Genomic Distance-based Rapid Uncovering of Microbial Population Structures (GRUMPS): a reference free genomic data cleaning methodology

Kaleb Z. Abram[1], Zulema Udaondo[1], Intawat Nookaew[1], Michael S. Robeson[1], Se-Ran Jun[1]

[1]Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA

Running title: Reference Independent Rapid Uncovering of Bacterial Population Structures

Keywords: Population structure, phylogenomics, Mash, comparative genomics, taxonomy, Average Nucleotide Identity, data cleaning, k-means clustering, machine learning, big data.

ABSTRACT

Accurate datasets are essential for rigorous large-scale sequence-based analyses such as phylogenomics and pangenomics. With the increasing rate of available sequence data and the varying quality of those sequences, reliable approaches to rapidly identify and automatically remove poor quality and misidentified genomes from datasets before performing sequence-based analyses, are greatly needed. Here we present a robust, controlled, computationally efficient method to obtain species level population structures of bacterial species regardless of the number of sequences present in the analysis. Genus level datasets can also be used with our pipeline to classify genomes into their species. This methodology can be leveraged to rapidly clean datasets of entire species of bacteria and analyze the sub-species population structures present in the genomes provided. These cleaned datasets can be further reduced by a variety of methods to obtain sets of sequences with various levels of diversity that are representative of entire species.

Introduction

With sequencing technologies becoming increasingly affordable and accessible worldwide the amount of publicly available whole genome sequences is skyrocketing. Heavily sequenced species such as *Escherichia coli*, now have tens of thousands of sequences publicly available which can, nevertheless, overwhelm accessible computational resources when utilizing a variety of analytical methods, such as alignment-based analytical methodologies[1]. Although there are now over 300,000 bacterial assembled genomes available in public repositories such as GenBank, large microbial datasets can have inherent sampling and sequencing biases that are virtually impossible to avoid[2,3]. These biases are due in part to the tendency to focus sequencing efforts on clinically relevant strains, due to their importance in human health, on model organisms, and also due to funding availability[3]. Additionally, the quality of the sequences available in these repositories can vary greatly among different samples[4].

Analyses performed using biased datasets can lead to conclusions about the genomic traits of bacterial species that are based on analyses that fail to fully capture the real sequenced diversity of the species. In the same manner, analyses that include poor-quality genome sequences can potentially produce incorrect conclusions or mask biologically meaningful results[5]. However, results from comparative genomic and taxonomic studies at the species level that ignore a large number of the available sequences in order to avoid these biases, would ultimately constrain the knowledge that could be gained by a more comprehensive analysis. Therefore, the ready availability of high-quality genome datasets that benefit from the large abundance of microbial sequences without including these biases, are of utmost importance for bioinformatic approaches, particularly for comparative analyses that rely on cutoffs based on the presence of genes in a certain percentage of genomes in the analyses such as comparative genomics, pangenome, taxonomy, and core genome multi-locus sequence typing analyses[6–10].

To assist with the performance of large-scale whole genome sequence (WGS) analyses, computationally efficient programs such as FastANI[11] and Mash[12] have been developed. These programs aim to overcome computational barriers involved in all-to-all comparisons of large datasets via rapid approximation of the inferred genetic similarity between any two bacterial genomes. However, these programs are solely designed to speed up the pairwise comparisons process by utilizing k-mers and MinHash (a technique that quickly estimates the similarity of texts), and do not directly produce high-quality genomic datasets. In the same way, while many methodologies, metrics, and tools exist to assess the quality of genomes (such as CheckM[13], estimating the completeness and contamination, presence/absence of gene markers, N50/NG50, and L50/LG50 ratios)[14] these methodologies were not designed to automatically handle the cleaning of large datasets and in the absence of reference genomes. Without robust methodologies that can fill this gap, the advantages provided by methods such as FastANI[11] and Mash[12] are ultimately constrained by low-quality or mislabeled genomic sequences.

To address these issues, we have developed GRUMPS (**G**enetic distance based **R**apid **U**ncovering of **M**icrobial **P**opulation **S**tructures), a program written in Python designed to provide a novel methodology to rapidly and reproducibly clean bacterial datasets of any size. GRUMPS is an improvement and expansion of the methodology recently utilized to perform the largest *E. coli* and *Pseudomonas aeruginosa* population structure analyses to date[10,15]. GRUMPS leverages a variety of automated and statistical methods to reproducibly remove outlier genomes from bacterial species datasets. Most of the cleaning modes provided by GRUMPS leverage statistics and unsupervised machine learning methods to accurately identify and remove genomes that are divergent from the whole set of genomes under investigation. Furthermore, GRUMPS provides a method to isolate multiple species from each other when studying datasets containing multiple bacterial species or genus level datasets. The resultant isolated species can then be fed back into GRUMPS to obtain high-quality

bacterial species datasets utilizing the cleaning modes designed for species level datasets as well as a reconstruction of the population structure of the species under investigation using the pair-wise genome comparisons between all genomes in the analysis.

RESULTS

Theory. The genetic structure of a microbial population, or population structure, provides the organized distribution of genetic variations among a set of strains from different geographical locations and ecological niches inside the species[16,17]. Thus, the population structure of a set of strains can be leveraged to better understand the biological context of a bacterial species and the unique traits of subgroups within a given species[10,15]. The genetic structure of a microbial population can also be leveraged to unveil existing biases (such as those from sequencing efforts) within a dataset.

In theory, a genomic dataset of any bacterial species is composed of a collection of genome sequences whose sequence similarity distribution could be sufficiently approximated by a normal distribution centered around an overall species level similarity value (where 0 indicates identical genomes and 1 represents genomes with no shared characteristics). Mislabeled genomes or genomes of low quality would deviate markedly from the similarity values possessed by the members of the species and fall on the right side of a histogram of similarity values for the species. Therefore, we hypothesize that statistics-based methods can reliably identify and exclude genomes that fall outside the normal distribution of similarity values within the dataset, leading to an increase in the overall quality of the dataset without compromising the overall sequence diversity of the species.

Hypothetical dataset. To illustrate this theory, a hypothetical bacterial species dataset was constructed containing 1,533 members that possess a normally distributed range of sequence similarities as

measured by Mash distances. This uncleaned hypothetical species dataset has an average Mash distance of 0.045419 which is below the 0.05 Mash distance species boundary[12] established by Ondov *et al*. This species boundary is approximately equal to a DNA-DNA hybridization value of 70% and an ANI value of 95%[12]. Based on the assumptions described in the previous section, the distribution of the Mash distances for this species should be Gaussian and centered slightly lower than 0.05 based on the average Mash distance of the species (Figure 1).

A histogram representation of the Mash distances for this hypothetical bacterial species shows a bimodal distribution (Figure 1a). The maximum Mash distance in Figure 1a is 0.087769, which lies outside the species boundary of 0.05 Mash distance and would be approximately equal to an ANI value of 91%. While a bimodal distribution is not explicitly indicative of one or more outlier genomes, the rightmost peak in Figure 1a is above the species boundary of 0.05 Mash distance and thus indicates the presence of one or more groups of genomes that are outside the species boundary. These outlier values could belong to mislabeled genomes or to draft genome assemblies that might have a low sequence quality due to the presence of contaminant sequences that belong to other species within their genome which results in lower sequence similarity values to the other members of this hypothetical species[18].

