

Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing

Karel Břinda^{1,2*}, Alanna Callendrello¹, Kevin C. Ma³, Derek R. MacFadden^{1,4}, Themoula Charalampous⁵, Robyn S. Lee^{1,6}, Lauren Cowley⁷, Crista B. Wadsworth⁸, Yonatan H. Grad⁹, Gregory Kucherov^{9,10}, Justin O'Grady^{11,5}, Michael Baym¹² and William P. Hanage¹

Surveillance of drug-resistant bacteria is essential for healthcare providers to deliver effective empirical antibiotic therapy. However, traditional molecular epidemiology does not typically occur on a timescale that could affect patient treatment and outcomes. Here, we present a method called 'genomic neighbour typing' for inferring the phenotype of a bacterial sample by identifying its closest relatives in a database of genomes with metadata. We show that this technique can infer antibiotic susceptibility and resistance for both *Streptococcus pneumoniae* and *Neisseria gonorrhoeae*. We implemented this with rapid *k*-mer matching, which, when used on Oxford Nanopore MinION data, can run in real time. This resulted in the determination of resistance within 10 min (91% sensitivity and 100% specificity for *S. pneumoniae* and 81% sensitivity and 100% specificity for *N. gonorrhoeae* from isolates with a representative database) of starting sequencing, and within 4 h of sample collection (75% sensitivity and 100% specificity for *S. pneumoniae*) for clinical metagenomic sputum samples. This flexible approach has wide application for pathogen surveillance and may be used to greatly accelerate appropriate empirical antibiotic treatment.

Infections pose multiple challenges to healthcare systems and contribute to higher mortality, morbidity and escalating cost. Clinicians must regularly make rapid decisions on empirical antibiotic treatment of infectious syndromes without knowing the causative pathogen (or pathogens) or whether they are drug-susceptible or drug-resistant. In some cases, this is directly linked to poor outcomes; in the case of septic shock, the risk of death increases by an estimated 10% with every 60 min of delay in initiating effective treatment¹.

The molecular epidemiology of infectious disease allows us to identify high-risk pathogens and to determine their patterns of spread on the basis of their genetics or (increasingly) genomics. Conventionally, such studies, including outbreak investigations and characterization of previously untested resistant strains, have been conducted in retrospect, but this has been changing with the availability of increasingly inexpensive sequencing technologies^{2,3}. The wealth of data generated by genomics are promising, but introduces a challenge because while many features of a sequence are correlated with the phenotype of interest, few are causative.

Prescription, however, has long been informed by correlative features when causative ones are difficult to measure; for example, whether the same syndrome or pathogen occurring in other patients from the same clinical environment have responded to a particular antibiotic. This has also been observed at the genetic level as a result of genetic linkage between resistance elements and the rest of the genome. An example is given by the pneumococcus *S. pneumoniae*. The Centers for Disease Control and Prevention (CDC) has rated

the threat level of drug-resistant pneumococcus as "serious"⁴. While resistance arises in pneumococci through a variety of mechanisms, approximately 90% of the variance in the minimal inhibitory concentration (MIC) for antibiotics of different classes can be explained by the loci determining the strain type⁵, even though none of these loci themselves causes resistance. Thus, in the overwhelming majority of cases, resistance and susceptibility can be inferred from coarse strain typing based on the population structure. This population structure could be leveraged to offer an alternative approach to detecting resistance whereby rather than detecting high-risk genes, we identify high-risk strains. While many approaches have been developed to identify whether a pathogen carries mutations or genes known to confer resistance^{6–21} (see ref. ²² for a comprehensive review), this is not equivalent to the clinical question of whether the pathogen is susceptible.

We present a method called genomic neighbour typing that can bring molecular epidemiology closer to the bedside and provide information relevant to treatment at a much earlier stage. Our method takes sequences generated from a sample in real time and matches them to a database of genomes to identify the closest relatives. Because closely related isolates usually have similar properties, this yields an informed heuristic regarding the phenotype of the pathogen. We demonstrate this by identifying drug-resistant and drug-susceptible clones for both *S. pneumoniae* (the pneumococcus) and *N. gonorrhoeae* (the gonococcus) within minutes after the start of sequencing using Oxford Nanopore Technology (ONT). The method has many potential applications depending on

¹Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA. ²Department of Biomedical Informatics and Laboratory of Systems Pharmacology, Harvard Medical School, Boston, MA, USA. ³Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, MA, USA. ⁴Division of Infectious Diseases, Department of Medicine, University of Toronto, Toronto, Ontario, Canada. ⁵Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich, UK. ⁶Epidemiology Division, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada. ⁷Department of Biology and Biochemistry, University of Bath, Bath, UK.

⁸Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, NY, USA. ⁹CNRS/LIGM Université Paris-Est, Marne-la-Vallée, France. ¹⁰Skolkovo Institute of Science and Technology, Moscow, Russia. ¹¹Quadram Institute Bioscience, Norwich Research Park, Norwich, UK.

*e-mail: kbrinda@hsph.harvard.edu

the specific pathogen and the quality of the databases available for matching, which we discuss together with its limitations.

Results

Resistance is associated with clones in *S. pneumoniae* and *N. gonorrhoeae*. To quantify the association of clones with antibiotic resistance of the pathogens *S. pneumoniae* and *N. gonorrhoeae*, we constructed optimal predictors of resistance from bacterial lineages and measured the associated area under the receiver operating characteristic curve (AUC) (Supplementary Fig. 1). First, we applied the method to 616 pneumococcal genomes from a carriage study of children in Massachusetts, USA^{23,24}. Second, we used 1,102 clinical gonococcal isolates collected from 2000 to 2013 by the CDC's Gonococcal Isolate Surveillance Project²⁵. In both cases, the datasets comprised draft genome assemblies from Illumina HiSeq reads, resistance data and lineages inferred from sequence clusters computed using Bayesian analysis of population structure²⁶. Lineages of *S. pneumoniae* were predictive for benzylpenicillin, ceftriaxone, trimethoprim-sulfamethoxazole, erythromycin and tetracycline resistance, with AUC values ranging from 0.90 to 0.97 (Supplementary Fig. 2), which is consistent with previous works⁵. For *N. gonorrhoeae*, ciprofloxacin, ceftriaxone and cefixime attained comparably large AUC values (from 0.93 to 0.98), whereas azithromycin demonstrated a lower association (AUC value of 0.80) (Supplementary Fig. 3), which is as previously observed²⁵.

Rapid identification of nearest known relatives from sequencing reads. Based on the observed associations, we developed an approach that we term genomic neighbour typing to predict the phenotype from sequencing data. Genomic neighbour typing is a two-step algorithm that first compares a provided sample to a database of reference genomes with a known phylogeny and phenotype, and then predicts the probable phenotype of the sample based on the best hits (nearest neighbours) and their matching quality. We apply this here to the detection of drug resistance.

To implement genomic neighbour typing, we developed software called resistance-associated sequence elements (RASE) (Fig. 1). RASE takes a stream of nanopore reads and compares their k -mer content to references using a modified version of ProPhyle^{27,28}, which is a metagenomic classifier that implements a fast and memory-efficient exact coloured de Bruijn graph data structure²⁹ using a Burrows-Wheeler transform (BWT) index³⁰ (Methods). Using ProPhyle RASE identifies which references are the most similar to the read and increases their similarity weights (this approach was inspired by but differs from other similar approaches such as Kraken³¹ and Kallisto³²). These weights are cumulative scores that capture sample-to-reference similarity; they are set to zero at the beginning and are increased on the fly as sequencing proceeds according to the 'information content' of each read (Methods). Generally speaking, longer reads, such as those covering multiple accessory genes, tend to be specific and have high scores, whereas short reads or reads from the core genome are found in many lineages, tend to be nonspecific and have low scores. Weights serve as a proxy to the inverted genetic distance between the sample and the references.

Resistance or susceptibility is predicted in two steps based on the computed weights, the population structure and the reference phenotypes. First, RASE identifies the lineage of the best-matching reference genome and estimates the confidence of lineage assignment by comparing the two best-matching lineages to compute a 'lineage score' (Methods). Subsequently, RASE identifies the best match within that lineage and predicts resistance from the nearest resistant and susceptible neighbours. A comparison of their weights provides a 'susceptibility score', which quantifies the risk of resistance (Methods). When the weights are too similar, the confidence of the call is considered low; this happens when resistant and

susceptible strains are insufficiently genetically distinct, which is often the case for resistance that has recently emerged in evolutionary history (Methods). The ability to pinpoint the closest relatives in the database offers further resolution, even in the case where the resistance phenotype varies within a lineage.

Results of RASE are reported in real time as the best match in the database, together with susceptibility scores to the antibiotics being tested and a proportion of matching k -mers for quality control. As the run progresses, the scores fluctuate and eventually stabilize (an example is shown in Fig. 2).

RASE databases for hundreds of *S. pneumoniae* and *N. gonorrhoeae* strains. We constructed RASE databases for *S. pneumoniae* and *N. gonorrhoeae* from the same data as described above (Methods). We assigned each pneumococcal and gonococcal strains to antibiotic-specific resistance categories using the European Committee on Antimicrobial Susceptibility Testing (EUCAST) breakpoints³³ and the CDC Gonococcal Isolate Surveillance Project (GISP) breakpoints³⁴, respectively (Methods). Where MIC data were unavailable or insufficiently specific, we estimated the probable resistance phenotype using ancestral state reconstruction (Methods; Supplementary Note 1; Extended Data Figs. 3 and 4). To verify the results, we tested eight pneumococcal isolates for which resistance phenotypes were not originally available (Methods), and the measured MICs by microdilution matched the expected phenotypes (shown in bold in Table 1). We constructed the ProPhyle k -mer indexes using a k -mer length that was optimized to minimize prediction delays ($k=18$; Methods). The obtained pneumococcal and gonococcal RASE databases occupy 321 MB and 242 MB of RAM ($\times 4.3$ and $\times 12$, respectively, compression rate) and can be further compressed for transmission to 47 MB and 32 MB ($\times 29$ and $\times 90$, respectively, compression rate), respectively (Extended Data Fig. 1). This would allow RASE to be used on portable devices and its databases to be easily transmitted to the point of care over links with a limited bandwidth.

RASE identifies strains in the database within minutes. We first examined two pneumococcal isolates that were used to build the RASE database (100% sensitivity and 100% specificity, $n=10$; Table 1a) to test whether RASE can function in ideal circumstances. In the case of a fully susceptible isolate (SP01), the correct lineage and sequenced strain were identified within 1 min and 7 min, respectively. A multidrug-resistant isolate (SP02) was predicted even faster, with both lineage and the sequenced strain correctly detected and stabilized within 1 min. To compare our method to gene-based approaches for detecting resistance²², we evaluated how long it took for resistance genes to be sequenced on the device. We observed that at least 25 min would be needed for single copies to be detected (Supplementary Note 2; Extended Data Fig. 2).

