



**HAL**  
open science

## Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing

Karel Brinda, Alanna Callendrello, Kevin C Ma, Derek R Macfadden, Robyn Lee, Themoula Charalampous, Lauren Cowley, Crista B Wadsworth, Yonatan H Grad, Gregory Kucherov, et al.

### ► To cite this version:

Karel Brinda, Alanna Callendrello, Kevin C Ma, Derek R Macfadden, Robyn Lee, et al.. Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing. *Nature Microbiology*, 2020, 10.1038/s41564-019-0656-6 . hal-02416034

**HAL Id: hal-02416034**

**<https://hal.science/hal-02416034>**

Submitted on 12 Dec 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Rapid heuristic inference of antibiotic resistance and susceptibility by**  
2 **genomic neighbor typing**

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

6  
7 1 Center for Communicable Disease Dynamic, Department of Epidemiology, Harvard T.H. Chan  
8 School of Public Health, Boston, USA

9 2 Department of Biomedical Informatics and Laboratory of Systems Pharmacology, Harvard  
10 Medical School, Boston, USA

11 3 Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public  
12 Health, Boston, USA

13 4 Division of Infectious Diseases, Department of Medicine, University of Toronto, Canada

14 5 Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich, UK

15 6 Department of Biology and Biochemistry, University of Bath, UK

16 7 Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, USA

17 8 CNRS/LIGM Université Paris-Est, Marne-la-Vallée, France

18 9 Skolkovo Institute of Science and Technology, Moscow, Russia

19 10 Quadram Institute Bioscience, Norwich Research Park, Norwich, UK

20

21 \* Correspondence to [kbrinda@hsph.harvard.edu](mailto:kbrinda@hsph.harvard.edu)

22 **Abstract**

23

24 Surveillance of drug-resistant bacteria is essential for healthcare providers to deliver effective  
25 empiric antibiotic therapy. However, traditional molecular epidemiology does not typically occur  
26 on a timescale that could impact patient treatment and outcomes. Here we present a method  
27 called ‘genomic neighbor typing’ for inferring the phenotype of a bacterial sample by identifying  
28 its closest relatives in a database of genomes with metadata. We show that this technique can  
29 infer antibiotic susceptibility and resistance for both *S. pneumoniae* and *N. gonorrhoeae*. We  
30 implemented this with rapid *k*-mer matching, which, when used on Oxford Nanopore MinION  
31 data, can run in real time. This resulted in determination of resistance within ten minutes  
32 (sens/spec 91%/100% for *S. pneumoniae* and 81%/100% *N. gonorrhoeae* from isolates with a  
33 representative database) of sequencing starting, and for clinical metagenomic sputum samples  
34 (75%/100% for *S. pneumoniae*), within four hours of sample collection. This flexible approach has  
35 wide application to pathogen surveillance and may be used to greatly accelerate appropriate  
36 empirical antibiotic treatment.

37 **Introduction**

38  
39 Infections pose multiple challenges to healthcare systems, contributing to higher mortality,  
40 morbidity, and escalating cost. Clinicians must regularly make rapid decisions on empiric  
41 antibiotic treatment of infectious syndromes without knowing the causative pathogen(s) or  
42 whether they are drug-susceptible or drug-resistant. In some cases, this is directly linked to poor  
43 outcomes; in the case of septic shock, the risk of death increases by an estimated 10% with every  
44 60 minutes delay in initiating effective treatment<sup>1</sup>.

45  
46 The molecular epidemiology of infectious disease allows us to identify high-risk pathogens and  
47 determine their patterns of spread, on the basis of their genetics or (increasingly) genomics.  
48 Conventionally such studies, including outbreak investigations and characterization of novel  
49 resistant strains, have been conducted in retrospect, but this has been changing with the  
50 availability of new and increasingly inexpensive sequencing technologies<sup>2,3</sup>. The wealth of data  
51 generated by genomics is promising but introduces a new challenge: while many features of a  
52 sequence are correlated with the phenotype of interest, few are causative.

53  
54 Prescription, however, has long been informed by correlative features when causative ones are  
55 difficult to measure, for example whether the same syndrome or pathogen occurring in other  
56 patients from the same clinical environment have responded to a particular antibiotic. This has  
57 also been observed at the genetic level as well, as a result of genetic linkage between resistance  
58 elements and the rest of the genome. An example is given by the pneumococcus (*Streptococcus*  
59 *pneumoniae*). The Centers for Disease Control have rated the threat level of drug-resistant  
60 pneumococcus as 'serious' <sup>4</sup>. While resistance arises in pneumococci through a variety of  
61 mechanisms, approximately 90% of the variance in the minimal inhibitory concentration (MIC)  
62 for antibiotics of different classes can be explained by the loci determining the strain type<sup>5</sup>, even  
63 though none of these loci themselves causes resistance. Thus, in the overwhelming majority of  
64 cases, resistance and susceptibility can be inferred from coarse strain typing based on population

65 structure. This population structure could be leveraged to offer an alternative approach to  
66 detecting resistance in which rather than detecting high-risk genes, we identify high-risk strains.  
67 While many approaches have been developed to identify whether a pathogen carries mutations  
68 or genes known to confer resistance<sup>6-21</sup> (see ref<sup>22</sup> for a comprehensive review), this is not  
69 equivalent to the clinical question of whether the pathogen is susceptible.

70  
71 We present a method called ‘genomic neighbor typing’ which can bring molecular epidemiology  
72 closer to the bedside and provide information relevant to treatment at a much earlier stage. Our  
73 method takes sequences generated from a sample in ‘real time’ and matches them to a database  
74 of genomes to identify the closest relatives. Because closely related isolates usually have similar  
75 properties, this yields an informed heuristic as to the pathogen’s phenotype. We demonstrate  
76 this by identifying drug-resistant and drug-susceptible clones for both *Streptococcus pneumoniae*  
77 (the pneumococcus) and *Neisseria gonorrhoeae* (the gonococcus), within minutes after the start  
78 of sequencing using Oxford Nanopore Technology. The method has many potential applications,  
79 depending on the specific pathogen and quality of the databases available for matching, which  
80 we discuss together with its limitations.

81 **Results**

82

83 ***Resistance is associated with clones in S. pneumoniae and N. gonorrhoeae***

84

85 To quantify the association of clones with antibiotic resistance of the pathogens *S. pneumoniae*  
86 and *N. gonorrhoeae*, we constructed optimal predictors of resistance from bacterial lineages and  
87 measured the associated Area under the Receiver Operation Characteristic Curve (AUC)  
88 (**Supplementary Document 1**). First, we applied the method to 616 pneumococcal genomes from  
89 a carriage study in Massachusetts children<sup>23,24</sup>. Second, we used 1102 clinical gonococcal isolates  
90 collected from 2000 to 2013 by the Centers for Disease Control and Prevention’s Gonococcal  
91 Isolate Surveillance Project<sup>25</sup>. In both cases, the datasets comprised draft genome assemblies  
92 from Illumina HiSeq reads, resistance data, and lineages inferred from sequence cluster  
93 computed using Bayesian Analysis of Population Structure (BAPS)<sup>26</sup>. Lineages of *S. pneumoniae*  
94 are predictive for benzylpenicillin, ceftriaxone, trimethoprim-sulfamethoxazole, erythromycin,  
95 and tetracycline resistance with AUC ranging from 0.90 to 0.97 (**Supplementary Document 1**),  
96 consistent with previous works<sup>5</sup>. In *N. gonorrhoeae*, ciprofloxacin, ceftriaxone, and cefixime  
97 attained comparably large AUCs (from 0.93 to 0.98) whereas azithromycin demonstrated lower  
98 association (AUC 0.80), as observed previously<sup>25</sup>.

99

100 ***Rapid identification of nearest known relative from sequencing reads***

101

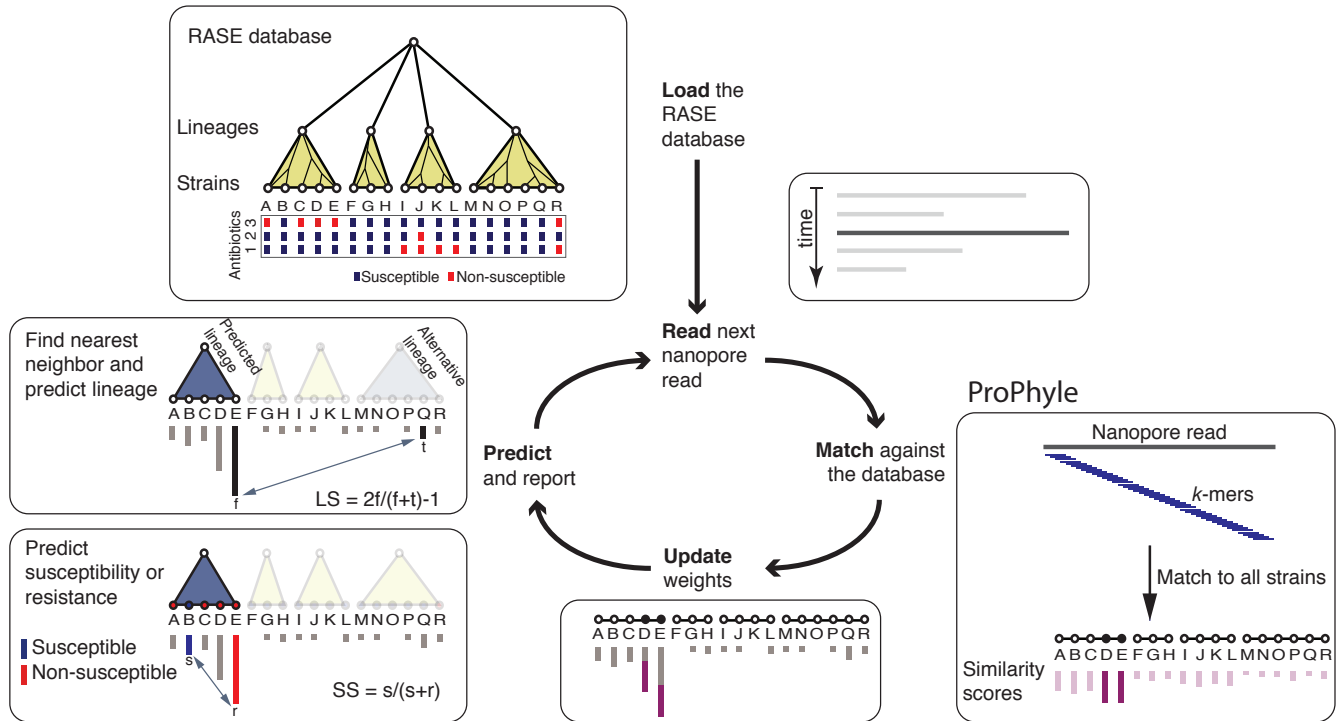
102 Based on the observed associations we developed an approach that we term ‘genomic neighbor  
103 typing’ to predict phenotype from sequencing data. Genomic neighbor typing is a two-step  
104 algorithm, which first compares a provided sample to a database of reference genomes with a  
105 known phylogeny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

108

109 To implement genomic neighbor typing we developed software called RASE (Resistance-  
110 Associated Sequence Elements) (**Figure 1**). RASE takes a stream of nanopore reads and compares  
111 their *k*-mer content to references using a modified version of ProPhyle<sup>27,28</sup>, a metagenomic  
112 classifier implementing a fast and memory-efficient exact colored de Bruijn graph data  
113 structure<sup>29</sup> using a BWT index<sup>30</sup> (Methods). Using ProPhyle RASE identifies which references are  
114 the most similar to the read and increases their similarity weights (this approach was inspired by  
115 but differs from other similar approaches such as Kraken<sup>31</sup> and Kallisto<sup>32</sup>). These weights are  
116 cumulative scores capturing sample-to-reference similarity; they are set to zero at the beginning  
117 and are increased on-the-fly as sequencing proceeds according to each read's 'information  
118 content' (Methods). Generally speaking, longer reads, such as those covering multiple accessory  
119 genes, tend to be specific and have high scores, whereas short reads or reads from the core  
120 genome are found in many lineages, tend to be non-specific and have low scores. Weights serve  
121 as a proxy to inverted genetic distance between the sample and the references.