The clustered heatmap of Mash distances from Figure 1b, is colored based on the pairwise Mash distances of the genomes from the dataset where white colors represent Mash distances near the species boundary of 0.05. Regions in the upper left corner of the clustered heatmap with a brownish color in Figure 1b, correspond to outlier genomes that are above the species boundary of 0.05 as defined by Mash. As can be observed in the dendrogram above the heatmap in Figure 1b, these outlier genomes are separated from the large cluster in the heatmap resulting in the bimodal distribution observed in Figure 1a. Both Figures 1a and 1b are based on the uncleaned distance matrix for this hypothetical species and indicate that the dataset contains genome sequences above the species boundary cut-off.

Therefore, utilizing this dataset containing outlier sequences in downstream WGS-based analyses could lead to erroneous biological conclusions [18–20].

The impact of removing the outlier genomes identified in the hypothetical bacterial dataset (n=383) is visualized in Figures 1c and 1d. As shown in Figure 1c, the right most peak of the histogram in Figure 1a has disappeared and the average Mash distances shown in Figure 1d have decreased considerably with a maximum distance for the whole hypothetical dataset of 0.0575621 and a lower mean of 0.0384772. The cleaned dataset also revealed a more detailed population structure compared to the uncleaned dataset shown in Figure 1b which is due to the automatic scaling of the heatmap colors to the new range of Mash distance values in the heatmap.

In order to apply the methodology used in the hypothetical bacterial dataset and test our hypothesis on real datasets, we developed the GRUMPS (**G**enetic distance based **R**apid **U**ncovering of **M**icrobial **P**opulation **S**tructures) pipeline. GRUMPS is written in Python and designed to be scalable, time efficient, and accessible to the entire scientific community using local computational resources. GRUMPS can be downloaded from the following GitHub repository: https://github.com/kalebabram/GRUMPS.

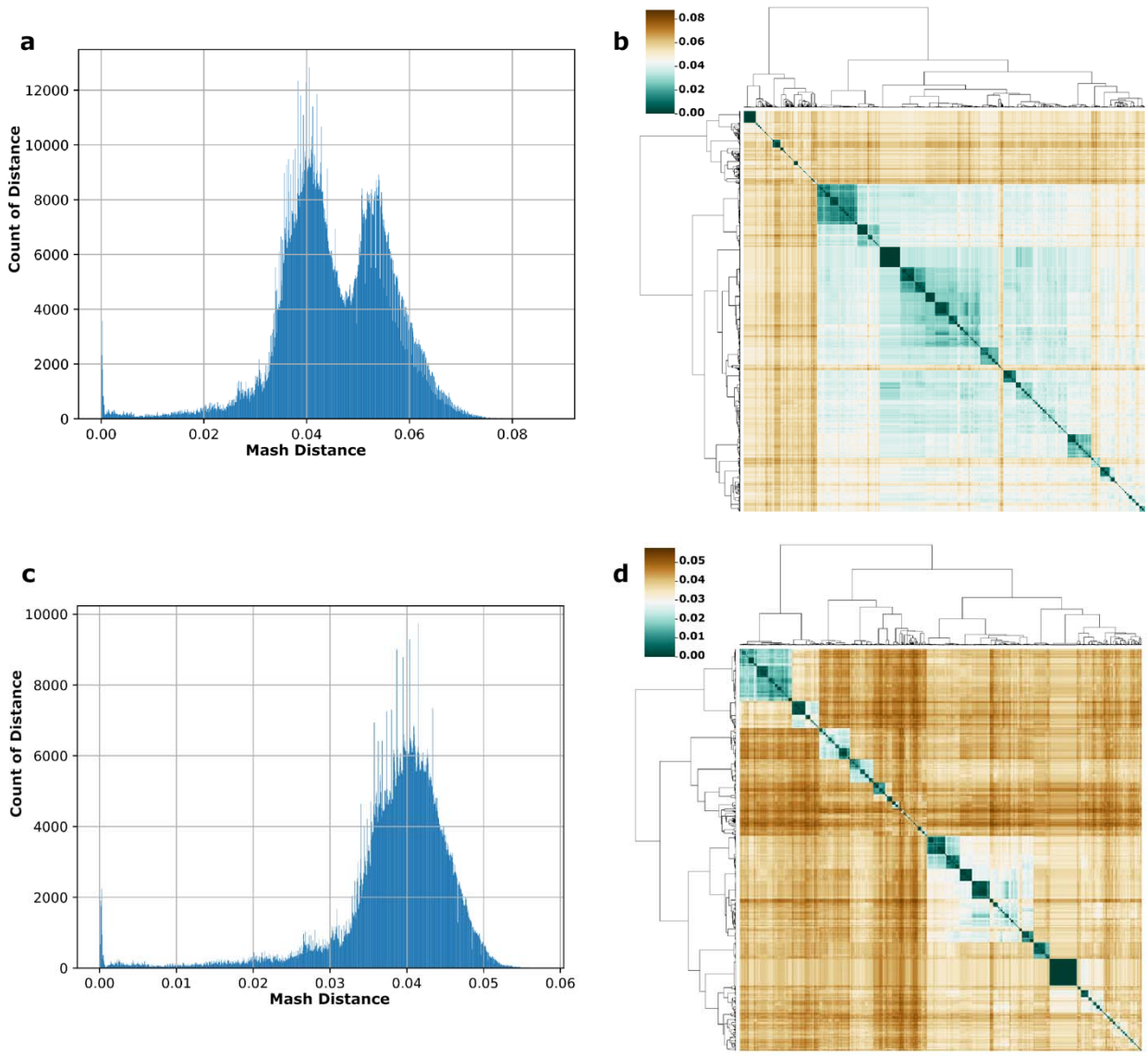


Figure 1 Hypothetical bacterial species before and after cleaning. Panels **a** and **b** show a set of 1,533 genomes belonging to a non-filtered hypothetical bacterial species. **a.** Histogram of Mash values for the entire hypothetical species. The bimodal distribution in panel **a** and the brown region of panel **b**, indicates that there are low-quality genomes present in the dataset. **b. Clustered** heatmap of Mash distances for the hypothetical species. Teal colors in heatmaps contained in panels b and d, represent genomes with higher similarity, and brown colors represent lower genetic similarity. Relative median similarities have the lightest color. The corresponding Mash distance to color is provided in the key in the upper left-hand corner of each heatmap panel. Panels **c** and **d** show a set of 1,170 genomes belonging to the hypothetical bacterial species after removing divergent genomes which can be caused by low-sequence quality or mislabeled genomes. **c.** Histogram of Mash values after cleaning and removing outlier genomes from the hypothetical dataset. The disappearance of the bimodal

distribution in panel **c** is due to the removal of the divergent genomes present in the initial dataset. **d.** Heatmap of Mash distances for the filtered hypothetical species.

General Summary of GRUMPS. GRUMPS is a novel methodology to rapidly and automatically clean datasets of bacterial genomes by identifying and removing outlier genomes using scalable and reproducible statistical methods in combination with unsupervised machine learning and user specified cutoffs. GRUMPS is intended to be used with distance matrices containing pairwise genome comparisons where column and row indexes are genome IDs (for example, Genbank Assembly Accessions). The values contained in the distance matrix are expected to range from 0 to 1, such as those provided by Mash[12], where a distance of 0 indicates identical genomes and a distance of 1 indicates no genetic features shared between genomes.

