panX: pan-genome analysis and exploration

Wei Ding[1], Franz Baumdicker[2], Richard A. Neher,[1*]

[1] *Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany,* [2] *Mathematisches Institut, Albert-Ludwigs University of Freiburg, 79104 Freiburg, Germany*

(Dated: August 29, 2016)

Horizontal transfer, gene loss, and gene duplication result in dynamic bacterial genomes subject to a complex mixture of different modes of evolution. Closely related strains can differ in the presence or absence of many genes, and the total number of distinct genes found in a set of related isolates – the pan-genome – is often many times larger than the genome of individual isolates. We have developed a pipeline for pan-genome analysis and a web-based visualization software for interactive exploration of the pan-genome. The visualization allows rapid filtering of and searching for genes. For each gene cluster, panX displays an alignment, a phylogenetic tree, maps mutations within that cluster to the branches of the tree, and infers loss and gain of genes on the core-genome phylogeny. PanX is available at `http://pangenome.de`. User generated pan-genomes can be visualized either using a custom web-server or by serving panX locally as stand-alone browser-based application.

In addition to vertically passing down their genome to offspring, bacteria have the capability to acquire genetic material from the environment via horizontal transfer (Soucy *et al.*, 2015). Genes are transferred among bacteria by a variety of mechanisms including active uptake, mobile genetic elements, and gene transfer by viruses (Thomas and Nielsen, 2005). In addition to gene gain, genes are frequently duplicated or lost. The mix of vertical transmission and horizontal transfer complicates phylogenetic analysis of bacterial genomes and results in patterns of genetic diversity that are difficult to interpret (Puigbò *et al.*, 2014).

A common approach when analyzing bacterial genomes is categorizing genes into the *core* or *accessory* genome (Lapierre and Gogarten, 2009; Tettelin *et al.*, 2008; Vernikos *et al.*, 2015). Core genes are shared by all strains in a group of isolates, accessory genes shared by two or more but not all strains, and unique genes are specific to a single strain. The core genome is often used to assess the relatedness among the genomes in the sample and to approximate the species tree, but extensive horizontal transfer has been documented in the core genome as well (Everitt *et al.*, 2014). The union of all genes found in a group of strains (e.g. strains from one species) is called the pan-genome, which is typically several times larger than the core genome.

By providing a repertoire of functional genes, gene gain from the pan-genome can facilitate the acquisition of new metabolic pathways (Fournier and Gogarten, 2008), the adaptation to new habitats, or the emergence of drug resistant variants (Ochman *et al.*, 2000). With the rapidly increasing number of sequenced bacterial genomes, it is now possible to detect associations between metadata such as habitats, phenotypes, clinical manifestations and the presence or absence of particular genes (Earle *et al.*, 2016; Méric *et al.*, 2014). Reliable detection of association, however, requires a good understanding of the pan-genome dynamics.

Pan-genome construction from a group of related bacterial genomes typically involves the identification of homologous regions by all-against-all comparisons followed by clustering orthologous genes (Vernikos *et al.*, 2015). Several software packages and pipelines have been developed to construct such pan-genomes that differ in the heuristics used to compare strains and generate clusters (Laing *et al.*, 2010; Page *et al.*, 2015; Zhao *et al.*, 2014, 2012). Roary by Page *et al.* (2015) is designed to handle hundreds to thousands of genomes by avoiding the costly all-against-all comparison through an initial clustering step.

One fundamental limitation, however, is the difficulty to interrogate, explore, and visualize the pan-genome and the evolutionary relationships between strains. In absence of recombination, the purely vertical evolutionary history of strains would be represented by a familiar phylogenetic tree. With horizontal transfer, the history of different loci in the genome is described by different trees resulting in a phylogenetic forest or network (Huson and Bryant, 2006; Koonin *et al.*, 2009). While phylogenetic networks can be visualized using consensus representations such as split networks (Huson, 1998), the history and distribution of individual proteins are often critical, for example when searching for associations with phenotypes like drug resistance. Individual clusters of orthologous sequences, however, can again be represented by a tree if genes are short enough that recombination within the gene can be ignored. Some gene trees are similar to the species tree, while others vary dramatically. The degree of incongruence of the gene tree with the species tree contains important information about the dynamics of gene gain and loss.

