Published in final edited form as:
Nat Microbiol. 2020 March ; 5(3): 455–464. doi:10.1038/s41564-019-0656-6.

Rapid inference of antibiotic resistance and susceptibility by genomic neighbor typing

Karel Břinda[1,2,*], **Alanna Callendrello**[1], **Kevin C Ma**[3], **Derek R MacFadden**[1,4], **Themoula Charalampous**[5], **Robyn S Lee**[1,6], **Lauren Cowley**[7], **Crista B Wadsworth**[8], **Yonatan H Grad**[3], **Gregory Kucherov**[9,10], **Justin O'Grady**[11,5], **Michael Baym**[2], **William P Hanage**[1]

[1]Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, USA

[2]Department of Biomedical Informatics and Laboratory of Systems Pharmacology, Harvard Medical School, Boston, USA

[3]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, USA

[4]Division of Infectious Diseases, Department of Medicine, University of Toronto, Canada

[5]Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich, UK

[6]Epidemiology Division, Dalla Lana School of Public Health, University of Toronto, Canada

[7]Department of Biology and Biochemistry, University of Bath, UK

[8]Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, USA

[*]Correspondence to kbrinda@hsph.harvard.edu.

Implementation and code availability

RASE was developed using Python, GNU Make, GNU Parallel[68], Snakemake[69], and the ETE 3[70] and PySam[52] libraries, and was based on ProPhyle[27,28] (commit b55e026). Bioconda[71] was used to ensure reproducibility of the software environments. The core RASE package and RASE documentation are available from https://github.com/c2-d2/rase. The prediction pipeline is available from https://github.com/c2-d2/rase-pipeline. Code for constructing the *S. pneumoniae* and *N. gonorrhoeae* RASE databases is available from 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.

Data availability

The analyses in the paper were performed with the following RASE databases: "*N. gonorrhoeae* GISP USA v1.4" (available from https://github.com/c2-d2/rase-db-ngonorrhoeae-gisp/releases) and "*S. pneumoniae* SPARC USA v1.3" (available from 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 (i.e., after removing the remaining human reads *in silico* in order to comply with privacy policies. Additional supplementary materials are available from https://github.com/c2-d2/rase-supplement.

Author contributions

KB, MB, WPH designed the study and developed the theory. KB, MB, GK, WPH developed the algorithms. KB developed the software. AC, KCM, TC, LC, CBW, YHG, JOG participated in the sequencing experiments. KB, KCM, DRM, TC, RSL, YHG, JOG, MB, WPH contributed to the interpretation of the results. KB, MB, WPH wrote the manuscript. All authors contributed to the final version of the manuscript.

Competing interests

JOG received financial support for attending ONT and other conferences and an honorarium for speaking at ONT headquarters. JOG received funding and consumable support from ONT for TC's PhD studentship. TC received financial support from ONT for attending an international conference.

[9]CNRS/LIGM Université Paris-Est, Marne-la-Vallée, France

[10]Skolkovo Institute of Science and Technology, Moscow, Russia

[11]Quadram Institute Bioscience, Norwich Research Park, Norwich, UK

Abstract

Surveillance of drug-resistant bacteria is essential for healthcare providers to deliver effective empiric antibiotic therapy. However, traditional molecular epidemiology does not typically occur on a timescale that could impact patient treatment and outcomes. Here we present a method called ‘genomic neighbor 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 *S. pneumoniae* and *N. 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 determination of resistance within ten minutes (sens/spec 91%/100% for *S. pneumoniae* and 81%/100% for *N. gonorrhoeae* from isolates with a representative database) of sequencing starting, and for clinical metagenomic sputum samples (75%/100% for *S. pneumoniae*), within four hours of sample collection. This flexible approach has wide application to pathogen surveillance and may be used to greatly accelerate appropriate empirical antibiotic treatment.

Introduction

Infections pose multiple challenges to healthcare systems, contributing to higher mortality, morbidity, and escalating cost. Clinicians must regularly make rapid decisions on empiric antibiotic treatment of infectious syndromes without knowing the causative pathogen(s) 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 minutes delay in initiating effective treatment[1].

The molecular epidemiology of infectious disease allows us to identify high-risk pathogens and 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 is promising but introduces a challenge: 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 well, as a result of genetic linkage between resistance elements and the rest of the genome. An example is given by the pneumococcus (*Streptococcus pneumoniae*). The Centers for Disease Control have 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 type[5], 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 population structure. This population structure could be leveraged to offer an alternative approach to detecting resistance in which 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[6,7,16–21,8–15] (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 neighbor typing' which 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 as to the pathogen's phenotype. We demonstrate this by identifying drug-resistant and drug-susceptible clones for both *Streptococcus pneumoniae* (the pneumococcus) and *Neisseria gonorrhoeae* (the gonococcus), within minutes after the start of sequencing using Oxford Nanopore Technology. The method has many potential applications, depending on the specific pathogen and 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 Operation Characteristic Curve (AUC) (SI Fig. 1). First, we applied the method to 616 pneumococcal genomes from a carriage study in Massachusetts children[23,24]. Second, we used 1102 clinical gonococcal isolates collected from 2000 to 2013 by the Centers for Disease Control and Prevention's Gonococcal Isolate Surveillance Project[25]. In both cases, the datasets comprised draft genome assemblies from Illumina HiSeq reads, resistance data, and lineages inferred from sequence cluster computed using Bayesian Analysis of Population Structure (BAPS)[26]. Lineages of *S. pneumoniae* are predictive for benzylpenicillin, ceftriaxone, trimethoprim-sulfamethoxazole, erythromycin, and tetracycline resistance with AUC ranging from 0.90 to 0.97 (SI Fig. 2), consistent with previous works[5]. In *N. gonorrhoeae*, ciprofloxacin, ceftriaxone, and cefixime attained comparably large AUCs (from 0.93 to 0.98) whereas azithromycin demonstrated lower association (AUC 0.80) (SI Fig. 3), as observed previously[25].

Rapid identification of nearest known relative from sequencing reads

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

To implement genomic neighbor typing we developed software called RASE (Resistance-Associated Sequence Elements) (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], a metagenomic classifier implementing a fast and memory-efficient exact colored de Bruijn graph data structure[29] using a BWT index[30] (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 Kraken[31] and Kallisto[32]). These weights are cumulative scores capturing sample-to-reference similarity; they are set to zero at the beginning and are increased on-the-fly as sequencing proceeds according to each read's 'information content' (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 non-specific and have low scores. Weights serve as a proxy to 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 neighbors. Comparison of their weights provides a 'susceptibility score', which quantifies the risk of resistance (Methods). When the weights are too similar, the call's confidence is considered low; this happens when resistant and susceptible strains are insufficiently genetically distinct, which is often the case for resistance emerging recently 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 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 an antibiotic-specific resistance categories using the European Committee on Antimicrobial Susceptibility Testing (EUCAST) breakpoints[33] and the CDC Gonococcal Isolate Surveillance Project (GISP) breakpoints[34], respectively (Methods). Where MIC data were unavailable or insufficiently specific, we estimated the likely resistance phenotype using ancestral state reconstruction (Methods, SI Note 1, Extended Data Fig. 3, and Extended Data

Fig. 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 with a *k*-mer length optimized to minimize prediction delays (*k*=18, Methods). The obtained pneumococcal and gonococcal RASE databases occupy 321 MB and 242 MB RAM (4.3× and 12× compression rate) and can be further compressed for transmission to 47 MB and 32 MB (29× and 90x compression rate), respectively (Extended Data Fig. 1). This would allow RASE to be used on portable devices and its databases 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 (Table 1a, sens/spec 100%/100%, n=10) to test 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 minute and 7 minutes respectively. A multidrug-resistant isolate (SP02) was predicted even faster, with both lineage and the sequenced strain correctly detected and stabilized within 1 minute. To compare with gene-based approaches for detecting resistance[22] we evaluated how long it took for resistance genes to be sequenced on the device, and observed that at least 25 minutes would be needed for single copies to be detected (SI Note 2, Extended Data Fig. 2).