While GRUMPS was primarily designed to utilize distance matrices constructed from the output of Mash, helper scripts are provided to assist researchers in transforming their own distance or similarity matrices into a supported format. Due to the fact that GRUMPS was designed with Mash as the intended input, the default cutoff utilized by GRUMPS is 0.05 which is the experimentally determined species boundary for Mash distances[12]. The cutoff used by GRUMPS in the statistical filtering of datasets can also be set by researchers as a user-specified cutoff. When Mash distances are not leveraged then it is encouraged to use pairwise comparison methodologies that have definitive species-level cutoffs already experimentally determined (such as Average Nucleotide Identity (ANI)[21–23] or Average Amino Acid Identity (AAI)[24,25]. GRUMPS also provides a variety of filtering methods to maximize its utility.

Different Modes Available in GRUMPS. The “summary” mode is the intended starting point for any cleaning effort leveraging GRUMPS. This mode provides multiple statistical summaries of the input dataset as well as a histogram of the comparison values in the dataset and enables the manual

identification of outlier genomes, if desired. Based on internal testing, if a cutoff value different from the default value of 0.05 is desired, we recommend using the sum of the mean and standard deviation output by the “summary” mode of GRUMPS as the cutoff value for a given dataset. This can be useful when analyzing species, such as *M, tuberculosis,* whose genomic variability has been documented to be relatively smaller than other species[26] or datasets with enough contamination that the overall average of the dataset is greater than 0.05.

The “regular” and “strict” modes are the primary cleaning methodologies provided by GRUMPS. Both modes function by adding a small number of artificial outlier comparisons (5% of the total input dataset size) to the dataset where the pairwise values of all artificial outlier comparisons to all initial members of the dataset to be cleaned, are set to slightly above the cutoff (102% of the specified cutoff value). This creates a very small set of controlled artificial outliers that are just outside the boundary defined by the cutoff value for the species level, thus increasing the utility and reliability of the subsequent clustering step. Once the artificial outliers are added, both, "regular" and "strict" cleaning modes utilize k-means clustering (k=2) to produce two clusters, one cluster containing inlier genomes and the remaining cluster containing outlier genomes as well as the artificial outliers inserted by GRUMPS. The "strict" cleaning mode of GRUMPS then removes any genome whose mean distance to all other genomes in the dataset is greater than the sum of the mean and three times the standard deviation (three-sigma rule). For both “regular” and “strict” cleaning methods, an additional optional cleaning step, described below, can be applied.

The “sigma” mode is an alternative cleaning method which leverages the three-sigma rule on the distribution of comparisons for each genome to remove genomes with divergent behavior on either side of their distribution of comparisons relative to the rest of the genomes in the dataset. In this cleaning step, the distribution of comparison values for each genome is isolated and the extreme left and right

(defined as 0.03% on the left and 99.7% on the right) tails of each distribution are identified and a cutoff for each side is obtained by summing the mean of that side plus three times the standard deviation. By default, this cleaning mode is automatically applied as an additional cleaning step at the end of the “regular”, “strict”, “target”, and “remover” cleaning modes.

The “target” mode is intended to be used with one or more reference genomes (such as type strain assemblies or other genomes of interest) and removes any genomes from the dataset that do not have a distance value less than or equal to the set cutoff to all the reference genomes supplied. The identifier for each reference genomes utilized for “target” mode must be in columns and rows of the distance matrix that is the input to GRUMPS.

The “remover” mode is intended to remove one or more specified genomes from the input dataset which might consist of either easily identified or known outlier genomes. These genomes can be manually identified using the output of “summary” mode or if the researcher has conducted additional analyses indicating that a genome in their initial dataset should not be included in the cleaned dataset.

The “clique” mode is the final primary mode provided by GRUMPS and uses a graph-based clustering approach to clean datasets. This mode is intended to be used on datasets that contain multiple species (such as an entire genus). The “clique” mode partitions the dataset into multiple smaller uncleaned species level datasets as well as a set of unconnected genomes. The initial dataset is converted into a graph where nodes are genomes and edges are weighted by the distance between pairs of genomes. Utilizing the specified cutoff, GRUMPS trims all edges that are greater than the cutoff and returns all subgraphs. The distance matrix for each subgraph (presumably, species) identified by this mode is output, as well as a list of genome sequences that have no connections to any genome in the dataset. These distance matrices should be considered uncleaned and should be fed back into GRUMPS

to be cleaned by one of the two primary cleaning modes of GRUMPS, “regular” or “strict”, before subsequent analyses are performed on these datasets.

When utilizing “clique” mode, it is possible that some resultant datasets are not large enough for “regular” or “strict” mode to partition using the k-means clustering approach implemented in these modes. This can occur when the “clique” mode returns a resultant dataset with a small number of members where 5% of the resultant dataset is less than 1 (for example a resultant dataset with less than 100 members). To address this issue, a “small” cleaning mode is provided by GRUMPS which uses the average distance of each genome (ignoring self-comparisons) to remove outlier genomes. For example, if a dataset containing 95 genomes has 5 genomes with an average distance above the specified cutoff, then those 5 genomes would be considered outliers and removed from the dataset.

A high-level overview of the intended GRUMPS workflow is provided in Figure 2a, including an alternative entry point to the workflow which have helper scripts provided to aid researchers in alternate workflows leveraging GRUMPS. A detailed summary of the steps used in each cleaning method provided by GRUMPS is shown in Figure 2b.