122  
123 Resistance or susceptibility is predicted in two steps based on the computed weights, the  
124 population structure, and the reference phenotypes. First, RASE identifies the lineage of the best  
125 matching reference genome and estimates the confidence of lineage assignment by comparing  
126 the two best matching lineages to compute a 'lineage score' (Methods). Subsequently, RASE  
127 identifies the best match within that lineage and predicts resistance from the nearest resistant  
128 and susceptible neighbors. Comparison of their weights provides a 'susceptibility score', which  
129 quantifies the risk of resistance (Methods). When the weights are too similar, the call's  
130 confidence is considered low; this happens when resistant and susceptible strains are  
131 insufficiently genetically distinct, which is often the case for resistance emerging recently in  
132 evolutionary history (Methods). The ability to pinpoint the closest relatives in the database offers  
133 further resolution, even in the case where the resistance phenotype varies within a lineage.

134  
135 Results of RASE are reported in real time as the best match in the database, together with  
136 susceptibility scores to the antibiotics being tested and a proportion of matching *k*-mers for



**Figure 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).



137 quality control. As the run progresses, the scores fluctuate and eventually stabilize (an example  
138 shown in **Figure 2**).

139

#### 140 ***RASE databases for hundreds of *S. pneumoniae* and *N. gonorrhoeae* strains***

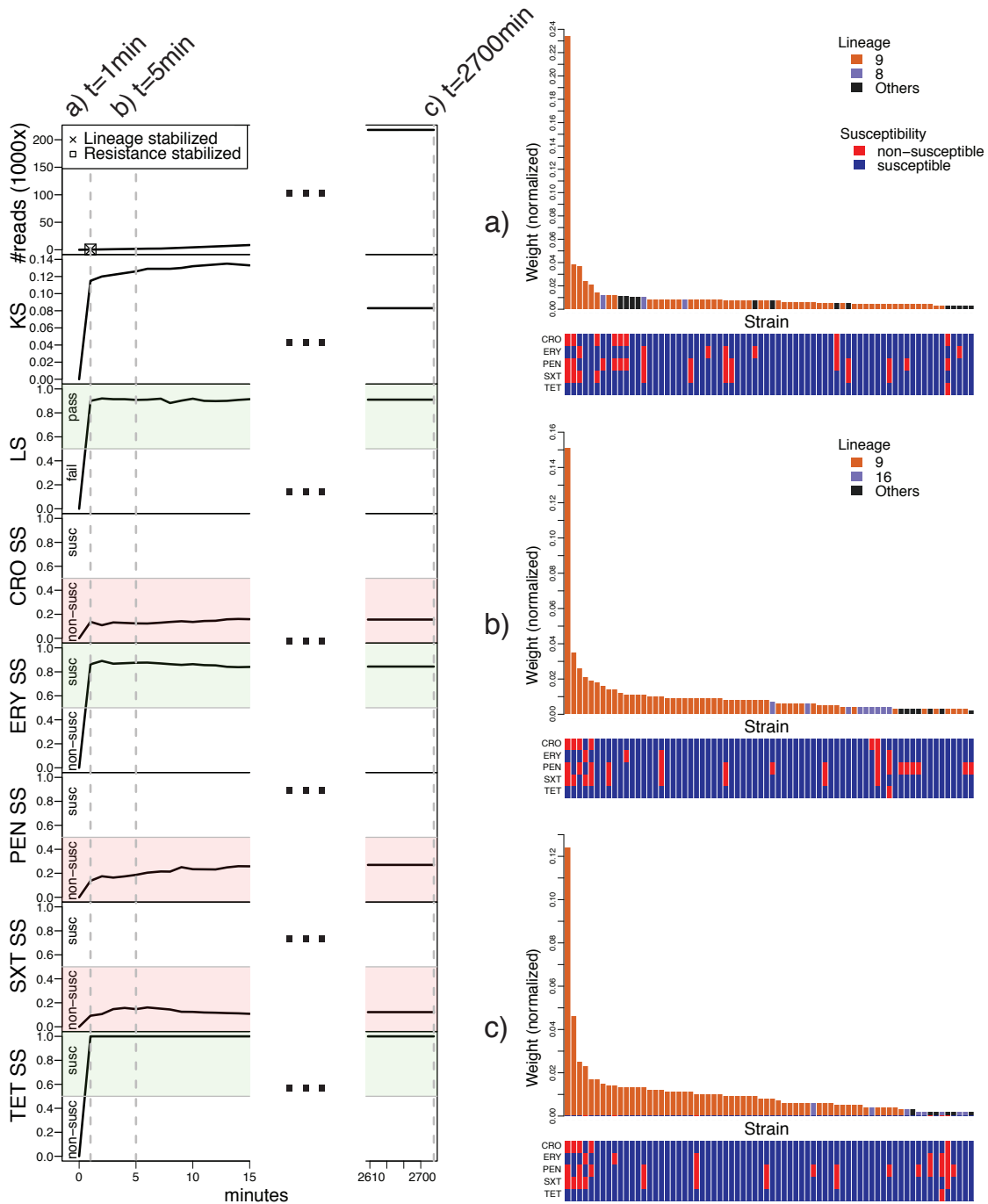
141

142 We constructed RASE databases for *S. pneumoniae* and *N. gonorrhoeae* from the same data as  
143 described above (Methods). We assigned each pneumococcal and gonococcal strains to an  
144 antibiotic-specific resistance categories using the European Committee on Antimicrobial  
145 Susceptibility Testing (EUCAST) breakpoints<sup>33</sup> and the CDC Gonococcal Isolate Surveillance  
146 Project (GISP) breakpoints<sup>34</sup>, respectively (Methods). Where MIC data were unavailable or  
147 insufficiently specific, we estimated the likely resistance phenotype using ancestral state  
148 reconstruction (Methods, **Supplementary Note 1**). To verify the results, we tested eight  
149 pneumococcal isolates for which resistance phenotypes were not originally available (Methods),  
150 and the measured MICs by microdilution matched the expected phenotypes (shown in bold in  
151 **Table 1**). We constructed the ProPhyle *k*-mer indexes with a *k*-mer length optimized to minimize  
152 prediction delays (*k*=18, Methods). The obtained pneumococcal and gonococcal RASE databases  
153 occupy 321 MB and 242 MB RAM (4.3× and 12× compression rate) and can be further  
154 compressed for transmission to 47 MB and 32 MB (29× and 90× compression rate), respectively  
155 (**Supplementary Figure 1**). This would allow RASE to be used on portable devices and its  
156 databases easily transmitted to the point of care over links with a limited bandwidth.

157

#### 158 ***RASE identifies strains in the database within minutes***

159 We first examined two pneumococcal isolates that were used to build the RASE database  
160 (**Table 1a**, sens/spec 100%/100%, n=10) to test RASE can function in ideal circumstances. In the  
161 case of a fully susceptible isolate (SP01), the correct lineage and sequenced strain were identified  
162 within 1 minute and 7 minutes respectively. A multidrug-resistant isolate (SP02) was predicted  
163 even faster, with both lineage and the sequenced strain correctly detected and stabilized within  
164 1 minute. To compare with gene-based approaches for detecting resistance<sup>22</sup> we evaluated how



**Figure 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. **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.

**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 <sup>(1)</sup>	S <sup>(1)</sup>	Yes	Yes
SP02	yes	9.6%	19A	19A	R	R	R	R	R	R	R	R	R <sup>(2)</sup>	R <sup>(2)</sup>	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 <sup>(2)</sup>	R	R	R	R	S	S	OoD	Yes
SP04	yes	12%	19A	19A	R	R	R	R	R	R	R	R	R	R <sup>(4)</sup>	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 <sup>(2)</sup>	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 <sup>(5)</sup>
SP08	no	2.5%	0.9%	S	S	S	S!	S	S <sup>(6)</sup>
SP09	no	4.0%	1.2%	NA	S	S	S	S	S <sup>(7)</sup>
SP10	yes	21%	5.2%	R	R	R	R	R	R <sup>(8)</sup>
SP11	yes	70%	14%	R	R	R	R	R	R <sup>(8)</sup>
SP12	yes	86%	17%	S	S	S	S	R	S <sup>(5)</sup>

**Legend**

S	Susceptible
R	Non-susceptible
!	Low confidence call
NA	Not available
OoD	Out-of-database
(...)	ID of a retested sample
SP	Fraction of <i>S. pneumoniae</i> reads

Correct prediction
Incorrect prediction
Cannot be evaluated

**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 3). Metagenomic samples are sorted by the estimated fraction of *S. pneumoniae* reads.

165 long it took for resistance genes to be sequenced on the device, and observed that at least 25  
166 minutes would be needed for single copies to be detected (**Supplementary Note 2**).

167  
168 We then performed a similar evaluation with five gonococcal isolates from the database  
169 (**Table 2a**, sens/spec 57%/100%, n=20); however, here we selected more complicated cases.  
170 First, we tested a susceptible isolate (GC01), for which RASE identified the correct strain and  
171 antibiogram within 3 minutes of sequencing. We then sequenced an isolate with a novel and  
172 uncommon mechanism of cephalosporin resistance that has emerged recently (GC02)<sup>35</sup>. Under  
173 such circumstances, the resistant strain and its susceptible neighbors tend to be genetically very  
174 similar, which could confound our analysis. However, RASE was still able to identify the correct  
175 resistance phenotypes in 9 minutes, with the delay being due to difficulty distinguishing between  
176 the close relatives, reflected also by a susceptibility score in the low-confidence range (Methods).  
177 This was repeated in further experiments with the same isolate (GC03) which consistently  
178 reported low confidence in resistance phenotype (Methods), which is a feature of our approach  
179 intended to draw operators' attention and indicate that further testing is necessary. In this  
180 experiment, RASE also resolved sample mislabeling (**Supplementary Note 3**). For a multidrug-  
181 resistant isolate (GC04) RASE predictions stabilized within 2 minutes but incorrectly predicted  
182 susceptibility to ceftriaxone. A subsequent analysis revealed that the ceftriaxone MIC of the  
183 sample was equal to the CDC GISP breakpoint (0.125 µg/mL), whereas the best match in the  
184 database had an MIC of 0.062 µg/mL, within a single doubling dilution. We further found that  
185 RASE performed well even with extremely poor data and low-quality reads (GC05,  
186 **Supplementary Note 4**). We also evaluated how genomic neighbor typing would perform if RASE  
187 used Kraken<sup>31</sup> instead of ProPhyle<sup>28</sup> (**Supplementary Note 5**).

188

### 189 ***RASE identifies the closest relative of novel isolates***

190 We next examined four novel pneumococcal isolates (**Table 1b**, sens/spec 89%/100%, n=20) for  
191 which the serotype and limited antibiogram and lineage data were known. We compared three  
192 characteristics of the sample to assess our performance: the serotype, the 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) Clinical 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

**Table 2: Predicted phenotypes of *N. gonorrhoeae* for a) database isolates and b) clinical isolates. The table is in the same format as Table 1.**

193 and the antibiograms (benzylpenicillin, ceftriaxone, trimethoprim-sulfamethoxazole,  
194 erythromycin, and tetracycline resistance according to the EUCAST breakpoints<sup>33</sup>).

195  
196 In all cases, the closest relative was identified within 5 minutes, even if the correct MLST  
197 sequence type was absent from the RASE database (an example shown in **Figure 2**). The two  
198 samples from the 23F clone (SP03 and SP06) were correctly called as being closely related to the  
199 Tennessee 23F-4 clone identified by PMEN, a clone strongly associated with macrolide  
200 resistance<sup>36</sup>. Consistent with this, the two samples were indeed resistant to erythromycin.  
201 However, the Tennessee 23F-4 clone was absent from the Massachusetts sample, with the best  
202 match being a comparatively distantly related strain that was penicillin resistant, but  
203 erythromycin susceptible. This illustrates the importance of a relevant database.

204  
205 We evaluated RASE with 14 clinical gonococcal isolates from the RaDAR-Go project<sup>37</sup>  
206 (Switzerland, 2015–2016) (**Table 2b**, sens/spec 93%/100%, n=56). These isolates were previously  
207 sequenced using nanopore and have full antibiograms available<sup>38</sup>. The 55/56 correct calls  
208 indicate the strength of the genomic neighbor typing in a clinical setting. The only incorrect call  
209 (susceptibility to azithromycin in GC15) was marked as being low-confidence call on the basis of a  
210 poor susceptibility score. It should be noted that the ranges for what is considered low-  
211 confidence could vary among settings and pathogens but can be empirically determined and  
212 modified by users. In this case our results suggest that informative results can be obtained even  
213 using a database from one region (the US) to predict phenotype in another (Europe). However,  
214 this may not be the case for all pathogens.