We then performed a similar evaluation with five gonococcal isolates from the database (57% sensitivity and 100% specificity, $n=20$; Table 2a); however, here, we selected more complicated cases. First, we tested a susceptible isolate (GC01), for which RASE identified the correct strain and antibiogram within 3 min of sequencing. We then sequenced an isolate (GC02) with an uncommon mechanism of cephalosporin resistance that has recently emerged³⁵. Under such circumstances, the resistant strain and its susceptible neighbours tend to be genetically very similar, which could confound our analysis. However, RASE was still able to identify the correct resistance phenotypes in 9 min, with the delay due to the difficulty in distinguishing between the close relatives, which was reflected by a susceptibility score in the low-confidence range (Methods). This was repeated in further experiments with the same isolate (GC03), which consistently reported low confidence in the resistance phenotype (Methods). This feature of our approach intends to alert the operators to indicate that further testing is necessary.

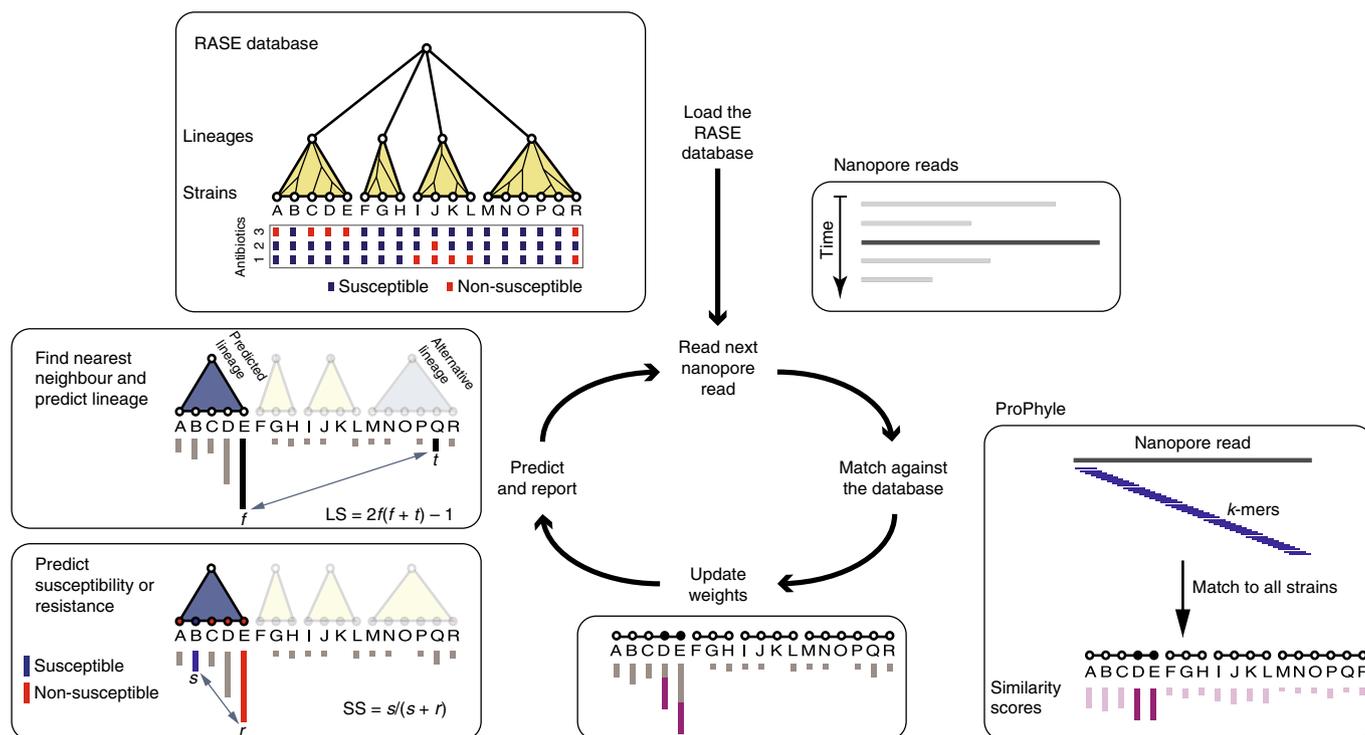


Fig. 1 | Overview of the RASE approach. In the first loading step, the precomputed RASE database is loaded into memory. As reads are generated, they are matched against the database using ProPhyle to calculate similarity to individual strains. The weights for the most similar strains (D and E in the figure) are increased proportionately to the number of matching *k*-mers. Finally, resistance is predicted from the obtained weights and from the resistance profiles of the database strains in the following manner. First, the best lineage is identified as the lineage of the best match (having the highest weight, E in the figure) and its score is calculated (lineage score (LS)). Second, for every antibiotic, a score quantifying the chance of susceptibility (susceptibility score (SS)) is calculated based on the most similar susceptible and resistant strains inside the identified lineage (B and E in the figure, respectively). The susceptibility or resistance to each of the antibiotics is predicted from their susceptibility scores by a comparison with a threshold (0.5 in the default setting) and reported together with the lineage, the best matching strain and the known properties of that strain (for example, the original antibiograms and the MLST-identified sequence type or serotype).

In this experiment, RASE also resolved sample mislabelling (Supplementary Note 3). For a multidrug-resistant isolate (GC04), RASE predictions stabilized within 2 min, but incorrectly predicted susceptibility to ceftriaxone. A subsequent analysis revealed that the ceftriaxone MIC of the sample was equal to the CDC GISP breakpoint ($0.125 \mu\text{g ml}^{-1}$), whereas the best match in the database had an MIC of $0.062 \mu\text{g ml}^{-1}$, within a single doubling dilution. We further found that RASE performed well even with extremely poor data and low-quality reads (for example, GC05; Supplementary Note 4). We also evaluated how genomic neighbour typing would perform if RASE used Kraken³¹ instead of ProPhyle^{27,28}, and results are presented in Supplementary Note 5.

RASE identifies the closest relative of previously untested isolates. We next examined four pneumococcal isolates (89% sensitivity and 100% specificity, $n = 20$; Table 1b) for which the serotype and limited antibiogram and lineage data were known. We compared the following three characteristics of the sample to assess our performance: the serotype, the sequence type as determined by multilocus sequence typing (MLST) and the antibiograms (benzylpenicillin, ceftriaxone, trimethoprim–sulfamethoxazole, erythromycin and tetracycline resistance according to the EUCAST breakpoints³³).

In all cases, the closest relative was identified within 5 min, even if the correct MLST sequence type was absent from the RASE database (an example is shown in Fig. 2). The two samples from the 23F clone (SP03 and SP06) were correctly called as being closely related to the Tennessee 23F-4 clone identified by the Pneumococcal

Molecular Epidemiology Network, a clone strongly associated with macrolide resistance³⁶. Consistent with this, the two samples were indeed resistant to erythromycin. However, the Tennessee 23F-4 clone was absent from the Massachusetts sample, with the best match being a comparatively distantly related strain that was resistant to penicillin, but susceptible to erythromycin. This illustrates the importance of a relevant database.

We evaluated RASE with 14 clinical gonococcal isolates from the RaDAR-Go project³⁷ (Switzerland, 2015–2016) (93% sensitivity and 100% specificity, $n = 56$; Table 2b). These isolates were previously sequenced using nanopore and have full antibiograms available³⁸. The 55 out of 56 correct calls indicate the strength of the genomic neighbour typing in a clinical setting. The only incorrect call (susceptibility to azithromycin for GC15) was marked as being low confidence on the basis of a poor susceptibility score. It should be noted that the ranges for what is considered low confidence could vary among settings and pathogens, but can be empirically determined and modified by users. In this case, our results suggest that informative results can be obtained even when using a database from one region (the United States) to predict phenotypes in another region (Europe). However, this may not be the case for all pathogens.

Inference is still informative but lower quality on highly divergent lineages. As noted above, an important precondition of genomic neighbour typing is a comprehensive and relevant reference database. To evaluate the performance of RASE in a setting with an incomplete database, we used the gonococcal WHO

