PANINI: Pangenome Neighbor Identification for Bacterial Populations

Khalil Abudahab[1,2], Joaquín M. Prada[3], Zhirong Yang[4], Stephen D. Bentley[5], Nicholas J. Croucher[2], Jukka Corander*[4,5,6], David M. Aanensen*[1,2]

*corresponding authors

[1]Centre for Genomic Pathogen Surveillance, Wellcome Genome Campus, Cambridgeshire, UK

[2]Department of Infectious Disease Epidemiology, Imperial College London, W2 1PJ, UK

[3]Mathematics Institute, University of Warwick, Coventry, CV4 7AL

[4]HIIT, Helsinki Institute of Information Technology, Department of Mathematics and Statistics, University of Helsinki, FI-00014 Helsinki, Finland

[5]Pathogen Genomics, Wellcome Trust Sanger Institute,

[6]Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, N-0317 Oslo, Norway

mailto: d.aanensen@imperial.ac.uk & jukka.corander@medisin.uio.no

ABSTRACT

The standard workhorse for genomic analysis of the evolution of bacterial populations is phylogenetic modelling of mutations in the core genome. However, in the current era of population genomics, a notable amount of information about evolutionary and transmission processes in diverse populations can be lost unless the accessory genome is also taken into consideration. Here we introduce PANINI, a computationally scalable method for

identifying the neighbours for each isolate in a data set using unsupervised machine learning with stochastic neighbour embedding. PANINI is browser-based and integrates with the Microreact platform for rapid online visualisation and exploration of both core and accessory genome evolutionary signals together with relevant epidemiological, geographic, temporal and other metadata. Several case studies with single- and multi-clone pneumococcal populations are presented to demonstrate ability to identify biologically important signals from gene content data. PANINI is available at http://panini.wgsa.net/ and code at http://gitlab.com/cgps/panini

BACKGROUND

In less than a decade, bacterial population genomics has progressed from sequencing of dozens to thousands of strains [1,2,3,4]. The biological insights enabled by population genomics are particularly important in evolutionary epidemiology, as the genome sequences provide high resolution data for the estimation of transmission and evolutionary dynamics, including horizontal transfer of virulence and resistance elements. Phylogenetic trees are the main framework utilised for visualisation and exploration of population genomic data, both in terms of the level of relatedness of strains and for mapping relevant metadata such as geographic locations and host characteristics [5]. While trees are highly useful, they are in general estimated using only core genome variation (i.e. those regions of the genome common to all members of a sample), which may represent only a fraction of the relevant differences present in genomes across the study population. Several recent studies highlight the importance of

considering variation in gene content when investigating the ecological and evolutionary processes leading to the observed data [6, 7].

The rapidly increasing size of population genomic datasets calls for efficient visualisation methods to explore patterns of relatedness based on core genomic polymorphisms, accessory gene content, epidemiological, geographical and other metadata. Here we introduce a framework that integrates within the web application Microreact [5], by utilising a popular unsupervised machine learning technique for big data to infer neighbors of bacterial strains from accessory gene content data and to efficiently visualize the resulting relationships. The machine learning method, called t-SNE, has already gained widespread popularity for exploring image, video and textual data [8,9], but has to our knowledge not yet been utilized for bacterial population genomics.

Since gene content may in general be rapidly altered in bacteria, it provides a high-resolution evolutionary marker of relatedness which can extend far beyond core genome mutations [7]. Different processes driving horizontal movement of DNA, such homologous recombination, conjugative transfer of plasmids and phage infections, all affect the gene content within and outside of a chromosome. By contrasting core and non-core gene content, one can investigate and draw conclusions about genome dynamics across a sample collection. Here we demonstrate the biological utility of such an approach by application to multiple population data sets.

METHODS AND RESULTS

Student t-distributed Stochastic Neighbor Embedding (t-SNE) is a machine learning algorithm which is widely used for data visualization [8]. It is suitable for embedding a set of high-dimensional data items in a 2- or 3-dimensional space. The embedding approximately preserves the pairwise similarities between the data items.

