Lineage calling can identify antibiotic resistant clones within minutes

Karel Břinda[1,2], Alanna Callendrello[1], Lauren Cowley[1], Themoula Charalampous[3], Robyn S Lee[1], Derek R MacFadden[1,4], Gregory Kucherov[5,6], Justin O'Grady[7,3], Michael Baym[2,8], and William P Hanage[1]

1 Center for Communicable Disease Dynamic, Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, USA
2 Department of Biomedical Informatics, Harvard Medical School, Boston, USA
3 University of East Anglia, Norwich Research Park, Norwich, UK
4 Division of Infectious Diseases, Department of Medicine, University of Toronto, Canada
5 CNRS/LIGM Université Paris-Est, Marne-la-Vallée, France
6 Skolkovo Institute of Science and Technology, Moscow, Russia
7 Quadram Institute Bioscience, Norwich Research Park, Norwich, UK
8 Laboratory of Systems Pharmacology, Harvard Medical School, Boston, USA

Correspondence to kbrinda@hsph.harvard.edu

Introductory Paragraph

Surveillance of circulating drug resistant bacteria is essential for healthcare providers to deliver effective empiric antibiotic therapy. However, the results of surveillance may not be available on a timescale that is optimal for guiding patient treatment. Here we present a method for inferring characteristics of an unknown bacterial sample by identifying the presence of sequence variation across the genome that is linked to a phenotype of interest, in this case drug resistance. We demonstrate an implementation of this principle using sequence *k*-mer content, matched to a database of known genomes. We show this technique can be applied to data from an Oxford Nanopore device in real time and is capable of identifying the presence of a known resistant strain in 5 minutes, even from a complex metagenomic sample. This flexible approach has wide application to pathogen surveillance and may be used to greatly accelerate diagnoses of resistant infections.

Introduction

Antibiotic-resistant 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 without knowing if a patient's clinical syndrome is due to a drug resistant organism. 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].

Hence, there is interest in developing rapid, point-of-care techniques to detect the presence of a resistant strain in a sample, for diagnostics and surveillance purposes. The continuing development of sequencing technologies suggests that genomic data are particularly promising for this purpose[2]. In principle, if a resistance gene or mutation can be detected in a sample, this could be sufficient to inform treatment decisions. However for this to be applicable in practice, several conditions must be satisfied: foremost, the resistance determinant must be already identified, it must be sufficiently different from susceptible variants, and the genomic context must be known, as loci with homology to known resistance determinants are also found in non-pathogens[3]. Furthermore, to make diagnosis truly point-of-care, one must sequence as directly as possible from clinical samples, without time-consuming culture steps. This implies a metagenomic sample containing sequences from many different taxa, and so the genomic context of the resistance locus may be obscured if we use short read technologies for sequencing. An ideal approach would not depend on access to expensive, sophisticated sequencing equipment, making it deployable close to the point of care and in resource-poor settings.

The clinical question of whether an antibiotic is likely to work, i.e. the pathogen is susceptible, is not equivalent to identifying whether a pathogen carries those mutations or genes that are known to confer resistance. Prescription has long been informed by correlative features when

causative ones are difficult to measure, for example whether the same syndrome (or ideally pathogen) occurring in other patients from the same clinical environment have responded (or were susceptible to) to a particular antibiotic. This also has 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*), a major pathogen, responsible for approximately 1.6 million deaths per year. 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 and genes, approximately 90% of the variance in the minimal inhibitory concentration (MIC) for multiple antibiotics of different classes, could be explained by the loci determining the strain type alone[5]. This is particularly interesting, as none of the loci used for strain classification themselves causes resistance. Thus, in the overwhelming majority of cases, resistance can be accurately predicted from coarse strain typing based on population structure.

This population structure can be leveraged to offer an alternative approach to detecting resistance in which rather than detecting high-risk genes, we identify high-risk lineages. The additional information available from genomic data allows a better definition of those closely related parts of the population associated with resistance or susceptibility, which we call 'phylogroups'. High-risk phylogroups can be readily determined by analysis of existing high-quality draft genome assemblies, together with suitable metadata on MICs. Thus, given sufficient correlation between the phylogroup and phenotype of interest (for example drug resistance), rapid identification of the phylogroup alone can be sufficient for diagnostic purposes.

An attractive option for this approach is to use long-read sequencing, such as nanopore technology (Oxford Nanopore Technology (ONT)), given its additional correlative structure. Although the ONT MinION device has a high (~10%) per base error rate[6], it is also highly portable and deployable in field conditions[7]. Furthermore, sequencing reads are streamed the

computer as they are produced, so the results can be analyzed and reported in real time. Recently, nanopore sequencing has been shown to provide rapid re-identification of human samples within minutes[8], predict antibiotic resistance of pathogens within hours[9–14], or predict sequence types of bacterial isolates within an hour[15].

Here we present a method to match data from bacterial isolate sequencing and clinical metagenomics against a genomic database of known isolates for which resistance has already been determined, and predict antibiotic resistance based on the resistance profiles of the matches. We demonstrate, using the example of pneumococcus and five antibiotics (benzylpenicillin, ceftriaxone, trimethoprim-sulfamethoxazole, erythromycin, and tetracycline), that we can identify known resistant clones, and their serotype, on a standard laptop within 5 minutes even from metagenomic data. Moreover, our solution is suitable for applications in resource-poor contexts, making it not only useful for diagnosing infections, but also for enhancing surveillance.

Results

A database of resistance-associated sequence elements

To predict resistance in isolates and clinical samples we built a database of Resistance Associated Sequence Elements (RASE). We generated a *k*-mer-based representation of lineages for use to predict resistance by approximate matching. Following an analysis of the *S. pneumoniae* genome and characteristics of nanopore reads, we set *k*-mer length to 18 (see Methods). Our method depends on the initial availability of good quality data, and so we used genomes of pneumococci sampled from a carriage study in Massachusetts children[16,17] as the main reference dataset; it consists of 616 carriage samples isolated from Massachusetts children and comprises resistance data together with high quality draft genome assemblies from Illumina HiSeq reads. These isolates have already been classified using Multi-Locus Sequence Typing[18,19] (MLST), which is the current 'gold standard' for defining clones and clonal complexes used by the Pneumococcal Molecular Epidemiology Network[20] (PMEN).

Based on the measured MICs, we assigned each isolate to an antibiotic-specific resistance category using standard breakpoints (see Methods). Where data on MICs were not available, we estimated the likely resistance phenotype of an isolate using ancestral state reconstruction (see Methods). This was the case for a total of 494 records, concentrated in the data for tetracycline (291 records) and ceftriaxone (176 records) susceptibility. A further advantage of the dataset we chose was that we had access to the original isolates, and so additional resistance testing was possible; in our subsequent experiments, if original MIC data were not available for the best match in the RASE database, the relevant isolate was tested to confirm resistance phenotype (see Methods). In all of 8 cases tested, ancestral state reconstruction provided the correct resistance phenotype (shown in bold in **Table 1**). Out of all 616 isolates, 341 were associated with susceptibility to benzylpenicillin, 485 to ceftriaxone, 480 to

trimethoprim-sulfamethoxazole, 484 to erythromycin, and 551 to tetracycline (**Supplementary Tables 1 and 2**).

The constructed database occupies 320 MB RAM (4.3× compression rate) and can be further compressed to 47 MB (29× compression rate) (**Supplementary Figure 1**). The RASE database can be therefore used on portable devices and easily transmitted to the point of care over links with a limited bandwidth.

Lineage calling using inexact matching

We developed an approach that we term 'lineage calling' (**Figure 1**) to match a nanopore read to the phylogroup from which it came – where, as described above, phylogroup is a clade associated with either resistance or susceptibility. We then used a modified version of ProPhyle[21], an accurate, resource-frugal and deterministic phylogeny-based DNA classification tool based on the Burrows-Wheeler Transform[22], to assign nanopore reads to positions on phylogenetic trees and identify the closest match. Reads were assigned scores based on their similarity to known sequences in the database. Generally speaking, longer reads, such as those covering multiple accessory genes, tend to be specific and have high scores; whereas short reads from the core genome, tend to be non-specific and have low scores, being found in many genomes. Cumulative scores, which we call weights, are then used to measure how similar a sample is to known genomes associated with resistance, already in the database. We compute two metrics: the 'phylogroup score' and the 'susceptibility score' (described in more detail in methods). These are ratios comparing the weights of the best match in the database, with the weight of the next best match of a different phylogroup or susceptibility category respectively. Intuitively the scores measure the confidence with which a sample is assigned to a given phylogroup and quantify the risk of resistance based on the matching samples in the RASE database.