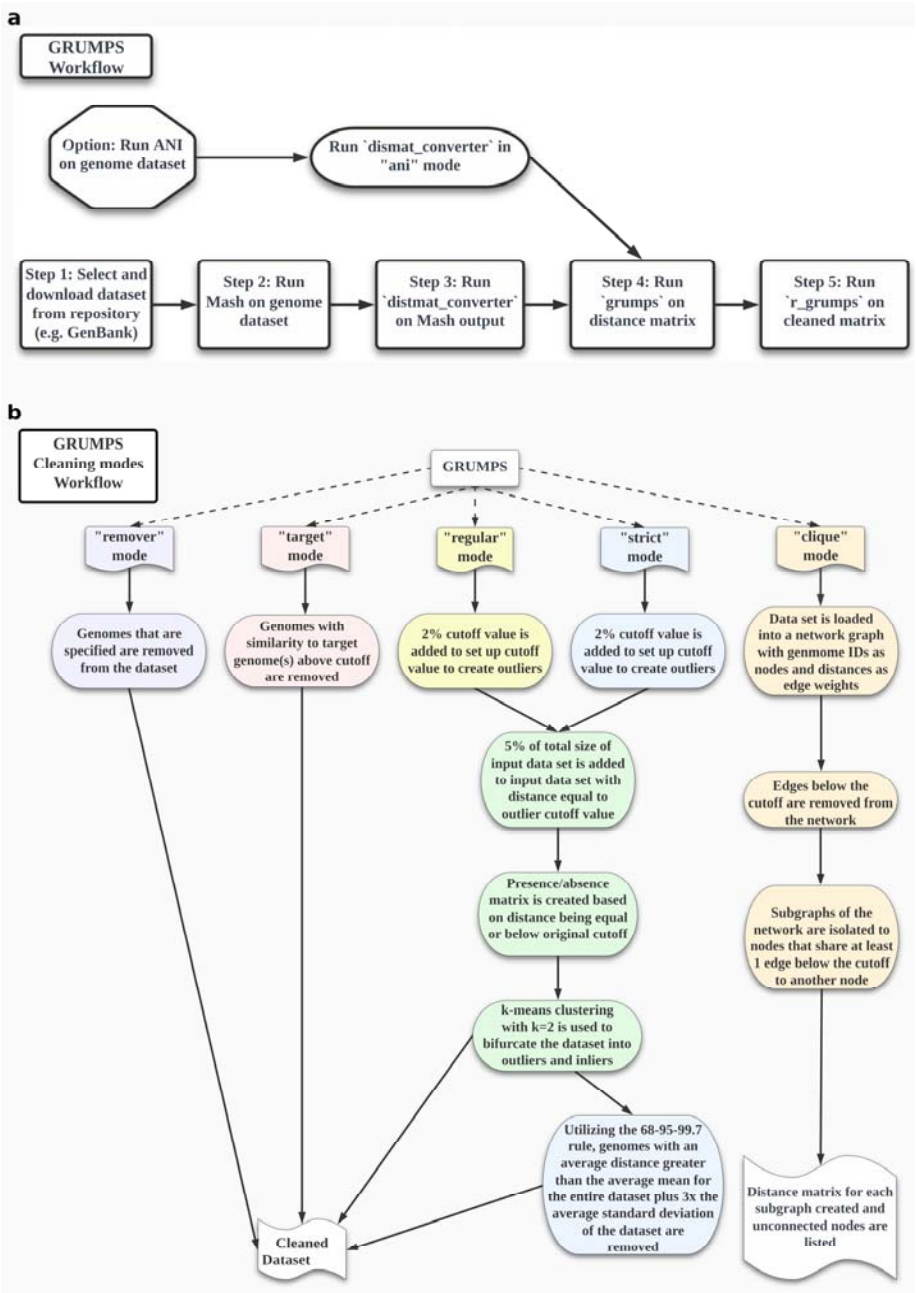
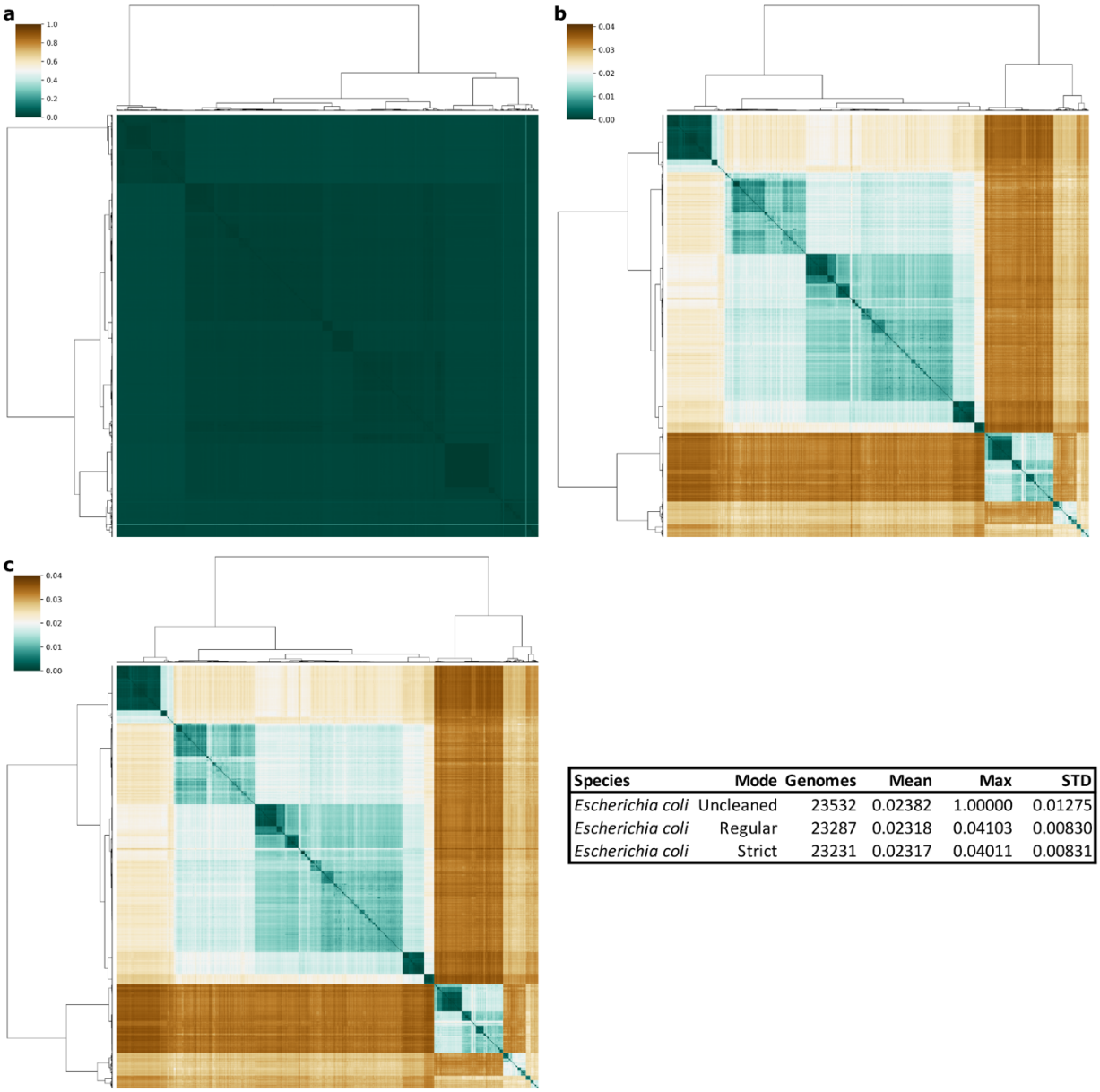


Figure 2. Graphical Summary of GRUMPS. Panel **a** contains a high-level overview of the GRUMPS workflow while panel **b** contains a summary of the major cleaning modes provided by GRUMPS. **a.** Each primary step is indicated by a square and move sequentially from left to right. Optional supported entry points in the workflow are represented by octagonal shapes. Oval shapes indicate the required steps needed to transition from the corresponding optional entry point to the primary GRUMPS workflow. **b.** For each mode, a generalized summary of the steps used in that mode are listed in order. Black arrows indicate the flow from start to finish for each mode. The steps of each mode are colored with a unique color except for the steps shared between “regular” and “strict” modes which are colored green.

Validation

While GRUMPS was developed and tested utilizing a variety of datasets containing multiple bacterial species from different phyla, a collection of three different bacterial species from the phyla Proteobacteria, Bacillota, and Actinomycetota (*Eschericia coli*, *Enterococcus faecium*, and *Mycobacterium abscessus* respectively) were used to illustrate the convergence between GRUMPS' results and previously described population structures.