The t-SNE algorithm consists of two main steps. First, it calculates the similarities between the data items in the high-dimensional space, which is typically based on normal distribution around each data item. The similarities are then normalized to be probabilities (i.e. they sum to one). Similarities in the low-dimensional space are analogously defined and normalized except that Student t-distribution replaces the Gaussians. Second, t-SNE minimizes Kullback-Leibler divergence between the two probability matrices over the embedding coordinates. Finally, the 2-D t-SNE result can be visualized as a scatter plot where each dot indicates a data item.

t-SNE as an unsupervised method is particularly useful for exploratory data analysis. It has a wide range of applications in music analysis, cancer research, computer security research, bioinformatics, and biomedical signal processing. In many cases, t-SNE is able to identify meaningful data structures such as clusters even without feature engineering or structural assumptions, e.g. about number of clusters underlying the data. Here, we use the latest version of the t-SNE projection method, adopting the Barnes-Hut algorithm for accelerating the divergence minimization [9].

To demonstrate utility within population genomics, firstly, we explore how the method performs in a simulated setting, where the relationship between all sequences is known and then extended our analysis to published bacterial population data sets, allowing us to uncover previously unseen relationships between data and to address important biological questions.

SIMULATED DATA

To our knowledge, this is the first time that the t-SNE projection method has been used to explore patterns of genetic relatedness between different bacterial isolates. We have therefore validated the methodology by assessing how well it identifies neighbours and clusters for simulated genetic sequences. Firstly, we randomly generated multiple synthetic datasets of related isolates, with each defined as a sequence of present/absent genes. Each dataset is generated using the following parameters:

1. There are 20 clusters as underlying subpopulations.
2. The number of isolates belonging to a cluster is drawn from a Poisson distribution with mean 15.
3. Each cluster is defined by a number of core genes, which ranges uniformly from 1 to 100.
4. Each isolate has a probability between 80% to 99% of independently carrying each of the core genes of the cluster it belongs to.
5. Conversely, each isolate has a probability (PN) to independently carry each of the non-core genes of its cluster. Non-core genes are composed of core genes of other clusters and "noise" genes which are not defining characteristics of any cluster (in total 300 genes).

126

127 Each generated dataset has on average 300 isolates with a gene content of 1300
128 genes present/absent on average. For each dataset, we estimated the genetic
129 pairwise Hamming distance (d_H) and the distance using the t-SNE algorithm (d_t).
130 The Hamming distance is simply the number of differences between two
131 sequences of equal length, which in this case refers to a gene being present in
132 one isolate but absent in the other. The implementation of the t-SNE algorithm
133 that we use yields a coordinate in a 2D plane for each isolate, and we calculate
134 the distance d_t simply as the Euclidean distance for each pair of isolates.

135

136 If a cluster is sufficiently differentiable in terms of its gene content, we expect the
137 Hamming distance within the cluster to be smaller than to any other isolate not
138 belonging to it. For the t-SNE algorithm to be considered valid, it should be able
139 to project the isolates from the same cluster on the 2D plane sufficiently close
140 together so that the Euclidean distance within the cluster is smaller than to any
141 other isolate. Given the conditions that were used to generate the synthetic
142 datasets, not all clusters are necessarily differentiable in terms of their gene
143 content, therefore we classified the t-SNE algorithm as performing erroneously
144 only when a pair of isolates belonging to different cluster are not identified as
145 such by the algorithm but are correctly identified using the Hamming distance.
146 For high levels of noise, i.e. a large value of PN, differentiating the clusters using
147 their gene content becomes increasingly difficult as the isolates may lack a
148 sufficiently stable signal of relatedness.

149

150 We analyzed the performance of the t-SNE algorithm for three levels of noise PN:

0.001, 0.005 and 0.01, which measures the average proportion of non-core genes in each isolate. We performed 100 repeats for each noise value, which for each repeat involves generating on average 300 sequences and comparing almost 45000 pairs of isolates. The average error for the three noise values was 0.5%, 1% and 4% respectively, with a small error representing a particular isolate mis-allocated (i.e. very close to a different cluster) and a large error representing two clusters which are not appropriately differentiated by the t-SNE algorithm, illustrated in Figure 1. The error of the t-SNE algorithm increases with the noise, as shown in Figure 1(iii), and with the total number of clusters (not shown).