215  
216 ***Phenotyping is still informative but lower quality on highly divergent lineages***  
217 As noted above, an important precondition of genomic neighbor typing is a comprehensive and  
218 relevant reference database. To evaluate RASE performance in a setting with an incomplete  
219 database, we used the gonococcal WHO 2016 reference strain collection<sup>39</sup>. This includes a global  
220 collection of 14 diverse isolates from Europe, Asia, North America, and Australia, collected over

221 two decades and exhibiting phenotypes ranging from pan-susceptibility to multidrug resistance,  
222 and as such the GISP database is expected to be non-representative in this study. The WHO  
223 strains are available from the National Collection of Type Cultures, and were previously  
224 sequenced using nanopore<sup>38</sup> and genetically and phenotypically characterized<sup>39</sup>. Surprisingly,  
225 RASE correctly identified all MLST sequence types represented in the database and in 7 cases it  
226 provided fully correct resistance phenotypes (**Supplementary Table 1**, sens/spec 67%/91%,  
227 n=56). In 6/7 cases where the complete resistance profile was not recovered, the closest  
228 relatives were identified correctly but were genetically divergent from the query isolates  
229 (**Supplementary Note 6**). In one case, the errors were due to a misidentification of the closest  
230 relatives by ProPhyle. Therefore, most prediction errors could be addressed with a more  
231 comprehensive database.

232

### 233 ***RASE can identify resistance in pneumococcus from sputum metagenomic samples***

234 Because bacterial culture and phenotyping via agar-dilution, Etest, or disk diffusion introduces  
235 significant delays in resistance profiling, direct metagenomic sequencing of clinical samples  
236 would be preferable for point-of-care use. We therefore analyzed metagenomic nanopore data  
237 from sputum samples obtained from patients suffering from lower respiratory tract infections<sup>40</sup>  
238 (UK, 2017), selecting 6 samples from the study that were already known to contain *S.*  
239 *pneumoniae* (**Table 1c**, sens/spec 75%/100%, n=16).

240

241 One sample (SP10) contained DNA from multiple bacterial species. However, within 5 minutes  
242 sequence was identified belonging to the Swedish 15A-25 clone (ST63) which is also known to be  
243 associated with resistance phenotypes including macrolides and tetracyclines<sup>41</sup>. This sample was  
244 confirmed to be resistant to erythromycin, as well as clindamycin, tetracycline and oxacillin  
245 according to the EUCAST breakpoints<sup>33</sup>. The original report of the Swedish 15A-25 clone did not  
246 report resistance to penicillin antibiotics<sup>41</sup>, which has subsequently emerged in this lineage.  
247 However, our database correctly identified the risk of penicillin resistance in this sample. The  
248 metagenomes SP11 and SP12 contain an estimated >20% reads that matched to *S. pneumoniae*,

249 and their serotypes were identified to be 15A and 3, respectively. The susceptibility scores of the  
250 best matches were fully consistent with the resistance profiles found in the samples, with the  
251 exception of tetracycline resistance in SP12 due to an incomplete database (**Supplementary**  
252 **Note 7**). The last remaining samples, SP07–SP09, contained less than 5% unambiguously  
253 pneumococcal reads. Despite the low proportions, all predicted phenotypes were concordant  
254 with phenotypic tests, with the exception of SP07, which matched the same strain as SP12  
255 (discussed above).



256 **Discussion**

257

258 This paper presents a method that we term genomic neighbor typing to pinpoint the closest  
259 relatives of a query genome within a suitable database and then to infer the phenotypic  
260 properties of the query strain on the basis of the reported properties of its relatives. At present,  
261 the precise lineage of a bacterial pathogen is often determined after most important clinical  
262 decisions have been made. However, incorporating genomic neighbor typing at an earlier stage  
263 offers a way of leveraging bacterial population structure to gain information on resistance and  
264 susceptibility, and inform antimicrobial therapy. The results from the metagenomic samples  
265 suggest that it is possible to apply this approach directly to clinical samples, and the success with  
266 both *S. pneumoniae* and *N. gonorrhoeae* indicates that it may have wide application.

267

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

276

277 The initial identification of the closest relative is consistently more robust in the pneumococcus  
278 than the gonococcus, as a result of the former having more *k*-mers that are specific to an  
279 individual lineage, reflecting greater sequence diversity. As a consequence of the much lower  
280 diversity in gonococcus, when multiple closely related genomes are present in the database,  
281 RASE fluctuates between them, even though it correctly identifies the region of the phylogeny. If  
282 these genomes vary in their resistance profile, this is properly reflected in an uncertain  
283 susceptibility score indicating that caution and further investigation are merited (e.g., GC03).

284  
285 As in all inference, the principle limitation of genomic neighbor typing is the representativeness  
286 of the database. While we have made use of relatively small samples from limited geographic  
287 areas to demonstrate proof of principle, in practice there are multiple examples of large genome  
288 databases generated by public health agencies, which could be combined with metadata on  
289 resistance for genomic neighbor typing. Such databases could, if necessary, be supplemented  
290 with local sampling. The relevant question for our approach therefore becomes whether the  
291 database contains a sufficiently high proportion of strains that will be encountered in the clinic  
292 and whether the resistance data are correct. Further work is required to determine the optimal  
293 structure and contents of databases for each application, but we emphasize the range of  
294 pathogens which appear to show promise for this approach. These include *E. coli*, in which data  
295 on MLST type supplemented with epidemiologic information can consistently produce AUCs in  
296 excess of 0.90 for multiple antibiotics<sup>42</sup>, suggesting great potential for neighbor typing to offer  
297 excellent resolution superior to MLST. However, genomic neighbor typing may be less suitable in  
298 the case where there is little genomic variation (e.g., *Mycobacterium tuberculosis*) or when  
299 resistance emerges rapidly on independent and diverse genomic backgrounds (e.g.,  
300 *Pseudomonas aeruginosa* or resistance elements on highly promiscuous plasmids).  
301  
302 In the case where the infectious agent is unknown this problem is significantly more challenging.  
303 *K*-mers from one pathogen can match others and produce false predictions, and so choice of the  
304 correct database for prediction is key. Doing this will likely require a two-step solution in which  
305 the reads are first passed through a metagenomic classifier such as Centrifuge<sup>43</sup> or MetaMaps<sup>44</sup>,  
306 which would be used to select the correct RASE database on which to make a resistance call.  
307  
308 Another limitation is the time required for sample preparation, which currently includes human  
309 DNA depletion, DNA isolation, and library preparation, taking a total of 4 hours. This is a rapidly  
310 evolving area of technology and automated rapid library preparation kits are already in

311 development<sup>45</sup>. Further advances in this space, in particular for the preparation of metagenomic  
312 samples, will be required to bring the method closer to the bedside.

313

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

318

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

329

330 Genomic neighbor typing can also be used to detect other phenotypes that are sufficiently tightly  
331 linked to a phylogeny, such as virulence. Further applications may include rapid outbreak  
332 investigations, as the closely related isolates involved in the outbreak would all be predicted to  
333 match to the same strain in the RASE database. The approach also lends itself to enhanced  
334 surveillance, including in the field; the 2014–2016 Ebola outbreak in West Africa, for example,  
335 saw MinION devices used in remote locations without advanced healthcare facilities<sup>2</sup>. Finally, at  
336 present empiric treatment decisions are made within successive ‘windows’<sup>46</sup>, in which increasing  
337 information becomes available, from initial Gram stain to full phenotypic characterization. The  
338 information from genomic neighbor typing is a natural complement to this process with the

339 potential to improve therapy long before it would become clinically apparent that the patient is  
340 not responding or before phenotypic susceptibility data were available. The combination of high-  
341 quality RASE databases with genomic neighbor typing offers an alternative forward-looking  
342 model for diagnostics and surveillance, with wide applications for the improved clinical  
343 management of infectious disease.

344 **Methods**

345

346 ***Overview***

347

348 RASE uses rapid approximate *k*-mer-based matching of long sequencing reads against a database  
349 of strains to predict resistance via neighbor typing. The database contains a highly compressed  
350 exact *k*-mer index, a representation of the tree population structure, and metadata such as  
351 lineage, resistance profiles, MLST sequence type and serotype. The RASE prediction pipeline  
352 iterates over reads from the nanopore sequencer and provides real-time predictions of lineage  
353 and resistance or susceptibility (**Figure 1**).

354

355 ***Resistance profiles***

356

357 For all antibiotics, RASE associates individual strains with a resistance category, ‘susceptible’ (S)  
358 or ‘non-susceptible’ (R). First, intervals of possible MIC values are extracted using regular  
359 expressions from the available textual antibiograms. For instance, ‘>=4’, ‘2’, and ‘NA’ would be  
360 translated to the intervals  $[4,+\infty)$ ,  $[2,2]$ , and  $[0,+\infty)$ , respectively. Then the acquired intervals are  
361 compared to the antibiotic-specific breakpoints (see below; **Supplementary Figures 3 and 4**). If a  
362 given breakpoint is above or below the interval, susceptibility or non-susceptibility is reported,  
363 respectively. However, no category can be assigned at this step if the breakpoint lies within the  
364 extracted interval, an antibiogram is entirely missing, it is insufficiently specific, or its parsing  
365 failed. Finally, missing categories are inferred using ancestral state reconstruction on the  
366 associated phylogenetic tree while maximizing parsimony (i.e., minimizing the number of nodes  
367 switching its resistance category; **Supplementary Figures 5 and 6**). When the solution for a node  
368 is not unique, non-susceptibility is assigned.

369

370 ***Genomic neighbor typing***

371  
372 All reference strains in the database are associated with similarity weights that are set to zero at  
373 the start of the run. Each time a new read is read from the stream,  $k$ -mer-based matching is  
374 applied to identify the strains with the maximum number of matching  $k$ -mers (see below). Such  
375 strains are read's nearest neighbors in the database according to the  $1/(\text{'number of matched } k\text{-mers'})$   
376 pseudodistance.

377  
378 The weights of the nearest neighbors are then increased according to the 'information content'  
379 of the read, calculated as the number of matched  $k$ -mers divided by the number of nearest  
380 neighbors. Reads that do not match (i.e., 0 matching  $k$ -mers in the database) are not used in  
381 subsequent analysis. The computed matches are also used for updating the  $k$ -mer score (KS),  
382 which is the proportion of matched  $k$ -mers in all reads. KS helps to assess whether a sample is  
383 truly matching the database and predicting resistance for the database species makes sense.

384  
385 The obtained weights serve as a proxy to inverted genetic distance and are used as a basis for the  
386 subsequent predictions of the lineage, and antibiotic resistance and susceptibility.

387  
388 ***Predicting lineage***

389  
390 A lineage is predicted as the lineage of the best matching reference strain, i.e., the one with the  
391 largest weight. The quality of lineage prediction is further quantified using a lineage score (LS),  
392 calculated as  $LS=2f/(f+t)-1$ , where  $f$  and  $t$  denote the weights of the best matches in the first  
393 ('predicted') and in the second best ('alternative') lineage, respectively. The values of LS can  
394 range from 0.0 to 1.0 with the following special cases:  $LS=1.0$  means that all reads were perfectly  
395 matching the predicted lineage, whereas  $LS=0.0$  means that the predicted and alternative  
396 lineages were matched equally well.

397

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

402

### 403 ***Predicting resistance and susceptibility***

404

405 Resistance or susceptibility are predicted for individual antibiotics independently, based the  
406 weights of the strains that belong to the predicted lineage. These are used to calculate a  
407 susceptibility score, which is further interpreted by comparing to pre-defined thresholds.

408

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

417

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

425 require a further re-calibration, based on the specific database, antibiotics, and application of  
426 RASE.

427

#### 428 ***S. pneumoniae* RASE database**

429

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

437

438 The draft assemblies were downloaded from the SRA FTP server using the accession codes  
439 provided in Table 1 in ref<sup>24</sup>. The phylogenetic tree was downloaded from DataDryad (accession:  
440 '10.5061/dryad.t55gq'). The pneumococcal ProPhyle index was constructed with the *k*-mer size  
441 *k*=18.