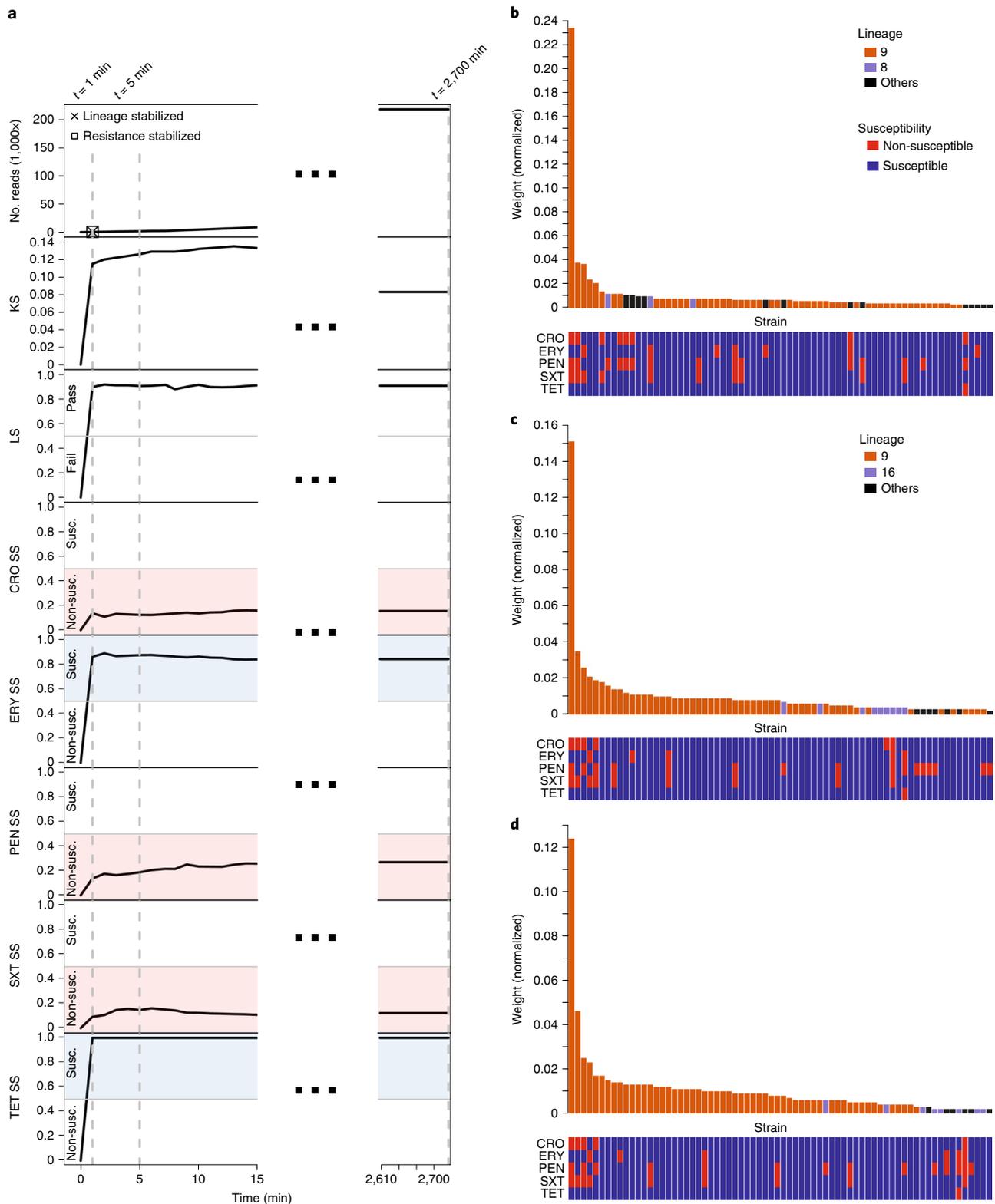


Fig. 2 | RASE obtains stable predictions of antibiotic resistance or susceptibility and lineage within minutes for an isolate of a pneumococcal 23F clone (SP06). **a**, The number of reads, LS, KS and SS for individual antibiotics as a function of time from the start of sequencing. In the top left plot, the times of stabilization are shown for the predicted lineage and susceptibility or resistance to all antibiotics (ceftriaxone (CRO), erythromycin (ERY), benzylpenicillin (PEN), trimethoprim-sulfamethoxazole (SXT) and tetracycline (TET)). Blue and red colours correspond to susceptibility (Susc.) and non-susceptibility (Non-susc.) calls, respectively. The dashed lines mark selected time points (1 min, 5 min and the end of sequencing (2,700 min)). **b–d**, Similarity rank plots for selected time points: 1 min (**b**), 5 min (**c**) and the end of sequencing (2,700 min; **d**). The bars correspond to the 70 best-matching strains in the database and display the normalized weights, which serve as a proxy to the inverted genetic distance. They are arranged by rank and coloured according to the presence in the predicted, alternative or another lineage. The panels underneath each chart display the resistance profiles of the strains.

Table 1 | Predicted phenotypes of *S. pneumoniae* for database isolates (a), non-database isolates (b) and metagenomes (c)

(a) Database isolates																
Sample	Lineage confidently detected	Matched <i>k</i> -mers (%)	Serotype		Antibiogram CRO		Antibiogram ERY		Antibiogram PEN		Antibiogram SXT		Antibiogram TET		MLST match	CC match
			Actual	Best match	Actual	Best match	Actual	Best match	Actual	Best match	Actual	Best match	Actual	Best match		
SP01	Yes	16	11D	11D ^a	S	S ^a	S	S ^a	S	S ^a	S	S ^a	S ⁽¹⁾	S ^{(1)a}	Yes ^a	Yes ^a
SP02	Yes	9.6	19A	19A ^a	R	R ^a	R	R ^a	R	R ^a	R	R ^a	R ⁽²⁾	R ^{(2)a}	Yes ^a	Yes ^a
(b) Non-database isolates																
Sample	Lineage confidently detected	Matched <i>k</i> -mers (%)	Serotype		Antibiogram CRO		Antibiogram ERY		Antibiogram PEN		Antibiogram SXT		Antibiogram TET		MLST match	CC match
			Actual	Best match	Actual	Best match	Actual	Best match	Actual	Best match	Actual	Best match	Actual	Best match		
SP03	Yes	3.1	23F	23F ^a	R	R ^a	R	S ^{(3)b}	R	R ^a	R	R ^a	S	S ^a	OoD ^c	Yes ^a
SP04	Yes	12	19A	19A ^a	R	R ^a	R	R ^a	R	R ^a	R	R ^a	R	R ^{(4)a}	OoD ^c	Yes ^a
SP05	No	1.8	19F	19F ^a	R	R ^a	R	R ^{a,d}	R	R ^a	R	R ^{a,d}	R	R ^{a,d}	OoD ^c	Yes ^a
SP06	Yes	8.3	23F	23F ^a	R	R ^a	R	S ^{(3)b}	R	R ^a	R	R ^a	S	S ^a	OoD ^c	Yes ^a
(c) Metagenomes																
Sample	Lineage confidently detected	SP (%)	Matched <i>k</i> -mers (%)	Antibiogram ERY		Antibiogram PEN		Antibiogram TET								
				Actual	Best match	Actual	Best match	Actual	Best match							
SP07	No	2.3	0.2	NA	S ^c	S	S ^a	R	S ^{(5)b}							
SP08	No	2.5	0.9	S	S ^a	S	S ^{a,d}	S	S ^{(6)a}							
SP09	No	4.0	1.2	NA	S ^c	S	S ^a	S	S ^{(7)a}							
SP10	Yes	21	5.2	R	R ^a	R	R ^a	R	R ^{(8)a}							
SP11	Yes	70	14	R	R ^a	R	R ^a	R	R ^{(8)a}							
SP12	Yes	86	17	S	S ^a	S	S ^a	R	S ^{(5)b}							

The table displays actual and predicted resistance phenotypes (where 'S' represents susceptible and 'R' represents non-susceptible) for individual experiments and information on the match of the predicted MLST-identified sequence type and the clonal complex (CC). Resistance categories in bold were inferred using ancestral reconstruction and were confirmed using phenotypic testing (see Methods and Supplementary Table 2). Metagenomic samples were sorted by the estimated fraction of *S. pneumoniae* reads (SP). ^aCorrect prediction. ^bIncorrect prediction. ^cCannot be evaluated. ^dLow confidence call. OoD, out of database; ⁽¹⁾, identity of a retested sample; NA, not available. For definitions of antibiotics see Fig. 2.

(World Health Organization) 2016 reference strain collection³⁹. This includes a global collection of 14 diverse isolates from Europe, Asia, North America and Australia, collected over two decades and exhibiting phenotypes ranging from pan-susceptibility to multi-drug resistance, and as such, the GISP database is expected to be non-representative in this study. The WHO strains are available from the National Collection of Type Cultures, and were previously sequenced using nanopore³⁸ and genetically and phenotypically characterized³⁹. Surprisingly, RASE correctly identified all MLST-identified sequence types represented in the database, and in seven cases it provided fully correct resistance phenotypes (67% sensitivity and 91% specificity, $n=56$; Supplementary Table 1). In six out of seven cases where the complete resistance profile was not recovered, the closest relatives were correctly identified, but were genetically divergent from the query isolates (Supplementary Note 6). In one case, the errors were due to a misidentification of the closest relatives by ProPhyle. Therefore, most prediction errors could be addressed with a more comprehensive database.

RASE can identify resistance in pneumococcus from sputum metagenomic samples. Because bacterial culture and phenotyping via agar-dilution, Etest or disk diffusion introduces significant delays in resistance profiling, direct metagenomic sequencing of clinical samples would be preferable for point-of-care use. We therefore analysed metagenomic nanopore data from sputum samples obtained from patients suffering from lower respiratory tract infections⁴⁰ (UK, 2017), selecting six samples from the study that were

already known to contain *S. pneumoniae* (75% sensitivity and 100% specificity, $n=16$; Table 1c).

One sample (SP10) contained DNA from multiple bacterial species. However, within 5 min, the sequence was identified to belong to the Swedish 15A-25 clone (ST63), which is also known to be associated with resistance phenotypes including macrolides and tetracyclines⁴¹. This sample was confirmed to be resistant to erythromycin and to clindamycin, tetracycline and oxacillin according to the EUCAST breakpoints³³. The original report of the Swedish 15A-25 clone did not report resistance to penicillin antibiotics⁴¹, which has subsequently emerged in this lineage. However, our database correctly identified the risk of penicillin resistance in this sample. The metagenomes SP11 and SP12 contained an estimated >20% reads that matched *S. pneumoniae*, and their serotypes were identified to be 15A and 153, respectively. The susceptibility scores of the best matches were fully consistent with the resistance profiles found in the samples, with the exception of tetracycline resistance in SP12 due to an incomplete database (Supplementary Note 7). The last remaining samples, SP07–SP09, contained <5% unambiguously pneumococcal reads. Despite the low proportions, all predicted phenotypes were concordant with phenotypic tests, with the exception of SP07, which matched the same strain as SP12 (discussed above).

Discussion

This paper presents a method, which we termed genomic neighbour typing, to pinpoint the closest relatives of a query genome within a suitable database and then to infer the phenotypic properties

Table 2 | Predicted phenotypes of *N. gonorrhoeae* for database isolates (a) and non-database isolates (b)

(a) Database isolates												
Sample	Lineage confidently detected	Matched <i>k</i> -mers (%)	Antibiogram AZM		Antibiogram CFM		Antibiogram CIP		Antibiogram CRO		MLST match	
			Actual	Best match	Actual	Best match	Actual	Best match	Actual	Best match		
GC01	Yes	27	S	S ^a	S	S ^a	S	S ^a	S	S ^a	Yes ^a	
GC02	Yes	27	S	S ^a	R	R ^{a,c}	S	S ^a	R	R ^{a,c}	Yes ^a	
GC03	Yes	33	S	S ^a	R	S ^{b,c}	S	S ^a	R	S ^{b,c}	Yes ^a	
GC04	Yes	21	S	S ^a	R	R ^a	R	R ^a	R	S ^b	Yes ^a	
GC05	Yes	7	R	R ^a	S	S ^a	S	S ^a	S	S ^a	Yes ^a	
(b) Non-database isolates												
Sample	Lineage confidently detected	Matched <i>k</i> -mers (%)	Antibiogram AZM		Antibiogram CFM		Antibiogram CIP		Antibiogram CRO		MLST match	
			Actual	Best match	Actual	Best match	Actual	Best match	Actual	Best match		
GC06	Yes	19	S	S ^a	R	R ^a	R	R ^a	S	S ^a	S ^a	
GC07	No	20	S	S ^a	S	S ^a	R	R ^a	S	S ^a	S ^a	
GC08	No	19	S	S ^a	S	S ^a	R	R ^a	S	S ^a	S ^a	
GC09	No	18	S	S ^a	S	S ^a	S	S ^a	S	S ^a	S ^a	
GC10	No	20	S	S ^a	S	S ^a	R	R ^a	S	S ^a	S ^a	
GC11	No	20	S	S ^a	S	S ^a	R	R ^a	S	S ^a	S ^a	
GC12	No	20	S	S ^a	S	S ^a	R	R ^a	S	S ^a	S ^a	
GC13	Yes	20	S	S ^a	S	S ^a	R	R ^a	S	S ^a	S ^a	
GC14	Yes	19	S	S ^a	S	S ^a	R	R ^a	S	S ^a	S ^a	
GC15	Yes	19	R	S ^{b,c}	S	S ^a	S	S ^a	S	S ^a	S ^a	
GC16	No	18	S	S ^a	S	S ^{a,c}	R	R ^a	S	S ^{a,c}	S ^{a,c}	
GC17	No	19	S	S ^a	S	S ^{a,c}	R	R ^a	S	S ^{a,c}	S ^{a,c}	
GC18	No	20	S	S ^a	S	S ^a	R	R ^a	S	S ^a	S ^a	
GC19	Yes	18	S	S ^a	S	S ^a	R	R ^a	S	S ^a	S ^a	