Here, we present panX, a web-based environment for microbial pan-genome data visualization and exploration based on an automated pan-genome identification pipeline. The pipeline breaks the annotated genomes of a large number of strain (your own genomes, NCBI Reference Sequences, etc) into genes and further clusters annotated genes into orthologous groups. From these clusters, we identify the core genome, build a strain-level phylogeny using SNPs in the core genome, construct multiple alignments of sequences in gene clusters, build trees for individual genes and map the gene presence/absence pattern onto the core genome tree. The interactive browser-based application then allows the exploration of the above features and provides flexible filter, sort, and search functionalities. This application is available at `pangenome.de` with a collection of pan-genomes prepared

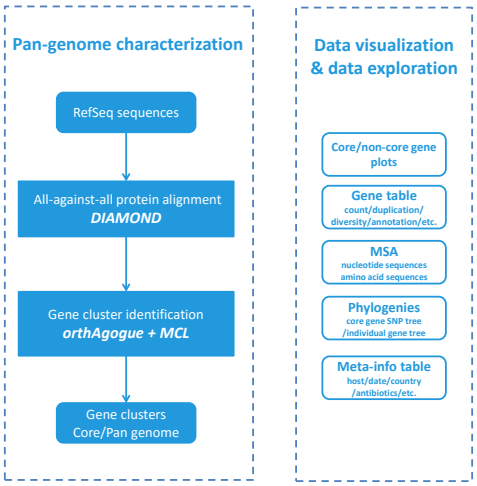

FIG. 1 **panX analysis pipeline:** panX uses DIAMOND (Buchfink *et al.*, 2015), orthAgogue (Ekseth *et al.*, 2013) and MCL (Dongen, 2000; Enright *et al.*, 2002) to determine clusters of orthologous genes from a collection of annotated genomes. Gene clusters are then analyzed phylogenetically and prepared for visualization.

by us, but can also be deployed on other servers with custom pan-genomes, or can be run locally as a browser-based desktop application.

Identification of orthologous gene clusters

The computational pipeline underlying panX (illustrated in Fig. 1) is broadly similar to other tools used to construct pan-genomes (Laing *et al.*, 2010; Page *et al.*, 2015; Zhao *et al.*, 2014, 2012). PanX uses annotated genome sequences in GenBank format or a list of accession numbers of sequences in the RefSeq database. To identify homologous proteins, panX performs an all-against-all similarity search using DIAMOND (Buchfink *et al.*, 2015). Next, panX constructs a sequence similarity matrix from the DIAMOND output using orthAgogue (Ekseth *et al.*, 2013), which filters DIAMOND results to exclude hits between non-orthologous genes. The similarly matrix serves as input for the Markov Clustering Algorithm (MCL) (Dongen, 2000; Enright *et al.*, 2002) to create the clusters of orthologous genes. The DIAMOND similarity search can be multi-threaded and panX uses 64 CPUs by default if run on a compute cluster.

In our experience, it is advisable to cluster proteins aggressively and split clusters with many paralogous pairs in a post-processing step. PanX transverses the tree of each gene cluster and calculates how many paralogs ρ would be separated if the tree was split at each branch of the tree. Having identified the branch with the maximal paralogy score ρ, panX splits the cluster if ρ and the branch length τ satisfy

$$\rho/\rho_0 + \tau/\tau_0 > 1\,, \qquad (1)$$

where ρ_0 is half the number of strains represented in the cluster and τ_0 is 100-fold the median branch length of the tree. The large threshold of 100-fold the medium branch length is chosen to only split diverged paralogs and can be reduced if desired. The post-processing typically only affects a small number of clusters and the results are insensitive to the exact choice of these parameters.

The above construction of orthologous cluster corresponds to steps 2-5 of the panX pipeline, which can be replaced by `roary`. The two approaches give similar results, but we found our pipeline to be preferable for diverse sets of sequences (see below).

Phylogenetic analysis of gene clusters

To reduce the computational burden of the subsequent visualization of the pan-genome, alignments, trees, and other properties of the gene clusters are precomputed. The input for this phylogenetic analysis is either the output of steps 1-5 of the pipeline presented above, or the output of `roary`.

Alignment. All protein sequences belonging to one cluster are aligned using mafft (Katoh *et al.*, 2002). The protein alignment is then used to construct a codon-alignment of the corresponding nucleotide sequences by inserting a gap of length three for every gap in the amino acid alignment.