442

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

445

#### 446 ***N. gonorrhoeae* RASE database**

447

448 The *N. gonorrhoeae* RASE database was constructed with the CDC GISP breakpoints<sup>34</sup> ([mg/L]):  
449 azithromycin (AZM): 2.0, cefixime (CFM): 0.25, ciprofloxacin (CIP): 1.0, and ceftriaxone (CRO):  
450 0.125. Before applying the breakpoints, azithromycin MICs for strains collected before 2005 were  
451 doubled in order to correct for the known inconsistencies of the phenotyping protocol due to a  
452 change in formulation of the commercial media<sup>47</sup>.



453

454 The draft assemblies and the phylogenetic tree were downloaded from Zenodo (accession:  
455 '10.5281/zenodo.2618836'). Three prevalent types of plasmids<sup>48</sup> were downloaded from  
456 GenBank, localized in the GISP database using BLAST<sup>49</sup>, and removed from the dataset: the  
457 cryptic plasmid ('pJD1', GenBank accession 'NC\_001377.1'), the beta-lactamase plasmid ('pJD4',  
458 GenBank accession 'NC\_002098.1'), and the conjugative plasmid ('pEP5289', GenBank accession  
459 'GU479466.1'). The gonococcal ProPhyle index was constructed with the  $k$ -mer size  $k=18$ .

460

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

463

#### 464 ***K-mer-based matching***

465

466 Reads were matched against the RASE databases using the ProPhyle classifier<sup>27,28</sup> (commit  
467 b55e026) and its ProPhex component<sup>50,51</sup>. ProPhyle index stores  $k$ -mers of all strains in a highly  
468 compressed form, reducing the required memory footprint. In the database construction phase,  
469 the strains'  $k$ -mers are first propagated along the phylogenetic tree and then greedily assembled  
470 to contigs. The obtained contigs are then placed into a single text file, for which a BWT index is  
471 constructed<sup>30</sup>.

472

473 In the course of sequencing, each read is decomposed into overlapping  $k$ -mers. The  $k$ -mers are  
474 then searched in the BWT index by ProPhex using BWT search using a sliding window<sup>50</sup>. For every  
475  $k$ -mer, the obtained matches are translated back on the tree. This provides a list of nodes whose  
476 descending leaves are the strains containing that  $k$ -mer. Finally, strains with maximum number of  
477 matched  $k$ -mers are identified for each read, and reported in the SAM/BAM format<sup>52</sup>.

478

#### 479 ***Optimizing k-mer length***

480

481 The  $k$ -mer length is the main parameter of the classification. First, the subword complexity  
482 function<sup>53</sup> of pneumococcus was calculated using JellyFish<sup>54</sup> (version 2.2.10) (**Supplementary**  
483 **Figure 7**). Then, based on the characteristics of the function and the  $k$ -mer range supported by  
484 ProPhyle, the possible range of  $k$  was determined as in [17, 32]. For these  $k$ -mer lengths, RASE  
485 indexes were constructed and their performance evaluated using the RASE prediction pipeline  
486 and selected experiments. While RASE showed robustness to  $k$ -mer length in terms of final  
487 predictions, prediction delays differed (**Supplementary Figure 8**). Based on the obtained timing  
488 data, we set  $k$  to 18.

489

### 490 ***Comparison to Kraken***

491

492 For each RASE database, a fake NCBI taxonomy was generated from the database tree. Then a  
493 library was built using Kraken<sup>31</sup> (v1.1.1, with default parameters) from the same FASTA files as  
494 used for building the RASE database. Finally, Kraken databases were constructed for both  $k=18$   
495 and  $k=31$ .

496

497 The obtained Kraken databases were used to classify reads from individual experiments. The  
498 obtained Kraken assignment were subsequently converted using an ad-hoc Python script to  
499 RASE-BAM (a subset of the BAM format<sup>52</sup> used by RASE). Finally, RASE prediction was applied on  
500 the BAM files, with the use of the RASE database metadata, and the results compared with the  
501 results of the standard RASE with ProPhyle.

502

### 503 ***Measuring time***

504

505 To determine how RASE works with nanopore data generated in real time, the timestamps of  
506 individual reads extracted were using regular expressions from the read names. These were then  
507 used for sorting the base-called nanopore reads by time. When the RASE pipeline was applied,  
508 the timestamps were used for expressing the predictions as a function of time. The times of

509 ProPhyle assignments were also compared to the original timestamps to ensure that the  
510 prediction pipeline was not slower than sequencing.

511  
512 When timestamps of sequencing reads were not available (i.e., the gonococcal WHO and clinical  
513 samples), RASE estimated the progress in time from the number of processed base pairs. This  
514 was done by dividing the cumulative base-pair count by the typical nanopore flow, which we had  
515 previously estimated from SP01 as 1.43Mbps per second. However, such an estimated progress  
516 is indicative only, as it does not follow the true order of reads in the course of sequencing. As the  
517 nanopore signal quality tends to decrease over time (see the decrease of KS in **Figure 2** after  
518 t=15mins), the randomized read order provides results of lower quality than true real-time  
519 sequencing.

520

#### 521 ***Lower time estimates on resistance gene detection***

522

523 A complete genome of the multidrug-resistant SP02 isolate was assembled from the nanopore  
524 reads using the CANU<sup>55</sup> (version 1.5, with default parameters). Prior to the assembly step, reads  
525 were filtered using SAMsift<sup>56</sup> based on the matching quality with the pneumococcal RASE  
526 database: only reads at least 1000bp long with at least 10% 18-mers shared with some of the  
527 reference draft assemblies were used. The obtained assembly was further corrected by Pilon<sup>57</sup>  
528 (version 1.2, default parameters) using Illumina reads from the same isolate (taxid '1QJAP' in the  
529 SPARC dataset<sup>24</sup>) mapped to the nanopore assembly using BWA-MEM<sup>58</sup> (version 0.7.17, with the  
530 default parameters) and sorted using SAMtools<sup>52</sup>.

531

532 The obtained assembly was searched for resistance-causing genes using the online CARD tool<sup>8</sup> (as  
533 of 2018/08/01). All of the original nanopore reads were then mapped using Minimap2<sup>59</sup> (version  
534 2.11, with '-x map-ont') to the corrected assembly and resistance genes in the reads identified  
535 using BEDtools-intersect<sup>60</sup> (version 2.27.1, with '-F 95'). Timestamps of the resistance-

536 informative reads were extracted and associated with the genes. Only reads longer than 2kbp  
537 were used in the analysis.

538

### 539 ***Evaluation of the N. gonorrhoeae WHO samples***

540

541 To evaluate the predictions of the WHO samples, we inferred a phylogenetic tree from a data set  
542 comprising both the GISP isolates and the WHO isolates. First, reads were downloaded for the  
543 GISP isolates (NCBI BioProject: 'PRJEB2999' and 'PRJEB7904') and for the WHO isolates F–P (NCBI  
544 BioProject: 'PRJEB4024'). For the WHO isolates U–Z, read data were simulated from the finished  
545 de-novo assemblies (NCBI BioProject: 'PRJEB14020') using Art-Illumina<sup>61</sup> (version 2.5.1). Reads  
546 were mapped to the NCCP11945 reference genome (GenBank accession: 'CP001050.1') using  
547 BWA-MEM<sup>58</sup> (version 0.7.17) and deduplicated using Picard<sup>62</sup> (version 2.8.0). Pilon<sup>57</sup> (version  
548 1.16, with '--mindepth 10 --minmq 20') was used to call variants and further filtered to include  
549 only 'pass' sites and sites where the alternate allele was supported with AF > 0.9. Gubbins<sup>63</sup>  
550 (version 2.3.4) with RAxML<sup>64</sup> (version 8.2.10) were run on the aligned pseudogenomes to  
551 generate the final recombination-corrected phylogeny (**Supplementary File 1**).

552

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

555

### 556 ***Library preparation***

557

558 For isolates SP01-SP06, cultures were grown in Todd–Hewitt medium with 0.5% yeast extract  
559 (THY; Becton Dickinson and Company, Sparks, MD) at 37°C in 5% CO<sub>2</sub> for 24 hrs. High-molecular-  
560 weight (>1 µg) genomic DNA was extracted and purified from cultures using DNeasy Blood and  
561 Tissue kit (QIAGEN, Valencia CA). DNA concentration was measured using Qubit fluorometer  
562 (Invitrogen, Grand Island NY). Library preparation was performed using the Oxford Nanopore  
563 Technologies 1D ligation sequencing kit SQK LSK108.

564

565 For experiments SP07-SP12, library preparation was performed using the ONT Rapid Low-Input  
566 Barcoding kit SQK-RLB001, with saponin-based host DNA depletion used for reducing the  
567 proportion of human reads. More details can be found in the original manuscript<sup>40</sup>.

568

569 For isolates GC01-GC05, cultures were grown on Chocolate-Agar media i.e., Difco GC base media  
570 containing 1% IsoVitaleX (Becton Dickinson Co., Franklin Lakes, NJ) and 1% Remel Hemoglobin  
571 (Thermo Fisher Scientific, Carlsbad, CA) at 37°C in 5% CO<sub>2</sub> for 20 hrs. For GC01-GC04 genomic  
572 DNA was extracted and purified from cultures using the PureLink Genomic DNA MiniKit (Thermo  
573 Fisher Scientific, Carlsbad, CA), and for GC05 DNA was extracted using the phenol-chloroform  
574 method<sup>65</sup>. Genomic DNA was extracted and purified from cultures using the PureLink Genomic  
575 DNA MiniKit (Thermo Fisher Scientific, Carlsbad, CA). DNA concentration was measured using the  
576 Qubit fluorometer (Invitrogen, Grand Island, NY). Library preparation was performed using the  
577 Oxford Nanopore Technologies 1D ligation sequencing kit SQK-LSK109.

578

### 579 ***MinION sequencing***

580

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

590

591 ***Testing resistance phenotype***

592

593 Additional retesting of SPARC isolates was done using microdilution. Organism suspensions were  
594 prepared from overnight growth on blood agar plates to the density of a 0.5 McFarland standard.  
595 This organism suspension was then diluted to provide a final inoculum of 10<sup>5</sup> to 10<sup>6</sup> CFU/mL.  
596 Microdilution trays were prepared according to the NCCLS methodology with cation-adjusted  
597 Mueller-Hinton broth (Sigma-Aldrich) supplemented with 5% lysed horse blood (Hemostat  
598 Laboratories)<sup>66,67</sup>. Penicillin (TRC Canada) and chloramphenicol (USB) concentrations ranged from  
599 0.016 to 16 µg/mL. Erythromycin (Enzo Life Sciences), tetracycline (Sigma-Aldrich), and  
600 trimethoprim-sulfamethoxazole (MP Biomedicals) concentrations ranged from 0.0625 to  
601 64 µg/mL. Ceftriaxone (Sigma-Aldrich) concentrations ranged from 0.007 to 8 µg/mL. The  
602 microdilution trays were incubated in ambient air at 35°C for 24 h. The MICs were then visually  
603 read and breakpoints applied. A list of individual microdilution measurements and the obtained  
604 resistance categories is provided in **Supplementary Table 2**.

605

606 Resistance of streptococcus in the metagenomic samples (SP07–SP12) was determined by agar  
607 diffusion using the EUCAST methodology and breakpoints<sup>33</sup>. First, the inoculated agar plates  
608 were incubated at 37 °C overnight and then examined for growth with the potential for re-  
609 incubation up to 48 hours. Then, the samples were screened to oxacillin: if the zone diameter  $r$   
610 was  $>20\text{mm}$ , the isolate was considered sensitive to benzylpenicillin, otherwise a full MIC  
611 measurement to benzylpenicillin was done. Finally, the isolate was screened for resistance to  
612 tetracycline ( $r \geq 25\text{mm}$  for sensitive,  $r < 22\text{mm}$  for resistant) and erythromycin ( $r \geq 22\text{mm}$  for  
613 sensitive,  $r < 19\text{mm}$  for resistant); when the isolate showed intermediate resistance, a full MIC  
614 measurement was done.