The table displays actual and predicted resistance phenotypes (S and R) for individual experiments and information on the match of the predicted MLST-identified sequence type. ^aCorrect prediction. ^bIncorrect prediction. ^cLow-confidence call. AZM, azithromycin; CFM, cefixime; CIP, ciprofloxacin.

of the query strain on the basis of the reported properties of its relatives. At present, the precise lineage of a bacterial pathogen is often determined after most important clinical decisions have been made. However, incorporating genomic neighbour typing at an earlier stage offers a way of leveraging bacterial population structure to gain information on resistance and susceptibility, and to guide antimicrobial therapy. The results from the metagenomic samples suggest that it is possible to apply this approach directly to clinical samples, and the success with both *S. pneumoniae* and *N. gonorrhoeae* indicates that it may have wide application.

The two pathogens studied here present contrasting features; the gonococcus is Gram-negative, harbours plasmids and has a strikingly uniform core genome, while the pneumococcus is Gram-positive, does not contain plasmids and is diverse in both its core and accessory genome. Both exhibit high rates of homologous recombination, which is expected to both spread chromosomally encoded resistance elements and to scramble the phylogenetic signal that we use to identify the lineages. Despite these differences and the large degree of recombination, our approach performs well with both pathogens, with some differences that indicate opportunities and limitations of the application.

The initial identification of the closest relative was consistently more robust in the pneumococcus than the gonococcus, which is a result of the former having more *k*-mers that are specific to an

individual lineage, thus reflecting greater sequence diversity. As a consequence of the much lower diversity in gonococcus, when multiple closely related genomes are present in the database, RASE fluctuated between them, even though it correctly identified the region of the phylogeny. If these genomes vary in their resistance profile, this is properly reflected in an uncertain susceptibility score that indicates caution and further investigation are merited (for example, GC03).

As in all inference, the principle limitation of genomic neighbour typing is the representativeness of the database. While we have made use of relatively small samples from limited geographical areas to demonstrate proof of principle, in practice, there are multiple examples of large genomic databases generated by public health agencies, which could be combined with metadata on resistance for genomic neighbour typing. Such databases could, if necessary, be supplemented with local sampling. The relevant question for our approach therefore becomes whether the database contains a sufficiently high proportion of strains that will be encountered in the clinic and whether the resistance data are correct. Further work is required to determine the optimal structure and contents of databases for each application, but we emphasize the range of pathogens that appear to show promise for this approach. These include *Escherichia coli*, in which data on MLST-identified sequence types supplemented with epidemiological information consistently

produced AUC values in excess of 0.90 for multiple antibiotics⁴², which suggests that there is great potential for neighbour typing to offer excellent resolution that is superior to MLST. However, we anticipate that genomic neighbour typing may be less suitable in cases when there is little within-species genomic variation (meaning it is hard to identify the nearest neighbour) or when resistance rapidly emerges on independent and diverse genomic backgrounds (meaning resistance is poorly correlated with those backgrounds) (Supplementary Note 8).

In cases where the infectious agent is unknown, this problem is significantly more challenging. *k*-mers from one pathogen can match others and produce false predictions, and so the choice of the correct database for prediction is key. Doing this will probably require a two-step solution in which the reads are first passed through a metagenomic classifier such as Centrifuge⁴³ or MetaMaps⁴⁴, which would be used to select the correct RASE database on which to make a resistance call.

Another limitation is the time required for sample preparation, which currently includes human DNA depletion, DNA isolation and library preparation, thereby taking a total of 4 h. However, this is a rapidly evolving area of technology and automated rapid library preparation kits are already in development⁴⁵. Further advances in this space, in particular for the preparation of metagenomic samples, will be required to bring the method closer to the bedside.

We have demonstrated that effectively predicting resistance and susceptibility from sequencing data does not require knowledge of causal resistance determinants. In fact, neighbour typing only requires that the phenotype be sufficiently strongly associated with the population structure to make reliable predictions.

A key advantage of this approach is that it requires very little genomic data, thus it is not limited by high error rates or low coverage. In particular, it is not attempting to define the exact genome sequence of the sample being tested, but merely which lineage it comes from. As a result, even when a small fraction of *k*-mers in the read are informative in matching to the RASE database, this is sufficient to call the lineage. This has the benefit of being faster than gene detection by virtue of the informative *k*-mers being distributed throughout the genome, and so more likely to appear in the first few reads sequenced by the nanopore. Therefore, the approach we present here can be seen as an application of compressed sensing; that is, by measuring a sparse signal distributed broadly across our data, we can identify it with comparatively few error-tolerant measurements.

Genomic neighbour typing can also be used to detect other phenotypes that are sufficiently tightly linked to a phylogeny, such as virulence. Further applications may include rapid outbreak investigations, as the closely related isolates involved in the outbreak would all be predicted to match to the same strain in the RASE database. The approach also lends itself to enhanced surveillance, including in the field; the 2014–2016 Ebola outbreak in West Africa, for example, saw MinION devices used in remote locations without advanced healthcare facilities⁷. Finally, at present, empirical treatment decisions are made within successive ‘windows’⁴⁶, in which increasing information becomes available, from initial Gram stain to full phenotypic characterization. The information from genomic neighbour typing is a natural complement to this process and has the potential to improve therapy long before it would become clinically apparent that the patient is not responding or before phenotypic susceptibility data were available. The combination of high-quality RASE databases with genomic neighbour typing offers an alternative forward-looking model for diagnostics and surveillance, with wide applications for the improved clinical management of infectious disease.

Methods

Overview. RASE uses rapid approximate *k*-mer-based matching of long sequencing reads against a database of strains to predict resistance via neighbour typing. The

database contains a highly compressed exact *k*-mer index, a representation of the tree population structure and metadata such as lineage, resistance profiles, MLST-based sequence type and serotype. The RASE prediction pipeline iterates over reads from the nanopore sequencer and provides real-time predictions of lineage and resistance or susceptibility (Fig. 1).

Resistance profiles. For all antibiotics, RASE associates individual strains with a resistance category: ‘susceptible’ (S) or ‘non-susceptible’ (R). First, intervals of possible MIC values are extracted using regular expressions from the available textual antibiograms. For instance, ‘≥4’, ‘2’ and ‘NA’ would be translated to the intervals $[4, +\infty)$, $[2, 2]$ and $[0, +\infty)$, respectively. Then the acquired intervals are compared to the antibiotic-specific breakpoints. If a given breakpoint is above or below the interval, susceptibility or non-susceptibility is reported, respectively. However, no category can be assigned at this step if the breakpoint lies within the extracted interval, an antibiogram is entirely missing, it is insufficiently specific or its parsing failed. Finally, missing categories are inferred using ancestral state reconstruction on the associated phylogenetic tree while maximizing parsimony (that is, minimizing the number of nodes switching its resistance category; Extended Data Figs. 3 and 4). When the solution for a node is not unique, non-susceptibility is assigned.

Genomic neighbour typing using nanopore sequencing. All reference strains in the database are associated with similarity weights that are set to zero at the start of the run. Each time a new read is read from the stream, *k*-mer-based matching is applied to identify the strains with the maximum number of matching *k*-mers (see below). Such strains are nearest neighbours to the read in the database according to the $1/(\text{‘number of matched } k\text{-mers’})$ pseudodistance.

The weights of the nearest neighbours are then increased according to the information content of the read, calculated as the number of matched *k*-mers divided by the number of nearest neighbours. Reads that do not match (that is, 0 matching *k*-mers in the database) are not used in subsequent analysis. The computed matches are also used for updating the *k*-mer score (KS), which is the proportion of matched *k*-mers in all reads. The KS helps to assess whether a sample is truly matching the database and whether predicting resistance for the database species makes sense.

The obtained weights serve as a proxy to the inverted genetic distance and are used as a basis for the subsequent predictions of the lineage and the antibiotic resistance and susceptibility.

Predicting lineage. A lineage is predicted as the lineage of the best-matching reference strain; that is, the one with the largest weight. The quality of lineage prediction is further quantified using a lineage score (LS), calculated as $LS = 2f/(f+t) - 1$, where *f* and *t* denote the weights of the best matches in the first (‘predicted’) and in the second best (‘alternative’) lineage, respectively. The values of the LS can range from 0.0 to 1.0 with the following special cases: LS = 1.0 means that all reads were perfectly matching the predicted lineage, whereas LS = 0.0 means that the predicted and alternative lineages were matched equally well.

The LS is used to measure how well a sample matches the identified lineage. If the LS is higher than a specified threshold (0.6 in default settings), the call is considered successful. If the score is lower than this, the sample cannot be securely assigned to a lineage, and this should draw the attention of the operators. Note that custom RASE databases may require a recalibration of the threshold.

Predicting resistance and susceptibility. Resistance or susceptibility is independently predicted for individual antibiotics based on the weights of the strains that belong to the predicted lineage. These are used to calculate a susceptibility score, which is further interpreted by comparing to predefined thresholds.

The susceptibility score (SS) is calculated as $SS = s/(s+r)$, where *s* and *r* denote the weights of the best-matching susceptible and best-matching non-susceptible strains within the lineage, respectively. The values of the SS can range from 0.0 to 1.0 with the following special cases: SS = 0.0 and SS = 1.0 mean that all reads match only resistant or susceptible strains in the lineage, respectively. In practice, this happens only if the lineage is entirely associated with resistance or susceptibility. SS = 0.5 means that the best matching resistant and susceptible strains are matched equally well. As follows from the score definition, if SS is greater than 0.5, then the best-matching strain is susceptible, otherwise it is non-susceptible.