Tree building and ancestral reconstruction. PanX extracts all variable positions from the nucleotide alignments of all single copy core genes (those gene clusters in which all strains are represented exactly once) to construct a core-genome SNP matrix. This SNP matrix is used to build a core genome phylogenetic tree using FastTree (Price *et al.*, 2010), which is further refined by RaxML (Stamatakis, 2014) following a similar strategy as implemented in nextflu (Neher and Bedford, 2015). Due to homologous recombination this core genome tree may not reflect the true history for each of the genes in the core genomes (Everitt *et al.*, 2014) and branch lengths do not reflect sequence similarity since only variable sites are used (Bertels *et al.*, 2014). However, it is still an useful approximation of the relationships of the different strains that can be used as a scaffold to investigate the evolution of the mobile genome and the distribution of phenotypes.

From the nucleotide sequence alignment of each gene cluster, we calculate an approximate maximum likelihood phylogenetic tree using FastTree. Refinement by Raxml is skipped due to excessive computational costs when computing trees on several thousand gene clusters. Once the tree for a gene cluster is inferred, ancestral sequences of internal nodes are reconstructed using a joint maximum likelihood approach (Felsenstein, 2004) as implemented in treetime (Sagulenko *et al.*, 2016). Likely mutations are mapped onto the branches of the tree using this ancestral reconstruction.

Then, we infer the presence or absence of each gene cluster on internal nodes of the core genome SNP tree using an analogous ancestral inference procedure. Individual gain and loss events are associated with branches based on this ancestral reconstruction. The gain and loss rates are chosen such that the

likelihood for the observed presence/absence pattern of genes is maximized (Felsenstein, 1992; Zamani-Dahaj *et al.*, 2016). We found that optimal loss rates are about twice the gene gain rate.

Gene clusters, trees, mutations, and metadata are stored as json files for the web visualization.

Web application for pan-genome exploration

To explore the pan-genome constructed by the pipeline described above, we developed a browser based visualization. The layout of the application is that of a large dashboard (see Fig. 2), on which multiple aspects of the pan-genome can be interrogated simultaneously.

At the top, three graphs provide basic statistics on the abundance and length distribution of all genes. In the middle row, a searchable table contains summary statistics and annotations for all gene clusters. The alignment viewer on the right shows the nucleotide or amino acid alignment of gene cluster selected in the table. Below the table, the core genome SNP tree is shown, along with a phylogenetic tree of the currently selected gene cluster. At the very bottom, a second searchable table allows rapid access to meta information available for different strains.

The hallmark of the panX web-application are the interconnected components that illustrate different properties of the gene clusters. The pan-genome statistic charts at the top allow rapid sub-setting of gene clusters. The left chart shows an inverse cumulative distribution of clusters sizes, i.e., clusters are sorted by decreasing number of strains represented in the cluster, such that all core genes present in all strains are shown on the left. The size of the core genes is then simply the length of the plateau of the curve to the first drop. The core genome is followed by gradual decline in gene number from common to rare accessory genes. Lastly, a long tail contains the strain-specific singletons. Subsets of genes can be easily defined by selecting a range of the graph with the mouse. Similarly, the center chart shows the distribution of gene length.

The pie chart on the right illustrates the relative proportion of core and accessory genome, each of which can be selected by clicking on the sectors in the chart. To allow for soft and strict definitions of the core genome, the cut-off delineating core and accessory genome can be adjusted with a slider.

Rapid and searchable access to alignment and gene trees

The table of all gene clusters is dynamically restricted to the range of gene abundances and gene lengths selected above. The table can be searched by gene name and annotation or sorted by gene count, diversity etc. Annotations of all input sequences (also discordant annotations of genes belonging to the same gene cluster) are accessible by expanding the annotation field. Similarly, the column *duplicated* specifies whether the gene cluster contains more than one gene per strain. The list of strains in which genes are duplicated and copy number of this gene can be accessed by expanding the row. Each row contains triggers to show the corresponding nucleotide or amino acid sequence alignment in the alignment viewer (MSA) from Biojs (Gómez *et al.*, 2013). This trigger also updates the phylogenetic tree viewers.

Interactive core genome tree and gene tree viewers

To facilitate the comparison between the core genome SNP tree and the gene tree, the two trees have connected interactive elements. When placing the mouse on a leaf node in one tree, the corresponding nodes are highlighted in both trees. Similarly, if the mouse is placed over an internal node, all nodes in the corresponding clades are highlighted with different colors for each strain. This gives a rapid impression if the core genome tree and the gene tree are compatible and whether the gene is duplicated in some of the strains, see Fig. 3.