WEB APPLICATION - https://panini.wgsa.net/

The t-SNE algorithm implemented in C++ (https://github.com/lvdmaaten/bhtsne) was wrapped as a Node.js native module and embedded within a web application. The application is written in JavaScript and utilises React (https://facebook.github.io/react) for front-end and the Vis.js library (http://visjs.org) for network visualisation.

1) Data are uploaded as a gene presence/absence matrix - Panini expects data in the .rtab format (the output from Roary: the pan genome pipeline [10]; https://sanger-pathogens.github.io/Roary) However, this is simply a data file containing gene rows and isolate columns with '1' or '0' indicating presence/absence of a particular gene for a particular isolate.

2) Genes present in all isolates are ignored (i.e. core genome) and non-core genes are clustered using t-SNE with default parameters (auto perplexity and theta=0.5 – parameters can be changed by users).

3) The Results (x, y coordinates, a '.dot' format file containing graph layout, csv

and JSON) are made available for download and reuse. Results are also visualised directly within the PANINI web application as a graph layout.

To interpret the data in an epidemiological, phylogeographic and geographic context, the estimated network can also be uploaded directly to the Microreact platform allowing a user to add other forms of data to relate to the resulting neighbor embedding, typically a phylogenetic tree, geographical locations of the isolates, and temporal data (Further information and instructions at https://microreact.org).

UTILITY WITH EXISTING PUBLISHED DATASETS

To demonstrate utility of t-SNE clustering we applied the method to three published datasets which used Whole Genome Sequencing (WGS) to study the evolution of the bacterium *Streptococcus pneumoniae*. The first, a population level dataset, detailed population-wide diversity of pneumococci within Massachusetts, USA pre- and post vaccine introduction [2], while the second and third detailed international collections of globally-disseminated multidrug-resistant lineages of *Streptococcus pneumoniae* [11, 12]. Additional biological insights made possible with PANINI are described, and links to the projects within Microreact for further exploration of associated metadata and download of raw data formats are provided.

ANALYSIS OF A DIVERSE PNEUMOCOCCAL POPULATION

Data Visualisation and download:

https://microreact.org/project/panini-sparc?ui=nt

Source data and .RTab file:

https://gitlab.com/cgps/panini/datasets/tree/master/SPARC

Video walkthrough for PANINI and Microreact creation/use:

https://vimeo.com/230416235

PANINI was applied to a collection of 616 systematically-sampled pneumococcal isolates from a vaccine and antimicrobial resistance surveillance project in Massachusetts [13]. The original analysis of gene content in this collection identified 5,442 'clusters of orthologous genes' (COGs) [2], the core set of which was used to define fifteen 'sequence clusters' with BAPS (http://www.helsinki.fi/bsg/software/BAPS) [18]. For most of the sequence clusters, the correspondence between a group in the PANINI output and the original sequence clusters was exact (Figure 2A), reflecting their similarity both in terms of the core and accessory genomes [14]. These sets of isolates therefore represent well-defined, distinct lineages. However, SC1, SC6, SC10 and SC12 all exhibited distinct substructuring in the PANINI output. This corresponded well with the diverse core genome observed in these clusters (Figure 2B), and in each case, these groups were consistent with clades within the sequence clusters. These sequence clusters are therefore likely to represent amalgams of genotypes that should be subdivided into multiple clusters. Conversely, PANINI revealed clear substructuring within the previously unclustered SC16, which was also

consistent with the core genome phylogeny. Hence PANINI can easily facilitate the division of a diverse population into discrete genotypes that are coherent in their accessory and core genome content.

EXTENSIVE PROPHAGE VARIATION IN A MULTIDRUG-RESISTANT LINEAGE

Data Visualisation and download:

https://microreact.org/project/panini-pmen2?ui=nt

Source data and .RTab file:

https://gitlab.com/cgps/panini/datasets/tree/master/PMEN2

PANINI was applied to an analysis of orthologous genes across a global collection of 190 isolates from the multidrug-resistant *Streptococcus pneumoniae* clone PMEN2 [11], which caused a large outbreak of disease in Iceland starting in the late 1980s (Figure 3A). Multiple distinct clusters were again evident in the output (Figure 3B). In some cases, these were consistent with the phylogeny. The original analysis identified two independent entries of the lineage into Iceland, clades IC1 and IC2, the latter of which contained many fewer isolates and was clustered as IcA in the annotated output. By contrast, IC1 was distributed across four clusters IcB-E, which did not correspond with clear clades in the phylogeny. The difference between IcB and IcC is technical, rather than biological: all IcB isolates were sequenced early in the project with 54 nt reads, whereas most IcC isolates were sequenced with 75 nt reads. Unusually for pneumococci, the isolates in both these groups were trilysogenic, carrying prophage similar to ϕ670-6B.1 and ϕ670-6B.2, found in the *S. pneumoniae* 670-6B genome inserted

between *dnaN* and *pth* (att_{670}), and within the *comYC* gene (att_{comYC}), respectively; and a prophage isolated from 0211+13275, inserted at SPN23F15280 - SPN23F15810 (att_{MM1}) [15]. The apparent rapid acquisition, and stable maintenance, of multiple viral loci may relate to the abrogation of these bacteria's competence system by the insertion of prophage ϕIC1 into *comYC* [11,16]. Group IcD, interspersed with IcB and IcC within clade IC1 in the phylogeny, differs in the absence of prophage similar to ϕ670-6B.2. IcE, also polyphyletic within clade IC1, differed in having lost the region of PPI-1 that encodes the *pia* iron transport operon, which plays a role in pneumococcal pathogenesis in animal models [17]. Hence it is not surprising to find these isolates were only recovered from sputum, otitis media samples or nasopharyngeal swabs.

Multiple distinct clusters of non-Icelandic isolates were also observed. These all represented cases where t-SNE grouped isolates that were disparate in terms of their country and year of isolation, , as well as having a polyphyletic distribution across the whole genome phylogeny. These groupings represented cases of convergent evolution through parallel acquisition very similar prophage. Group IntA lacked any prophage similar to those shown in Figure 3C; group IntB had prophage with some similarity to both prophage in the reference genome; group IntC only had a prophage with similarity to ϕ0211+13275, whereas group IntD had prophage similar to ϕ0211+13275 and ϕ670-6B.1 as well. Hence the rapid movement of prophage sequences within lineages [14] clearly substantially contributes to the changes in gene content observed over short timescales. PANINI facilitates rapid analysis of these diverse elements, and their complex

relationship with bacterial population structure.

MOBILE ELEMENT AND SEROTYPE VARIATION IN A VACCINE-ESCAPE LINEAGE

Data Visualisation and download:

https://microreact.org/project/panini-pmen14?ui=nt

Source data and .RTab file:

https://gitlab.com/cgps/panini/datasets/tree/master/PMEN14

PANINI was similarly applied to 176 isolates of the multidrug-resistant *S. pneumoniae* PMEN14 lineage [11]. Although the sequences came from many countries, the collection was strongly enriched for bacteria from the Maela refugee camp in Thailand, which fell into five clades (ML1-5), of which ML2 was the largest. The groups identified by PANINI were again polyphyletic (Figure 4A), with ML2 split up in a similar manner to the PMEN2 clade IC1. This was again driven by the distribution of prophage sequence: group 1 isolates were free of prophage, whereas group 2 isolates were infected with a 'group 2-type' prophage, and group 3 isolates were infected with a similar, but distinct, 'group 3-type' prophage (Figure 4B). Clade ML2 isolates in group 4 were distinguished by variation in another mobile genetic element, a phage-related chromosomal island (PRCI), shared by most of the isolates. This PRCI was absent from these assemblies, either because at least part of the element had been lost through deletion, replacement with a related sequence (isolate 6259_1-15), or the acquisition of a second, highly similar PRCI that prevented effective assembly of

either (isolates 6237_8-12, 6237_8-13 and 6237_8-18). In this latter case, mapping to the element was still evident.

A fifth group, which did not include any Maela isolates, corresponded to the antibiotic-susceptible outgroup isolates. These differed through the absence of a third type of mobile element, the Tn*916* integrative and conjugative element, an antibiotic resistance-encoding genomic island that was absent from these 'outgroup' isolates. Additionally, these bacteria shared two smaller genomic islands, encoding putative lantibiotic biosynthesis and restriction-modification operons, which were absent from the multidrug-resistant isolates. Variation in other non-mobile element islands was also detectable. The group 1-19A subcluster contained isolates of serotype 19A, produced through two independent serotype switching recombinations at the capsule polysaccharide synthesis (*cps*) locus that resulted in genotypes '19A ST320' and '19A ST236'. These changes were responsible for allowing isolates to evade the seven valent polysaccharide conjugate vaccine, which targeted the lineage's ancestral serotype 19F, expressed by almost all the rest of the collection [12]. A smaller serotype switching recombination, which did not replace the entire serotype-determining *cps* locus, generated the '19A ST271' isolates [12]. The smaller associated change in gene content meant this isolate was not clearly distinguished from the rest of group 1 (Figure 4A).

DISCUSSION

The rapid increase in sampling density of bacterial populations for epidemiological and evolutionary studies highlights the need of combining

326 traditional genomic markers, such as SNP loci and small insertions or deletions
327 in coding regions, with measures of difference in terms of gene content. As many
328 bacteria have varied accessory genomes, changes in the gene content can offer a
329 way to identify epidemiologically or evolutionarily important clues about the
330 evolutionary processes affecting a pathogen's spread. As we illustrated here,
331 such information is most useful when clustering is combined within a
332 phylogeographical approach, and visualized jointly in a seamless fashion
333 enabling the rapid interpretation of core and non-core clustering in the context
334 of where and when data were collected.
335
336 The t-SNE algorithm is a very efficient approach to cluster isolates based on their
337 gene content. In the simulated scenarios considering synthetic data, the errors in
338 clustering always remained small, either representing an isolate allocated to a
339 wrong cluster, or two clusters which were not appropriately differentiated.
340 However, this only occurred in simulations with the "noise" level much higher
341 than expected in nature. In general, what we defined as "core" genes in a cluster
342 rarely appear in isolates not belonging to the cluster, and if they do, it is typically
343 at much lower frequencies than those we considered. Furthermore, in our
344 synthetic datasets we formed clusters defined by as few as a single core gene.
345 These clusters with a limited number of core genes, combined with relatively
346 high levels of "noise", are in practice almost completely indistinguishable from
347 others, as illustrated in Figure 1 (iii - clusters K, L, O and Q). Overall, our
348 simulated datasets are conservative, as the gene absence and presence variation
349 is higher than expected in natural populations, and therefore indicate that the t-
350 SNE is a promising approach for rapidly and accurately clustering bacteria based

on gene content.

When applied to a population-wide genomic dataset, the algorithm was clearly able to identify distinct lineages within a diverse collection. This analysis could highlight which clusters, defined using the core genome, could be sensibly subdivided, and which small groups of within a diverse set of strains could be justifiably regarded as new clusters. Within lineages, the same congruence between core and accessory genomes across clades was not observed. Instead, clusters were distinguished by rapidly occurring, homoplasic alterations, such as phage infection. In this context, PANINI provides an intuitive way in which to understand the distribution of rapidly-evolving aspects of the genome, which are difficult to analyse in a conventional phylogenetic framework. PANINI is therefore a promising platform through which biologically-important changes in bacterial gene content can be uncovered at all levels of evolutionary, ecological and epidemiological analyses.

FUNDING

The Development of PANINI was funded by The Centre for Genomic Pathogen Surveillance and Wellcome Trust Grant 099202. J.C. was supported by the ERC grant no. 742158. Z.Y. was supported by the COIN Centre of Excellence. J.M.P. gratefully acknowledge funding of the NTD Modelling Consortium by the Bill & Melinda Gates Foundation in partnership with the Task Force for Global Health. The views, opinions, assumptions, or any other information set out in this report should not be attributed to the Bill & Melinda Gates Foundation and the Task Force for Global Health or any person connected with them. NJC is funded by a

376 Sir Henry Dale Fellowship, jointly funded by the Wellcome Trust and Royal
377 Society (Grant Number 104169/Z/14/Z).

378

379 **REFERENCES**

380 1. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, et al.
381 Evolution of MRSA during hospital transmission and intercontinental spread.
382 Science. 2010 Jan 22;327(5964):469-74

383 2. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, et al.
384 Population genomics of post-vaccine changes in pneumococcal epidemiology.
385 Nat Genet. 2013 Jun;45(6):656-63

386 3. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, et al
387 Dense genomic sampling identifies highways of pneumococcal recombination.
388 Nat Genet. 2014 Mar;46(3):305-309

389 4. Aanensen DM, Feil EJ, Holden MT, Dordel J, Yeats CA, et al. Whole-Genome
390 Sequencing for Routine Pathogen Surveillance in Public Health: a Population
391 Snapshot of Invasive Staphylococcus aureus in Europe. MBio. 2016 May 5;7(3)

392 5. Argimón S, Abudahab K, Goater RJ, Fedosejev A, Bhai J, et al. Microreact:
393 visualizing and sharing data for genomic epidemiology and phylogeography.
394 Microb Genom. 2016 Nov 30;2(11):e000093

395 6. Marttinen P, Croucher NJ, Gutmann MU, Corander J, Hanage WP.
396 Recombination produces coherent bacterial species clusters in both core and
397 accessory genomes. Microb Genom. 2015 Nov 5;1(5)

398 7. McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, et al. Combined Analysis of
399 Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-
400 Resolution View into the Evolution of Bacterial Populations. PLoS Genet. 2016

401 Sep 12;12(9):e1006280.

402 8. van der Maaten L & Hinton G. Visualizing Data using t-SNE. JMLR 2008

403 9(Nov):2579--2605.

404 9. van der Maaten L. Accelerating t-SNA using Tree-Based Algorithms. JMLR

405 2014 15(Oct):3221-3245.

406 10. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, et al. Roary: rapid

407 large-scale prokaryote pan genome analysis. Bioinformatics. 2015 Nov

408 15;31(22):3691-3

409 11. Croucher NJ, Hanage WP, Harris SR, McGee L, van der Linden M, de

410 Lencastre H, et al. Variable recombination dynamics during the emergence,

411 transmission and “disarming” of a multidrug-resistant pneumococcal clone. BMC

412 Biol. 2014;12: 49.

413 12. Croucher NJ, Chewapreecha C, Hanage WP, Harris SR, McGee L, et al.

414 Evidence for soft selective sweeps in the evolution of pneumococcal multidrug

415 resistance and vaccine escape. Genome Biol Evol. 2014 Jun 10;6(7):1589-602.

416 13. Finkelstein JA, Huang SS, Daniel J, Rifas-Shiman SL, Kleinman K, et al.

417 Antibiotic-resistant Streptococcus pneumoniae in the heptavalent pneumococcal

418 conjugate vaccine era: predictors of carriage in a multicommunity sample.

419 Pediatrics. 2003 Oct;112(4):862-9

420 14. Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD,

421 Hanage WP. Diversification of bacterial genome content through distinct

422 mechanisms over different timescales. Nat Commun. Nature Publishing Group;

423 2014;5: 5471.

424 15. Romero P, Croucher NJ, Hiller NL, Hu FZ, Ehrlich GD, Bentley SD, et al.

425 Comparative genomic analysis of ten Streptococcus pneumoniae temperate

426 bacteriophages. J Bacteriol. 2009;191: 4854–4862.

427 16. Croucher NJ, Mostowy R, Wymant C, Turner P, Bentley SD, Fraser C.
428 Horizontal DNA Transfer Mechanisms of Bacteria as Weapons of Intragenomic
429 Conflict. PLOS Biol. 2016;14: e1002394.