Results of prediction are reported in real time as the best matching genomes in the database, together with the phylogroup score and the susceptibility scores to the antibiotics being tested (examples shown in **Figures 2 and 3**). As the run progresses, these scores fluctuate and eventually stabilize.

Testing isolates present in the RASE database

We examined two isolates that were used to build the RASE database (SP01 and SP02 in **Table 1A**). They were selected to test whether we can correctly assign phylogroup even under the best circumstances, given the relatively high error rate of nanopore sequencing[6]. The profile obtained from the fully susceptible isolate is shown in **Figure 2.** Due to errors in the nanopore sequence, only 20% of the bases matched *k*-mers in the RASE database – yet despite this, the correct phylogroup was assigned within 1 minute. The best match stabilized within 7 minutes, and this matched the isolate used in the test. The second tested isolate was predicted even faster, with phylogroup and best match correctly detected and stabilized within 1 minute. These experiments provide a proof of principle that lineage calling can be accurate and fast even using sequence data with a relatively high per base error rate.

We also evaluated how long it took for resistance genes to be reliably detected in nanopore reads. For SP02 we observed that at least 15 minutes was needed to detect resistance, assuming that the genes in question can be unambiguously identified in nanopore data despite the high per base error rate, and that the presence of the loci is directly linked to the resistance phenotype (**Supplementary Figure 2**). If this is not the case, further delays would be expected. Thus, lineage calling can offer a time advantage compared to methods based on identifying the presence of resistance genes even in a sample of DNA from a purified isolate as opposed to a metagenome, potentially allowing for more rapid changes to antimicrobial therapy.

Testing isolates not present in the RASE database

We then examined four additional isolates (SP03–SP06 in **Table 1A**) for which the serotype and limited antibiogram data were known, but the lineage was unknown. To identify the lineages of these isolates we sequenced them by Illumina Miniseq, and confirmed the antibiogram of the antibiotics being tested in this study. We compared three characteristics of the sample to assess our performance: the serotype, the sequence type (ST) and the antibiograms (benzylpenicillin, ceftriaxone, trimethoprim-sulfamethoxazole, erythromycin, and tetracycline resistance according to EUCAST breakpoints[23]). Multi-locus Sequence Typing[18] (MLST) is the gold standard for strain assignment and divides the pathogen population into clonal complexes (equivalent to lineages).

In all cases, the correct clonal complex was identified within five minutes, even if the correct ST was absent from the RASE database, indicating the strength of the lineage calling method in rapidly detecting similarity. However, this also illustrates the importance of a high quality and suitable database for comparison, which contains the clones that are likely to be encountered in disease. The two 23F samples (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[20]. Consistent with this, the two samples were indeed resistant to erythromycin, as was the closest match in the RASE database constructed from the Massachusetts sample. In the case of SP05, the phylogroup score was borderline, reflecting divergence of the sample undertest from the database, even though the susceptibility scores were accurate for the antibiotics tested.

Metagenomic sample testing

Because culture introduces significant delays, direct metagenomic sequencing of clinical samples would be preferable. We therefore analyzed nanopore metagenomics data from sputum samples obtained from patients suffering from lower respiratory tract infections[2], selecting 6 samples from the study that were already known to contain *Streptococcus pneumoniae* (**Table 1B**, sorted by the estimated proportion of *S. pneumoniae* reads).

The sample displayed in **Figure 3** (SP10) contains DNA from multiple bacterial species, and as a result, few of the reads match to the *k*-mers in the RASE database (7% in contrast with 20% for the sample used for proof of principle above). However, the sample was still inferred, again within 5 minutes, to contain DNA identified as belonging to the Swedish 15A-25 clone (ST63) which is also known to be associated with resistance phenotypes including macrolides and tetracyclines[24]. This sample was confirmed to be resistant to the erythromycin, as well as clindamycin, tetracycline and oxacillin[2] according to EUCAST[23]. The result for oxacillin is especially noteworthy, as the initial report of this clone did not report resistance to penicillin antibiotics[24]. However, resistance to this class has subsequently emerged in this lineage, and so the database used in this work 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 susceptibility profiles found in the samples, with the exception of tetracycline resistance of SP12. Further analysis of the reads from SP12 using Krocus[15] suggested that the pneumococcal DNA present was from the ST180 clonal complex, and matched specifically either to the sequence type ST180 or ST3798. This is consistent with identification as serotype 3, because this clonal complex contains the great majority of isolates with this capsule type, which historically has not been associated with resistance[25]. However, improved sampling and study of this lineage has recently found highly divergent subclades that are associated with resistance. These lineages

were previously rare, and thus were less likely to be included in our database, but now are increasing in frequency[26]. In this case, ST 3798 is found to be in clade 1B, which is notable for exhibiting sporadic tetracycline resistance. Again, the failure to match to this is a result of the original database not containing a suitable example for comparison.

The last remaining samples, SP07–SP09, contained less than 5% unambiguously pneumococcal reads, and as a result the phylogroup was not securely identified in these. Nevertheless, all predicted phenotypes were concordant with phenotypic tests, with the exception of SP07 which matches the same isolate as SP12 (discussed above).

Discussion

Effective methods for detecting resistance, or susceptibility from gene sequences do not need to perform GWAS in reverse – using lineage calling, there is no requirement to detect the variation that causes the phenotype, only that it be sufficiently strongly associated with the phenotype to make reliable predictions. The results presented here show that if an identical genome is present in the database, ProPhyle accurately matches it in 5 minutes and accurately predicts resistance/susceptibility, and if the genome is not present the closest relative is identified within a similar time span. Moreover, ProPhyle can be used successfully with metagenomic data, here identifying the presence of the Sweden 15A-23 clone in a sputum sample taken from a patient with lower respiratory tract infection in the UK. Together, these results suggest that we can achieve robust lineage calling, even from complex data, within minutes of nanopore sequencing.

A key advantage of this approach is that it is not limited by the relatively high error rate of nanopore sequencing; 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.

Lineage calling has several advantages over methods that aim to detect the presence of the specific sequences that confer resistance. Most importantly, we can identify clones that are associated with susceptibility as well as resistance. The relevant loci need not be known in advance, and because we are seeking to identify the lineage rather than the loci, it is much

quicker. In our experiments it consistently took longer for a single copy of a resistance gene of interest to pass through the pore and be identified than to identify the lineage. This is particularly important when detecting mutational resistance that requires high genome coverage (>30x). Finally, when resistance is plasmid-borne, identifying the lineage may be more reliable at predicting susceptibility/resistance by lineage calling in metagenomic data, as the source organism of plasmids in a metagenome is hard to identify.

These results suggest a two-step model for resistance diagnostics, in which the first is to characterize the important pathogens in the population with highly accurate, high quality draft genomes together with metadata on resistance or other phenotypes of interest, and then to analyze clinical samples directly using nanopore-based metagenomics and the RASE software. The importance of a high quality and representative database is shown by the failure to accurately call erythromycin resistance for SP03 and SP06; the closest match to these two in the RASE database was relatively distantly related to them and had diverged in its antibiogram. Given the value and importance of an appropriate database, which is evident from our results, it is notable that health laboratories are increasingly collecting datasets suitable for use with RASE. The US Centers for Disease Control and Prevention have started using WGS to characterize samples from their Active Bacterial Core Surveillance system, which obtains isolates and MIC data from all isolates of *S. pneumoniae* causing invasive disease in a population of more than 23 million. As a result of this initiative, raw reads and resistance data for 1781 isolates collected from 2015 already exists[27,28]. While it is unlikely that a random patient presenting with disease would be infected by a lineage not present in this sample, it is possible. In the event that the sequenced isolate belongs to a clade that is absent from the database or the confidence in cluster assignment to the studied species is not sufficiently strong, RASE reports comparable similarity for multiple different phylogroups and the phylogroup score drops accordingly (see experiments SP05 and SP07–SP09 in supplementary online material). This will allow attention to rapidly be concentrated on any examples of bacteria that are not present in the database. If we are to move away from culture towards

metagenomic-based infection diagnosis in future, this feature of RASE will be extremely valuable, pointing us toward clinical samples containing unusual lineages that can be cultured and characterized.