***Escherichia coli* validation case.** A set of 23,532 *E. coli* genomes was used to illustrate the utility of both "regular" and "strict" cleaning modes of GRUMPS with the default cutoff of 0.05. The heatmaps shown in Figure 3 represent three different states of the dataset. Figure 3a contains a clustered heatmap of Mash pairwise distances of the 23,532 *E. coli* assemblies without being processed by GRUMPS (unclean dataset). Figure 3b contains the same dataset after being processed with the "regular" mode of GRUMPS with the optional "sigma" cleaning step applied, and Figure 3c after using "strict" mode of GRUMPS (using the default cutoff) with the optional "sigma" cleaning step applied. The values of the mean, maximum, and standard deviation are provided for each mode in the table of Figure 3. After using "regular" mode on the initial dataset, 245 genomes were removed which reduced the mean and maximum Mash pairwise distances values by 0.00064 and 0.95897 respectively. After using "strict" mode on the initial dataset, 301 genomes were removed which reduced the mean and maximum of Mash pairwise distances values by 0.00065 and 0.95989 respectively. The "strict" mode removed an additional 56 genomes than the "regular" mode and has a mean value 0.00001 and a maximum value 0.00092 less than those of "regular" mode. The results obtained by both "regular" and "strict" modes of GRUMPS are consistent with the results published by Abram et al in 2021[15] (Supplemental Figure 1a), despite utilizing a dataset twice the size. The heatmap in Supplementary Figure 1b shows the distribution of the 23,287 *E. coli* genomes in their previously described 14 phylogroups.



Species	Mode	Genomes	Mean	Max	STD
Escherichia coli	Uncleaned	23532	0.02382	1.00000	0.01275
Escherichia coli	Regular	23287	0.02318	0.04103	0.00830
Escherichia coli	Strict	23231	0.02317	0.04011	0.00831

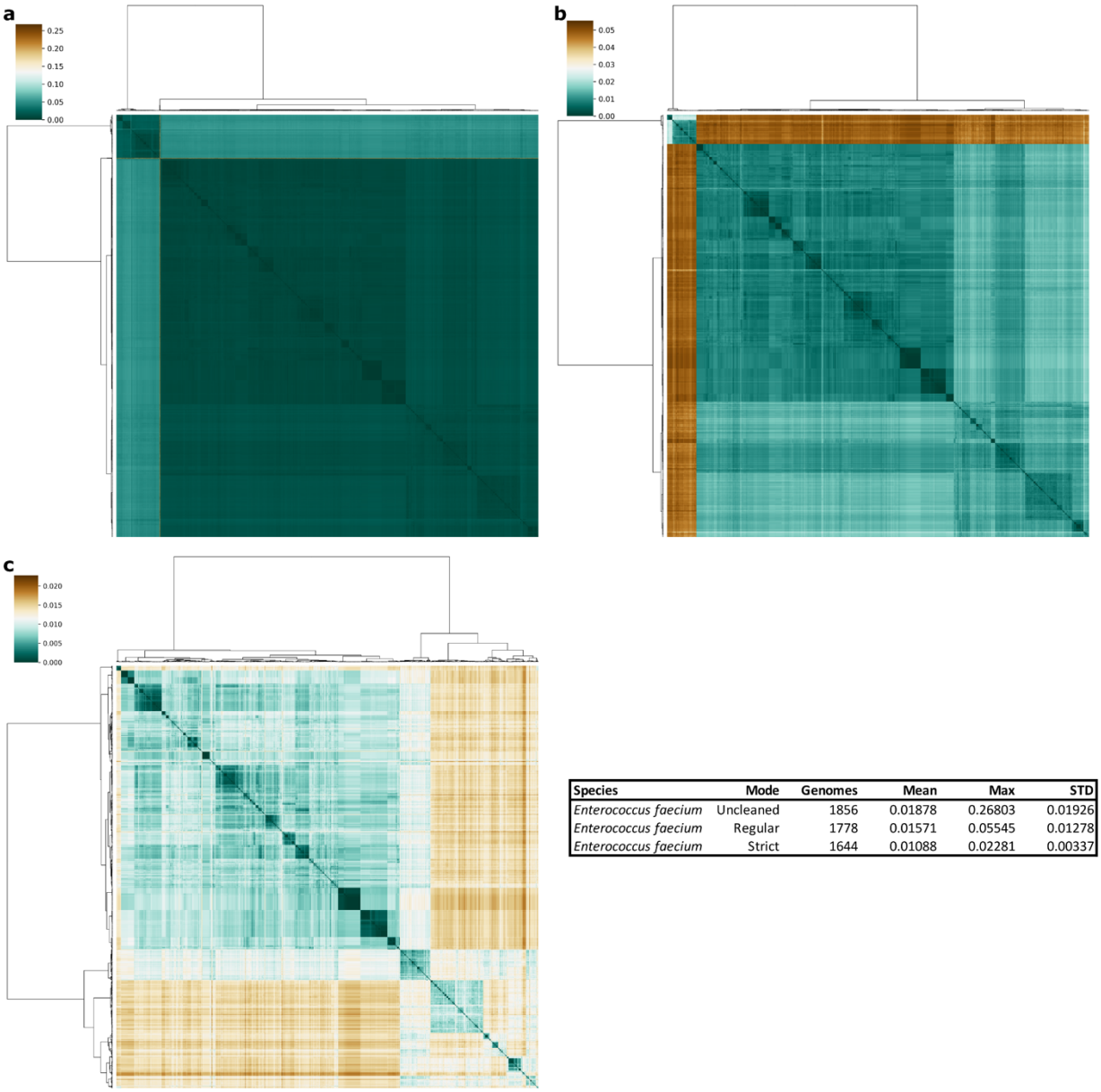
Figure 3. *Escherichia coli* genome datasets before and after cleaning with GRUMPS. Panel **a** contains a heatmap of pairwise Mash distance values for a set of 23,532 *E. coli* genomes. This dataset has not been cleaned using GRUMPS and includes many outlier genomes outside the species cutoff (0.05). Panel **b** contains a heatmap for a set of 23,287 *E. coli* genomes. This dataset has been cleaned using GRUMPS in "regular" mode with the optional "sigma" cleaning step and a majority of the outlier genomes removed. The population structure of the species is visible. Panel **c** contains a heatmap for a set of 23,231 *E. coli* genomes. This dataset has been cleaned using GRUMPS in "strict" mode with the optional "sigma" cleaning step and has removed all outlier genomes from the dataset. The population structure of the *E. coli* species is well defined and is similar to the results obtained using a set of 10,667 *E. coli* genomes[15].

***Enterococcus faecium* validation case.** As a means to expand the scope of the validation process, a set of 1,856 *E. faecium* genomes were used to illustrate the utility of both the "regular" and "strict" cleaning modes of GRUMPS with the optional "sigma" cleaning step using the default cutoff of 0.05 for Mash distance (Figure 4). After using "regular" mode on the initial dataset, 78 genomes were removed which reduced the mean and maximum values by 0.00307 and 0.21258 respectively. After using "strict" mode, an additional 134 genomes were removed from the processed dataset and reduced the mean and maximum values by 0.00483 and 0.03264 respectively.

The population structure *E. faecium* is typically divided into three clades: A1, A2, and B. A recently published study[27] by Udaondo *et al.* investigated the population structure of *E. faecium* using FastANI and a set of 2,273 genomes downloaded from GenBank, to classify genomes to their respective clade. Leveraging the clade classifications from this study, the described population structure of *E. faecium* was compared to the results obtained by GRUMPS. Genomes identified as clade B are contained in the small square of genomes visible in the upper left-hand portion of Figure 4b. Genomes identified as clades A1 and A2 correspond to the left and right sides, respectively, of the bifurcated branch of the dendrogram of Figure 4b and 4c. The removal of the genomes belonging to clade B when using the "strict" cleaning mode of GRUMPS is due to the effect of the heavy sequencing bias towards genomes belonging to clades A1 and A2 which are in turn more genetically similar, illustrated by the color differences of the heatmap in Figure 4b. This sequencing bias combined with the greater genetic similarity of clade A1 and A2 strains relative to clade B strains, shifts the mean Mash distance of the dataset towards clades A1 and A2 and causes the clade B genomes to be removed by the three-sigma rule employed in the "strict" cleaning mode of GRUMPS.