615

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

618

619 **Data, implementation and availability**

620  
621 RASE was developed using Python, GNU Make, GNU Parallel<sup>68</sup>, Snakemake<sup>69</sup>, and the ETE 3<sup>70</sup> and  
622 PySam<sup>52</sup> libraries, and was based on ProPhyle (commit b55e026). Bioconda<sup>71</sup> was used to ensure  
623 reproducibility of the software environments. All code, the generated databases and other  
624 supplementary materials are available under the MIT license from [https://github.com/c2-](https://github.com/c2-d2/rase-supplement)  
625 [d2/rase-supplement](https://github.com/c2-d2/rase-supplement). The analyses in the paper were performed with the following versions of  
626 the RASE databases: “*N. gonorrhoeae* GISP USA v1.4” and “*S. pneumoniae* SPARC USA v1.3”.  
627 Sequencing data for all experiments can be downloaded from Zenodo (accession:  
628 ‘10.5281/zenodo.3346055’); for the metagenomic experiments, only the filtered datasets (i.e.,  
629 after removing the remaining human reads *in silico*) were made publicly available.

630

631 **Acknowledgements**

632  
633 This work was supported by the Bill & Melinda Gates Foundation (GCGH GCE OPP1151010, KB  
634 and WPH), NIH – National Institute of Allergy and Infectious Diseases (R01 AI106786-05, KB), the  
635 Canadian Institutes of Health Research (MFE 152448, RSL), the Canadian Institutes for Health  
636 Research (a fellowship grant, DRM), and the David and Lucile Packard Foundation (MB). This  
637 paper presents independent research funded by the National Institute for Health Research  
638 (NIHR) under its Programme Grants for Applied Research Programme (Reference Number RP-PG-  
639 0514-20018, JOG), the UK Antimicrobial Resistance Cross Council Initiative (MR/N013956/1,  
640 JOG), Rosetrees Trust (A749, JOG), the University of East Anglia (JOG, TC), and Oxford Nanopore  
641 Technologies (JOG, TC). Portions of this research were conducted on the O2 and Odyssey high-  
642 performance compute clusters, supported by the Research Computing Groups at Harvard  
643 Medical School and at the Harvard Faculty of Arts and Sciences, respectively. The authors thank  
644 Joshua Metlay for providing the test isolates for experiments SP03–SP06, which were collected as  
645 part of a population-wide surveillance study done in the Philadelphia region, supported by NIH

646 (R01 AI46645), and to Brian J Arnold, Taj Azarian and Cristina M Herren for useful comments in  
647 various stages of this project.

648

649 **Transparency declarations**

650

651 JOG received financial support for attending ONT and other conferences and an honorarium for  
652 speaking at ONT headquarters. JOG received funding and consumable support from ONT for TC's

653 PhD studentship.

654



655 **Bibliography**

- 656 1. Kumar, A. *et al.* Duration of hypotension before initiation of effective antimicrobial  
657 therapy is the critical determinant of survival in human septic shock. *Crit. Care Med.* **34**,  
658 1589–1596 (2006).
- 659 2. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**,  
660 228–232 (2016).
- 661 3. Faria, N. R. *et al.* Mobile real-time surveillance of Zika virus in Brazil. *Genome Med.* **8**, 2–5  
662 (2016).
- 663 4. CDC. Antibiotic resistance threats in the United States, 2013. *Current* 114 (2013).  
664 doi:CS239559-B
- 665 5. Li, Y. *et al.* Penicillin-Binding Protein Transpeptidase Signatures for Tracking and Predicting  
666  $\beta$ -Lactam Resistance Levels in *Streptococcus pneumoniae*. *MBio* **7**, e00756-16 (2016).
- 667 6. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J. Antimicrob.*  
668 *Chemother.* **67**, 2640–2644 (2012).
- 669 7. Steiner, A., Stucki, D., Coscolla, M., Borrell, S. & Gagneux, S. KvarQ: Targeted and direct  
670 variant calling from fastq reads of bacterial genomes. *BMC Genomics* **15**, 1–12 (2014).
- 671 8. Jia, B. *et al.* CARD 2017: expansion and model-centric curation of the comprehensive  
672 antibiotic resistance database. *Nucleic Acids Res.* **45**, D566–D573 (2017).
- 673 9. Zankari, E. *et al.* PointFinder: a novel web tool for WGS-based detection of antimicrobial  
674 resistance associated with chromosomal point mutations in bacterial pathogens. *J.*  
675 *Antimicrob. Chemother.* **72**, 2764–2768 (2017).
- 676 10. Hunt, M. *et al.* ARIBA: rapid antimicrobial resistance genotyping directly from sequencing  
677 reads. *Microb. Genomics* **3**, 1–21 (2017).
- 678 11. Votintseva, A. A. *et al.* Same-day diagnostic and surveillance data for tuberculosis via  
679 whole genome sequencing of direct respiratory samples. *J. Clin. Microbiol.* JCM.02483-16  
680 (2017). doi:10.1128/JCM.02483-16
- 681 12. Rowe, W. P. M. & Winn, M. D. Indexed variation graphs for efficient and accurate  
682 resistome profiling. *Bioinformatics* **34**, 3601–3608 (2018).

- 683 13. Feldgarden, M. *et al.* Using the NCBI AMRFinder Tool to Determine Antimicrobial  
684 Resistance Genotype-Phenotype Correlations Within a Collection of NARMS Isolates.  
685 *bioRxiv* (2019). doi:10.1101/550707
- 686 14. Gupta, S. K. *et al.* ARG-ANNOT, a New Bioinformatic Tool To Discover Antibiotic Resistance  
687 Genes in Bacterial Genomes. *Antimicrob. Agents Chemother.* **58**, 212–220 (2014).
- 688 15. Rowe, W. *et al.* Search Engine for Antimicrobial Resistance: A Cloud Compatible Pipeline  
689 and Web Interface for Rapidly Detecting Antimicrobial Resistance Genes Directly from  
690 Sequence Data. *PLoS One* **10**, e0133492 (2015).
- 691 16. Kaminski, J. *et al.* High-Specificity Targeted Functional Profiling in Microbial Communities  
692 with ShortBRED. *PLOS Comput. Biol.* **11**, e1004557 (2015).
- 693 17. de Man, T. J. B. & Limbago, B. M. SSTAR, a Stand-Alone Easy-To-Use Antimicrobial  
694 Resistance Gene Predictor. *mSphere* **1**, 1–10 (2016).
- 695 18. Clausen, P. T. L. C., Zankari, E., Aarestrup, F. M. & Lund, O. Benchmarking of methods for  
696 identification of antimicrobial resistance genes in bacterial whole genome data. *J.*  
697 *Antimicrob. Chemother.* **71**, 2484–2488 (2016).
- 698 19. Yang, Y. *et al.* ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection  
699 from metagenomic data using an integrated structured ARG-database. *Bioinformatics* **32**,  
700 2346–2351 (2016).
- 701 20. Garner, E., Pruden, A., Heath, L. S. & Vikesland, P. DeepARG : A deep learning approach for  
702 predicting antibiotic resistance genes from metagenomic data. (2017).
- 703 21. Antonopoulos, D. A. *et al.* PATRIC as a unique resource for studying antimicrobial  
704 resistance. *Brief. Bioinform.* 1–9 (2017). doi:10.1093/bib/bbx083
- 705 22. Boolchandani, M., D’Souza, A. W. & Dantas, G. Sequencing-based methods and resources  
706 to study antimicrobial resistance. *Nat. Rev. Genet.* **2019** 1 (2019). doi:10.1038/s41576-  
707 019-0108-4
- 708 23. Croucher, N. J. *et al.* Population genomics of post-vaccine changes in pneumococcal  
709 epidemiology. *Nat. Genet.* **45**, 656–63 (2013).
- 710 24. Croucher, N. J. *et al.* Population genomic datasets describing the post-vaccine evolutionary

- 711 epidemiology of *Streptococcus pneumoniae*. *Sci. data* **2**, 150058 (2015).
- 712 25. Grad, Y. H. *et al.* Genomic Epidemiology of Gonococcal Resistance to Extended-Spectrum  
713 Cephalosporins, Macrolides, and Fluoroquinolones in the United States, 2000–2013. *J.*  
714 *Infect. Dis.* **214**, 1579–1587 (2016).
- 715 26. Cheng, L., Connor, T. R., Siren, J., Aanensen, D. M. & Corander, J. Hierarchical and Spatially  
716 Explicit Clustering of DNA Sequences with BAPS Software. *Mol. Biol. Evol.* **30**, 1224–1228  
717 (2013).
- 718 27. Břinda, K., Salikhov, K., Pignotti, S. & Kucherov, G. ProPhyle: An accurate, resource-frugal  
719 and deterministic DNA sequence classifier. (2017). doi:10.5281/zenodo.1045429
- 720 28. Břinda, K. Novel computational techniques for mapping and classifying Next-Generation  
721 Sequencing data. *PhD Thesis, Université Paris-Est* (2017).
- 722 29. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and  
723 genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44**, 226–232 (2012).
- 724 30. Ferragina, P. & Manzini, G. Opportunistic data structures with applications. in *Proceedings*  
725 *41st Annual Symposium on Foundations of Computer Science* 390–398 (IEEE Comput. Soc,  
726 2000). doi:10.1109/SFCS.2000.892127
- 727 31. Wood, D. E. & Salzberg, S. L. Kraken: Ultrafast metagenomic sequence classification using  
728 exact alignments. *Genome Biol.* **15**, (2014).
- 729 32. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq  
730 quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
- 731 33. *The European Committee on Antimicrobial Susceptibility Testing. Breakpoint tables for*  
732 *interpretation of MICs and zone diameters. Version 7.0.* (2017).
- 733 34. CDC. 0. (2018).
- 734 35. Palace, S. *et al.* RNA polymerase mutations cause cephalosporin resistance in clinical  
735 *Neisseria gonorrhoeae* isolates. *bioRxiv* (2019). doi:10.1101/626457
- 736 36. McGee, L. *et al.* Nomenclature of Major Antimicrobial-Resistant Clones of *Streptococcus*  
737 *pneumoniae* Defined by the Pneumococcal Molecular Epidemiology Network. *J. Clin.*  
738 *Microbiol.* **39**, 2565–2571 (2001).

- 739 37. Donà, V. *et al.* Mismatch Amplification Mutation Assay-Based Real-Time PCR for Rapid  
740 Detection of *Neisseria gonorrhoeae* and Antimicrobial Resistance Determinants in Clinical  
741 Specimens. *J. Clin. Microbiol.* **56**, 1–10 (2018).
- 742 38. Golparian, D. *et al.* Antimicrobial resistance prediction and phylogenetic analysis of  
743 *Neisseria gonorrhoeae* isolates using the Oxford Nanopore MinION sequencer. *Sci. Rep.* **8**,  
744 17596 (2018).
- 745 39. Unemo, M. *et al.* The novel 2016 WHO *Neisseria gonorrhoeae* reference strains for global  
746 quality assurance of laboratory investigations: phenotypic, genetic and reference genome  
747 characterization. *J. Antimicrob. Chemother.* **71**, 3096–3108 (2016).
- 748 40. Charalampous, T. *et al.* Nanopore metagenomics enables rapid clinical diagnosis of  
749 bacterial lower respiratory infection. *Nat. Biotechnol.* 387548 (2019). doi:10.1038/s41587-  
750 019-0156-5
- 751 41. Sá-Leão, R. *et al.* Carriage of internationally spread clones of *Streptococcus pneumoniae*  
752 with unusual drug resistance patterns in children attending day care centers in Lisbon,  
753 Portugal. *J. Infect. Dis.* **182**, 1153–60 (2000).
- 754 42. MacFadden, D. R. *et al.* Comparing Patient Risk Factor-, Sequence Type-, and Resistance  
755 Locus Identification-Based Approaches for Predicting Antibiotic Resistance in *Escherichia*  
756 *coli* Bloodstream Infections. *J. Clin. Microbiol.* **57**, 1–9 (2019).
- 757 43. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive  
758 classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
- 759 44. Dilthey, A., Jain, C., Koren, S. & Phillippy, A. MetaMaps - Strain-level metagenomic  
760 assignment and compositional estimation for long reads. *bioRxiv* 372474 (2018).  
761 doi:10.1101/372474
- 762 45. Quick, J. *Ultra-long read sequencing protocol for RAD004 (Version 3)*. (2018).  
763 doi:10.17504/protocols.io.mrxc57n
- 764 46. MacFadden, D. R., Leis, J. A., Mubareka, S. & Daneman, N. The Opening and Closing of  
765 Empiric Windows: The Impact of Rapid Microbiologic Diagnostics. *Clin. Infect. Dis.* **59**,  
766 1199–1200 (2014).