A more serious issue, which we have not encountered in this study, but which may limit the application of our approach to other pathogen-drug combinations, is the degree of linkage between resistance and a specific lineage. If this is low, such that there is very weak association between lineage and resistance phenotype, then we would not expect our approach to work. This is particularly the case if resistance can arise from a single mutation during the course of treatment (e.g., porin mutations which confer diminished susceptibility to carbapenems[27]). Such an eventuality would not be detectable by any sequence-based method, but we note this would also mislead conventional gold standard susceptibility testing if the mutation has not already arisen at time of sample collection. In the case of the pneumococcus the degree of linkage between resistance and the rest of the genome is high, as shown by the success of ancestral state reconstruction in inferring the resistance status of isolates for which MIC data were not originally reported. This suggests that perfect resistance data for all isolates may not be necessary in all circumstances, however this will require further work to fully define, as will how the RASE approach scales with increasing database size.

Another limitation of this approach for point-of-care use is the complexity and time required for sample preparation, which currently includes human DNA depletion, DNA isolation and library preparation, taking a total of 4 hours. However, we note that ONT Voltrax technology can be used for automated library preparation and, potentially in the future, host depletion and DNA extraction. Automation will simplify and speed up the sample preparation turnaround time. It should be noted that this has been further reduced, with a Rapid Sequencing Kit offering library preparation in 10 minutes[29]. Further advances in this space, including reduced costs, will be required to bring the method closer to the bedside. For instance, the ONT Flongle flowcell ($100 as of August 2018) may help to address this issue.

The benefits of lineage calling are in identifying high-risk clones earlier. It is easy to see how our approach may be extended to include calling specific resistance loci, where they are known, but a key advantage of our approach is that it is not limited by the requirement to know them in advance. Lineage calling can be used to detect any phenotype that is sufficiently tightly linked to a phylogeny, for instance to identify highly virulent strains that might merit closer attention. Further applications may include rapid outbreak investigations, as the closely related isolates involved in the outbreak will all be predicted to match to the same strain in the RASE database. The approach also lends itself to enhanced surveillance, including field work situations; the recent Ebola outbreak in West Africa, for example, saw MinION devices used in remote locations without centralized and advance healthcare facilities. Finally, this approach is not at present intended to supplant empiric therapies. Given the urgency of instituting appropriate therapies, prescriptions should be made as early as possible. However, we may be able, through lineage calling of samples taken when the tentative diagnosis is made, to institute effective therapy at the second dose when the initial therapy is inadequate, long before it would become clinically apparent the patient is not responding. The combination of high quality RASE databases with lineage calling hence offers an alternative model for diagnostics and surveillance, with wide applications for the management of infectious disease.

Methods

Overview

RASE uses rapid approximate *k*-mer-based matching of long sequencing reads against a database of genomes to predict resistance via lineage calling, using two key components: a database containing genomic data and associated antibiograms, and a prediction pipeline. The database contains a highly compressed lossless *k*-mer index, a representation of the tree population structure, and metadata such as a phylogroup, serotype, sequence type and resistance profiles (see 'Resistance profiles'). The pipeline iterates over reads from the nanopore sequencer and provides real-time predictions of phylogroup and resistance (**Figure 1**).

Resistance profiles

For all antibiotics, RASE associates individual isolates with a resistance category, susceptible or non-susceptible. First, MIC values are mined using regular expressions from the available textual antibiograms, i.e., strings describing an interval of possible MIC values. Second, the acquired intervals are compared to the antibiotic-specific breakpoints (**Supplementary Figure 3**). 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, or an antibiogram is present, but parsing failed. Third, 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) breakpoints (**Supplementary Figure 4**). When the solution is not unique, non-susceptibility is assigned.

The RASE database was constructed with the standard EUCAST breakpoints[23] ([g/ml]): benzylpenicillin (PEN): 0.06, ceftriaxone (CRO): 0.25, trimethoprim-sulfamethoxazole (TMP): 1.00, Erythromycin (ERY): 0.25, 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[23]).

K-mer-based matching

RASE uses the ProPhyle classifier[21] (version 0.3.1.0) and its ProPhex component[30] to identify the most similar genomes in the database for every sequencing read. Its index stores *k*-mers of all isolates' assemblies in a highly compressed form, reducing the required memory footprint. The database *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[31] is constructed. The index can be searched for individual *k*-mers, retrieving a list of nodes whose descending leaves correspond to isolates containing that *k*-mers.

In course of sequencing, every read is matched against the index and matches for all read's *k*-mers retrieved. These matches are then propagated to the level of leaves and isolates with the highest number of shared *k*-mers identified.

Predicting resistance from phylogroups

All isolates 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 matched against the DB, the weights for the best match are increased according to the read's 'information content', calculated as the number of shared *k*-mers between a genome and the read, divided by the number of best hits.

Predictions are calculated based on the current state of the weights and the lineage or phylogroup in which the best-matched isolate is found. First, a phylogroup is predicted as the phylogroup of the best matching isolate. Then, a phylogroup score is calculated PGS=2f/(f+t)-1, where f and t denote the scores of the best matches in the first ('predicted') and second best ('alternative') phylogroup respectively. If PGS is higher than a specified threshold (0.6 in default settings), the call is considered successful. If the score is lower than this, the read cannot be securely assigned to a phylogroup, and this counts as a failure. Reads that do not match are not used in subsequent analysis to predict resistance.

Resistance is predicted for individual antibiotics independently, using weights within the predicted phylogroup. While certain phylogroups are certainly associated with susceptibility, some others are not. For the latter, we propose the use of the susceptibility scores that combine the resistance characteristics of the most similar strains in the RASE database. A susceptibility score is calculated as SUS=s/(s+r), where s and r denote the score of the best susceptible and non-susceptible strains within the predicted phylogroup. If SUS is greater than a specified threshold (0.6 in default settings), susceptibility to the antibiotic is reported, non-susceptibility otherwise. In most of cases, this algorithm predicts non-susceptibility or susceptibility as the one of the best match. Nevertheless, when two genomes with different resistance categories are of similar weights, non-susceptibility may be reported even though the best match is susceptible.

To determine how RASE works with nanopore data generated in real time, the timestamps of individual reads were first extracted and then used for sorting the base-called nanopore reads. 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.

Optimizing k-mer length

First, the subword complexity function[32] of pneumococcus was calculated using JellyFish[33] (version 2.2.10) (**Supplementary Figure 5**). Then, based on the characteristics of the function and technical limitations of ProPhyle, the possible range of *k* was determined as [17, 32]. For these *k*-mer lengths, RASE indexes were constructed and their performance evaluated using the RASE prediction pipeline and selected experiments. All these lengths *k*-mer lengths led to similar predictions, but different prediction delays (**Supplementary Figure 6**). Based on the obtained timing data, we set *k* to 18.

Lower time bounds on resistance gene detection

A complete genome assembly of the multidrug resistant SP02 isolate was computed from the Nanopore reads using the CANU[34] (version 1.5, with default parameters). Prior to the assembly step, reads were filtered using SAMsift[35] based on the matching quality with the 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[36] (version 1.2, default parameters) using Illumina reads from the same isolate (taxid '1QJAP' in the SPARC dataset[17]) mapped to the nanopore assembly using BWA-MEM[37] (version 0.7.17, with the default parameters) and sorted using SAMtools[38].

The obtained assembly was searched for resistance-causing genes using the online CARD tool[39] (as of 2018/08/01). All of the original nanopore reads were then mapped using Minimap2[40] (version 2.11, with '-x map-ont') to the corrected assembly and resistance genes in the reads identified using BEDtools–intersect[41] (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.

Library preparation

For experiments 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 (>1ug) 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[2].

MinION sequencing

Sequencing was performed on the MinION MK1 device using R9.4/FLO-MIN106 flowcells, 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-SP09, 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.

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)[42,43]. 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 3**.

Resistance of streptococcus in the metagenomic samples (SP07–SP12) was determined by agar diffusion using the EUCAST methodology[23]. 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≥25mm for sensitive, r<22mm for resistant) and erythromycin (r≥22mm for sensitive, r<19mm 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 4**.