We then performed a similar evaluation with five gonococcal isolates from the database (Table 2a, sens/spec 57%/100%, n=20); 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 minutes of sequencing. We then sequenced an isolate with an uncommon mechanism of cephalosporin resistance that has emerged recently (GC02)[35]. Under such circumstances, the resistant strain and its susceptible neighbors tend to be genetically very similar, which could confound our analysis. However, RASE was still able to identify the correct resistance phenotypes in 9 minutes, with the delay being due to difficulty distinguishing between the close relatives, reflected also 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 resistance phenotype (Methods), which is a feature of our approach intended to draw operators' attention and indicate that further testing is necessary. In this experiment, RASE also resolved sample mislabeling (SI Note 3). For a multidrug-resistant isolate (GC04) RASE predictions stabilized within 2 minutes 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 µg/mL), whereas the best match in the database had an MIC of 0.062 µg/mL, within a single doubling dilution. We further found that RASE performed well even with extremely poor data and low-quality reads (GC05, SI Note 4). We also evaluated how genomic neighbor typing would perform if RASE used Kraken[31] instead of ProPhyle[27,28] (SI Note 5).

RASE identifies the closest relative of previously untested isolates

We next examined four pneumococcal isolates (Table 1b, sens/spec 89%/100%, n=20) for which the serotype and limited antibiogram and lineage data were known. We compared three characteristics of the sample to assess our performance: the serotype, the MLST sequence type, and the antibiograms (benzylpenicillin, ceftriaxone, trimethoprim-sulfamethoxazole, erythromycin, and tetracycline resistance according to the EUCAST breakpoints[33]).

In all cases, the closest relative was identified within 5 minutes, even if the correct MLST sequence type was absent from the RASE database (an example 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 PMEN, a clone strongly associated with macrolide resistance[36]. 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 penicillin resistant, but erythromycin susceptible. This illustrates the importance of a relevant database.

We evaluated RASE with 14 clinical gonococcal isolates from the RaDAR-Go project[37] (Switzerland, 2015–2016) (Table 2b, sens/spec 93%/100%, n=56). These isolates were previously sequenced using nanopore and have full antibiograms available[38]. The 55/56 correct calls indicate the strength of the genomic neighbor typing in a clinical setting. The only incorrect call (susceptibility to azithromycin in 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 using a database from one region (the US) to predict phenotype in another (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 neighbor typing is a comprehensive and relevant reference database. To evaluate RASE performance in a setting with an incomplete database, we used the gonococcal WHO 2016 reference strain collection[39]. 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 multidrug 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 nanopore[38] and genetically and phenotypically characterized[39]. Surprisingly, RASE correctly identified all MLST sequence types represented in the database and in 7 cases it provided fully correct resistance phenotypes (Supplementary Table 1, sens/spec 67%/91%, n=56). In 6/7 cases where the complete resistance profile was not recovered, the closest relatives were identified correctly but were genetically divergent from the query isolates (SI 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 analyzed metagenomic nanopore data from sputum samples obtained from patients suffering from lower respiratory tract infections[40] (UK, 2017), selecting 6 samples from the study that were already known to contain *S. pneumoniae* (Table 1c, sens/spec 75%/100%, n=16).

One sample (SP10) contained DNA from multiple bacterial species. However, within 5 minutes sequence was identified belonging to the Swedish 15A-25 clone (ST63) which is also known to be associated with resistance phenotypes including macrolides and tetracyclines[41]. This sample was confirmed to be resistant to erythromycin, as well as clindamycin, tetracycline and oxacillin according to the EUCAST breakpoints[33]. The original report of the Swedish 15A-25 clone did not report resistance to penicillin antibiotics[41], 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 contain an estimated >20% reads that matched to *S. pneumoniae*, and their serotypes were identified to be 15A and 3, 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 (SI Note 7). The last remaining samples, SP07–SP09, contained less than 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 that we term genomic neighbor typing to pinpoint the closest relatives of a query genome within a suitable database and then to infer the phenotypic properties 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 neighbor typing at an earlier stage offers a way of leveraging bacterial population structure to gain information on resistance and susceptibility, and inform 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, harbors 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 for the application.

The initial identification of the closest relative is consistently more robust in the pneumococcus than the gonococcus, as a result of the former having more *k*-mers that are specific to an individual lineage, 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 fluctuates between them, even though it correctly identifies the region of the phylogeny. If these genomes vary in their resistance profile, this is properly reflected in an uncertain susceptibility score indicating that caution and further investigation are merited (e.g., GC03).

As in all inference, the principle limitation of genomic neighbor typing is the representativeness of the database. While we have made use of relatively small samples from limited geographic areas to demonstrate proof of principle, in practice there are multiple examples of large genome databases generated by public health agencies, which could be combined with metadata on resistance for genomic neighbor 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 which appear to show promise for this approach. These include *E. coli*, in which data on MLST type supplemented with epidemiologic information can consistently produce AUCs in excess of 0.90 for multiple antibiotics[42], suggesting great potential for neighbor typing to offer excellent resolution superior to MLST. However, we anticipate that genomic neighbor typing may be less suitable in case of a little within-species genomic variation (meaning it is hard to identify the nearest neighbor) or when resistance emerges rapidly on independent and diverse genomic backgrounds (meaning resistance is poorly correlated with those backgrounds) (SI Note 8).

In the case 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 choice of the correct database for prediction is key. Doing this will likely require a two-step solution in which the reads are first passed through a metagenomic classifier such as Centrifuge[43] or MetaMaps[44], 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, taking a total of 4 hours. This is a rapidly evolving area of technology and automated rapid library preparation kits are already in development[45]. 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, neighbor 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: by measuring a sparse signal distributed broadly across our data we can identify it with comparatively few error-tolerant measurements.

Genomic neighbor 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 facilities[2]. Finally, at present empiric 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 neighbor typing is a natural complement to this process with 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 neighbor 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 neighbor 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 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 (i.e., minimizing the number of nodes switching its resistance category; Extended Data Fig. 3 and Extended Data Fig. 4). When the solution for a node is not unique, non-susceptibility is assigned.

Genomic neighbor 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 read's nearest neighbors in the database according to the 1/('number of matched k-mers') pseudodistance.

The weights of the nearest neighbors 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 neighbors. Reads that do not match (i.e., 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. KS helps to assess whether a sample is truly matching the database and predicting resistance for the database species makes sense.

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

Predicting lineage

A lineage is predicted as the lineage of the best matching reference strain, i.e., 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 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.

LS is used to measure how well a sample matching the identified lineage. If 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 operators' attention. Note that custom RASE databases may require a re-calibration of the threshold.

Predicting resistance and susceptibility

Resistance or susceptibility are predicted for individual antibiotics independently, based 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 pre-defined thresholds.

The susceptibility score 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. The values of 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.