- 767 47. Kersh, E. N. *et al.* Rationale for a *Neisseria gonorrhoeae* Susceptible Only Interpretive  
768 Breakpoint for Azithromycin. *Clin. Infect. Dis.* **30329**, 1–7 (2019).
- 769 48. Cehovin, A. & Lewis, S. B. Mobile genetic elements in *Neisseria gonorrhoeae*: movement  
770 for change. *Pathog. Dis.* **75**, 1–12 (2017).
- 771 49. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment  
772 search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
- 773 50. Salikhov, K. Efficient algorithms and data structures for indexing DNA sequence data. *PhD*  
774 *Thesis, Université Paris-Est* (2017).
- 775 51. Břinda, K., Salikhov, K., Pignotti, S. & Kucherov, G. ProPhex: A lossless k-mer index based  
776 on the Burrows-Wheeler Transform. (2018). doi:10.5281/zenodo.1247431
- 777 52. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9  
778 (2009).
- 779 53. Lothaire, M. *Algebraic Combinatorics on Words*. (Cambridge University Press, 2002).  
780 doi:10.1017/CBO9781107326019
- 781 54. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of  
782 occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
- 783 55. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer  
784 weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
- 785 56. Břinda, K., Baym, M. & Hanage, W. P. SAMsift: advanced filtering and tagging of SAM/BAM  
786 alignments using Python expressions. (2018). doi:10.5281/zenodo.1048211
- 787 57. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection  
788 and genome assembly improvement. *PLoS One* **9**, (2014).
- 789 58. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.  
790 *arXiv* 3 (2013).
- 791 59. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 1–3 (2018).  
792 doi:10.1093/bioinformatics/bty191
- 793 60. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic  
794 features. *Bioinformatics* **26**, 841–842 (2010).

- 795 61. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read  
796 simulator. *Bioinformatics* **28**, 593–594 (2012).
- 797 62. Broad Institute, G. repository. Picard Tools.
- 798 63. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial  
799 whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).
- 800 64. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large  
801 phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
- 802 65. Green, M. R. & Sambrook, J. Isolation of High-Molecular-Weight DNA Using Organic  
803 Solvents. *Cold Spring Harb. Protoc.* **2017**, pdb.prot093450 (2017).
- 804 66. CLSI. *Susceptibility Tests for Bacteria That Grow Aerobically; Approved Standard—Ninth*  
805 *Edition. CLSI document M07-A9* (2012).
- 806 67. CLSI. *Performance Standards for Antimicrobial Susceptibility Testing; Twenty-Second*  
807 *Informational Supplement. CLSI document M100-S22* (2012).
- 808 68. Tange, O. GNU Parallel: the command-line power tool. *login USENIX Mag.* **36**, 42–47  
809 (2011).
- 810 69. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine.  
811 *Bioinformatics* **28**, 2520–2522 (2012).
- 812 70. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, analysis and visualization of  
813 phylogenomic data. *Mol. Biol. Evol.* **33**, msw046 (2016).
- 814 71. Grüning, B. *et al.* Bioconda: sustainable and comprehensive software distribution for the  
815 life sciences. *Nat. Methods* **15**, 475–476 (2018).
- 816 72. Unemo, M., Fath, O., Fredlund, H., Limnios, A. & Tapsall, J. Phenotypic and genetic  
817 characterization of the 2008 WHO *Neisseria gonorrhoeae* reference strain panel intended  
818 for global quality assurance and quality control of gonococcal antimicrobial resistance  
819 surveillance for public health purposes. *J. Antimicrob. Chemother.* **63**, 1142–1151 (2009).
- 820 73. Ohnishi, M. *et al.* Is *Neisseria gonorrhoeae* Initiating a Future Era of Untreatable  
821 Gonorrhea?: Detailed Characterization of the First Strain with High-Level Resistance to  
822 Ceftriaxone. *Antimicrob. Agents Chemother.* **55**, 3538–3545 (2011).

- 823 74. Unemo, M. *et al.* High-Level Cefixime- and Ceftriaxone-Resistant *Neisseria gonorrhoeae* in  
824 France: Novel penA Mosaic Allele in a Successful International Clone Causes Treatment  
825 Failure. *Antimicrob. Agents Chemother.* **56**, 1273–1280 (2012).
- 826 75. Vincent, L. R. *et al.* In Vivo -Selected Compensatory Mutations Restore the Fitness Cost of  
827 Mosaic penA Alleles That Confer Ceftriaxone Resistance in *Neisseria gonorrhoeae*. *MBio* **9**,  
828 1–18 (2018).
- 829 76. Page, A. J. & Keane, J. A. Rapid multi-locus sequence typing direct from uncorrected long  
830 reads using Krocus. *PeerJ* **6**, e5233 (2018).
- 831 77. Croucher, N. J. *et al.* Dominant Role of Nucleotide Substitution in the Diversification of  
832 Serotype 3 Pneumococci over Decades and during a Single Infection. *PLoS Genet.* **9**,  
833 e1003868 (2013).
- 834 78. Azarian, T. *et al.* Global emergence and population dynamics of divergent serotype 3  
835 CC180 pneumococci. *PLOS Pathog.* **14**, e1007438 (2018).
- 836

837 **Supplementary notes**

838

839 **Supplementary Note 1.** Out of all 616 pneumococcal strains (**Supplementary Table 4a**), after the  
840 ancestral reconstruction step 485 were associated with susceptibility to ceftriaxone, 484 to  
841 erythromycin, 341 to benzylpenicillin, 480 to trimethoprim-sulfamethoxazole, and 551 to  
842 tetracycline (**Supplementary Table 5a**). In case of gonococcus, ancestral reconstruction was  
843 needed only for cefixime (62 records affected). Out of all 1102 gonococcal strains  
844 (**Supplementary Table 4b**), 808 were associated with susceptibility to azithromycin, 833 to  
845 cefixime, 508 to ciprofloxacin, and 1033 to ceftriaxone (**Supplementary Table 5b**). In our  
846 subsequent experiments, if original MIC data were not available for the best match in the RASE  
847 database, the relevant strain was tested to confirm resistance phenotype (Methods).

848

849 **Supplementary Note 2.** We evaluated how long it took for resistance genes to be reliably  
850 detected in nanopore reads. For SP02 we observed that at least 25 minutes were needed to  
851 detect resistance (i.e., to observe all resistance genes at least once), assuming that the genes in  
852 question can be unambiguously identified in nanopore data despite the high per-base error rate,  
853 and that the presence of the loci is directly linked to the resistance phenotype (**Supplementary**  
854 **Figure 2**). If this is not the case (for example if resistance is conferred by a single SNP, requiring  
855 coverage with multiple reads), further delays would be expected. Thus, genomic neighbor typing  
856 can offer a time advantage compared to methods based on identifying the presence of resistance  
857 genes even in a sample of DNA from a purified isolate as opposed to a metagenome, potentially  
858 allowing for more rapid changes to antimicrobial therapy.

859

860 **Supplementary Note 3.** We originally attempted to evaluate a multidrug-resistant isolate  
861 (GCGS0938 in the GISP collection); however, RASE placed it onto a distant part of the phylogeny  
862 and identified it as GCGS0324 or GCGS1095. A subsequent analysis revealed that the sample was  
863 mislabeled and that it was indeed GCGS1095, i.e., the same strain as in GC02, although from a  
864 different stock.



865  
866 **Supplementary Note 4.** We evaluated how RASE performs in extremely unfavorable sequencing  
867 conditions; we sequenced an isolate (GC05) from the GISP collection with the use of an expired  
868 flow cell (purchased in October 2017, expired in December 2017, and the sequencing done in  
869 April 2018). In consequence, we obtained only 3.5 Mbps of low-quality reads (only 7% of  
870 matching  $k$ -mers compared to 20% obtained in the other isolates) (GC05 in **Table 2a**). An  
871 experiment with such a low yield would normally be discarded; despite that RASE provided  
872 correct and stabilized predictions (once the first long read was obtained from the sequencer at  
873  $t=21$ mins).

874  
875 **Supplementary Note 5.** We evaluated how genomic neighbor typing would perform if RASE used  
876 Kraken<sup>31</sup> instead of ProPhyle<sup>28</sup> for the read-to-strain comparison (the matching step in **Figure 1**).  
877 Both tools use  $k$ -mer-based matching to assign sequencing reads to a phylogenetic tree, but with  
878 several key differences. Whereas Kraken stores for each  $k$ -mer the lowest common ancestor  
879 (LCA) only, assigns reads to the LCA of the best hits and ignores low-complexity  $k$ -mers, ProPhyle  
880 indexes all  $k$ -mers using an exact index and can thus resolve ambiguities both on the level of  
881 individual  $k$ -mers and read assignments.

882  
883 To compare both tools, we implemented a RASE wrapper for Kraken (Methods) and applied that  
884 to the same read and database data. We then compared the final inference results obtained with  
885 Kraken (with  $k=18$  and  $k=31$ ) with the results obtained from the standard RASE pipeline  
886 (**Supplementary File 2**).

887  
888 For *S. pneumoniae* and *N. gonorrhoeae*, the number of inference errors increased more than 1.5x  
889 and 1.7x, respectively (in case of both  $k$ -mer sizes). In the case of *N. gonorrhoeae*, RASE-Kraken  
890 showed large systematic biases in neighbor typing, assigning 16 ( $k=18$ ) and 18 ( $k=31$ ) out of the  
891 gonococcal 33 samples to a single strain (GCGS1028), whereas RASE-ProPhyle identified this  
892 strain only once. While in the WHO dataset the numbers of RASE-ProPhyle and RASE-Kraken

893 errors were comparable (10 vs. 12 and 11), in the RaDAR-Go dataset it increased from 1 to 8 and  
894 10. Overall, the obtained results suggest that Kraken is less suited for the use in genomic  
895 neighbor typing than ProPhyle.

896  
897 **Supplementary Note 6.** We analyzed the results of the WHO gonococcal samples  
898 (**Supplementary Table 1**). First, we evaluated the RASE ability to predict MLST sequence types. In  
899 all cases, either RASE predicted the correct sequence type (n=9), or the true sequence type was  
900 not present in the reference database (n=5). The latter was the case only in the samples F  
901 through P, which belonged to the initial 2008 WHO reference panel and were collected primarily  
902 in the late 1990s, with the majority of specimens isolated from the Eastern Hemisphere<sup>72</sup>. The  
903 GISP database, comprising strains collected in the US from 2000–2013, may not be  
904 representative then of the circulating lineages in those regions during that time span, which  
905 could result in both sequence type and antibiogram prediction errors. However, we observed  
906 perfect prediction of sequence types in the additional 2016 WHO reference strains comprising U  
907 through Z that were collected in 2007 and onwards<sup>39</sup>.