Data, implementation and availability

RASE was developed using Python, GNU Make, GNU Parallel[44], Snakemake[45], and the ETE3[46] and PySam libraries. Bioconda[47] was used to ensure reproducibility of the software

environments. The code for constructing databases, together with the default RASE DB, is available from http://github.com/c2-d2/rase-db; the RASE prediction pipeline is located at http://github.com/c2-d2/rase-pipeline; and additional material to this paper can be found on http://github.com/c2-d2/rase-supplement. All code is available under the MIT license. Sequencing data for all experiments from this study can be downloaded from http://doi.org/10.5281/zenodo.1405173; for the metagenomic experiments, only the filtered datasets (i.e., after removing the remaining human reads *in silico*) were made publicly available.

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 (FRN 152448, RSL), and the Canadian Institutes for Health Research (a fellowship grant, DRM). 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 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). The authors also thank Yonatan H Grad, Brian J Arnold, Taj Azarian, and Cristina M Herren for useful comments in various stages of this project.

Transparency declarations

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.

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.* **34,** 1589–1596 (2006).
2. Charalampous, T. *et al.* Rapid Diagnosis of Lower Respiratory Infection using Nanopore-based Clinical Metagenomics. *bioRxiv* 387548 (2018). doi:10.1101/387548
3. D'Costa, V. M. *et al.* Antibiotic resistance is ancient. *Nature* **477,** 457–461 (2011).
4. CDC. Antibiotic resistance threats in the United States, 2013. *Current* 114 (2013). doi:CS239559-B
5. 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).
6. Wick, R., Judd, L. M. & Holt, K. E. *Comparison of Oxford Nanopore basecalling tools*. doi:10.5281/zenodo.1188469
7. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530,** 228–232 (2016).
8. Zaaijer, S. *et al.* Rapid re-identification of human samples using portable DNA sequencing. *Elife* **6,** 1–29 (2017).
9. Votintseva, A. A. *et al.* Same-day diagnostic and surveillance data for tuberculosis via whole genome sequencing of direct respiratory samples. *J. Clin. Microbiol.* JCM.02483-16 (2017). doi:10.1128/JCM.02483-16
10. Schmidt, K. *et al.* Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *J. Antimicrob. Chemother.* **72,** 104–114 (2017).
11. Cao, M. D. *et al.* Streaming algorithms for identification of pathogens and antibiotic resistance potential from real-time MinION™ sequencing. *Gigascience* **5,** 32 (2016).
12. Quick, J. *et al.* Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol.* **16,** 114 (2015).

13. Leggett, R. M. *et al.* Rapid MinION metagenomic profiling of the preterm infant gut microbiota to aid in pathogen diagnostics. *Bioarxiv* 180406 (2017). doi:10.1101/180406
14. Bradley, P. *et al.* Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat. Commun.* **6,** 10063 (2015).
15. Page, A. J. & Keane, J. A. Rapid multi-locus sequence typing direct from uncorrected long reads using Krocus. *PeerJ* **6,** e5233 (2018).
16. Croucher, N. J. *et al.* Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat. Genet.* **45,** 656–63 (2013).
17. Croucher, N. J. *et al.* Population genomic datasets describing the post-vaccine evolutionary epidemiology of Streptococcus pneumoniae. *Sci. data* **2,** 150058 (2015).
18. Maiden, M. C. J. *et al.* Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci.* **95,** 3140–3145 (1998).
19. Enright, M. C. & Spratt, B. G. A multilocus sequence typing scheme for Streptococcus pneumoniae: identification of clones associated with serious invasive disease. *Microbiology* **144,** 3049–3060 (1998).
20. 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).
21. Břinda, K., Salikhov, K., Pignotti, S. & Kucherov, G. ProPhyle: An accurate, resource-frugal and deterministic DNA sequence classifier. (2017). doi:10.5281/zenodo.1045429
22. Burrows, M. & Wheeler, D. J. *A Block-sorting Lossless Data Compression Algorithm*. *Digital SRC Research Report* (1994).
23. *The European Committee on Antimicrobial Susceptibility Testing. Breakpoint tables for interpretation of MICs and zone diameters. Version 7.0.* (2017).
24. 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,** 1153–60 (2000).

25. Croucher, N. J. *et al.* Dominant Role of Nucleotide Substitution in the Diversification of Serotype 3 Pneumococci over Decades and during a Single Infection. *PLoS Genet.* **9,** e1003868 (2013).
26. Azarian, T. *et al.* Global emergence and population dynamics of divergent serotype 3 CC180 pneumococci. *bioRxiv* 314880 (2018). doi:10.1101/314880
27. Metcalf, B. J. *et al.* Using whole genome sequencing to identify resistance determinants and predict antimicrobial resistance phenotypes for year 2015 invasive pneumococcal disease isolates recovered in the United States. *Clin. Microbiol. Infect.* **22,** 1002.e1-1002.e8 (2016).
28. Li, Y. *et al.* Validation of β-lactam minimum inhibitory concentration predictions for pneumococcal isolates with newly encountered penicillin binding protein (PBP) sequences. *BMC Genomics* **18,** 621 (2017).
29. Quick, J. *Ultra-long read sequencing protocol for RAD004 (Version 3)*. (2018). doi:10.17504/protocols.io.mrxc57n
30. 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
31. Ferragina, P. & Manzini, G. Opportunistic data structures with applications. in *Proceedings 41st Annual Symposium on Foundations of Computer Science* 390–398 (IEEE Comput. Soc, 2000). doi:10.1109/SFCS.2000.892127
32. Lothaire, M. *Algebraic Combinatorics on Words*. (Cambridge University Press, 2002). doi:10.1017/CBO9781107326019
33. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27,** 764–770 (2011).
34. 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).
35. Břinda, K., Baym, M. & Hanage, W. P. SAMsift: advanced filtering and tagging of SAM/BAM alignments using Python expressions. (2018). doi:10.5281/zenodo.1048211
36. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant

detection and genome assembly improvement. *PLoS One* **9,** (2014).

37. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 3 (2013).
38. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–9 (2009).
39. Jia, B. *et al.* CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45,** D566–D573 (2017).
40. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 1–3 (2018). doi:10.1093/bioinformatics/bty191
41. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
42. CLSI. *Susceptibility Tests for Bacteria That Grow Aerobically; Approved Standard—Ninth Edition*. *CLSI document M07-A9* (2012).
43. CLSI. *Performance Standards for Antimicrobial Susceptibility Testing; Twenty-Second Informational Supplement*. *CLSI document M100-S22* (2012).
44. Tange, O. GNU Parallel: the command-line power tool. *;login USENIX Mag.* **36,** 42–47 (2011).
45. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **28,** 2520–2522 (2012).
46. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, analysis and visualization of phylogenomic data. *Mol. Biol. Evol.* **33,** msw046 (2016).
47. Dale, R. *et al.* Bioconda: A sustainable and comprehensive software distribution for the life sciences. *bioRxiv* 207092 (2017). doi:10.1101/207092
48. Croucher, N. J. *et al.* Role of Conjugative Elements in the Evolution of the Multidrug-Resistant Pandemic Clone Streptococcus pneumoniae Spain 23F ST81. *J. Bacteriol.* **191,** 1480–1489 (2009).

A)

Sample	PG detected	Serotype		Antibiogram - PEN		Antibiogram - CRO		Antibiogram - TMP		Antibiogram - ERY		Antibiogram - TET		ST match	CC match
		Actual	Best match	Actual	Best match	Actual	Best match	Actual	Best match	Actual	Best match	Actual	Best match		
sp01	yes	11D	11D	S	S	S	S	S	S	S	S	**S**(1)	**S**(1)	Yes	Yes
sp02	yes	19A	19A	R	R	R	R	R	R	R	R	**R**(2)	**R**(2)	Yes	Yes
sp03	yes	23F	23F	R	R	R	R	R	R	R	**S**(3)	S	S	OoD	Yes
sp04	yes	19A	19A	R	R	R	R	R	R	R	R	R	**R**(4)	OoD	Yes
sp05	no (borderline)	19F	19F	R	R	R	R	R	R	R	R	R	R	OoD	Yes
sp06	yes	23F	23F	R	R	R	R	R	R	R	**S**(3)	S	S	OoD	Yes

B)

Sample	PG detected	SP	Antibiogram - PEN		Antibiogram - ERY		Antibiogram - TET	
sp07	no	2.3%	S	S	NA	S	R	**S**(5)
sp08	no|yes	2.5%	S	S(!)	S	S	S	**S**(6)
sp09	no	4.0%	S	S	NA	S	S	**S**(7)
sp10	yes	21%	R	R	R	R	R	**R**(8)
sp11	yes	70%	R	R	R	R	R	**R**(8)
sp12	yes	86%	S	S	S	S	R	**S**(5)

Legend

S Susceptible
S(!) Best match S, but high risk of R reported
R Non-susceptible
NA Not available
OoD Out-of-database
(...) ID of a retested sample
...|... Original | filtered sample (when the results differ)
SP Fraction of S. pneumoniae reads

Correct prediction
Incorrect prediction
Cannot be evaluated

Table 1: Predicted phenotypes for (A) isolates and (B) 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 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 3**).