The SS is used for predicting resistance or susceptibility and for evaluating the confidence of the prediction. If the SS is greater than 0.5, susceptibility to the antibiotic is reported, otherwise non-susceptibility is reported. Hence, resistance is predicted as the resistance of the best match. However, when the SS is within the [0.4, 0.6] range, it is considered a low-confidence call, and as such it should draw the attention of the operators; this usually indicates that resistance or susceptibility has recently emerged in the evolutionary history, and genomic neighbour typing may not be able to confidently distinguish between these similar, but phenotypically distinct, strains. Note that the thresholds above might require a further recalibration based on the specific database, antibiotics and application of RASE.

***S. pneumoniae* RASE database.** The *S. pneumoniae* RASE database was constructed using the EUCAST breakpoints³³ for the following antibiotics (mg l⁻¹):

ceftriaxone (0.25), erythromycin (0.25), benzylpenicillin (0.06), trimethoprim-sulfamethoxazole (1.00) and tetracycline (1.00). While we used the above values in the present work, others may be readily defined and the database rapidly updated. This is especially useful in cases where breakpoints may vary depending on the site of infection (as is the case with pneumococcal meningitis and otitis media, for which lower MICs are considered to be resistant³³).

The draft assemblies were downloaded from the SRA FTP server using the accession codes provided in table 1 in ref. ²⁴. The phylogenetic tree was downloaded from DataDryad (accession no.: <https://doi.org/10.5061/dryad.t55gq>). The pneumococcal ProPhyle index was constructed with the *k*-mer size of *k* = 18.

The obtained *S. pneumoniae* RASE database, including the source code and data, is available at <https://github.com/c2-d2/rase-db-s pneumoniae-sparc>.

***N. gonorrhoeae* RASE database.** The *N. gonorrhoeae* RASE database was constructed with the CDC GISP breakpoints³⁴ for the following antibiotics (mg⁻¹): azithromycin (2.0), cefixime (0.25), ciprofloxacin (1.0) and ceftriaxone (0.125). Before applying the breakpoints, azithromycin MICs for strains collected before 2005 were doubled to correct for the known inconsistencies of the phenotyping protocol, which were due to a change in formulation of the commercial media⁴⁷.

The draft assemblies and the phylogenetic tree were downloaded from Zenodo (accession no.: <https://doi.org/10.5281/zenodo.2618836>). The following three prevalent types of plasmids⁴⁸ were downloaded from GenBank, localized in the GISP database using BLAST⁴⁹ and removed from the dataset: the cryptic plasmid pJD1 (GenBank accession no.: NC_001377.1), the beta-lactamase plasmid pJD4 (GenBank accession no.: NC_002098.1) and the conjugative plasmid pEP5289 (GenBank accession no.: GU479466.1). The gonococcal ProPhyle index was constructed with the *k*-mer size of *k* = 18.

The obtained *N. gonorrhoeae* RASE database, including the source code and data, is available at <https://github.com/c2-d2/rase-db-ngonorrhoeae-gisp>.

***k*-mer-based matching.** Reads were matched against the RASE databases using the ProPhyle classifier^{27,28} (commit b55e026) and its ProPhex component^{50,51}. The ProPhyle index stores *k*-mers of all strains in a highly compressed form, reducing the required memory footprint. In the database construction phase, the *k*-mers of the strains are first propagated along the phylogenetic tree and then greedily assembled to simplitigs⁵². The obtained simplitigs are then placed into a single text file, for which a BWT index is constructed³⁰.

In the course of sequencing, each read is decomposed into overlapping *k*-mers. The *k*-mers are then searched in the BWT index by ProPhex using BWT search using a sliding window³⁰. For every *k*-mer, the obtained matches are translated back on the tree. This provides a list of nodes whose descending leaves are the strains containing that *k*-mer. Finally, strains with maximum number of matched *k*-mers are identified for each read and reported in the SAM/BAM format⁵³.

Optimizing the *k*-mer length. The *k*-mer length is the main parameter of the classification. First, the subword complexity function⁵⁴ of pneumococcus was calculated using JellyFish⁵⁵ (v.2.2.10) (Extended Data Fig. 5). Then, based on the characteristics of the function and the *k*-mer range supported by ProPhyle, the possible range of *k* was determined as previously described^{17,32}. For these *k*-mer lengths, RASE indexes were constructed and their performance evaluated using the RASE prediction pipeline and selected experiments. While RASE showed robustness to *k*-mer length in terms of final predictions, prediction delays differed (Extended Data Fig. 6). Based on the obtained timing data, we set the *k*-mer length to *k* = 18.

Comparison to Kraken. For each RASE database, a fake NCBI taxonomy was generated from the database tree. Then, a library was built using Kraken³¹ (v.1.1.1, default parameters) from the same FASTA files as used for building the RASE database. Finally, Kraken databases were constructed for both *k* = 18 and *k* = 31.

The obtained Kraken databases were used to classify reads from individual experiments. The obtained Kraken assignments were subsequently converted using an ad hoc Python script to RASE-BAM (a subset of the BAM format⁵³ used by RASE). Finally, RASE prediction was applied on the BAM files, with the use of the RASE database metadata, and the results compared with the results of the standard RASE with ProPhyle.

Measuring time. To determine how RASE works with nanopore data generated in real time, the timestamps of individual reads were extracted using regular expressions from the read names. These were then used for sorting the base-called nanopore reads by time. When the RASE pipeline was applied, the timestamps were used for expressing the predictions as a function of time. The times of ProPhyle assignments were also compared to the original timestamps to ensure that the prediction pipeline was not slower than sequencing.

When timestamps of sequencing reads were not available (that is, the gonococcal WHO data and clinical samples), RASE estimated the progress in time from the number of processed base pairs. This was done by dividing the cumulative base-pair count by the typical nanopore flow, which we had previously estimated from SP01 as 1.43 Mbps per s. However, such an estimated progress is indicative only, as it does not follow the true order of reads in the course of

sequencing. As the nanopore signal quality tends to decrease over time (see the decrease of KS in Fig. 2 after *t* = 15 mins), the randomized read order provides results of lower quality than true real-time sequencing.

Lower time estimates on resistance gene detection. A complete genome of the multidrug-resistant SP02 isolate was assembled from the nanopore reads using CANU⁵⁶ (v.1.5, default parameters). Before the assembly step, reads were filtered using SAMsift⁵⁷ based on the matching quality with the pneumococcal RASE database: only reads at least 1,000-bp long with at least 10% 18-mers shared with some of the reference draft assemblies were used. The obtained assembly was further corrected using Pilon⁵⁸ (v.1.2, default parameters) and Illumina reads from the same isolate (taxid 1QJAP in the SPARC dataset²⁴) mapped to the nanopore assembly using BWA-MEM⁵⁹ (v.0.7.17, default parameters) and sorted using SAMtools⁵³.

The obtained assembly was searched for resistance-causing genes using the online CARD tool¹⁶ (as of 1 Aug 2018). All of the original nanopore reads were then mapped using Minimap2 (v.2.11, with '-x map-ont')⁶⁰ to the corrected assembly, and resistance genes in the reads were identified using BEDtools-intersect⁶¹ (v.2.27.1, with '-F 95'). Timestamps of the resistance-informative reads were extracted and associated with the genes. Only reads longer than 2 kbp were used in the analysis.

Evaluation of the *N. gonorrhoeae* WHO samples. To evaluate the predictions of the WHO samples, we inferred a phylogenetic tree from a dataset comprising both the GISP isolates and the WHO isolates. First, reads were downloaded for the GISP isolates (NCBI BioProject no.: PRJEB2999 and PRJEB7904) and for the WHO isolates F–P (NCBI BioProject no.: PRJEB4024). For the WHO isolates U–Z, read data were simulated from the finished de novo assemblies (NCBI BioProject no.: PRJEB14020) using Art-Illumina⁶² (v.2.5.1). Reads were mapped to the NCCP11945 reference genome (GenBank accession no.: CP001050.1) using BWA-MEM⁵⁹ (v.0.7.17) and deduplicated using Picard⁶³ (v.2.8.0). Pilon⁵⁸ (v.1.16, with '-mindepth 10–minmq 20') was used to call variants and further filtered to include only 'pass' sites, and sites where the alternative allele was supported with AF > 0.9. Gubbins⁶⁴ (v.2.3.4) with RAXML⁶⁵ (v.8.2.10) were run on the aligned pseudogenomes to generate the final recombination-corrected phylogeny (Supplementary Data 1).

The closest relatives identified by RASE were verified using the obtained tree. For every WHO isolate, the obtained RASE prediction was compared to the closest GISP isolate on the tree.

Library preparation. For isolates SP01–SP06, cultures were grown in Todd–Hewitt medium with 0.5% yeast extract (THY; Becton Dickinson) at 37 °C in 5% CO₂ for 24 h. High-molecular-weight (>1 µg) genomic DNA was extracted and purified from cultures using a DNeasy Blood and Tissue kit (Qiagen). DNA concentration was measured using a Qubit fluorometer (Invitrogen). Library preparation was performed using an ONT 1D ligation sequencing kit (SQK LSK108).

For isolates SP07–SP12, library preparation was performed using an ONT Rapid Low-Input Barcoding kit (SQK-RLB001), with saponin-based host DNA depletion used for reducing the proportion of human reads. More details can be found in the original manuscript⁴⁰.

For isolates GC01–GC05, cultures were grown on Chocolate-Agar media (that is, Difco GC base media containing 1% IsoVitaleX (Becton Dickinson) and 1% Remel Hemoglobin (Thermo Fisher Scientific)) at 37 °C in 5% CO₂ for 20 h. For GC01–GC04, genomic DNA was extracted and purified from cultures using a PureLink Genomic DNA mini kit (Thermo Fisher Scientific). For GC05, DNA was extracted using the phenol–chloroform method⁶⁶. Genomic DNA was extracted and purified from cultures using the PureLink Genomic DNA mini kit (Thermo Fisher Scientific). DNA concentration was measured using a Qubit fluorometer (Invitrogen). Library preparation was performed using the ONT 1D ligation sequencing kit (SQK-LSK109).

MinION sequencing. Sequencing was performed on a MinION MK1 device using R9.4/FLO-MIN106 flow cells according to the manufacturer's instructions. For experiments SP01–SP06, base calling was performed using ONT Metrichor (v.1.6.11 (SP01), v.1.7.3 (SP02), v.1.7.14 (SP03–SP06)) simultaneously with sequencing, and all reads passing Metrichor quality check were used in the further analysis. For experiments SP07–SP12, the ONT MinKNOW software (v.1.4-1.13.1) was used to collect raw sequencing data, and ONT Albacore (v.1.2.2-2.1.10) was used for local base-calling of the raw data after sequencing runs were completed. For experiments GC01–GC05, ONT MinKNOW software was used to collect raw sequencing data, and ONT Albacore (v.2.3.4) was used for local base-calling.