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

S. pneumoniae RASE database

The *S. pneumoniae* RASE database was constructed with the EUCAST breakpoints[33] ([mg/L]): ceftriaxone (CRO): 0.25, erythromycin (ERY): 0.25, benzylpenicillin (PEN): 0.06, trimethoprim-sulfamethoxazole (SXT): 1.00, and tetracycline (TET): 1.00. While we have used the above values in the present work, others may be readily defined and the database rapidly updated. This is especially useful in the case where breakpoints may vary depending on the site of infection (as is the case with pneumococcal meningitis and otitis media, where lower MICs are considered to be resistant[33]).

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: '10.5061/dryad.t55gq'). The pneumococcal ProPhyle index was constructed with the *k*-mer size *k*=18.

The obtained *S. pneumoniae* RASE database including the source code and data is available from https://github.com/c2-d2/rase-db-spneumoniae-sparc.

N. gonorrhoeae RASE database

The *N. gonorrhoeae* RASE database was constructed with the CDC GISP breakpoints[34] ([mg/L]): azithromycin (AZM): 2.0, cefixime (CFM): 0.25, ciprofloxacin (CIP): 1.0, and ceftriaxone (CRO): 0.125. Before applying the breakpoints, azithromycin MICs for strains collected before 2005 were doubled in order to correct for the known inconsistencies of the phenotyping protocol due to a change in formulation of the commercial media[47].

The draft assemblies and the phylogenetic tree were downloaded from Zenodo (accession: '10.5281/zenodo.2618836'). Three prevalent types of plasmids[48] were downloaded from GenBank, localized in the GISP database using BLAST[49], and removed from the dataset:

the cryptic plasmid (‘pJD1’, GenBank accession ‘NC_001377.1’), the beta-lactamase plasmid (‘pJD4’, GenBank accession ‘NC_002098.1’), and the conjugative plasmid (‘pEP5289’, GenBank accession ‘GU479466.1’). The gonococcal ProPhyle index was constructed with the *k*-mer size *k*=18.

The obtained *N. gonorrhoeae* RASE database including the source code and data is available from 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]. ProPhyle index stores *k*-mers of all strains in a highly compressed form, reducing the required memory footprint. In the database construction phase, the strains’ *k*-mers are first propagated along the phylogenetic tree and then greedily assembled to contigs. The obtained contigs are then placed into a single text file, for which a BWT index is constructed[30].

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 window[50]. 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 format[52].

Optimizing *k*-mer length

The *k*-mer length is the main parameter of the classification. First, the subword complexity function[53] of pneumococcus was calculated using JellyFish[54] (version 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 in [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 *k* to 18.

Comparison to Kraken

For each RASE database, a fake NCBI taxonomy was generated from the database tree. Then a library was built using Kraken[31] (version 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 assignment were subsequently converted using an ad-hoc Python script to RASE-BAM (a subset of the BAM format[52] 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 extracted were 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 (i.e., the gonococcal WHO 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 1.43Mbps per second. 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=15mins), 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 CANU[55] (version 1.5, default parameters). Prior to the assembly step, reads were filtered using SAMsift[56] based on the matching quality with the pneumococcal RASE database: only reads at least 1000bp long with at least 10% 18-mers shared with some of the reference draft assemblies were used. The obtained assembly was further corrected by Pilon[57] (version 1.2, default parameters) using Illumina reads from the same isolate (taxid '1QJAP' in the SPARC dataset[24]) mapped to the nanopore assembly using BWA-MEM[58] (version 0.7.17, default parameters) and sorted using SAMtools[52].

The obtained assembly was searched for resistance-causing genes using the online CARD tool[16] (as of 2018/08/01). All of the original nanopore reads were then mapped using Minimap2[59] (version 2.11, with '-x map-ont') to the corrected assembly and resistance genes in the reads identified using BEDtools–intersect[60] (version 2.27.1, with '-F 95'). Timestamps of the resistance-informative reads were extracted and associated with the genes. Only reads longer than 2kbp 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 data set comprising both the GISP isolates and the WHO isolates. First, reads were downloaded for the GISP isolates (NCBI BioProject: 'PRJEB2999' and 'PRJEB7904') and for the WHO isolates F–P (NCBI BioProject: 'PRJEB4024'). For the WHO isolates U–Z, read data were simulated from the finished de-novo assemblies (NCBI BioProject: 'PRJEB14020') using Art-Illumina[61] (version 2.5.1). Reads were mapped to the NCCP11945 reference genome (GenBank accession: 'CP001050.1') using BWA-MEM[58] (version 0.7.17) and deduplicated using Picard[62] (version 2.8.0). Pilon[57] (version 1.16, with '--mindepth 10 --minmq 20') was used to call variants and further filtered to include only 'pass' sites and sites where the alternate allele was supported with AF > 0.9. Gubbins[63] (version 2.3.4) with RAxML[64] (version 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 and Company, Sparks, MD) at 37°C in 5% CO2 for 24 hrs. High-molecular-weight (>1 μg) genomic DNA was extracted and purified from cultures using DNeasy Blood and Tissue kit (QIAGEN, Valencia CA). DNA concentration was measured using Qubit fluorometer (Invitrogen, Grand Island NY). Library preparation was performed using the Oxford Nanopore Technologies 1D ligation sequencing kit SQK LSK108.

For experiments SP07-SP12, library preparation was performed using the 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 manuscript[40].

For isolates GC01-GC05, cultures were grown on Chocolate-Agar media i.e., Difco GC base media containing 1% IsoVitaleX (Becton Dickinson Co., Franklin Lakes, NJ) and 1% Remel Hemoglobin (Thermo Fisher Scientific, Carlsbad, CA) at 37°C in 5% CO2 for 20 hrs. For GC01-GC04 genomic DNA was extracted and purified from cultures using the PureLink Genomic DNA MiniKit (Thermo Fisher Scientific, Carlsbad, CA), and for GC05 DNA was extracted using the phenol-chloroform method[65]. Genomic DNA was extracted and purified from cultures using the PureLink Genomic DNA MiniKit (Thermo Fisher Scientific, Carlsbad, CA). DNA concentration was measured using the Qubit fluorometer (Invitrogen, Grand Island, NY). Library preparation was performed using the Oxford Nanopore Technologies 1D ligation sequencing kit SQK-LSK109.

MinION sequencing

Sequencing was performed on the 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 (versions 1.6.11 (SP01), 1.7.3 (SP02), 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 (versions 1.4-1.13.1) was used to collect raw sequencing data and ONT Albacore (versions 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 (version 2.3.4) was used for local base-calling.