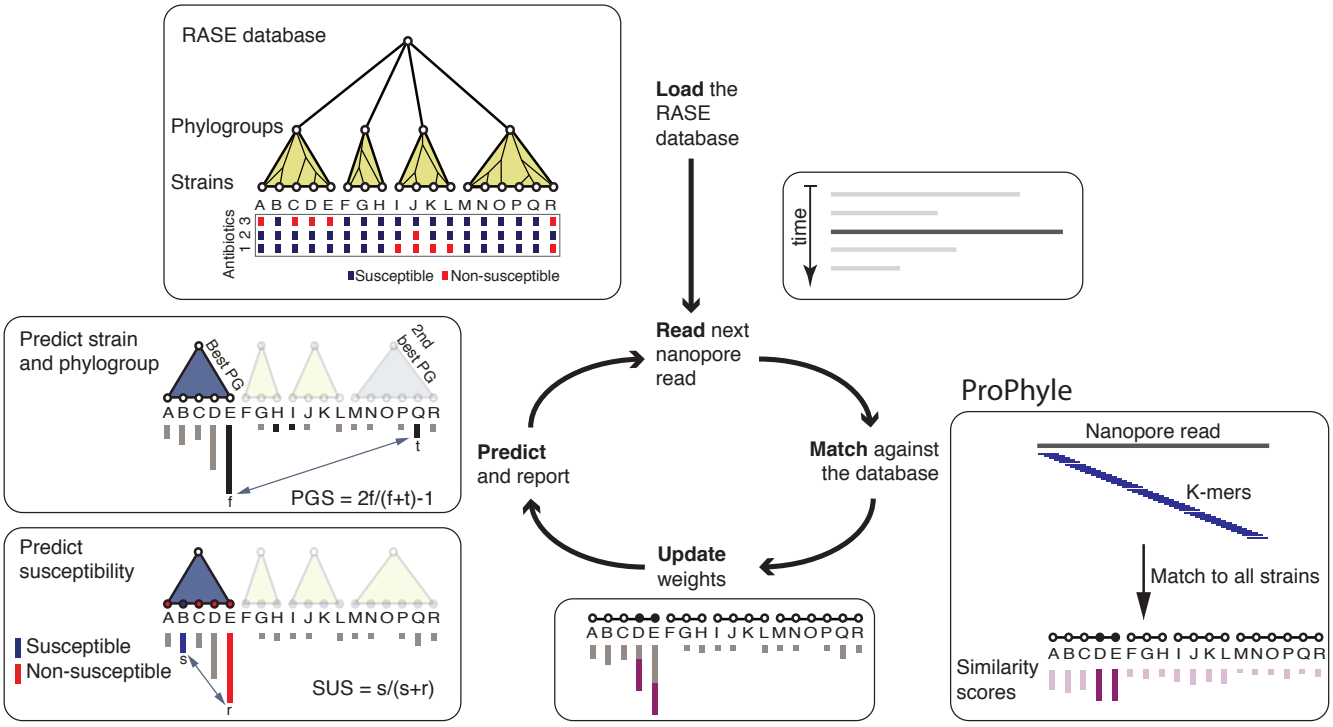


Figure 1: Overview of the RASE approach. The RASE approach uses three components: the RASE database, an approximate k-mer-based matching component based on ProPhyle, and a prediction component interpreting the risk based on the resistance of strains of the assigned phylogroup. In the load step, the precomputed RASE database is loaded into memory. The RASE pipeline iterates over reads streamed from the nanopore sequencer. Each read is matched against the database using ProPhyle. Retrieved assignments are propagated to the leaves and similarity scores computed. These are used to identify best-matching strains (possibly many) and to update weights associated with these strains. Indeed, a single read is rarely specific, it typically matches equally scored multiple nodes. The best phylogroup is identified and a phylogroup score calculated (PGS). Based on the resistance profiles of strains in this phylogroup, susceptibility to each of the antibiotics is predicted from the best match and reported together with a susceptibility score quantifying the risk of resistance.

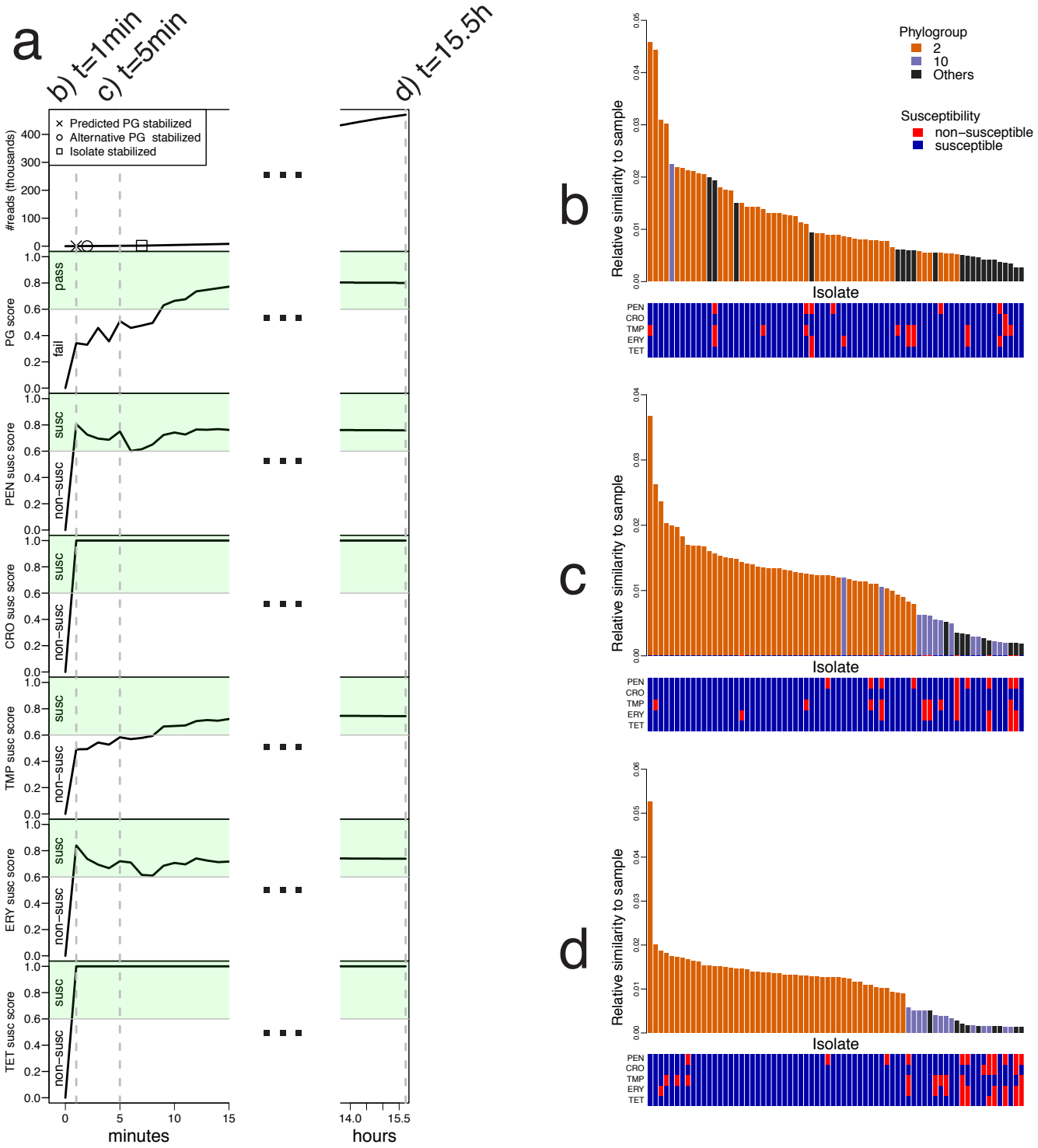


Figure 2: Timeline and rank plots for an isolate. a) Number of reads, phylogroup score, and susceptibility scores for individual antibiotics as a function of time from the start of sequencing. The point markers depict the times of stabilization for the predicted phylogroup, the alternative phylogroup and the most similar isolate, respectively. b)-d) Similarity rank plots for selected time points (1 minute, 5 minutes, and the end of sequencing). The bars correspond to 70 best matching isolates in the database and display the predicted level of sample-to-strain relative similarity (i.e., normalized weights). They are arranged by rank and colored according to the presence in the predicted, alternative or another phylogroup. The bottom panels display the susceptibility profiles of the isolates.

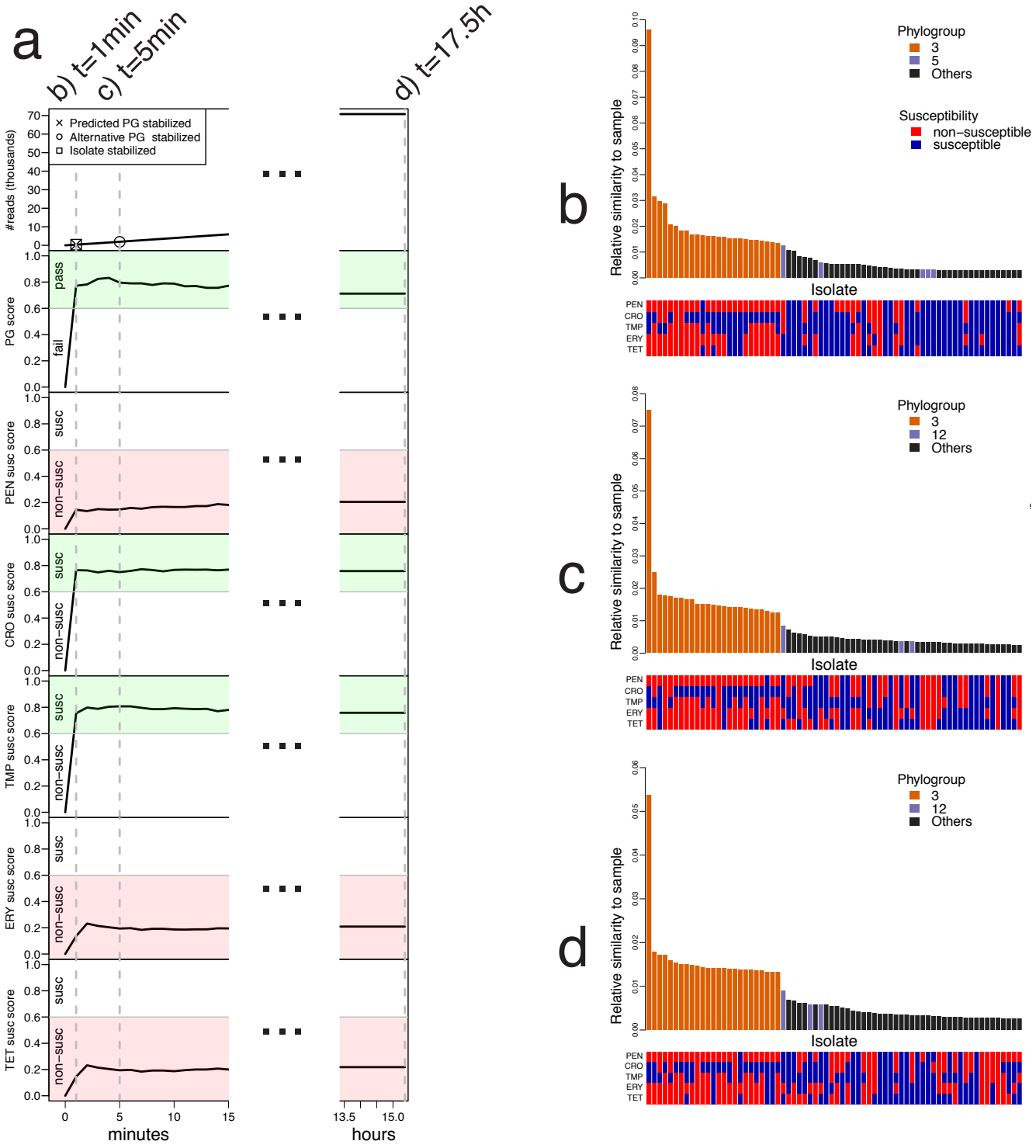
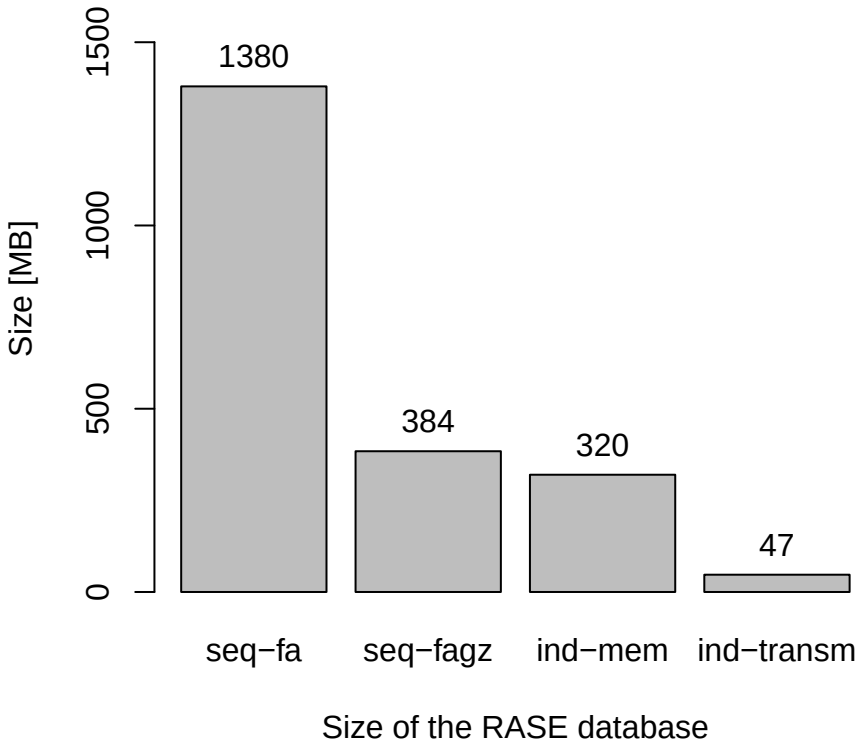
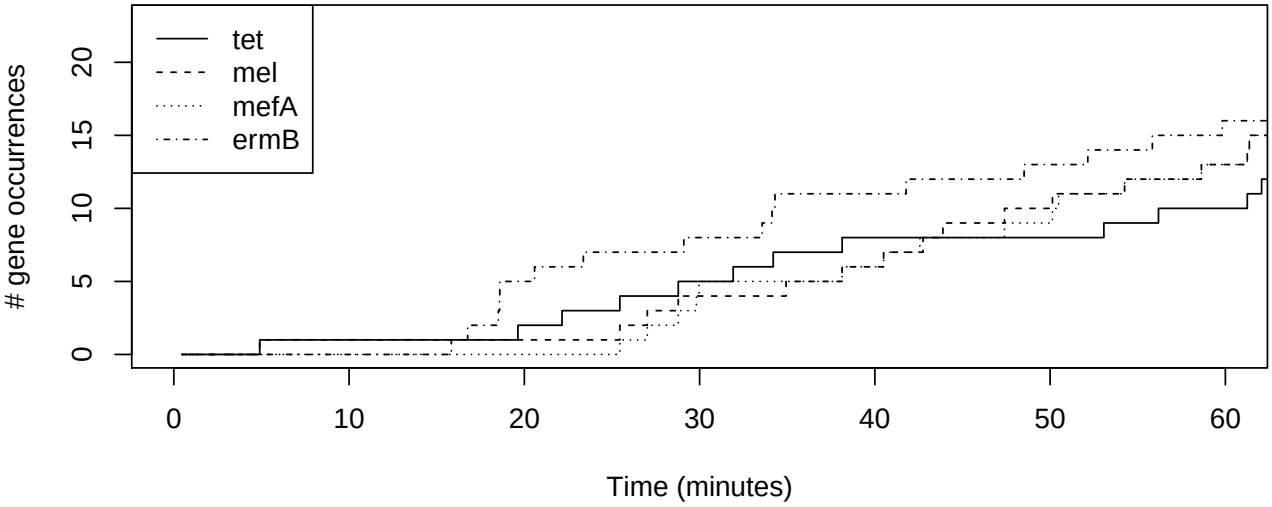