Testing resistance phenotype. Additional retesting of SPARC isolates was done using microdilution. Organism suspensions were prepared from overnight growths on blood agar plates to the density of a 0.5 McFarland standard. This organism suspension was then diluted to provide a final inoculum of 105 to 106 colony-forming units per ml. Microdilution trays were prepared according to the NCCLS methodology with cation-adjusted Mueller–Hinton broth (Sigma-Aldrich) supplemented with 5% lysed horse blood (Hemostat Laboratories)^{67,68}. Penicillin

(TRC Canada) and chloramphenicol (USB) concentrations ranged from 0.016 to 16 µg ml⁻¹. Erythromycin (Enzo Life Sciences), tetracycline (Sigma-Aldrich) and trimethoprim-sulfamethoxazole (MP Biomedicals) concentrations ranged from 0.0625 to 64 µg ml⁻¹. Ceftriaxone (Sigma-Aldrich) concentrations ranged from 0.007 to 8 µg ml⁻¹. The microdilution trays were incubated in ambient air at 35 °C for 24 h. The MICs were then visually read and breakpoints applied. A list of individual microdilution measurements and the obtained resistance categories is provided in Supplementary Table 2.

Resistance of streptococcus in the metagenomic samples (SP07–SP12) was determined by agar diffusion using the EUCAST methodology and breakpoints³³. First, the inoculated agar plates were incubated at 37 °C overnight and then examined for growth, with the potential for reincubation up to 48 h. Then, the samples were screened to oxacillin: if the zone diameter *r* was >20 mm, the isolate was considered to be sensitive to benzylpenicillin, otherwise a full MIC measurement to benzylpenicillin was done. Finally, the isolate was screened for resistance to tetracycline (*r* ≥ 25 mm for sensitive, *r* < 22 mm for resistant) and erythromycin (*r* ≥ 22 mm for sensitive, *r* < 19 mm for resistant); when the isolate showed intermediate resistance, a full MIC measurement was done.

Results for all tested samples, isolates and metagenomes are summarized in Supplementary Table 3.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The analyses in the paper were performed using the following RASE databases: “*N. gonorrhoeae* GISP USA v1.4” (available at <https://github.com/c2-d2/rase-db-ngonorrhoeae-gisp/releases>) and “*S. pneumoniae* SPARC USA v1.3” (available at <https://github.com/c2-d2/rase-db-spneumoniae-sparc/releases>). Nanopore reads for all experiments from this study have been deposited in Zenodo with the accession code 10.5281/zenodo.3346055; for the metagenomic experiments (SP07–SP12), only the filtered datasets were made publicly available (that is, after removing the remaining human reads in silico to comply with privacy policies). Additional supplementary materials are available at <https://github.com/c2-d2/rase-supplement>.

Code availability

RASE was developed using Python, GNU Make, GNU Parallel⁶⁹, Snakemake⁷⁰ and the ETE 3 (ref. ⁷¹) and PySam⁵³ libraries, and was based on ProPhyle^{27,28} (commit b55e026). Bioconda⁷² was used to ensure reproducibility of the software environments. The core RASE package and RASE documentation are available at <https://github.com/c2-d2/rase>. The prediction pipeline is available at <https://github.com/c2-d2/rase-pipeline>. Codes for constructing the *S. pneumoniae* and *N. gonorrhoeae* RASE databases are available at <https://github.com/c2-d2/rase-db-spneumoniae-sparc> and <https://github.com/c2-d2/rase-db-ngonorrhoeae-gisp>, respectively. A skeleton for creating new databases is available from <https://github.com/c2-d2/rase-db-skeleton>.

Received: 6 August 2019; Accepted: 6 December 2019;

Published online: 10 February 2020

References

- Kumar, A. et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit. Care Med.* **34**, 1589–1596 (2006).
- Quick, J. et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).
- Faria, N. R. et al. Mobile real-time surveillance of Zika virus in Brazil. *Genome Med.* **8**, 2–5 (2016).
- Antibiotic Resistance Threats in the United States, 2013* (Centers for Disease Control and Prevention, 2013).
- Li, Y. et al. Penicillin-binding protein transpeptidase signatures for tracking and predicting β-lactam resistance levels in *Streptococcus pneumoniae*. *mBio* **7**, e00756-16 (2016).
- Zankari, E. et al. Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67**, 2640–2644 (2012).
- Steiner, A., Stucki, D., Coscollá, M., Borrell, S. & Gagneux, S. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics* **15**, 881 (2014).
- Gupta, S. K. et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob. Agents Chemother.* **58**, 212–220 (2014).
- Rowe, W. et al. Search engine for antimicrobial resistance: a cloud compatible pipeline and web interface for rapidly detecting antimicrobial resistance genes directly from sequence data. *PLoS ONE* **10**, e0133492 (2015).
- Kaminski, J. et al. High-specificity targeted functional profiling in microbial communities with ShortBRED. *PLoS Comput. Biol.* **11**, e1004557 (2015).
- de Man, T. J. B. & Limbago, B. M. SSTAR, a stand-alone easy-to-use antimicrobial resistance gene predictor. *mSphere* **1**, e00050-15 (2016).
- Clausen, P. T. L. C., Zankari, E., Aarestrup, F. M. & Lund, O. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. *J. Antimicrob. Chemother.* **71**, 2484–2488 (2016).
- Yang, Y. et al. ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database. *Bioinformatics* **32**, 2346–2351 (2016).
- Garner, E., Pruden, A., Heath, L. S. & Vikesland, P. DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* **6**, 23 (2018).
- Antonopoulos, D. A. et al. PATRIC as a unique resource for studying antimicrobial resistance. *Brief. Bioinform.* **20**, 1094–1102 (2017).
- Jia, B. et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45**, D566–D573 (2017).
- Zankari, E. et al. PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *J. Antimicrob. Chemother.* **72**, 2764–2768 (2017).
- Hunt, M. et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb. Genomics* **3**, e000131 (2017).
- Votintseva, A. A. et al. Same-day diagnostic and surveillance data for tuberculosis via whole genome sequencing of direct respiratory samples. *J. Clin. Microbiol.* **55**, 1285–1298 (2017).
- Rowe, W. P. M. & Winn, M. D. Indexed variation graphs for efficient and accurate resistome profiling. *Bioinformatics* **34**, 3601–3608 (2018).
- Feldgarden, M. et al. Using the NCBI AMRFinder tool to determine antimicrobial resistance genotype-phenotype correlations within a collection of NARMS isolates. Preprint at <https://www.biorxiv.org/content/10.1101/550707v1> (2019).
- Boolchandani, M., D’Souza, A. W. & Dantas, G. Sequencing-based methods and resources to study antimicrobial resistance. *Nat. Rev. Genet.* **20**, 356–370 (2019).
- Croucher, N. J. et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat. Genet.* **45**, 656–663 (2013).
- Croucher, N. J. et al. Population genomic datasets describing the post-vaccine evolutionary epidemiology of *Streptococcus pneumoniae*. *Sci. Data* **2**, 150058 (2015).
- Grad, Y. H. et al. Genomic epidemiology of gonococcal resistance to extended-spectrum cephalosporins, macrolides, and fluoroquinolones in the United States, 2000–2013. *J. Infect. Dis.* **214**, 1579–1587 (2016).
- Cheng, L., Connor, T. R., Siren, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* **30**, 1224–1228 (2013).
- Břinda, K., Salikhov, K., Pignotti, S. & Kucherov, G. karel-brinda/prophyle: ProPhyle 0.3.1.0. Zenodo <https://doi.org/10.5281/zenodo.1045429> (2017).
- Břinda, K. *Novel Computational Techniques for Mapping and Classification of Next-Generation Sequencing Data*. PhD thesis, Univ. Paris-Est. (2016); <https://hal.archives-ouvertes.fr/tel-01484198v3>
- Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44**, 226–232 (2012).
- Ferragina, P. & Manzini, G. Opportunistic data structures with applications. In *Proc. 41st Annual Symposium on Foundations of Computer Science* 390–398 (IEEE Computer Society, 2000).
- Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
- Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
- Breakpoint Tables for Interpretation of MICs and Zone Diameters. Version 7.0.* (The European Committee on Antimicrobial Susceptibility Testing, 2017).
- Kirkcaldy, R. D. et al. *Neisseria gonorrhoeae* antimicrobial susceptibility surveillance—The Gonococcal Isolate Surveillance Project, 27 sites, United States, 2014. *MMWR. Surveill. Summ.* **65**, 1–19 (2016).
- Palace, S. et al. RNA polymerase mutations cause cephalosporin resistance in clinical *Neisseria gonorrhoeae* isolates. Preprint at <https://www.biorxiv.org/content/10.1101/626457v2> (2019).
- McGee, L. et al. Nomenclature of major antimicrobial-resistant clones of *Streptococcus pneumoniae* defined by the Pneumococcal Molecular Epidemiology Network. *J. Clin. Microbiol.* **39**, 2565–2571 (2001).
- Donà, V. et al. Mismatch amplification mutation assay-based real-time PCR for rapid detection of *Neisseria gonorrhoeae* and antimicrobial resistance determinants in clinical specimens. *J. Clin. Microbiol.* **56**, e00365-18 (2018).
- Golparian, D. et al. Antimicrobial resistance prediction and phylogenetic analysis of *Neisseria gonorrhoeae* isolates using the Oxford Nanopore MinION sequencer. *Sci. Rep.* **8**, 17596 (2018).
- Unemo, M. et al. The novel 2016 WHO *Neisseria gonorrhoeae* reference strains for global quality assurance of laboratory investigations: phenotypic, genetic and reference genome characterization. *J. Antimicrob. Chemother.* **71**, 3096–3108 (2016).