For *E. faecium*, the "regular" cleaning mode of GRUMPS with the default cutoff obtained the results that best agree with the currently accepted population structure of the species in which the three

known clades (A1, A2 and B) are visible and delineated from each other. However, it is worth noting that recent genome-based studies have indicated that strains classified as *E. faecium* clade B should be reclassified as members of the species *Enterococcus lactis* due to the closer genomic proximity of clade B strains to the type strain of *E. lactis* and the lack of hospital-associated markers that are present in *E. faecium* clade A1 and A2 strains[28–30]. To further investigate the interactions between clade B *E. faecium* strains and *E. lactis*, we utilized two type strain genomes for both species (GenBank Accessions: GCA_001544255.1, GCA_900447735.1, GCA_015904215.1, and GCA_015751045.1) and compared them to all clade B strains. All clade B strains were found to be more genetically similar to *E. lactis* and had Mash distances ranging from 0.00947898 to 0.0232342 which is noticeably less than the Mash distances to the type strain genomes of *E. faecium* which had Mash distances to all clade B genomes ranging from 0.040365 to 0.0488843.



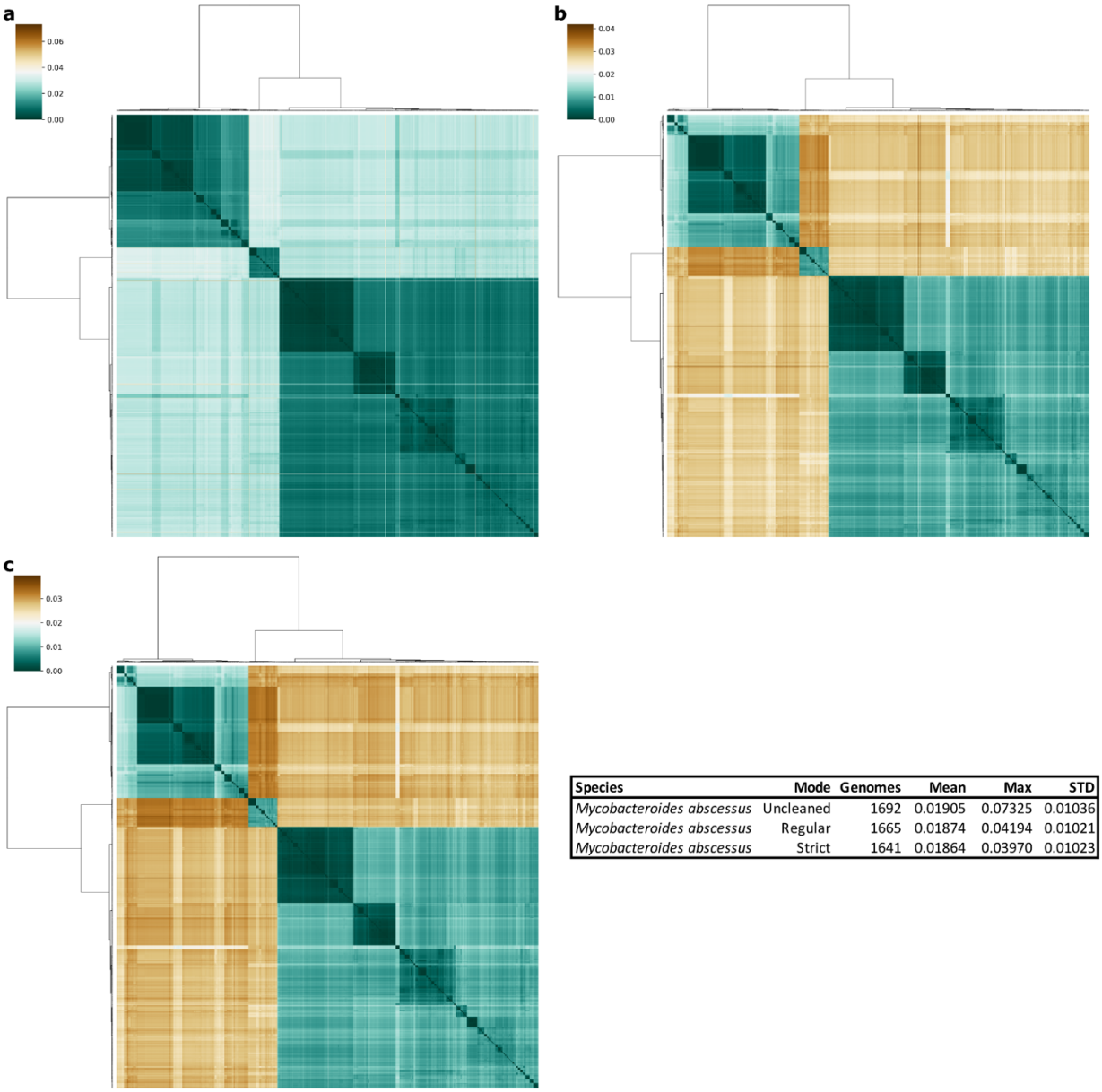
Species	Mode	Genomes	Mean	Max	STD
Enterococcus faecium	Uncleaned	1856	0.01878	0.26803	0.01926
Enterococcus faecium	Regular	1778	0.01571	0.05545	0.01278
Enterococcus faecium	Strict	1644	0.01088	0.02281	0.00337

Figure 4 *Enterococcus faecium* genome datasets before and after cleaning with GRUMPS. Panel **a** contains a heatmap for a set of 1,856 *E. faecium* genomes. This dataset has not been processed and includes many outlier genomes. Panel **b** contains a heatmap for a set of 1,778 *E. faecium* genomes. This dataset has been cleaned using GRUMPS "regular" mode with the optional "sigma" cleaning step and removed 78 outlier genomes. The population structure of the species is more apparent than the uncleaned dataset with clade B clearly differentiated from clades A1 and A2. Panel **c** contains a heatmap for a set of 1,644 *E. faecium* genomes. This dataset has been cleaned using GRUMPS "strict" mode with the optional "sigma" cleaning step. While the genomes belonging to clade B have been removed, the population structure of *E. faecium*

clades A1 and A2 is clearly visible. The values of the mean and maximum Mash distances, as well as the standard deviation are provided for each mode in the table.

***Mycobacterium abscessus* validation case.** A set of 1,692 *M. abscessus* genomes was used to illustrate the utility of both "regular" and "strict" cleaning modes of GRUMPS with the default cutoff of 0.05 and applying the optional "sigma" cleaning step (Figure 5). After using "regular" mode on the initial dataset, 27 genomes were removed. This reduced the mean and maximum values by 0.00031 and 0.03131 respectively. After using "strict" mode on the initial dataset, 51 genomes were removed leaving 1,641 genomes in the cleaned dataset which reduced the mean and maximum values by 0.00041 and 0.03355 respectively. The results obtained by both the "regular" and "strict" modes of GRUMPS clearly reveal the population structure of *M. abscessus*, with each of its three recognized subspecies (*M. abscessus* subsp. abscessus, *M. abscessus subsp.bolletii and M. abscessus subsp. massiliense)* forming their own distinct cluster within the heatmaps of Figure 5b and 5c. The first square in the upper left-hand corner of the clustered heatmap is composed of approximately 500 genomes belonging to *M. abscessus* subsp. *massiliense*. The middle square of Figure 5c is composed of approximately 200 genomes belonging to *M. abscessus* subsp. *bolletii*. The largest square in the bottom right corner of the clustered heatmaps is composed of approximately 1,000 genomes belonging to *M. abscessus* subsp. *abscessus*. Subspecies information was validated for these genomes by extracting the subspecies metadata for each genome. We also leveraged the classification provided by a recent study[31] by Matsumoto *et al.* published in 2019 which utilized a multilocus sequence type (MLST) based methodology, mlstverse, to differentiate these three subspecies from each other. An important difference between the results by Matsumoto *et al.* and those produced by GRUMPS is the absence of an intermixed subpopulation conformed by 43 genomes (28 *M. abscessus* subsp. *massiliens* and 15 *M. abscessus* subsp. *abscessus*) that was observed in the results produced by mlstverse[31]. The higher

resolution of whole genome sequence methods in comparison with typing methodologies that leverages only a set of genes has been demonstrated in other studies[10,15,27]. In this case, GRUMPS produced a seamless classification of *M. abscessus* genomes in their three recognized subspecies (Figure 5).



Species	Mode	Genomes	Mean	Max	STD
Mycobacteroides abscessus	Uncleaned	1692	0.01905	0.07325	0.01036
Mycobacteroides abscessus	Regular	1665	0.01874	0.04194	0.01021
Mycobacteroides abscessus	Strict	1641	0.01864	0.03970	0.01023

Figure 5. *Mycobacterium abscessus* genome datasets before and after cleaning with GRUMPS. Panel **a** contains a heatmap of pairwise Mash distance values for a set of 1,692 *M. abscessus* genomes. This dataset has not been cleaned using GRUMPS and includes genomes that are outside the species cutoff (0.05). Panel **b** contains a heatmap for a set of 1,665 *M. abscessus* genomes. This dataset has been cleaned using GRUMPS in "regular" mode with the optional "sigma" cleaning step and removed 27 outlier genomes. Panel **c** contains a heatmap for a set of 1,641 *M. abscessus* genomes. This dataset has been cleaned using GRUMPS in "strict" mode with the optional "sigma" cleaning step and has removed 51 outlier genomes from the dataset. The population structure of the species in both panels **b** and **c** is much clearer than the uncleaned dataset

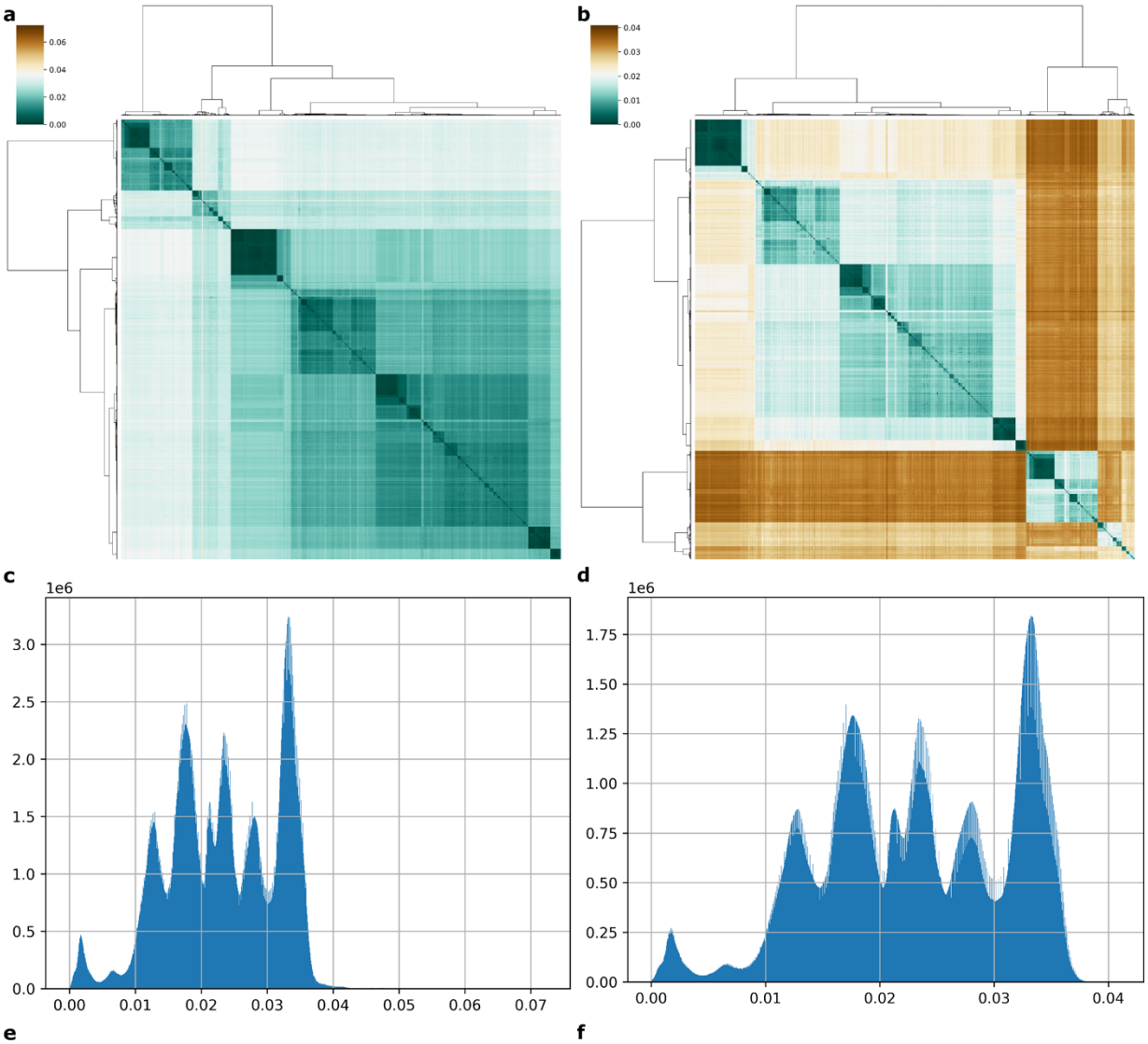
with the three subspecies of *M. abscessus* clearly visible. The values of the mean and maximum Mash distances, as well as the standard deviation are provided for each mode in the table.

Precision, Recall, and Accuracy Analyses. To assess the performance of the two primary cleaning methods provided by GRUMPS, we created a distance matrix of Mash distances using 45,558 genomes from *E. coli*, *Salmonella enterica*, and *Staphylococcus aureus* to create a set of confusion distance matrices to determine the precision, recall, and accuracy of both the 'regular' and 'strict' cleaning modes of GRUMPS. For each of these three species, 10 confusion distance matrices were constructed that contained all the genomes of that species in GenBank (as of April 6, 2020) and a set of confusion data composed of randomly selected genomes from the other two species. The size of the confusion data is equal to 10% of the total number of genomes of the species under investigation where 5% of the genomes come from each of the other two species not under investigation. Each confusion distance matrix was then cleaned using both the 'regular' and 'strict' cleaning modes of GRUMPS.