The most likely gene loss and gain events inferred by the ancestral reconstruction algorithm are indicated on the tree by dashed or thick lines, respectively. Mutations in the amino acid or nucleotide sequence of the gene are mapped onto the gene tree and can be inspected using the tool-tips associated with branches in the tree.

In addition to mutation and gain/loss events, the tree can be colored with metadata associated with different strains. Such metadata would typically include collection dates, sampling location, host species or resistance phenotypes.

Pan-genomes of 25 bacterial groups

We ran the panX on collections of NCBI reference genomes of 25 bacterial groups including many human pathogens, see Tab. I. Most of these collections are closely related genomes, but we also included a diverse group of genomes of *Prochlorococcus*. *Prochlorococcus* is a marine cyanobacterium that is responsible for a significant fraction of the marine primary production and serves as a model system in marine microbial ecology (Biller *et al.*, 2014b).

While we relied on annotations available in NCBI for most species, we re-annotated the genomes of 40 *Prochlorococcus* sequences (Biller *et al.*, 2014a) using prokka (Seemann, 2014). The annotation was derived from a custom database based on the 12 annotated *Prochlorococcus* strains CCMP1375, MED4, MIT9313, NATL2A, MIT9312, AS9601, MIT9515, NATL1A, MIT9303, MIT9301, MIT9215 and MIT9211. *Prochlorococcus* is a much more diverse population than the other species we investigated, see Tab. I, which makes it a challenging case for pan-genome analysis.

We used *P. aeruginosa* and *Prochlorococcus* pan-genomes to compare the results of the panX pan-genome identification pipeline to that of roary. Clustering in panX relies on filtering by orthAgogue, while roary splits paralogs based on the neighboring genes. Both methods use MCL to produce the final clustering. We computed the size distribution and the fraction of clusters that match between pipelines, see Fig. 4. The size distributions are similar, but panX reports a smaller number of core gene clusters (3790 vs 3905 for *P. aeruginosa*)

FIG. 2 **Interconnected components of the panX web application:** The top panels provide a statistical characterization of the pan-genome and allow filtering of gene clusters by abundance and gene length. The gene cluster table on the center left is searchable and sortable and allows the user to select individual gene clusters for closer inspection. Upon selection in the table, the alignment of gene cluster is loaded into the viewer on the center right, the gene tree is loaded into the tree viewer at the bottom right, and presence/absence patterns of this gene cluster are mapped onto the core genome tree at the bottom left.

and fewer singleton gene (1193 vs 2276). More than 90% of core gene clusters found by one pipeline are also found by the other, see Fig. 4C&D. Concordance of clustering decreases for less common genes, see Fig. 4. We used standard parameters for Roary but adjusted the minimal blastp identity cut-off to `-i 70`.

For the diverse *Prochlorococcus* dataset, roary (Page *et al.*, 2015) separates nearly all *Prochlorococcus* genes and identifies only 10 core genes when using standard parameters. After lowering the minimum percentage identity for blastp in roary to 10% (`-i 0.1`), roary identified 1104 core genes vs 1130 identified by panX. While roary warns that it has not been designed to support such diverse datasets, the majority of the resulting clusters agree with those identified by panX. The clustering by panX uses the DIAMOND hits that pass filtering by orthagogue and is thus less dependent on the mean sequence diversity.

From the *Prochlorococcus* 40 genomes, panX and roary identified to 6,444 and 6,944 clusters of orthologous genes, respectively – not too far from the estimated pan-genome size of >8,500 genes present in more than one percent of the population (Baumdicker *et al.*, 2012).

Availability

The web application panX is available at `pangenome.de` and features precomputed pan-genomes of the species listed in Tab. I. The computational pipeline to identify the pan-genome consists of a collection of python scripts and a master script that runs desired analysis steps in series. The visualization is build on node.js server and makes extensive use of `bio.js` (Gómez *et al.*, 2013), `d3` (Bostock, 2016), `dc.js` (Woodhull *et al.*, 2016), and other javascript libraries. The analysis pipeline and the code for the web application is made available under the GPL3 license at `https://github.com/`

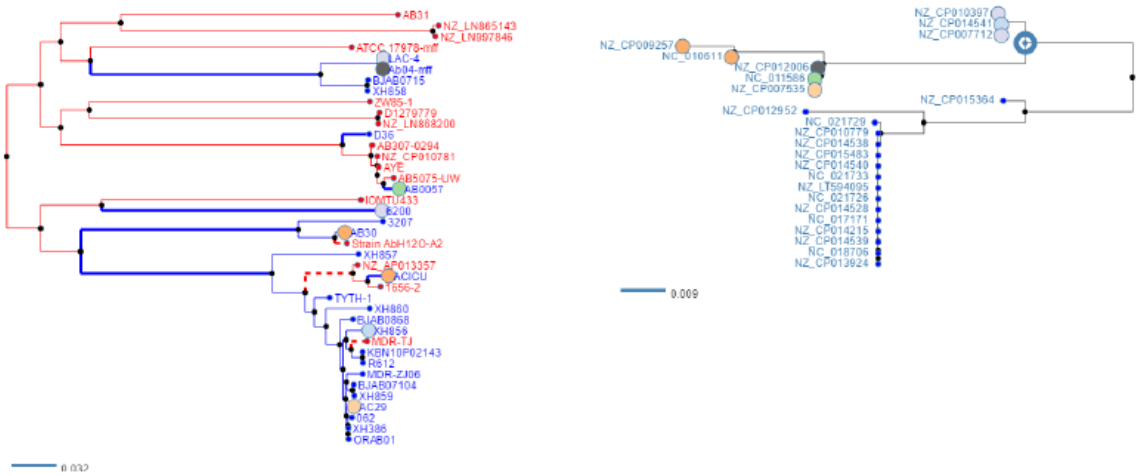

FIG. 3 **Linked core genome and gene trees.** The core genome tree shows the strains in which the current gene is present or absent in blue and red, respectively. Placing the mouse over an internal node in one of the trees (upper clade of the gene tree on the right in this example) highlights all strains in the corresponding clade in both trees. This gives the user a rapid impression of phylogenetic incongruence and likely gene gain and loss events.

`neherlab` as repositories `pan-genome-analysis` and `pan-genome-visualization`.

The web application can be used locally to inspect and explore pan-genomes produced by the panX pipeline.

Conclusions

Being able to visualize and explore high dimensional data is often the key to developing insight into the mechanisms driving complex dynamics. PanX is meant to enable such exploration of large sets of bacterial genomes, which are characterized by the evolution of individual genes as well as the the gain and loss of genes. Several other projects to visualize bacterial pan-genomes are actively developed, including *phandango* by Hadfield (2016), *PanViz* by Pedersen (2016), and *FriPan* by Powell (2016). The design of panX focused on combined breadth and depth: Besides summary statistics and species trees, panX allows to select interesting sets of genes or search for individual genes. Alignments and phylogenetic trees of genes can then be analyzed in detail with individual mutations and gain/loss events mapped to the gene tree and the core tree, respectively. The evolutionary patterns of genes can then be compared to meta-information such as resistance phenotypes associated with the individual strains. In contrast to other tools, panX does not rely on synteny information and does not require closed genomes as input. It can therefore be used on draft genomes assembled from short reads that remain fragmented into dozens of contigs.

We hope that panX will inform future investigations of the dynamics of pan-genomes and adaptation of bacteria to new habitats and environmental challenges. Horizontal transfer is pivotal for many aspects of bacterial adaptation (Ochman *et al.*, 2000), but at the same time it remains much more difficult to analyze than evolution by vertical descent (Puigbò *et al.*, 2014). The ability to interactively explore such pan-genomes might help to grasp the complexity of this dynamics.

On the other hand, a web-based tool that can be readily kept up-to-date by addition of newly sequenced isolates would be useful in pathogen surveillance. When paired with meta-information such as resistance, pathogenicity, sampling date, location and comorbidities, panX can elucidate adaptation, spread, and transmission chains of pathogens. Similar approaches have proved useful at tracking spread and evolution of seasonal influenza virus or ebola virus during the recent outbreak in West Africa (Gardy *et al.*, 2015; Neher and Bedford, 2015). The panX can fulfill a similar role for bacterial pathogens and could, for example, be deployed in individual hospitals, track food-borne outbreaks or monitor the global spread of MRSA (Holden *et al.*, 2013).

It becomes increasingly clear that horizontal evolution is not limited to bacteria. Many eukaryotes have been shown to gain genetic material from vectors or the environment (Boto, 2014; Rödelsperger and Sommer, 2011). Even in absence of obvious horizontal transfer, diverse loci with frequent deletions, duplications, and insertions such as the VDJ domains used to assemble B- and T-cell receptors are difficult to analyze phylogenetically and exploratory tools like panX might help to understand their dynamic evolution.

Acknowledgements

We gratefully acknowledge stimulating discussions with Matthias Willmann and Erik van Nimwegen as well as advice on `dc.js` from Gordon Woodhull. This work was supported by the Max Planck Society.

TABLE I **Summary statistics of pan-genomes available at `pangenome.de`.**
*Average hamming distance in nucleotide core gene alignments.

Species	genomes	core genes	all genes	singletons	diversity*
Acinetobacter baumannii	40	1933	7044	972	0.0136
Bacillus anthracis	42	3994	6073	329	0.0001
Bacillus cereus	30	3208	11921	2409	0.0487
Burkholderia pseudomallei	51	4193	9971	1742	0.0031
Bacillus subtilis	40	2113	7024	1399	0.0576
Bacillus thuringiensis	35	2350	15575	4291	0.0485
Clostridium botulinum	19	984	8471	1627	0.1643
Corynebacterium pseudotuberculosis	40	1205	2246	86	0.0045
Chlamydia trachomatis	67	818	975	6	0.0049
Campylobacter	107	842	3049	380	0.0431
Enterobacter cloacae	17	3136	8950	2006	0.0894
Escherichia coli	183	1410	15313	3247	0.0154
Francisella tularensis	29	854	2301	164	0.0061
Helicobacter pylori	83	708	2479	316	0.041
Klebsiella pneumoniae	63	3006	11532	2038	0.0098
Listeria monocytogenes	60	2066	4380	371	0.0328
Neisseria meningitidis	69	1166	3200	395	0.0143
Pseudomonas aeruginosa (panX)	37	3790	8951	1193	0.0055
Pseudomonas aeruginosa (roary)	37	3905	10553	2278	0.0053
Prochlorococcus marinus (panX)	40	1130	6444	1544	0.2777
Prochlorococcus marinus (roary)	40	1104	6944	1938	0.2795
Salmonella enterica	206	1548	10277	1627	0.0093
Streptococcus pneumoniae	29	1213	3904	929	0.0099
Streptococcus pyogenes	49	983	2632	286	0.0082
Staphylococcus aureus	74	1627	4522	580	0.0094
Vibrio cholerae	16	2615	5694	1613	0.0066
Xanthomonas citri	21	3971	4639	38	0.0012

References

Baumdicker, F., W. R. Hess, and P. Pfaffelhuber, 2012, Genome Biology and Evolution **4**(4), 443.

Bertels, F., O. K. Silander, M. Pachkov, P. B. Rainey, and E. v. Nimwegen, 2014, Mol Biol Evol **31**(5), 1077.

Biller, S. J., P. M. Berube, J. W. Berta-Thompson, L. Kelly, S. E. Roggensack, L. Awad, K. H. Roache-Johnson, H. Ding, S. J. Giovannoni, G. Rocap, L. R. Moore, and S. W. Chisholm, 2014a, Scientific Data **1**, 140034.

Biller, S. J., P. M. Berube, D. Lindell, and S. W. Chisholm, 2014b, Nature Reviews Microbiology **13**(1), 13.

Bostock, M., 2016, D3: Data-driven documents, URL `http://d3js.org`.

Boto, L., 2014, Proc Biol Sci **281**(1777), ISSN 0962-8452.

Buchfink, B., C. Xie, and D. H. Huson, 2015, Nature methods **12**(1), 59.

Dongen, S. v., 2000.

Earle, S. G., C.-H. Wu, J. Charlesworth, N. Stoesser, N. C. Gordon, T. M. Walker, C. C. A. Spencer, Z. Iqbal, D. A. Clifton, K. L. Hopkins, N. Woodford, E. G. Smith, *et al.*, 2016, Nature Microbiology **1**(5), 16041, ISSN 2058-5276.

Ekseth, O. K., M. Kuiper, and V. Mironov, 2013, Bioinformatics , btt582.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis, 2002, Nucleic acids research **30**(7), 1575.

Everitt, R. G., X. Didelot, E. M. Batty, R. R. Miller, K. Knox, B. C. Young, R. Bowden, A. Auton, A. Votintseva, H. Larner-Svensson, J. Charlesworth, T. Golubchik, *et al.*, 2014, Nat Commun **5**, 3956.

Felsenstein, J., 1992, Evolution **46**, 159.

Felsenstein, J., 2004, *Inferring Phylogenies* (Sinauer).

Fournier, G. P., and J. P. Gogarten, 2008, J. Bacteriol. **190**(3), 1124, ISSN 0021-9193, 1098-5530.

Gardy, J., N. J. Loman, and A. Rambaut, 2015, Genome Biology **16**, 155.

Gómez, J., L. J. García, G. A. Salazar, J. Villaveces, S. Gore, A. García, M. J. Martín, G. Launay, R. Alcántara, N. D. T. Ayllón, *et al.*, 2013, Bioinformatics , btt100.

Hadfield, J., 2016, Phandango: Interactive visualization of genome phylogenies, URL `https://jameshadfield.github.io/phandango/`.

Holden, M. T. G., L.-Y. Hsu, K. Kurt, L. A. Weinert, A. E. Mather, S. R. Harris, B. Strommenger, F. Layer, W. Witte, H. de Lencastre, R. Skov, H. Westh, *et al.*, 2013, Genome Res. **23**(4), 653, ISSN 1549-5469.

Huson, D. H., 1998, Bioinformatics **14**(1), 68.

Huson, D. H., and D. Bryant, 2006, Molecular biology and evolution **23**(2), 254.

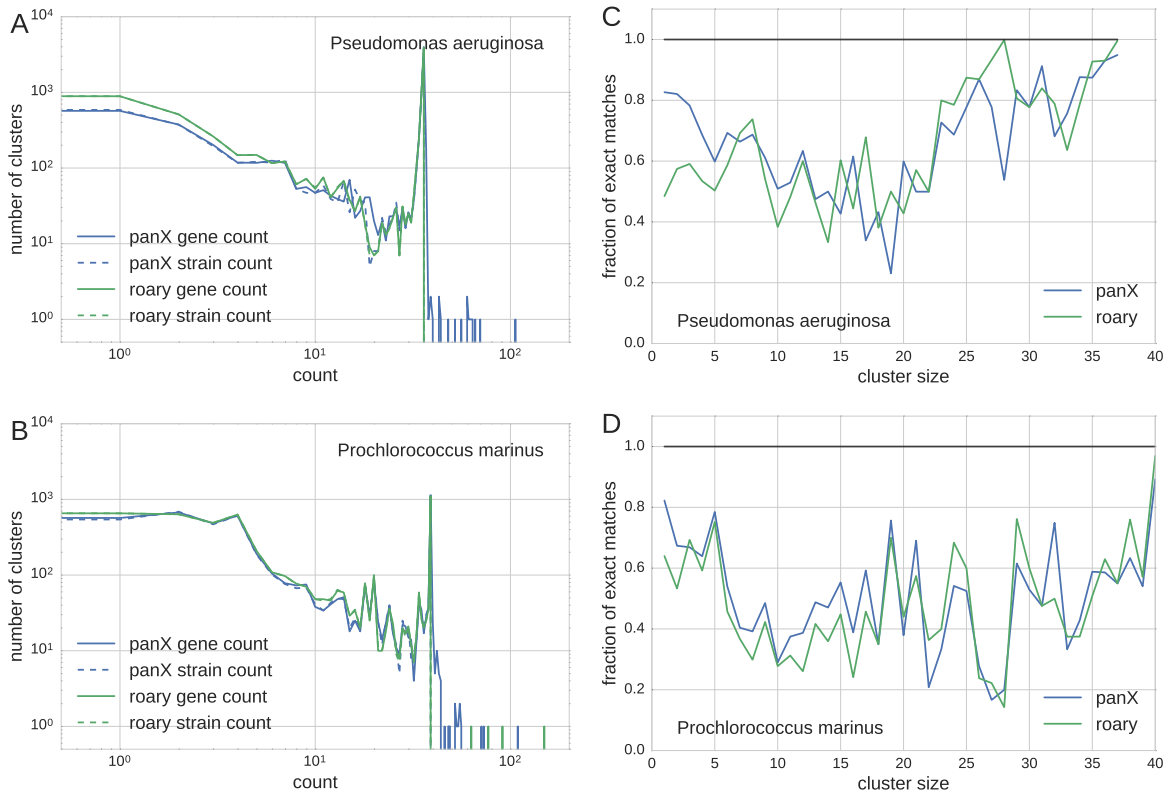

FIG. 4 **Pan-genome statistics:** We constructed a pan-genome of 37 *P. aeruginosa* strains from NCBI RefSeq using the panX pipeline and roary (similarity cutoff 70%). (A& B) shows the cluster size distribution of the pan-genomes of *P. aeruginosa* and *Prochlorococcus* obtained with both pipelines. Cluster size distributions are measured as the number of strains represented in a cluster (dashed lines) and as the number of genes (including close paralogs – solid lines). (C&D) shows the fraction of clusters identified by either pipeline, that have an exact match among the clusters identified by the other pipeline.

Katoh, K., K. Misawa, K.-i. Kuma, and T. Miyata, 2002, Nucleic acids research **30**(14), 3059.

Koonin, E., Y. Wolf, and P. Puigb, 2009, Cold Spring Harb Symp Quant Biol **74**, 205, ISSN 0091-7451.

Laing, C., C. Buchanan, E. N. Taboada, Y. Zhang, A. Kropinski, A. Villegas, J. E. Thomas, and V. P. Gannon, 2010, BMC bioinformatics **11**(1), 1.

Lapierre, P., and J. P. Gogarten, 2009, Trends Genet. **25**(3), 107, ISSN 0168-9525.

Méric, G., K. Yahara, L. Mageiros, B. Pascoe, M. C. Maiden, K. A. Jolley, and S. K. Sheppard, 2014, PloS one **9**(3), e92798.

Neher, R. A., and T. Bedford, 2015, Bioinformatics **31**(21), 3546, ISSN 1367-4811.

Ochman, H., J. G. Lawrence, and E. A. Groisman, 2000, Nature **405**(6784), 299, ISSN 0028-0836.

Page, A. J., C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. Holden, M. Fookes, D. Falush, J. A. Keane, and J. Parkhill, 2015, Bioinformatics **31**(22), 3691.

Pedersen, T. L., 2016, Panviz: D3 based visualisation for comparative genomics, URL `https://github.com/thomasp85/PanViz`.

Powell, D., 2016, Fripan: Explore your bacterial pan-genome, URL `http://drpowell.github.io/FriPan/`.

Price, M. N., P. S. Dehal, and A. P. Arkin, 2010, PLoS ONE **5**(3), e9490, ISSN 1932-6203.

Puigbò, P., A. E. Lobkovsky, D. M. Kristensen, Y. I. Wolf, and E. V. Koonin, 2014, BMC Biology **12**, 66, ISSN 1741-7007.

Rödelsperger, C., and R. J. Sommer, 2011, BMC Evolutionary Biology **11**(1), 239, ISSN 1471-2148.

Sagulenko, P., V. Puller, and R. A. Neher, 2016, treetime, URL `https://github.com/neherlab/treetime`.

Seemann, T., 2014, Bioinformatics **30**(14), 2068.

Soucy, S. M., J. Huang, and J. P. Gogarten, 2015, Nat Rev Genet **16**(8), 472, ISSN 1471-0056.

Stamatakis, A., 2014, Bioinformatics **30**(9), 1312, ISSN 1367-4803.

Tettelin, H., D. Riley, C. Cattuto, and D. Medini, 2008, Current Opinion in Microbiology **11**(5), 472, ISSN 1369-5274.

Thomas, C. M., and K. M. Nielsen, 2005, Nat Rev Micro **3**(9), 711, ISSN 1740-1526.

Vernikos, G., D. Medini, D. R. Riley, and H. Tettelin, 2015, Current Opinion in Microbiology **23**, 148, ISSN 1369-5274.

Woodhull, G., N. Zhu, *et al.*, 2016, dc.js - dimensional charting javascript library, URL `https://dc-js.github.io/dc.js/`.

Zamani-Dahaj, S. A., M. Okasha, J. Kosakowski, and P. G. Higgs, 2016, Molecular Biology and Evolution **33**(7), 1843.

Zhao, Y., X. Jia, J. Yang, Y. Ling, Z. Zhang, J. Yu, J. Wu, and J. Xiao, 2014, Bioinformatics **30**(9), 1297.

Zhao, Y., J. Wu, J. Yang, S. Sun, J. Xiao, and J. Yu, 2012, Bioinformatics **28**(3), 416.