40. Charalampous, T. et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat. Biotechnol.* **37**, 783–792 (2019).
41. Sá-Leão, R. et al. Carriage of internationally spread clones of *Streptococcus pneumoniae* with unusual drug resistance patterns in children attending day care centers in Lisbon, Portugal. *J. Infect. Dis.* **182**, 1153–1160 (2000).
42. MacFadden, D. R. et al. Comparing patient risk factor-, sequence type-, and resistance locus identification-based approaches for predicting antibiotic resistance in *Escherichia coli* bloodstream infections. *J. Clin. Microbiol.* **57**, e01780-18 (2019).
43. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
44. Dilthey, A.T., Jain, C., Koren, S. & Phillippy, A. M. Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. *Nat. Commun.* **10**, 3066 (2019).
45. Quick, J. Ultra-long read sequencing protocol for RAD004 V.3. *Protocols.io* <https://doi.org/10.17504/protocols.io.mrxc57n> (2018).
46. MacFadden, D. R., Leis, J. A., Mubareka, S. & Daneman, N. The opening and closing of empiric windows: the impact of rapid microbiologic diagnostics. *Clin. Infect. Dis.* **59**, 1199–1200 (2014).
47. Kersh, E. N. et al. Rationale for a *Neisseria gonorrhoeae* susceptible only interpretive breakpoint for azithromycin. *Clin. Infect. Dis.* <https://doi.org/10.1093/cid/ciz292> (2019).
48. Cehovin, A. & Lewis, S. B. Mobile genetic elements in *Neisseria gonorrhoeae*: movement for change. *Pathog. Dis.* **75**, ftx071 (2017).
49. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
50. Salikhov, K. *Efficient Algorithms and Data Structures for Indexing DNA Sequence Data*. PhD thesis, Univ. Paris-Est. (2017); <https://hal.archives-ouvertes.fr/tel-01762479v1>
51. Břinda, K., Salikhov, K., Pignotti, S. & Kucherov, G. prophyle/prophex: ProPhex 0.1.1. *Zenodo* <https://doi.org/10.5281/zenodo.1247431> (2018).
52. Břinda, K., Baym, M. & Kucherov, G. Simplitigs as an efficient and scalable representation of de Bruijn graphs. Preprint at <https://www.biorxiv.org/content/10.1101/2020.01.12.903443v1> (2020).
53. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
54. Lothaire, M. *Algebraic Combinatorics on Words* (Cambridge Univ. Press, 2002).
55. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
56. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
57. Břinda, K. karel-brinda/samsift: SAMSift 0.2.5. *Zenodo* <https://doi.org/10.5281/zenodo.1048211> (2018).
58. Walker, B. J. et al. An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
59. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://arxiv.org/abs/1303.3997> (2013).
60. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
61. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
62. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
63. Picard Tools (Broad Institute, 2020); <http://broadinstitute.github.io/picard/>
64. Croucher, N. J. et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).
65. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
66. Green, M. R. & Sambrook, J. Isolation of high-molecular-weight DNA using organic solvents. *Cold Spring Harb. Protoc.* **2017**, pdb.prot093450 (2017).
67. *Susceptibility Tests for Bacteria That Grow Aerobically; Approved Standard—Ninth Edition* (The Clinical & Laboratory Standards Institute, 2012).
68. *Performance Standards for Antimicrobial Susceptibility Testing; Twenty-Second Informational Supplement* (The Clinical & Laboratory Standards Institute, 2012).
69. Tange, O. GNU Parallel: the command-line power tool. *login* **36**, 42–47 (2011).
70. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
71. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, msw046 (2016).
72. Grüning, B. et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* **15**, 475–476 (2018).

Acknowledgements

This work was supported by the Bill & Melinda Gates Foundation (GCGH GCE OPP1151010 to K.B. and W.P.H.), the NIH—National Institute of Allergy and Infectious Diseases (R01 AI106786-05 to K.B.), the Canadian Institutes of Health Research (MFE 152448, a fellowship grant, to R.S.L. and D.R.M.), the NSF (GRFP to K.C.M.), and the David and Lucile Packard Foundation (to M.B.). This paper presents independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (reference no. RP-PG-0514-20018 to J.O.), the UK Antimicrobial Resistance Cross Council Initiative (MR/N013956/1 to J.O.), the Rosetrees Trust (A749 to J.O.), the Biotechnology and Biological Sciences Research Council (BBSRC) Institute Strategic Programme Microbes in the Food Chain (BB/R012504/1 and its constituent projects BBS/E/F/000PR10348 and BBS/E/F/000PR10349 to J.O.), the University of East Anglia (to J.O. and T.C.), and Oxford Nanopore Technologies (to J.O. and T.C.). Portions of this research were conducted on the O2 and Odyssey high-performance compute clusters, supported by the Research Computing Groups at Harvard Medical School and at the Harvard Faculty of Arts and Sciences, respectively. We thank J. Metlay for providing the test isolates for experiments SP03–SP06, which were collected as part of a population-wide surveillance study done in the Philadelphia region, supported by the NIH (R01 AI46645), and to B. J. Arnold, T. Azarian and C. M. Herren for useful comments during various stages of this project. This paper is dedicated to the memory of Nicholas Greenfield.

Author contributions

K.B., M.B. and W.P.H. designed the study and developed the theory. K.B., M.B., G.K. and W.P.H. developed the algorithms. K.B. developed the software. A.C., K.C.M., T.C., L.C., C.B.W., Y.H.G. and J.O. participated in the sequencing experiments. K.B., K.C.M., D.R.M., T.C., R.S.L., Y.H.G., J.O., M.B. and W.P.H. contributed to the interpretation of the results. K.B., M.B. and W.P.H. wrote the manuscript. All authors contributed to the final version of the manuscript.

Competing interests

J.O. received financial support for attending ONT and other conferences and a honorarium for speaking at ONT headquarters. J.O. received funding and consumable support from ONT for T.C.'s PhD studentship. T.C. received financial support from ONT for attending an international conference. All other authors have no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41564-019-0656-6>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41564-019-0656-6>.

Correspondence and requests for materials should be addressed to K.B.

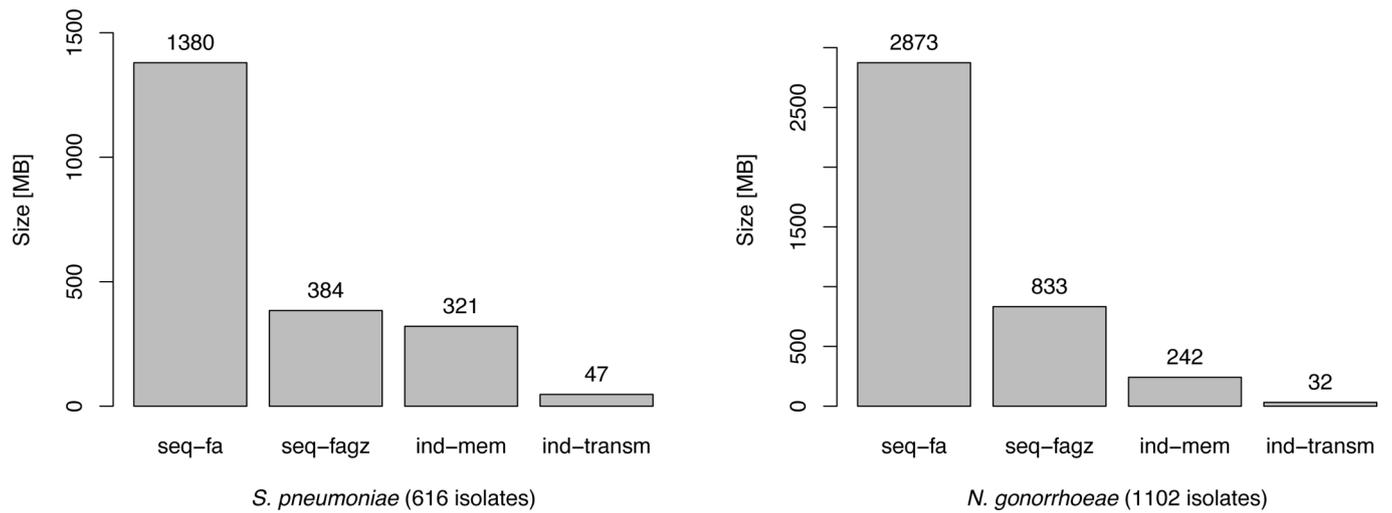
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