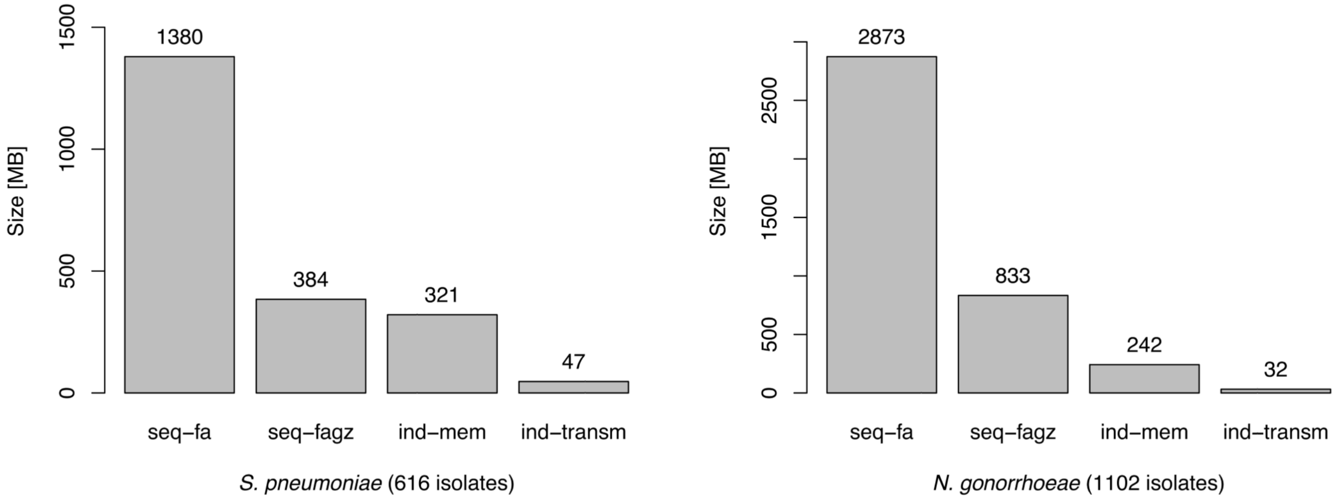
Testing resistance phenotype

Additional retesting of SPARC isolates was done using microdilution. Organism suspensions were prepared from overnight growth 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 CFU/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)[66,67]. Penicillin (TRC Canada) and chloramphenicol (USB) concentrations ranged from 0.016 to 16 μg/mL. Erythromycin (Enzo Life Sciences), tetracycline (Sigma-Aldrich), and trimethoprim-sulfamethoxazole (MP Biomedicals) concentrations ranged from 0.0625 to 64 μg/mL. Ceftriaxone (Sigma-Aldrich) concentrations ranged from 0.007 to 8 μg/mL. 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 breakpoints[33]. First, the inoculated agar plates were incubated at 37 °C overnight and then examined for growth with the potential for re-incubation up to 48 hours. Then, the samples were screened to oxacillin: if the zone diameter r was >20mm, the isolate was considered sensitive to benzylpenicillin, otherwise a full MIC measurement to benzylpenicillin was done. Finally, the isolate was screened for resistance to tetracycline ($r \geq 25$mm for sensitive, $r < 22$mm for resistant) and erythromycin ($r \geq 22$mm for sensitive, $r < 19$mm for resistant); when the isolate showed intermediate resistance, a full MIC measurement was done.

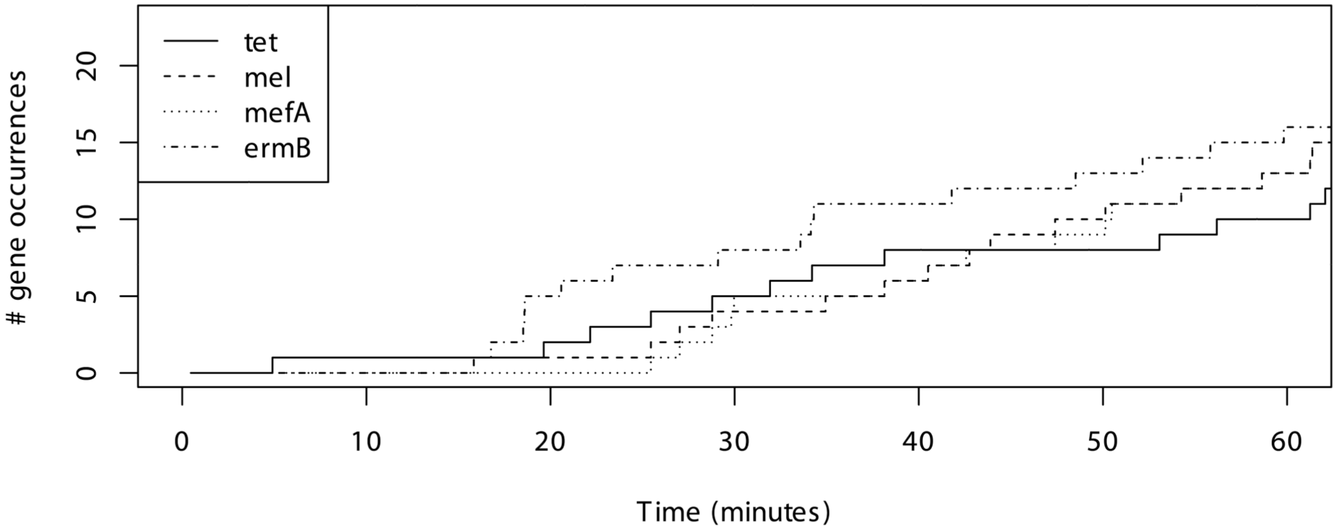
Results for all tested samples – isolates and metagenomes – are summarized in Supplementary Table 3.

Extended Data



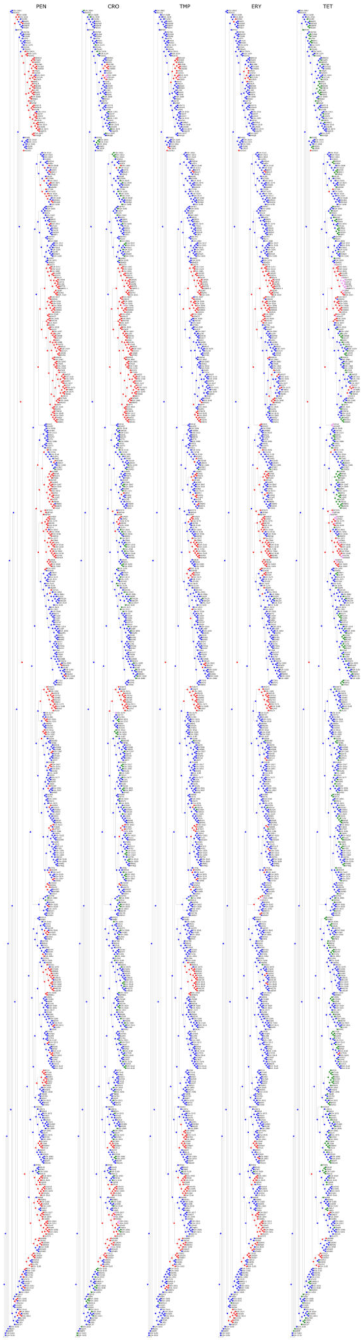
Extended Data Fig. 1. Size and memory footprint of the *S. pneumoniae* and *N. gonorrhoeae* RASE databases

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)



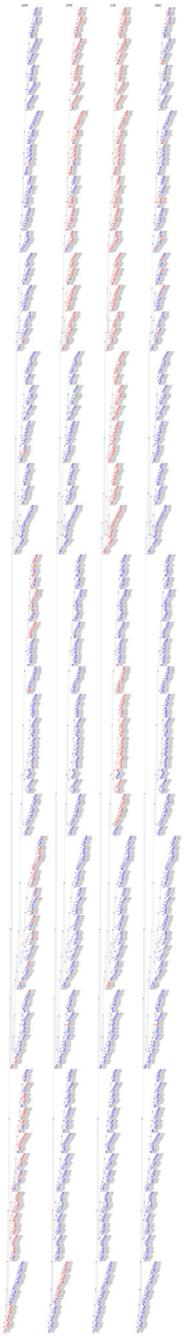
Extended Data Fig. 2. Timeline of resistance genes

Number of occurrences of individual resistance genes in reads of SP02, as a function of time for the first hour of nanopore sequencing.



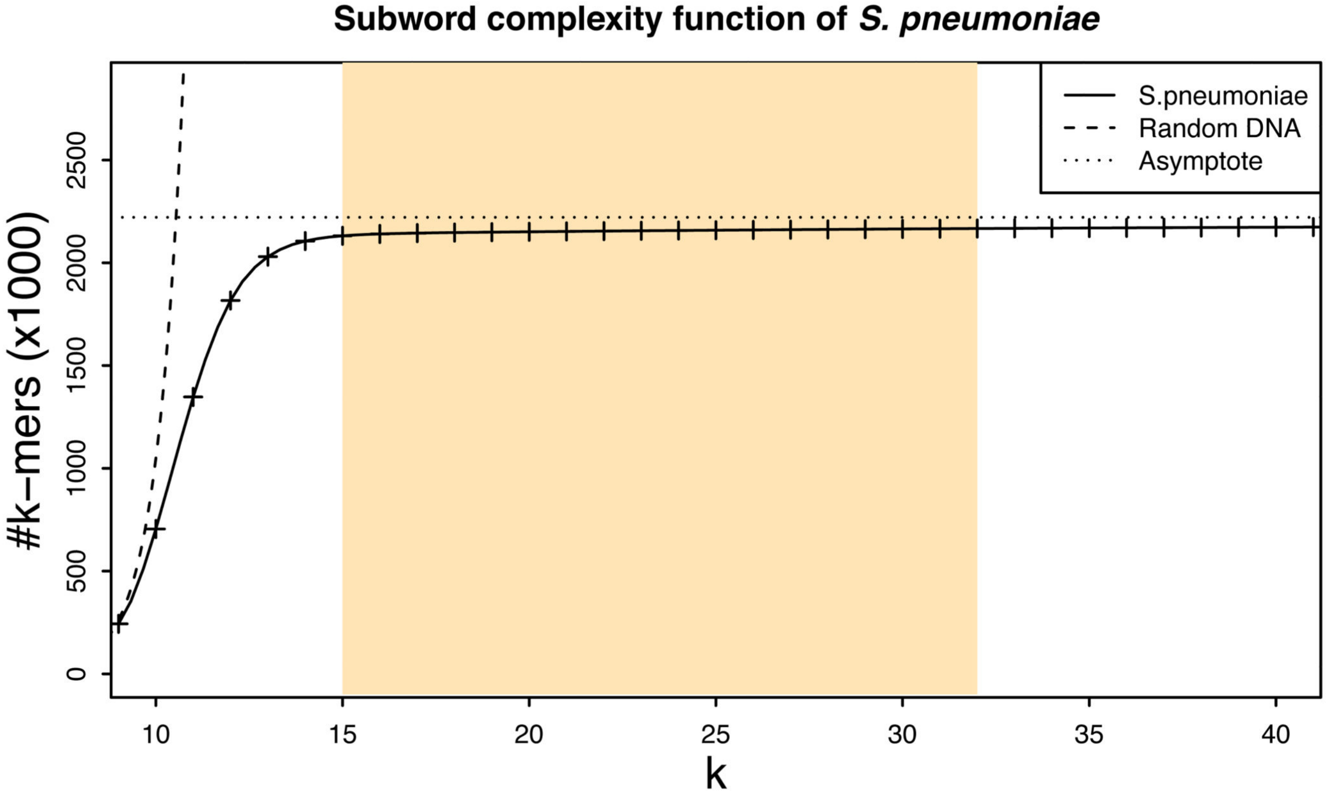
Extended Data Fig. 3. Ancestral state reconstruction of resistance categories in the *S. pneumoniae* RASE database

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).



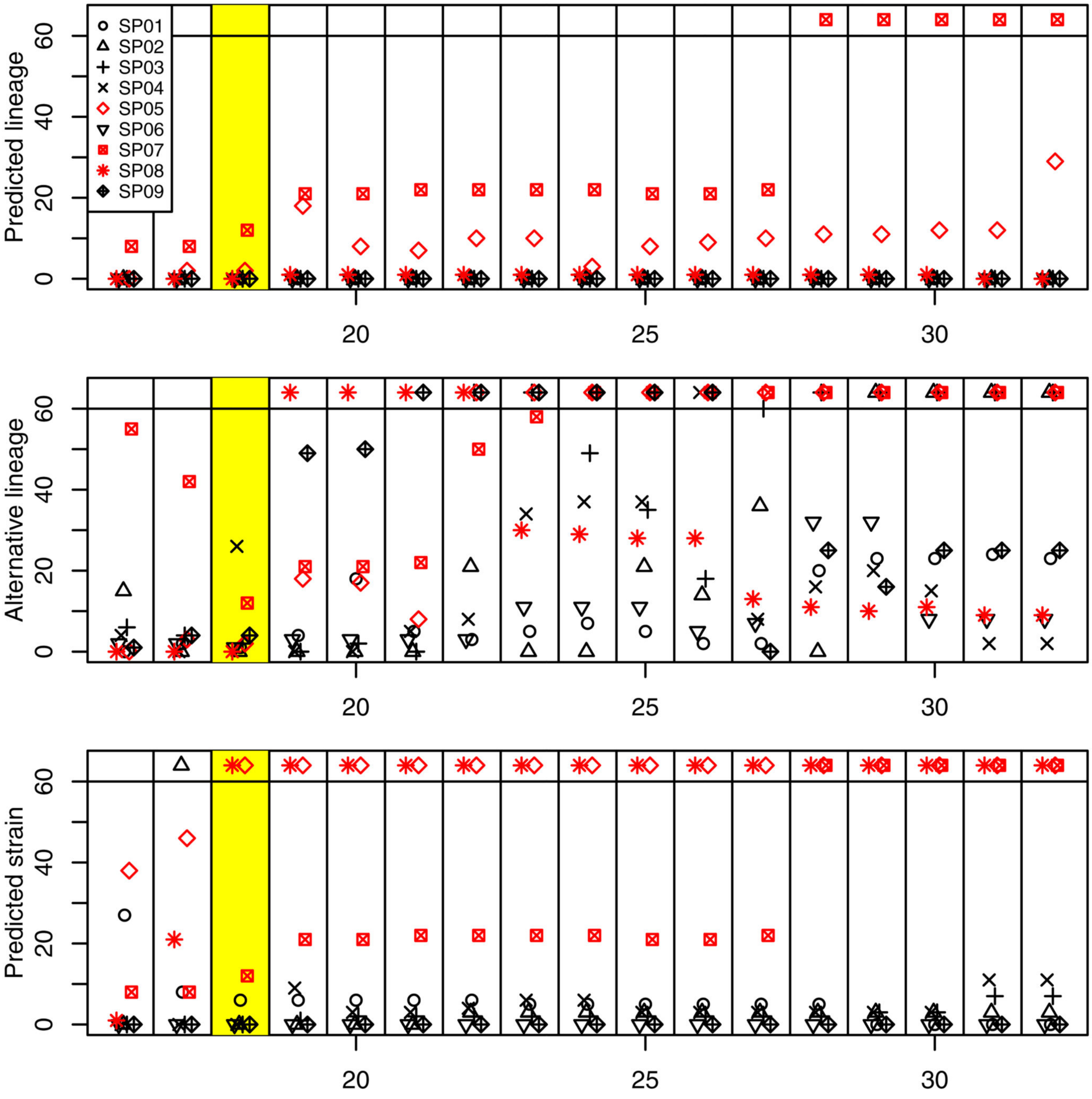
Extended Data Fig. 4. Ancestral state reconstruction of resistance categories in the *N. gonorrhoeae* RASE database

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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the Bill & Melinda Gates Foundation (GCGH GCE OPP1151010, KB and WPH), NIH – National Institute of Allergy and Infectious Diseases (R01 AI106786-05, KB), the Canadian Institutes of Health Research (MFE 152448, RSL; a fellowship grant, DRM), NSF (GRFP, KCM), and the David and Lucile Packard Foundation (MB). This paper presents independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (Reference Number RP-PG-0514-20018, JOG), the UK Antimicrobial Resistance Cross Council Initiative (MR/N013956/1, JOG), Rosetrees Trust (A749, JOG), 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, JOG), the University of East Anglia (JOG, TC), and Oxford Nanopore Technologies (JOG, TC). 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. The authors thank Joshua 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 NIH (R01 AI46645), and to Brian J Arnold, Taj Azarian and Cristina M Herren for useful comments in various stages of this project.