908  
909 We next sought to evaluate the resistance predictions. In 7 cases (F, K, N, O, P, U, W), the  
910 antibiograms were identified fully correctly; in 4 (G, V, X, Z) and 3 cases (L, M, Y) one and two  
911 mistakes were made, respectively. To explain these discrepancies, we inferred a recombination-  
912 corrected phylogenetic tree comprising the GISP database isolates as well as the WHO samples  
913 (**Supplementary File 1**). With the exception of G and Y, the WHO isolates and their respective  
914 RASE-predicted best matches were the closest GISP isolates, indicative of accurate matching by  
915 RASE. While branch lengths of L, M and V on the tree reveal that the corresponding parts of the  
916 phylogeny are not well sampled in the database, the X, Y, and Z samples emerged from lineages  
917 that are well-represented but have acquired an atypically high level of cephalosporin resistance.  
918 Whereas X and Z acquired a novel resistance-conferring mosaic penA allele<sup>73</sup>, Y acquired a novel  
919 active site mutation in the context of a pre-existing mosaic penA allele<sup>74</sup>. While both of these  
920 adaptations resulted in high-level resistance, these mutations also appear to incur fitness costs in

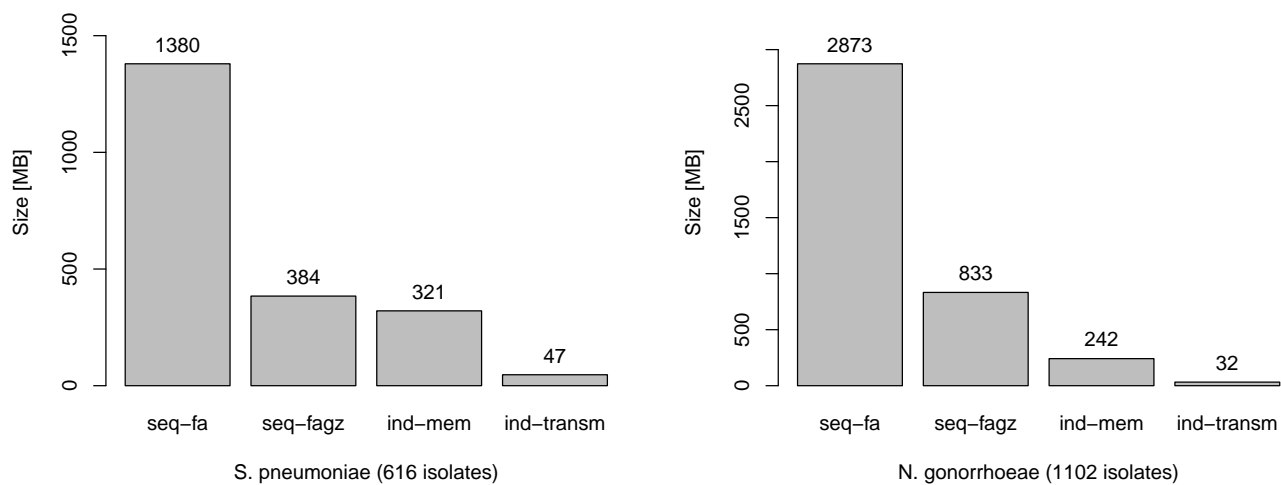
921 vitro and in the gonococcal mouse model<sup>75</sup>. In line with this, these strains have only been  
922 sporadically observed in genomic surveillance of clinical isolates. These results highlight how  
923 ancestral or emerging resistant lineages may not be well-captured by sequence-based methods  
924 including RASE and emphasize the value of continuous updating of the RASE database for public  
925 health.

926

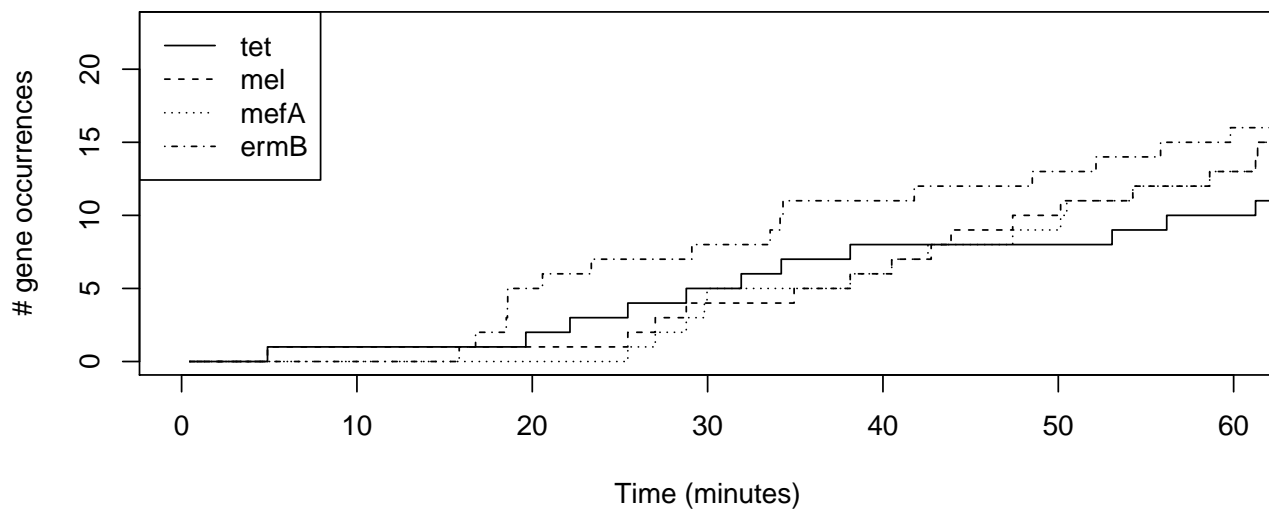
927 **Supplementary Note 7.** Further analysis of the reads from SP12 using Krocus<sup>76</sup> suggested that  
928 the pneumococcal DNA present was from the ST180 clonal complex, and matched specifically  
929 either to the sequence type ST180 or ST3798. This is consistent with identification as serotype 3,  
930 because this clonal complex contains the great majority of strains with this capsule type, which  
931 historically has not been associated with resistance<sup>77</sup>. However, improved sampling and study of  
932 this lineage has recently found highly divergent subclades that are associated with resistance.  
933 These lineages were previously rare, and thus were less likely to be included in our database, but  
934 now are increasing in frequency<sup>78</sup>. In this case, ST3798 is found to be in clade 1B, which is  
935 notable for exhibiting sporadic tetracycline resistance. Again, the failure to match to this is a  
936 result of the original database not containing a suitable example for comparison.

937

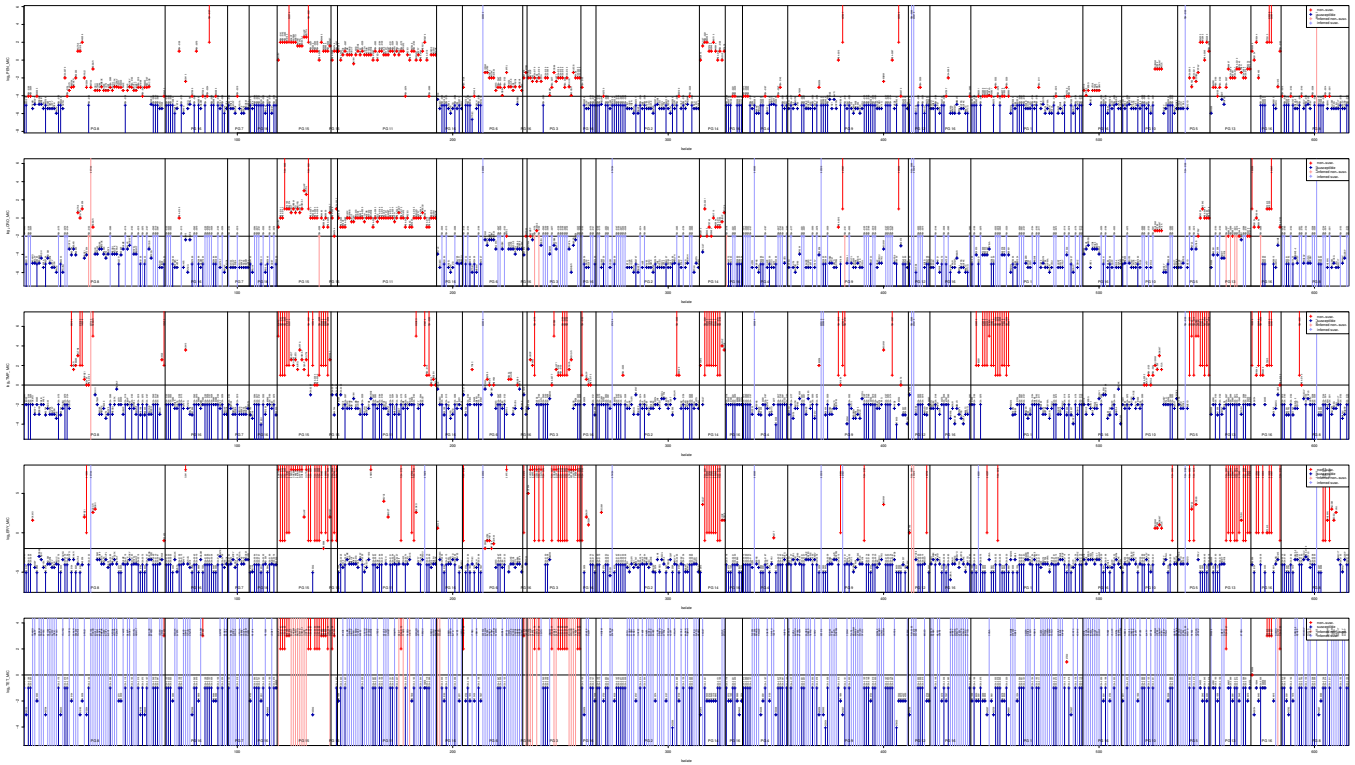
938



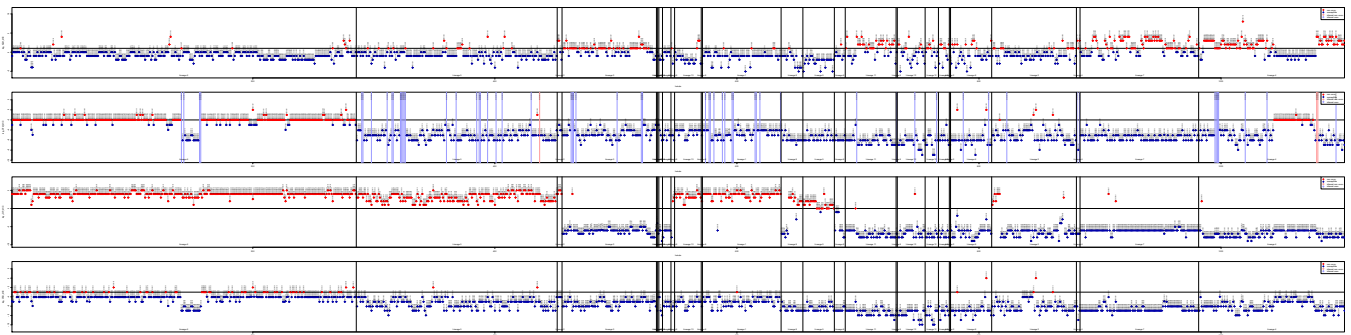
**Supplementary Figure 1: Size and memory footprint of the *S. pneumoniae* and *N. gonorrhoeae* RASE databases.** The graph compares the size of the ProPhyle RASE index to the size of the original sequences: original draft assemblies (seq-fa), original draft assemblies compressed using gzip (seq-fagz),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



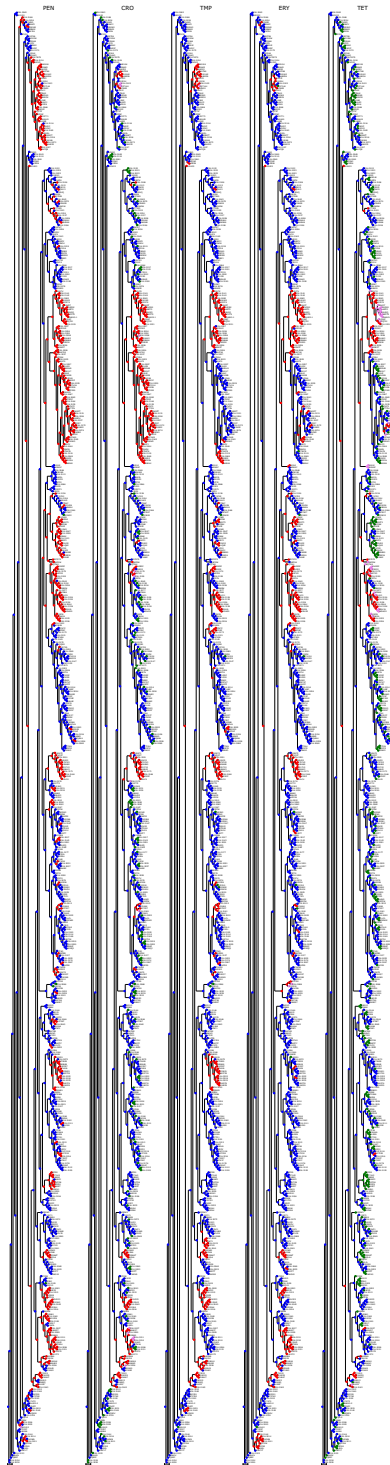
**Supplementary Figure 2: Timeline of resistance genes.** Number of occurrences of individual resistance genes in reads of SP02, as a function of time for the first hour of nanopore sequencing.



**Supplementary Figure 3: MIC intervals for individual strains in the *S. pneumoniae* 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 strains. 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.

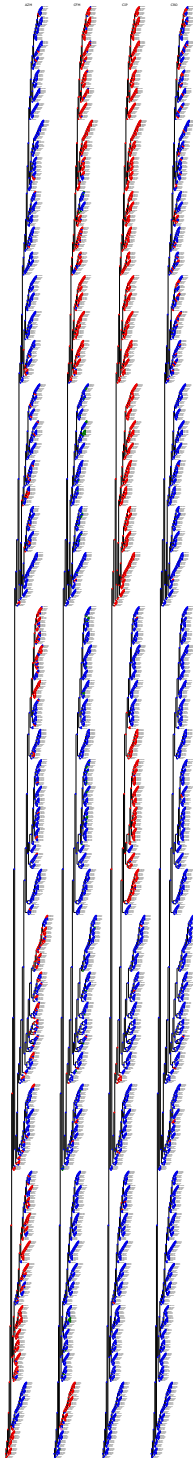


**Supplementary Figure 4: MIC intervals for individual strains in the *N. gonorrhoeae* RASE database.** The figure is of the same format as Supplementary Figure 3.



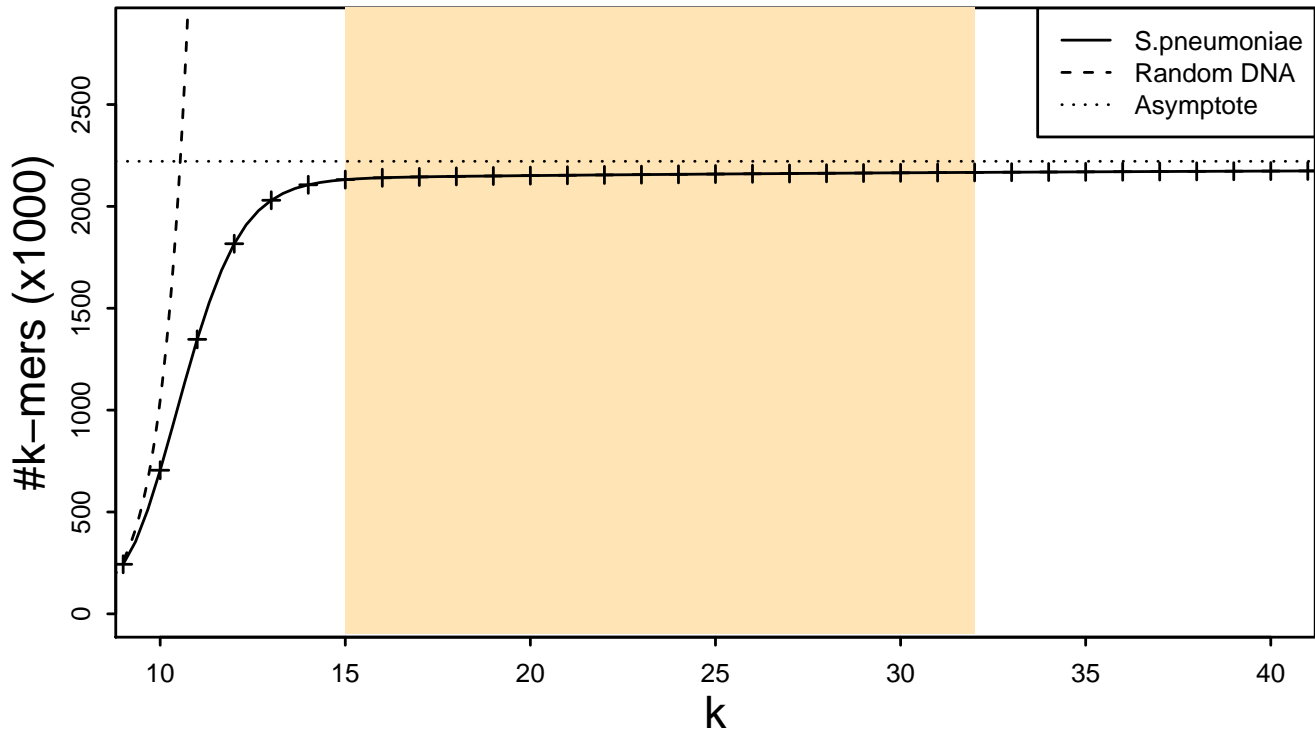
**Supplementary Figure 5: Ancestral state reconstruction of resistance categories in the *S. pneumoniae* RASE database.** Each panel corresponds to a single antibiotic and displays the database phylogenetic tree, colored according to the reconstructed resistance categories for the antibiotic (blue, green, red, violet correspond to 'susceptible', 'unknown – inferred susceptible', 'non-susceptible', 'unknown – inferred non-susceptible', respectively).



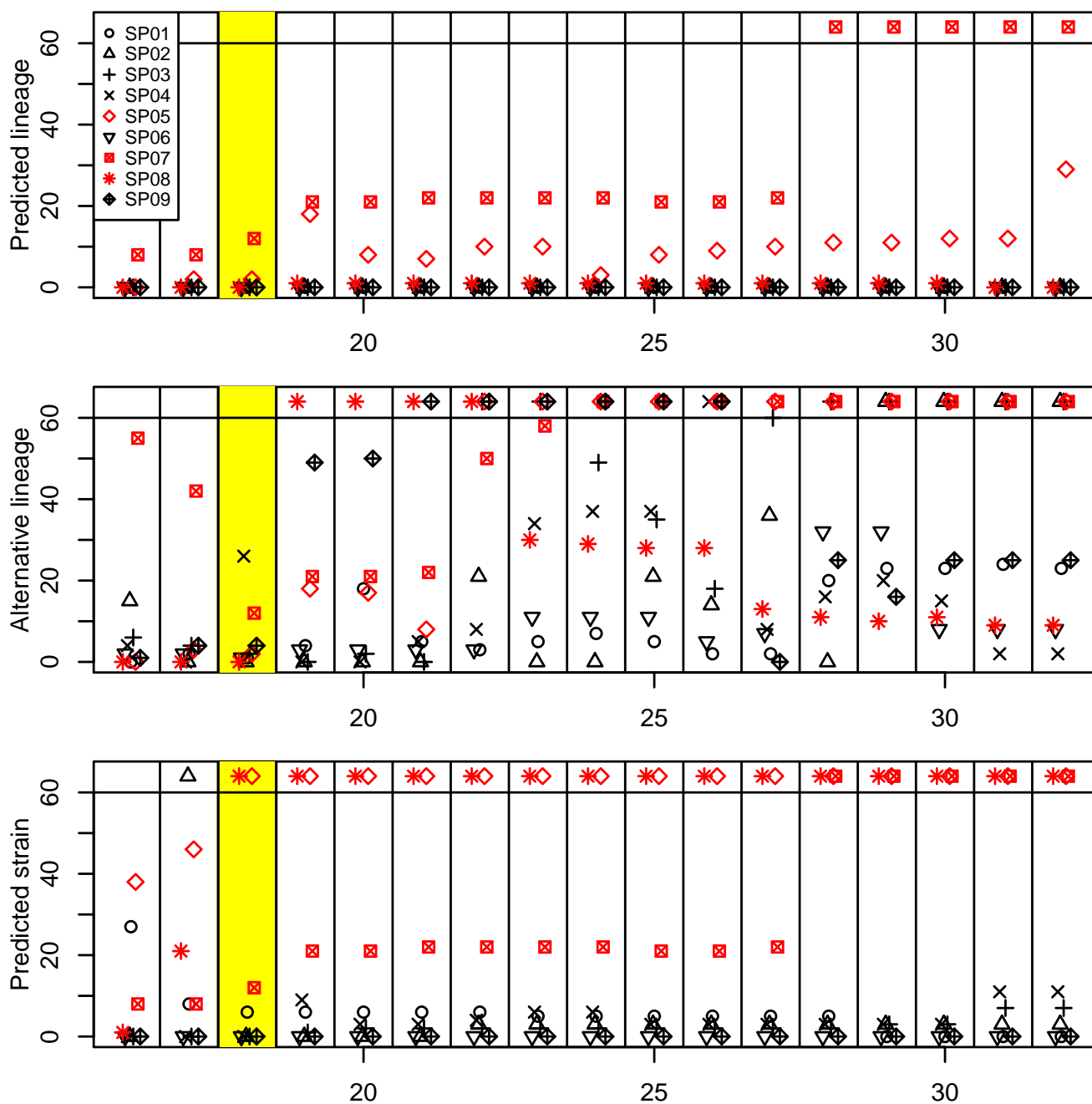


**Supplementary Figure 6: Ancestral state reconstruction of resistance categories in the *N. gonorrhoeae* RASE database.** The figure is of the same format as Supplementary Figure 5.

**Subword complexity function of *S. pneumoniae***



**Supplementary Figure 7: 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.



**Supplementary Figure 8: 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 File 1: Comprehensive phylogenetic tree for *N. gonorrhoeae*.** A recombination-corrected tree in the Newick format comprising the GISP database isolates and the WHO samples.

**Supplementary File 2: Comparison of ProPhyle- and Kraken-powered genomic neighbor typing.** The spreadsheet shows the final resistance and susceptibility inference calls for the ProPhyle (k=18) and Kraken (k=18 and k=31) classifiers plugged into RASE; erroneous calls are highlighted in red.

### WHO lineages

Sample	Region	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	
WHO F (2008)	Canada	no	17%	S	S!	S	S	S	S	S	S	OoD
WHO G (2008)	Thailand	no	14%	S	S	S	S	S	R	S	S	OoD
WHO K (2008)	Japan	yes	20%	S	S	R	R	R	R	S	S	yes
WHO L (2008)	Asia	yes	20%	S	S	S	R	R	S	R	R	OoD
WHO M (2008)	Philippines	yes	21%	S	R	S	S	R	S	S	S	yes
WHO N (2008)	Australia	no	19%	S	S	S	S	R	R	S	S	OoD
WHO O (2008)	Canada	yes	20%	S	S	S	S	S	S	S	S	yes
WHO P (2008)	USA	yes	19%	R	R	S	S	S	S	S	S	OoD
WHO U (2016)	Sweden	yes	20%	R	R	S	S	S	S	S	S	yes
WHO V (2016)	Sweden	yes	19%	R	S	S	S	R	R	S	S	yes
WHO W (2016)	Hong Kong	yes	20%	S	S	R	R	R	R	S	S	yes
WHO X (2016)	Japan	yes	21%	S	S	R	R	R	R	R	S	yes
WHO Y (2016)	France	no	18%	S	S	R	S	R	R	R	S	yes
WHO Z (2016)	Australia	yes	19%	S	S	R	R	R	R	R	S	yes

**Supplementary Table 1: Predicted phenotypes of *N. gonorrhoeae* for the WHO lineages.** The table is in the same format as Table 1.

**Supplementary Table 2: Additional MIC measurements for selected strains.** The table displays results from strain retesting. Each record contains date when the retesting was done, the antibiotic, the measured MIC, and the corresponding resistance category.

**Supplementary Table 3: Overview of performed resistance tests.** For all sequencing experiments, the table displays the best matching strains, their MICs, 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 4: Metadata for all strains included in the a) *S. pneumoniae* and b) *N. gonorrhoeae* RASE database.** Each record contains the strain's taxid, lineage, serotype (for *S. pneumoniae* only), MLST sequence type, 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 5: Prevalence of resistance phenotypes across lineages in the a) *S. pneumoniae* and b) *N. gonorrhoeae* RASE database.** Statistics on prevalence of resistance phenotypes across lineages before and after the ancestral state reconstruction step.

**Supplementary Table 6: Sensitivity and specificity of resistance and susceptibility inference in all the datasets.** The table shows the number of true positive (TP), true negative (TN), false negative (FN), and false positive (FP) calls for resistance/susceptibility in individual datasets and the resulting sensitivity and specificity.