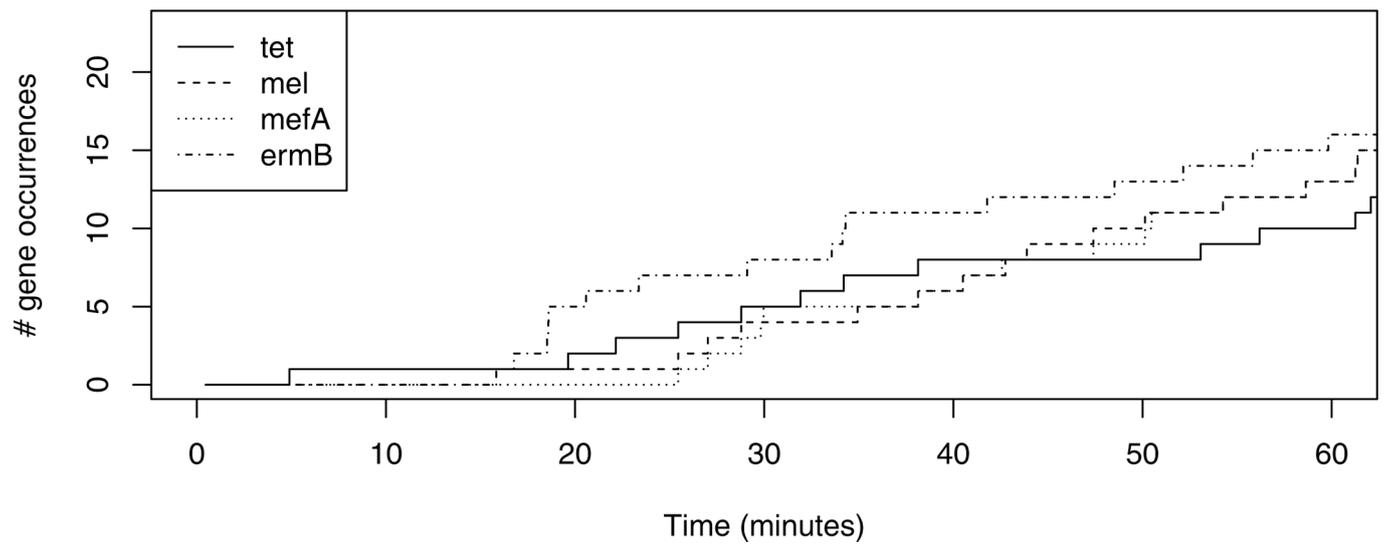


Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

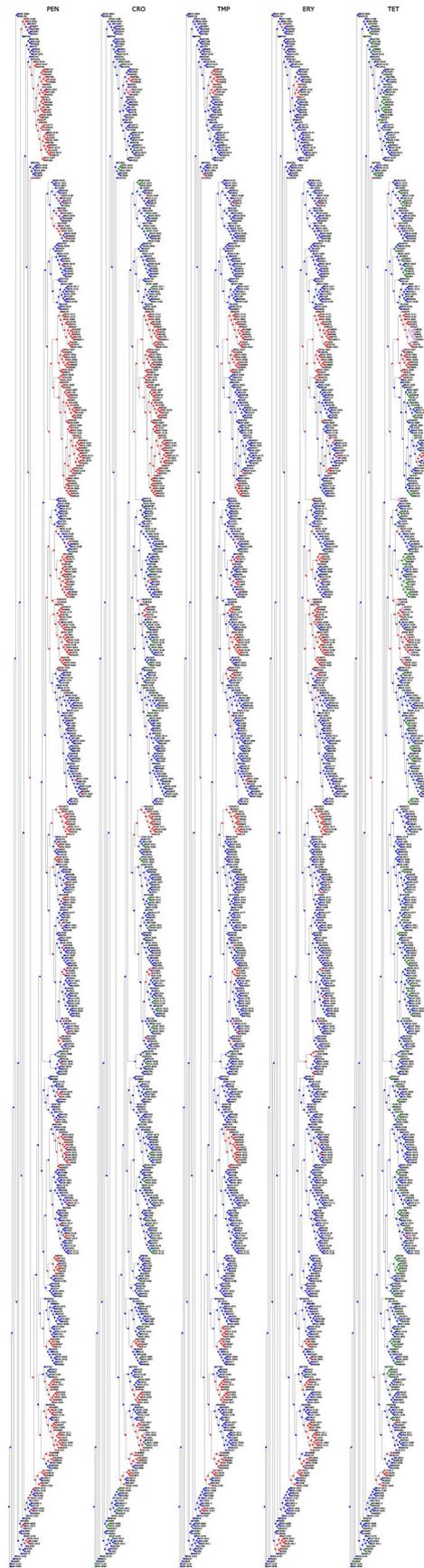
© The Author(s) 2020



Extended Data Fig. 1 | Size and memory footprint of the *S. pneumoniae* and *N. gonorrhoeae* RASE databases. The graph compares the size of the ProPhyle RASE index to the size of the original sequences: original draft assemblies (seq-fa), original draft assemblies compressed using gzip (seq-fagz), memory footprint of ProPhyle with the RASE index (ind-mem), and size of the ProPhyle RASE index compressed for transmission (ind-transm).



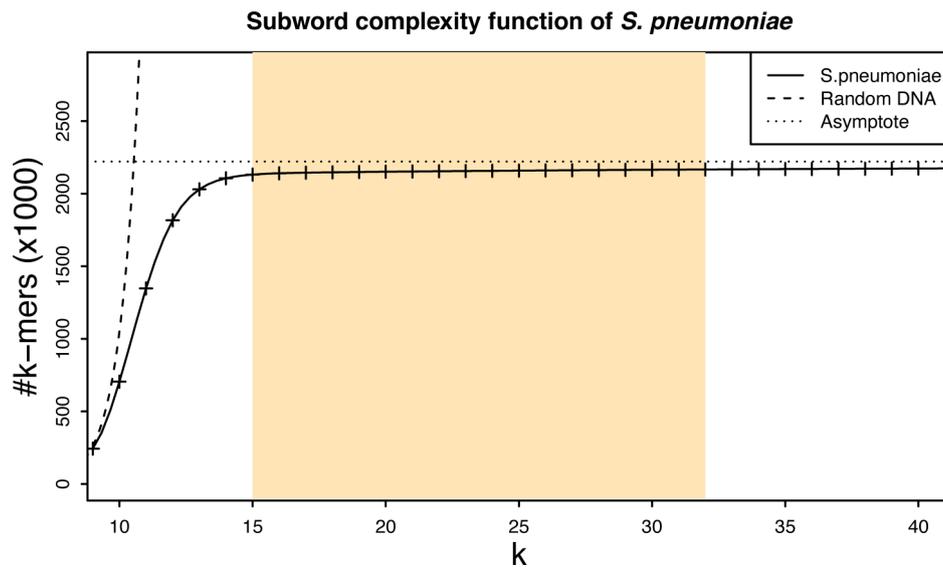
Extended Data Fig. 2 | Timeline of resistance genes. Number of occurrences of individual resistance genes in reads of SP02, as a function of time for the first hour of nanopore sequencing.



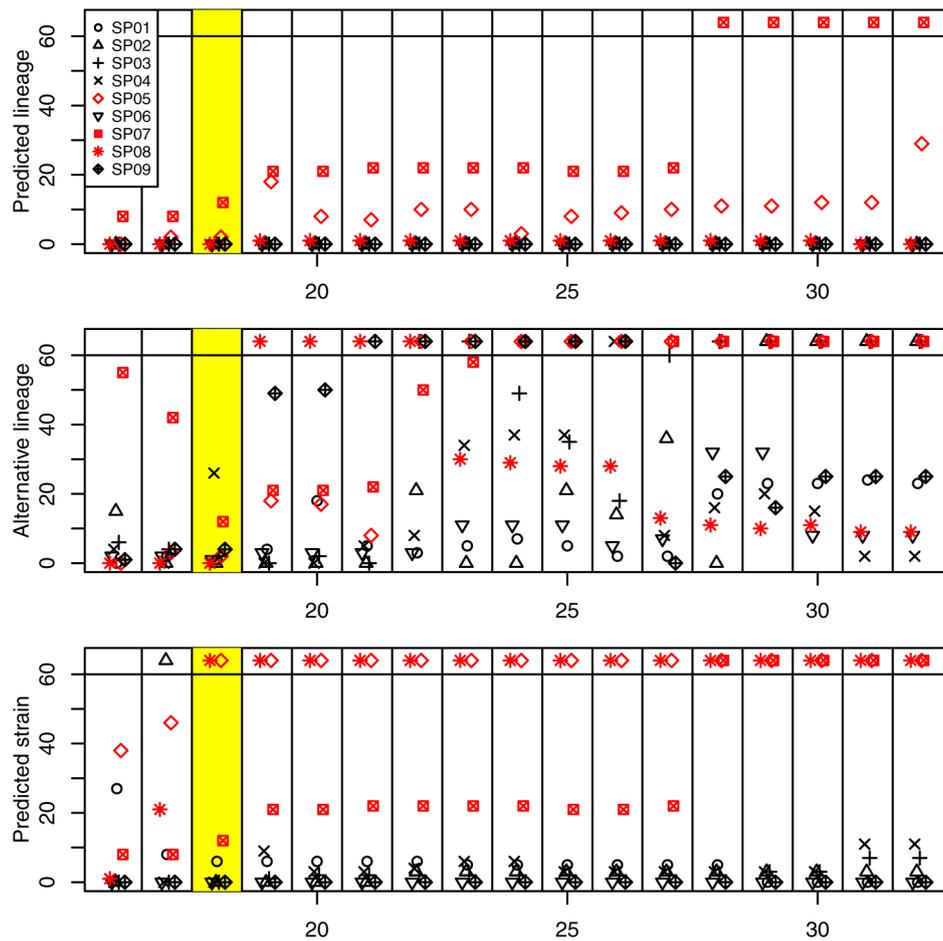
Extended Data Fig. 3 | Ancestral state reconstruction of resistance categories in the *S. pneumoniae* RASE database. Each panel corresponds to a single antibiotic and displays the database phylogenetic tree, colored according to the reconstructed resistance categories for the antibiotic (blue, green, red, violet correspond to 'susceptible', 'unknown - inferred susceptible', 'non-susceptible', 'unknown - inferred non-susceptible', respectively).



Extended Data Fig. 4 | Ancestral state reconstruction of resistance categories in the *N. gonorrhoeae* RASE database. Each panel corresponds to a single antibiotic and displays the database phylogenetic tree, colored according to the reconstructed resistance categories for the antibiotic (blue, green, red, violet correspond to 'susceptible', 'unknown - inferred susceptible', 'non-susceptible', 'unknown - inferred non-susceptible', respectively).



Extended Data Fig. 5 | Subword complexity of pneumococcus. The plot depicts the number of canonical k -mers as a function of k for *S. pneumoniae* ATCC 700669 (GenBank accession: 'NC_011900.1') and for a random DNA text containing all possible k -mers. For $k < 10$, the pneumococcus k -mer composition is similar to the one of random text. For $k > 14$, the k -mer sets are almost saturated and the complexity grows very slowly. Since the genome length is finite and bacterial chromosomes are circular, the function attains its maximum at the genome size (2,221,315 in this case). The highlighted region corresponds to the range of values of k , which are suitable for use in RASE.



Extended Data Fig. 6 | Delays in prediction based on the k -mer length. The plot displays delays in prediction as a function of the used k -mer length, for selected experiments and all possible k -mer lengths. Each horizontal panel displays times required for stabilization of one of the three predictions: the lineage, the alternative lineage, and the closest strain. Every column within a panel corresponds to a single k -mer length. When the required time exceeded 1 hour, the point is displayed at the top. Experiments where lineage could not be identified are plotted in red. The highlighted column corresponds to the k -mer length used for constructing the RASE databases in this paper.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- | | | |
|-------------------------------------|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | A description of all covariates tested |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

wget, curl

Data analysis

ONT data processing: ONT Metrichor (versions 1.6.11 (SP01), 1.7.3 (SP02), 1.7.14 (SP03–SP06)); ONT MinKNOW (versions 1.4-1.13.1) and ONT Albacore (versions 1.2.2-2.1.10) for SP07-SP12. ONT MinKNOW and Albacore (version 2.3.4) for GC01–GC05.

Other tools: ProPhyle (commit b55e026), GNU Make (version 4.2.1), GNU Parallel (version 20190322), Snakemake (version 5.4.5), ETE3 (version 3.1.1), PySAM (version 0.15.2), SAMtools (version 1.9), JellyFish (version 2.2.10), Kraken (version 1.1.1), CANU (version 1.5), Pilon (versions 1.2 and 1.16), BWA-MEM (version 0.7.17), CARD tool (as of 2018/08/01), BEDtools–intersect (version 2.27.1), Art-Illumina (version 2.5.1), Picard (version 2.8.0), Pilon (version 1.16), Gubbins (version 2.3.4), RAXML (version 8.2.10), and R (version 3.5.1).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequencing data for all experiments from this study can be downloaded from <http://doi.org/10.5281/zenodo.3346055>. Other supplementary materials are available from <https://github.com/c2-d2/rase-supplement/>. The pneumococcal and gonococcal RASE databases are available from <https://github.com/c2-d2/rase-db-spnemoniae-sparc> and <https://github.com/c2-d2/rase-db-ngonorrhoeae-gisp>.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size does not affect the statistical validity of a conclusion. The RASE databases were constructed from 616 isolates (SP) and 1102 isolates (GC). We evaluated the method with 12 (SP) and 33 (GC) sequencing experiments, where individual experiments were selected in such a way that we could test the method in several different scenarios (database isolates, non-database isolates with a representative database, non-database isolates with a non-representative database, and metagenomes).
Data exclusions	No reference strains were excluded from the databases. No sequencing experiments were excluded from the analysis.
Replication	No attempts at replication of individual sequencing experiments have been done.
Randomization	There were no experimental groups
Blinding	There were no experimental groups

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging