
38. Byington CL, Hulten KG, Ampofo K, Sheng X, Pavia AT *et al*. Molecular epidemiology of pediatric pneumococcal empyema from 2001 to 2007 in Utah. *J Clin Microbiol* 2010;48:520–525.
39. Duarte C, Sanabria O, Moreno J. Molecular characterization of Streptococcus pneumoniae serotype 1 invasive isolates in Colombia. *Rev Panam Salud Publica* 2013;33:422–426.
40. Traore Y, Tameklo TA, Njanpop-Lafourcade B-M, Lourd M, Yaro S *et al*. Incidence, seasonality, age distribution, and mortality of pneumococcal meningitis in Burkina Faso and Togo. *Clin Infect Dis* 2009;48 Suppl 2:S181–S189.
41. Byington CL, Spencer LY, Johnson TA, Pavia AT, Allen D *et al*. An epidemiological investigation of a sustained high rate of pediatric parapneumonic empyema: risk factors and microbiological associations. *Clin Infect Dis* 2002;34:434–440.
42. Lees JA, Kendall M, Parkhill J, Colijn C, Bentley SD *et al*. Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. *Wellcome Open Res* 2018;3:33.