430 17. Brown JS, Gilliland SM, Ruiz-Albert J, Holden DW. Characterization of pit, a
431 Streptococcus pneumoniae iron uptake ABC transporter. Infect Immun. 2002;70:
432 4389–4398. doi:10.1128/IAI.70.8.4389-4398.2002

433 18. Corander J, Marttinen P, Sirén J, Tang J. Enhanced Bayesian modelling in
434 BAPS software for learning genetic structures of populations. BMC
435 bioinformatics. 2008 Dec 16;9(1):539. doi: 10.1186/1471-2105-9-539

436

437 **FIGURE LEGENDS**

438

439 **Figure 1** Illustration of a simulated dataset, with the isolates' gene content
440 (left), black dots indicate presence of a gene, x-axis represent all the considered
441 genes (total of 1213 genes in this simulation). The right panels show the
442 embedded locations in the 2D plane as estimated by the t-SNE algorithm, with
443 each colour representing a cluster in the underlying simulation model. Clusters
444 are named using the alphabet (A, B, C…). From top to bottom, plots indicate
445 simulations generated with 0.1% (i), 0.5% (ii) and 1%(iii) noise, respectively.

446

447 **Figure 2** A) Annotated output of the PANINI algorithm applied to 616 *S.*
448 *pneumoniae* isolates from a diverse population in Massachusetts. Each node
449 represents an isolate, each of which is coloured according to its sequence cluster,
450 as defined using the core genome. Clusters of isolates belonging to the same

451 sequence cluster are circled and annotated. Where sequence clusters are divided
452 into multiple groups in the PANINI network, the circles are joined by dashed
453 lines. B) Core genome phylogeny based on comparison of conserved clusters of
454 orthologous genes adapted from [2] and displayed within Microreact. Sequence
455 clusters are annotated for comparison with non-core clustering.
456
457 **Figure 3** Analysis of the *S. pneumoniae* PMEN2 lineage. A) i) Core genome
458 phylogeny with tree leaves coloured by country of origin and ii) geographic
459 origin of isolates. B) Annotated output of the PANINI algorithm applied to 189
460 isolates from an international collection of representatives of the *S. pneumoniae*
461 PMEN2 lineage. Each point is coloured according to its region of origin. Groups
462 defined by the structure of the PANINI output are circled and annotated. Clusters
463 containing primarily Icelandic isolates (coloured orange) are labelled with 'Ic'
464 prefixes, whereas those containing isolates from multiple countries are labelled
465 with 'Int' prefixes. C) Variation in accessory loci associated with differential
466 classification of isolates into groups. The orange and brown bands across the top
467 of the figure indicate the extent of the three prophage and pneumococcal
468 pathogenicity island 1 (PPI-1) sequences, against which the short read data from
469 the isolates were mapped. The heatmap below includes one row per isolate,
470 which were ordered according to their grouping in panel A. The heatmap is
471 coloured blue where mapping coverage was low, indicating a locus is absent, and
472 red were mapping coverage was high, indicating a sequence was present.
473 Horizontal dashed lines indicate the boundaries between the groups of isolates,
474 which vertical dashed lines indicate the boundaries between loci.
475

Figure 4 Analysis of the *S. pneumoniae* PMEN14 lineage. A) Annotated output of the PANINI algorithm applied to 176 isolates from an international collection of representatives of the *S. pneumoniae* PMEN14 lineage. The main groups 1-5 are circled with solid lines and named; the subgroups within group 1 are circled by dashed lines. (B) Variation in accessory loci associated with differential classification of isolates into groups. This heatmap is displayed as in Figure 3. In this case, the sequence loci across the top are more functionally diverse. The first is the *neuB* coding sequence with an IS*Spn*8 element inserted into it. The lack of mapping to the middle of this column indicates the absence of this insertion sequence anywhere in the chromosome. The next loci are alternative alleles of the capsule polysaccharide synthesis locus, one encoding for the biosynthesis of the 7-valent polysaccharide conjugate vaccine (PCV7) type 19F polysaccharide, the other for the non-PCV7 type 19A polysaccharide. These are followed by two similar prophage, one associated with group 2 isolates, the other with group 3 isolates; the similarity between these two viruses means there is extensive mapping to both, even when an isolate only contains one of them. The PRCI absent from the assemblies of group 4 isolates is next; mapping suggests this is actually present in some, but PANINI nevertheless included them in this group because the acquisition of a further, related PRCI prevented either assembling accurately. This is followed by the Tn*916* conjugative element, absent from the group 5 isolates, which possess genomic islands encoding for the biosynthesis of a lantibiotic and a restriction-modification system, included at the right-hand end of the panel.