Bibliography

1. 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. 2006; 34:1589–1596. [PubMed: 16625125]
2. Quick J, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016; 530:228–232. [PubMed: 26840485]
3. Faria NR, et al. Mobile real-time surveillance of Zika virus in Brazil. Genome Med. 2016; 8:2–5. [PubMed: 26739746]
4. CDC. Antibiotic resistance threats in the United States, 2013. Current. 2013:114.
5. Li Y, et al. Penicillin-Binding Protein Transpeptidase Signatures for Tracking and Predicting β-Lactam Resistance Levels in Streptococcus pneumoniae. MBio. 2016; 7:e00756–16. [PubMed: 27302760]
6. Zankari E, et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012; 67:2640–2644. [PubMed: 22782487]
7. Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S. KvarQ: Targeted and direct variant calling from fastq reads of bacterial genomes. BMC Genomics. 2014; 15:1–12. [PubMed: 24382143]
8. Gupta SK, et al. ARG-ANNOT, a New Bioinformatic Tool To Discover Antibiotic Resistance Genes in Bacterial Genomes. Antimicrob Agents Chemother. 2014; 58:212–220. [PubMed: 24145532]
9. 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. 2015; 10:e0133492. [PubMed: 26197475]
10. Kaminski J, et al. High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. PLOS Comput Biol. 2015; 11:e1004557. [PubMed: 26682918]
11. de Man TJB, Limbago BM. SSTAR, a Stand-Alone Easy-To-Use Antimicrobial Resistance Gene Predictor. mSphere. 2016; 1:1–10.
12. Clausen PTLC, Zankari E, Aarestrup FM, Lund O. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. J Antimicrob Chemother. 2016; 71:2484–2488. [PubMed: 27365186]
13. Yang Y, et al. ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database. Bioinformatics. 2016; 32:2346–2351. [PubMed: 27153579]
14. Garner E, Pruden A, Heath LS, Vikesland P. DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. 2017

15. Antonopoulos DA, et al. PATRIC as a unique resource for studying antimicrobial resistance. Brief Bioinform. 2017; :1–9. DOI: 10.1093/bib/bbx083 [PubMed: 26868358]

16. Jia B, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. 2017; 45:D566–D573. [PubMed: 27789705]

17. 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. 2017; 72:2764–2768. [PubMed: 29091202]

18. Hunt M, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Microb Genomics. 2017; 3:1–21.

19. Votintseva AA, et al. Same-day diagnostic and surveillance data for tuberculosis via whole genome sequencing of direct respiratory samples. J Clin Microbiol. 2017; doi: 10.1128/JCM.02483-16

20. Rowe WPM, Winn MD. Indexed variation graphs for efficient and accurate resistome profiling. Bioinformatics. 2018; 34:3601–3608. [PubMed: 29762644]

21. Feldgarden M, et al. Using the NCBI AMRFinder Tool to Determine Antimicrobial Resistance Genotype-Phenotype Correlations Within a Collection of NARMS Isolates. bioRxiv. 2019; doi: 10.1101/550707

22. Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. Nat Rev Genet. 2019; doi: 10.1038/s41576-019-0108-4

23. Croucher NJ, et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet. 2013; 45:656–63. [PubMed: 23644493]

24. Croucher NJ, et al. Population genomic datasets describing the post-vaccine evolutionary epidemiology of Streptococcus pneumoniae. Sci data. 2015; 2

25. Grad YH, et al. Genomic Epidemiology of Gonococcal Resistance to Extended-Spectrum Cephalosporins, Macrolides, and Fluoroquinolones in the United States, 2000–2013. J Infect Dis. 2016; 214:1579–1587. [PubMed: 27638945]

26. Cheng L, Connor TR, Siren J, Aanensen DM, Corander J. Hierarchical and Spatially Explicit Clustering of DNA Sequences with BAPS Software. Mol Biol Evol. 2013; 30:1224–1228. [PubMed: 23408797]

27. Břinda K, Salikhov K, Pignotti S, Kucherov G. ProPhyle: An accurate, resource-frugal and deterministic DNA sequence classifier. 2017; doi: 10.5281/zenodo.1045429

28. Břinda, K. Novel computational techniques for mapping and classifying Next-Generation Sequencing data. PhD Thesis, Université Paris-Est; 2016.

29. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet. 2012; 44:226–232. [PubMed: 22231483]

30. Ferragina, P; Manzini, G. Opportunistic data structures with applications. Proceedings 41st Annual Symposium on Foundations of Computer Science; IEEE Comput. Soc; 2000. 390–398.

31. Wood DE, Salzberg SL. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014; 15

32. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016; 34:525–527. [PubMed: 27043002]

33. The European Committee on Antimicrobial Susceptibility Testing. Breakpoint tables for interpretation of MICs and zone diameters. Version 7.0. 2017.

34. Kirkcaldy RD, et al. Neisseria gonorrhoeae Antimicrobial Susceptibility Surveillance — The Gonococcal Isolate Surveillance Project, 27 Sites, United States, 2014. MMWR Surveill Summ. 2016; 65:1–19.

35. Palace S, et al. RNA polymerase mutations cause cephalosporin resistance in clinical Neisseria gonorrhoeae isolates. bioRxiv. 2019; doi: 10.1101/626457

36. McGee L, et al. Nomenclature of Major Antimicrobial-Resistant Clones of Streptococcus pneumoniae Defined by the Pneumococcal Molecular Epidemiology Network. J Clin Microbiol. 2001; 39:2565–2571. [PubMed: 11427569]

37. 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. 2018; 56:1–10.

38. Golparian D, et al. Antimicrobial resistance prediction and phylogenetic analysis of Neisseria gonorrhoeae isolates using the Oxford Nanopore MinION sequencer. Sci Rep. 2018; 8

39. 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. 2016; 71:3096–3108. [PubMed: 27432602]

40. Charalampous T, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat Biotechnol. 2019; doi: 10.1038/s41587-019-0156-5

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. 2000; 182:1153–60. [PubMed: 10979912]

42. MacFadden DR, 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. 2019; 57:1–9.

43. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016; 26:1721–1729. [PubMed: 27852649]

44. Dilthey A, Jain C, Koren S, Phillippy A. MetaMaps - Strain-level metagenomic assignment and compositional estimation for long reads. bioRxiv. 2018; doi: 10.1101/372474

45. Quick, J. Ultra-long read sequencing protocol for RAD004 (Version 3). 2018.

46. MacFadden DR, Leis JA, Mubareka S, Daneman N. The Opening and Closing of Empiric Windows: The Impact of Rapid Microbiologic Diagnostics. Clin Infect Dis. 2014; 59:1199–1200. [PubMed: 24965352]

47. Kersh EN, et al. Rationale for a Neisseria gonorrhoeae Susceptible Only Interpretive Breakpoint for Azithromycin. Clin Infect Dis. 2019; 30329:1–7.