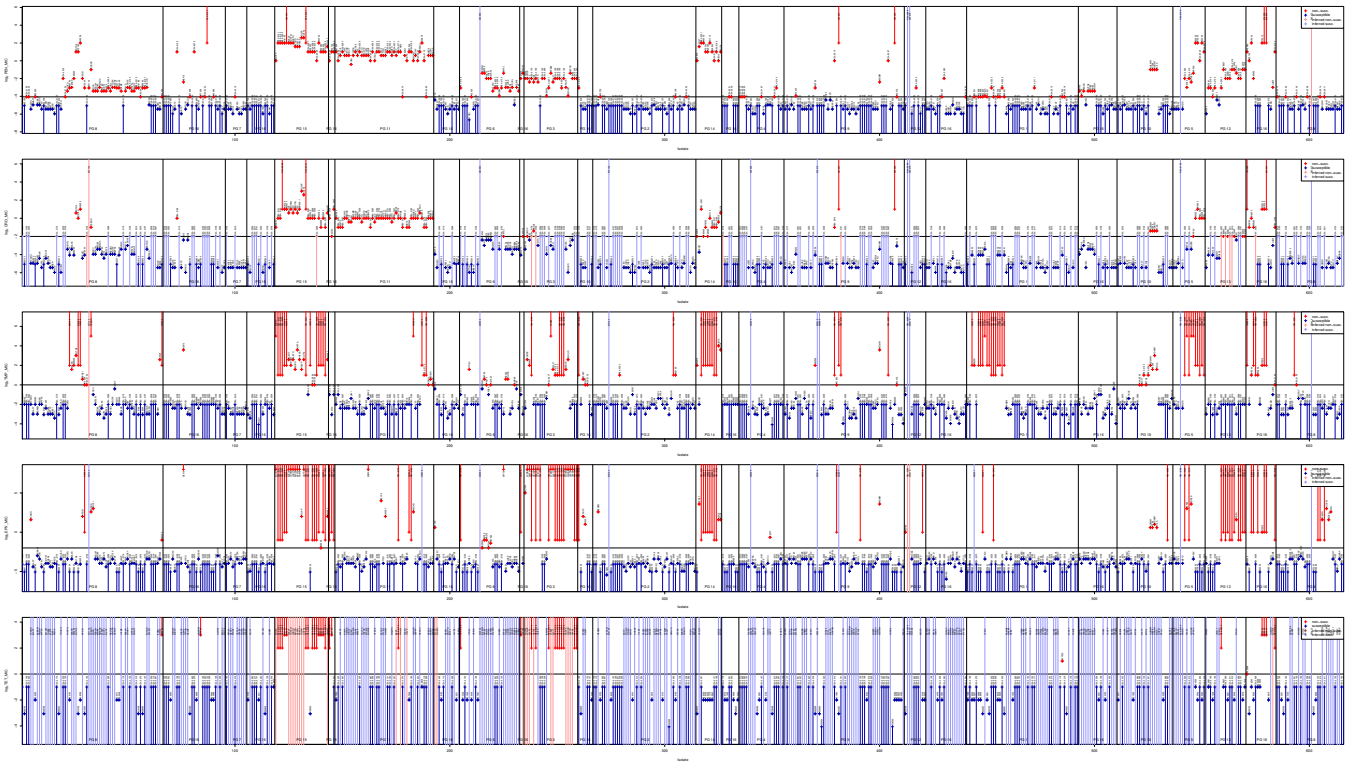
Figure 3: Timeline and rank plots for a metagenome. The figure is of the same format as **Figure 2**.



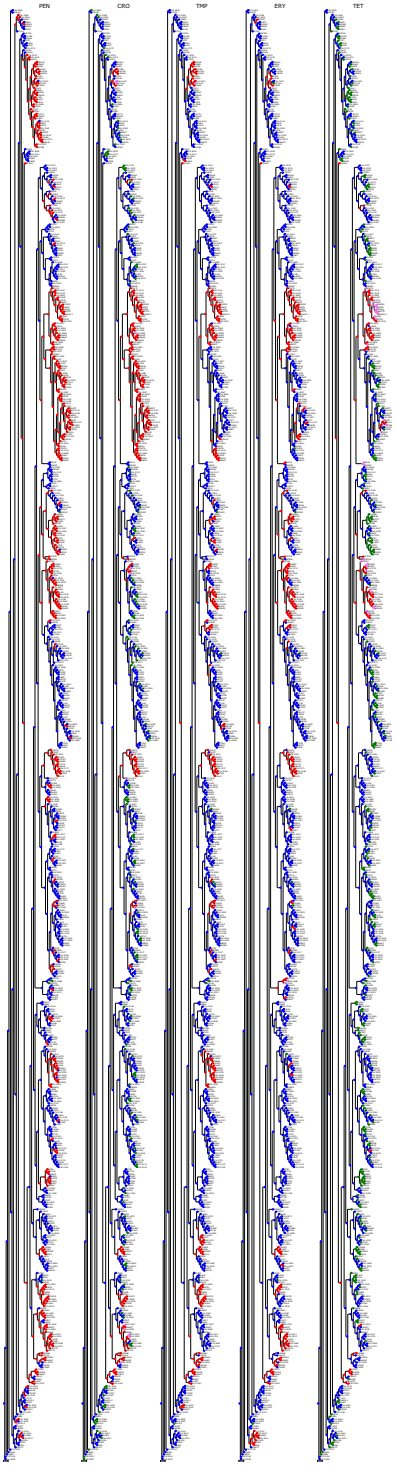
Supplementary Figure 1: Size and memory footprint of the RASE database and index. The graph compares the size of the ProPhyle RASE index to the size of the original sequences: original draft assemblies (seq−fa), original draft assemblies compressed using gzip (seq−fagz), memory footprint of ProPhyle with the RASE index (ind−mem), and size of the ProPhyle RASE index compressed for transmission (ind−transm).