48. Cehovin A, Lewis SB. Mobile genetic elements in Neisseria gonorrhoeae: movement for change. Pathog Dis. 2017; 75:1–12.

49. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215:403–410. [PubMed: 2231712]

50. Salikhov, K. Efficient algorithms and data structures for indexing DNA sequence data. PhD Thesis, Université Paris-Est; 2017.

51. Břinda K, Salikhov K, Pignotti S, Kucherov G. ProPhex: A lossless k-mer index based on the Burrows-Wheeler Transform. 2018; doi: 10.5281/zenodo.1247431

52. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–9. [PubMed: 19505943]

53. Lothaire, M. Algebraic Combinatorics on Words. Cambridge University Press; 2002.

54. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011; 27:764–770. [PubMed: 21217122]

55. Koren S, et al. Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation. Genome Res. 2017; 27:722–736. [PubMed: 28298431]

56. Břinda K, Baym M, Hanage WP. SAMsift: advanced filtering and tagging of SAM/BAM alignments using Python expressions. 2018; doi: 10.5281/zenodo.1048211

57. Walker BJ, et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014; 9

58. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013:3.

59. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. :1–3.2018; DOI: 10.1093/bioinformatics/bty191

60. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842. [PubMed: 20110278]

61. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012; 28:593–594. [PubMed: 22199392]

62. Broad Institute; GitHub: 2018. Picard Toolkit. http://broadinstitute.github.io/picard/

63. Croucher NJ, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015; 43:e15. [PubMed: 25414349]

64. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30:1312–1313. [PubMed: 24451623]

65. Green MR, Sambrook J. Isolation of High-Molecular-Weight DNA Using Organic Solvents. Cold Spring Harb Protoc. 2017; 2017

66. CLSI. Susceptibility Tests for Bacteria That Grow Aerobically; Approved Standard—Ninth Edition CLSI document M07-A9. 2012.

67. CLSI. Performance Standards for Antimicrobial Susceptibility Testing; Twenty-Second Informational Supplement CLSI document M100-S22. 2012.

68. Tange O. GNU Parallel: the command-line power tool. login USENIX Mag. 2011; 36:42–47.

69. Köster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics. 2012; 28:2520–2522. [PubMed: 22908215]

70. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, analysis and visualization of phylogenomic data. Mol Biol Evol. 2016; 33

71. Grüning B, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018; 15:475–476. [PubMed: 29967506]

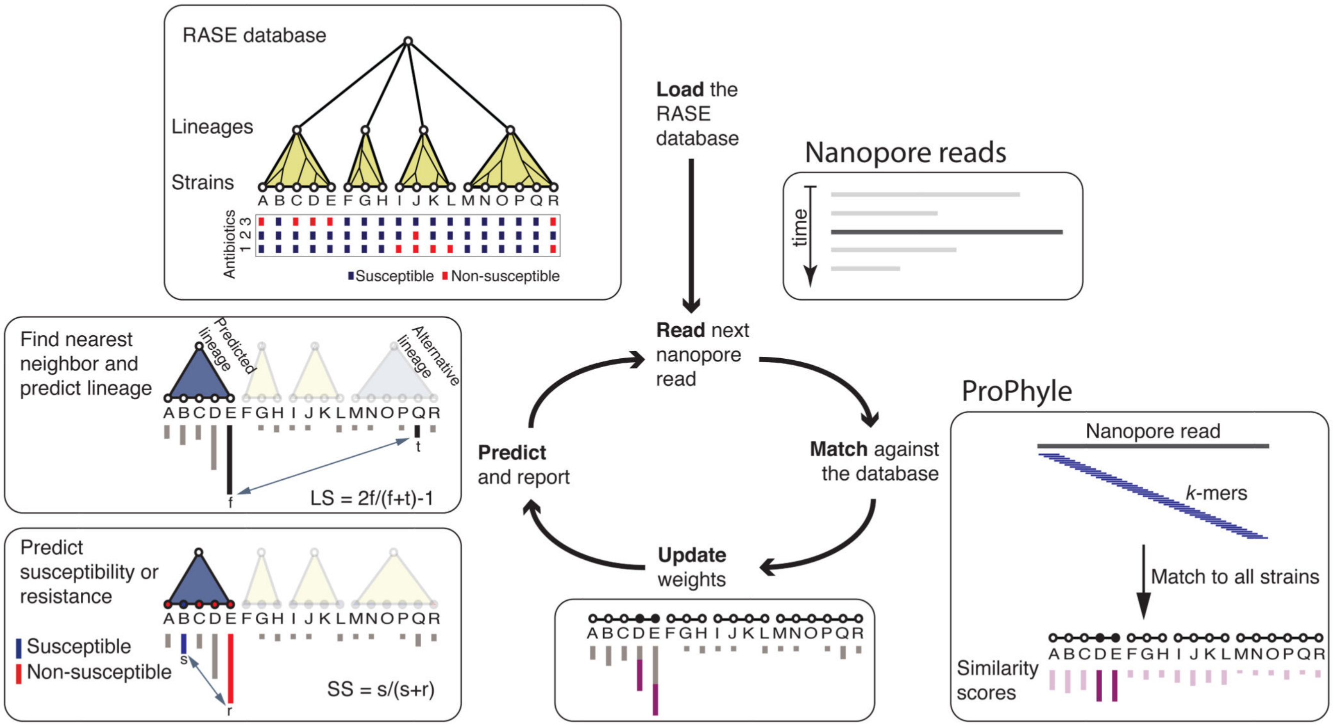


Fig. 1. Overview of the RASE approach.

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 proportionate to the number of matching *k*-mers. Finally, resistance is predicted from the obtained weights and the resistance profiles of the database strains as follows: 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 that strain's known properties (e.g., the original antibiograms, MLST sequence type, or serotype).