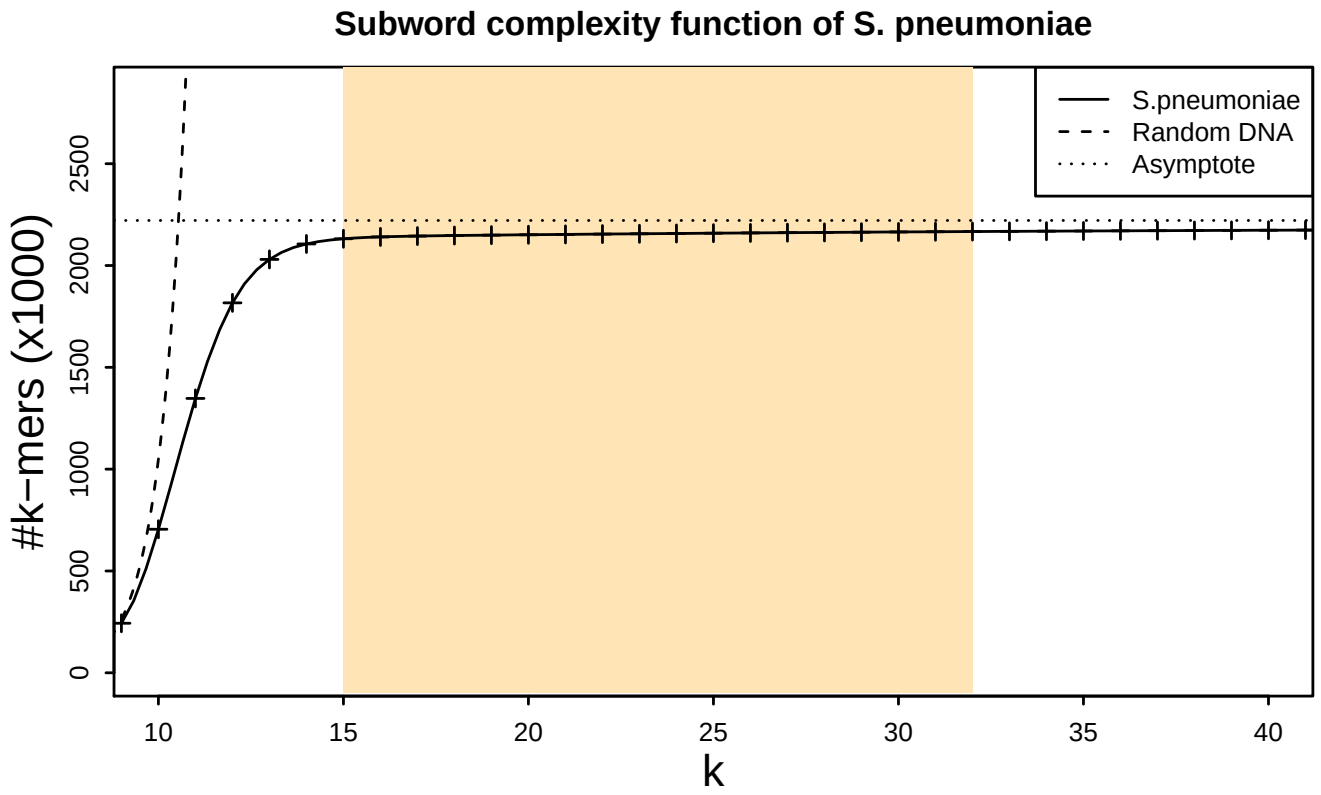
Supplementary Figure 2: Timeline of resistance genes. Number of occurrences of individual resistance genes in reads of SP02, as a function of time for the first hour of nanopore sequencing.



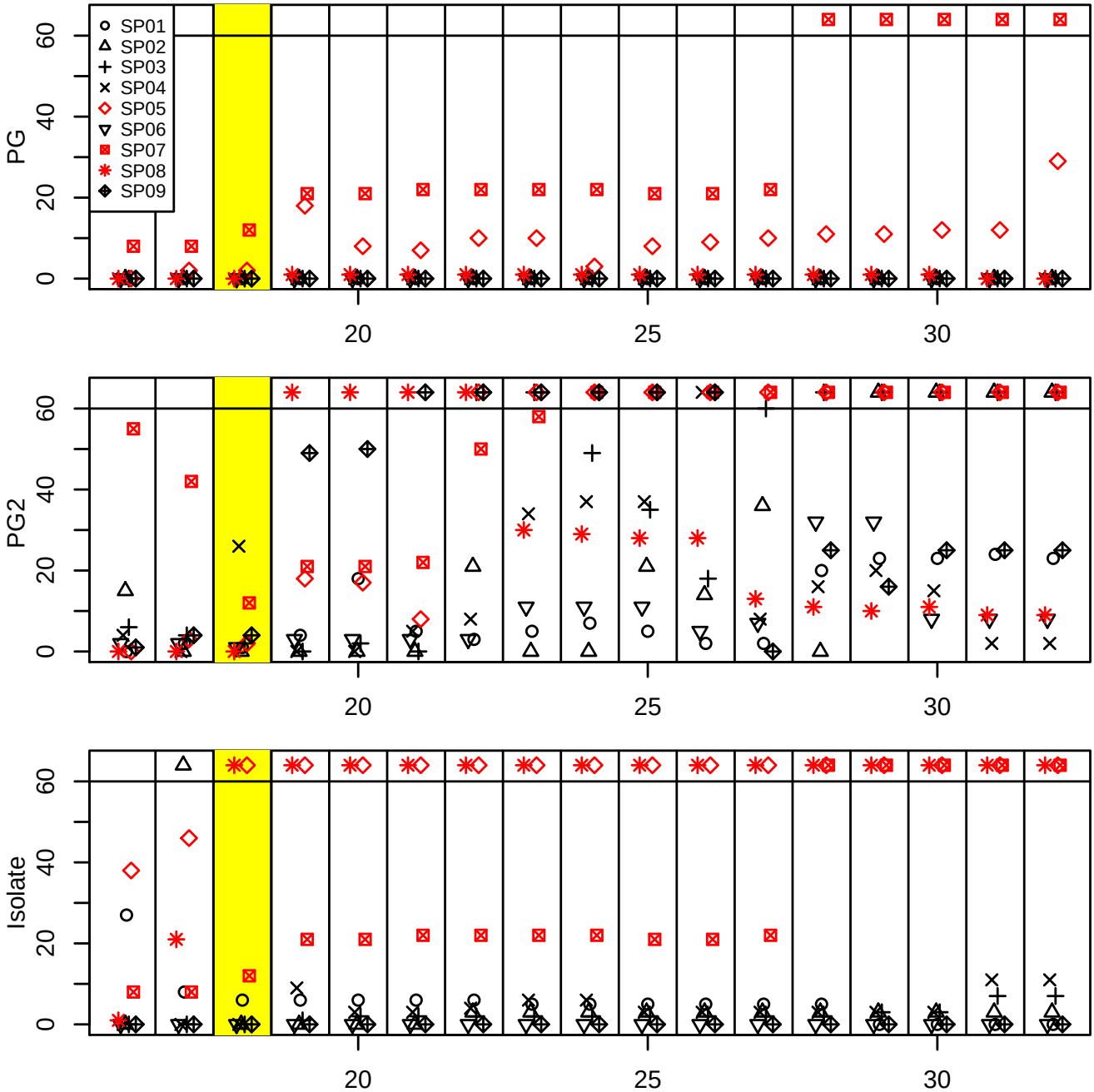
Supplementary Figure 3: MIC intervals for individual isolates in the RASE database. The plot illustrates MIC intervals and point values extracted from. Each panel corresponds to a single antibiotic, while vertical lines and points correspond to individual isolates. Their colors correspond to the resistance category after applying a breakpoint (horizontal lines). When a resistance category could not be assigned directly (i.e., in case of an interval crossing the breakpoint line), then it was inferred using ancestral state reconstruction.



Supplementary Figure 4: Ancestral state reconstruction of resistance categories in the 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).



Supplementary Figure 5: Subword complexity of pneumococcus. The plot depicts the number of canonical k-mers as a function of k for S.pneumoniae ATCC 700669 (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 has a finite length and is circular, the function has an asymptote, which would be attained for k equal to the length of the genome (2,221,315). The highlighted region corresponds to the range of k values, which are suitable for use in RASE. Note that k-mers longer than 32 are not currently supported by ProPhyle.



Supplementary Figure 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 all experiments and all possible k-mer lengths. Each horizontal panel displays times required for stabilization of one of the three predictions: phylogroup (PG), alternative phylogroup (PG2), and closest isolate (Isolate). 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 phylogroup could not be identified are plotted in red. The highlighted column corresponds to the k-mer length used for constructing RASE.

Supplementary Table 1: Prevalence of resistance phenotypes across phylogroups. For all sequencing experiments, the table displays the best matching isolates, the strain MIC and all measurements of database MICs (the original reported values or categories inferred using ancestral state reconstruction when not available, retested values, and the resulting resistance categories).

Supplementary Table 2: Metadata for all isolates included in the RASE database. Each record contains the strain's taxid, phylogroup, serotype, sequence type (ST), order in the phylogenetic tree, and three fields related to resistance for every antibiotics: the '_mic', '_int', '_cat' fields contain the original published MIC information (possibly corrected after retesting), the extracted MIC interval, and the resulting category after ancestral state reconstruction (S = susceptible, R = non-susceptible, s = unknown but reconstructed susceptible, r = unknown but reconstructed non-susceptible), respectively.

Supplementary Table 3: Additional MIC measurements for selected strains. The table contains results from strain retesting. Each record contains date when the retesting was done, the antibiotic, the measured MIC, and the corresponding resistance category according the RASE breakpoints.

Supplementary Table 4: Overview of performed resistance tests. For all sequencing experiments, the table displays the best matching isolates, the strain MIC and all measurements of database MICs (the original reported values or categories inferred using ancestral state reconstruction when not available, retested values, and the resulting resistance categories).