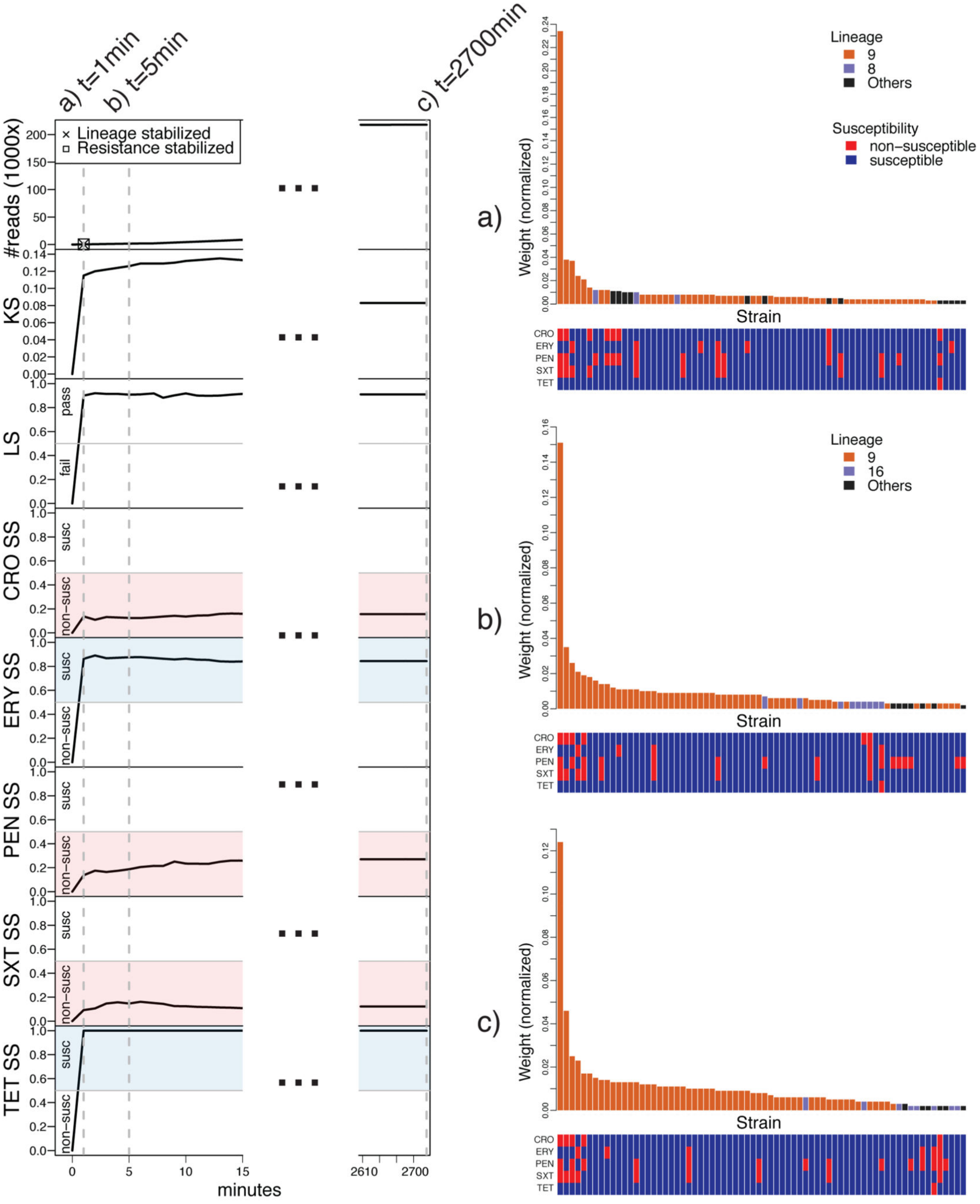


Fig. 2. RASE obtains stable predictions of antibiotic resistance or susceptibility and lineage within minutes for an isolate of a pneumococcal 23F clone (SP06).

Left: Number of reads, lineage score (LS), *k*-mer score (KS), and susceptibility scores (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. Blue and red colors correspond to susceptibility and non-susceptibility calls, respectively. The dashed lines mark selected time points (1 minute, 5 minutes, and the end of sequencing). **Right:** a)-c) Similarity rank plots for selected time points (1 minute, 5 minutes, and the end of sequencing). The bars correspond to 70 best matching strains in the database and display the normalized weights, which serve as a proxy to inverted genetic distance. They are arranged by rank and colored according to the presence in the predicted, alternative or another lineage. The bottom panels display the resistance profiles of the strains.

Table 1

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

The table displays actual and predicted resistance phenotypes (S = susceptible, R = non-susceptible) for individual experiments, as well as information on match of the predicted MLST sequence type and clonal complex. Resistance categories in bold were inferred using ancestral reconstruction and were also confirmed using phenotypic testing (see Methods and Supplementary Table 2). Metagenomic samples are sorted by the estimated fraction of *S. pneumoniae* reads.

a) Database isolates

Sample	Lineage confidently detected	Matched k-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	S	S	S	S	S	S	S	S	**S**[1]	**S**[1]	Yes	Yes
SP02	yes	9.6%	19A	19A	R	R	R	R	R	R	R	R	**R**[2]	**R**[2]	Yes	Yes

b) Non-database isolates

Sample	Lineage confidently detected	Matched k-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	R	R	R	**S**[3]	R	R	R	R	S	S	OoD	Yes
SP04	yes	12%	19A	19A	R	R	R	R	R	R	R	R	R	**R**[4]	OoD	Yes
SP05	no	1.8%	19F	19F	R	R	R	R!	R	R	R	R!	R	R!	OoD	Yes
SP06	yes	8.3%	23F	23F	R	R	R	**S**[3]	R	R	R	R	S	S	OoD	Yes

c) Metagenomes

Sample	Lineage confidently detected	SP	Matched k-mers	Antibiogram ERY		Antibiogram PEN		Antibiogram TET	
				Actual	Best match	Actual	Best match	Actual	Best match
SP07	no	2.3%	0.2%	NA	S	S	S	R	**S**[5]
SP08	no	2.5%	0.9%	S	S	S	S!	S	**S**[6]
SP09	no	4.0%	1.2%	NA	S	S	S	S	**S**[7]
SP10	yes	21%	5.2%	R	R	R	R	R	**R**[8]
SP11	yes	70%	14%	R	R	R	R	R	**R**[8]
SP12	yes	86%	17%	S	S	S	S	R	**S**[5]

Legend
🟩 Correct prediction
🟧 Incorrect prediction
🟦 Cannot be evaluated

S Susceptible

R Non-susceptible

! Low-confidence call

NA Not available

OoD Out-of-database

(…) ID of a retested sample

SP Fraction of S. pneumoniae reads

Table 2

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

The table displays actual and predicted resistance phenotypes (S = susceptible, R = non-susceptible) for individual experiments, as well as information on match of the predicted MLST sequence type.

a) Database isolates

Sample	Lineage confidently detected	Matched k-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	S	S	S	S	S	S	Yes
GC02	yes	27%	S	S	R	R!	S	S	R	R!	Yes
GC03	yes	33%	S	S	R	S!	S	S	R	S!	Yes
GC04	yes	21%	S	S	R	R	R	R	R	S	Yes
GC05	yes	7%	R	R	S	S	S	S	S	S	Yes

b) Non-database isolates

Sample	Lineage confidently detected	Matched k-mers	Antibiogram AZM		Antibiogram CFM		Antibiogram CIP		Antibiogram CRO	
			Actual	Best match	Actual	Best match	Actual	Best match	Actual	Best match
GC06	yes	19%	S	S	R	R	R	R	S	S
GC07	no	20%	S	S	S	S	R	R	S	S
GC08	no	19%	S	S	S	S	R	R	S	S
GC09	no	18%	S	S	S	S	S	S	S	S
GC10	no	20%	S	S	S	S	R	R	S	S
GC11	no	20%	S	S	S	S	R	R	S	S
GC12	no	20%	S	S	S	S	R	R	S	S
GC13	yes	20%	S	S	S	S	R	R	S	S
GC14	yes	19%	S	S	S	S	R	R	S	S
GC15	yes	19%	R	S!	S	S	S	S	S	S
GC16	no	18%	S	S	S	S!	R	R	S	S!
GC17	no	19%	S	S	S	S!	R	R	S	S!
GC18	no	20%	S	S	S	S	R	R	S	S
GC19	yes	18%	S	S	S	S	R	R	S	S

Legend
Correct prediction
Incorrect prediction

S Susceptible
R Non-susceptible
! Low-confidence call