To determine the accuracy, specificity, and sensitivity of each cleaning mode for each species, the output of GRUMPS for each species dataset and the respective cleaning mode was considered the ground truth. Thus, any genome that is present in both the output of GRUMPS for the species and the confusion datasets for that species are considered true positives. Genomes present in only the output from the confusion distance matrix were considered false positives. For both cleaning methods and all three species the accuracy, sensitivity, and specificity was 100%. Since all 30 confusion distance matrices resulted in the same accuracy, sensitivity, and specificity for both cleaning methods, we conclude that the primary cleaning methods provided by GRUMPS are highly reliable and replicable.

Validation of Sigma Mode. To illustrate the effect of the additional "sigma" cleaning option available in both "regular" and "strict" cleaning modes, we utilized the "regular" cleaning mode of GRUMPS with and without the optional "sigma" cleaning step on the *E. coli* dataset utilized in the validation of

GRUMPS. The resultant heatmaps are presented in Figure 6 with 6a containing the heatmap generated by “regular” mode without “sigma” cleaning and 6b containing the heatmap generated by “regular” mode with “sigma” cleaning. The population structure of *E. coli* as illustrated by the clustered heatmaps is more pronounced when “sigma” cleaning is performed despite the “sigma” cleaning only removing 0.8% more genomes. As the color scale for the cluster heatmap is automatically generated based on the minimum and maximum values present in the dataset, the removal of genomes that deviate from the expected distribution of values for the species removes comparisons that obscure the population structure in the clustered heatmap. The distribution of pairwise comparisons for both datasets are shown in Figure 6c and 6d respectively. Comparing the topology of the histogram for both datasets, it can be observed that the “sigma” cleaning step removes outlier genomes without modifying the overall population structure of the species which is represented in the histogram by peaks and valleys. Moreover, the “sigma” cleaning step reduces the maximum value for the *E. coli* dataset to 0.04103, which is below the species boundary of 0.05, while the “regular” cleaning mode without the “sigma” cleaning has a maximum value of 0.07234 which is above the species boundary.



e

Species	Mode	Genomes	Mean	Max	STD
Escherichia coli	Uncleaned	23532	0.02382	1.00000	0.01275
Escherichia coli	Regular	23487	0.02332	0.07234	0.00835
Escherichia coli	Strict	23426	0.02323	0.04797	0.00828

f

Species	Mode	Genomes	Mean	Max	STD
Escherichia coli	Uncleaned	23532	0.02382	1.00000	0.01275
Escherichia coli	Regular	23287	0.02318	0.04103	0.00830
Escherichia coli	Strict	23231	0.02317	0.04011	0.00831

Figure 6 Comparison of GRUMPS "regular" cleaning mode with and without "sigma" filtering using *Escherichia coli*. Panel **a** contains a heatmap of pairwise Mash distance values for a set of 23,487 *E. coli* genomes. This dataset was produced by the "regular" mode of GRUMPS without "sigma" cleaning and includes pairwise comparisons that are outside the species boundary (0.05). Panel **b** contains a heatmap generated by using the "regular" cleaning mode of GRUMPS with "sigma" cleaning and does not include any pairwise comparisons above the species boundary. Panels **c** and **d** contain histograms of the pairwise comparisons utilized to produce panels **a** and **b** respectively. Panel **e** contains the number of genomes, mean, maximum value, and standard deviation for the initial dataset (uncleaned), "regular" mode, and "strict" mode without "sigma" cleaning while panel **f** contains the same information with "sigma" cleaning.

Effect of plasmid sequences. While GRUMPS is a methodology to rapidly clean datasets, it should be noted that the method (ANI or Mash distance) utilized to obtain the genetic similarity between a pair of genomes ultimately defines the effect of inclusion or exclusion of plasmid sequences on the output of GRUMPS. Since GRUMPS utilizes statistical cleaning methods and unsupervised machine leaning on genetic similarity values, the similarity values input to GRUMPS determine whether a genome is included or excluded from the output of GRUMPS.

To investigate the effects of plasmid sequences on our methodology, we created a dataset of 877 complete genomes of *E. coli* that included plasmids and created a duplicate set of genomic sequences that had all plasmid sequences removed. For both 'regular' and 'strict' cleaning modes of GRUMPS, the removal of plasmid sequences resulted in a 0.228% variation between the two datasets. Of the 877 original sequences, only two genomes (GCA_900088105.2 and GCA_900184875.1) were not present in the plasmid excluded data but were present in the plasmid included data after applying either cleaning mode. Based on this test case, the exclusion of plasmid sequences results in a slight variation in the output of the cleaning methods provided by GRUMPS but ultimately should have little effect as the size of the dataset increases. However, it should be emphasized that GRUMPS only leverages the similarity values provided by other tools and ultimately the sensitivity of a given tool to inclusion or exclusion of plasmid sequences is outside the scope of GRUMPS.

Computational Performance. To benchmark the speed of GRUMPS' "summary", "regular", and "strict" modes, we utilized a collection of distance matrices composed of the *E. coli* genomes used in the validation section (ranging from 500 to 23,532 with a step size of 500). For both "regular" and "strict" cleaning modes, benchmarking was performed with and without the optional features available for these modes (see Methods). The walltime for "summary" mode ranged from 2.7 to 332.9 seconds. The walltime for "regular" mode without the heatmap and sigma options ranged from 2.4 to 1081.7

seconds. The walltime for “strict” mode without the heatmap and “sigma” options ranged from 2.4 to 1082.4 seconds. The walltime required for the heatmap and “sigma” options ranged from 4.5 to 1506 seconds and 2.4 to 1083.9 seconds respectively. In summary, the “sigma” option has little impact on the walltime while the heatmap option increased walltime from approximately 2 to 424 seconds depending on the size of the dataset. GRUMPS was able to process the *E. coli* dataset containing 23,533 genomes with both the heatmap and sigma options enabled in approximately 25 mins. Benchmarking was performed using a workstation with the following configuration: Ubuntu 22.04.1, 2.1 GHz Intel Xeon Gold 6230, 200GB of Micron DDR4 2933 RAM, and a 1TB Micron 2200S NVMe.

Discussion

GRUMPS is a robust data cleaning platform in which statistical analyses and machine leaning are leveraged to identify and remove low-quality and misidentified genomes providing a novel methodology to rapidly and reproducibly clean bacterial genomic datasets of any size. The precision, recall, and accuracy analyses as well as the comparable performance of GRUMPS on real world datasets to previously published results are indicative of the reliability of the methodology.

To analyze GRUMPS’ ability to accurately identify and remove low-quality genomes from species level datasets several validations analyses were performed. A selection of three bacterial species belonging to three unique phyla, Proteobacteria, Bacillota, and Actinomycetota were utilized to investigate GRUMPS’ performance on real world datasets.

First, the results obtained by GRUMPS’ “regular” cleaning mode were compared with the results obtained in the largest *E. coli* analysis conducted to date[15] using a dataset twice the size. The output from GRUMPS reproduced the results shown in Abram et al., 2021 (Figure 3 and Supplementary Figure 1) distributing the *E. coli* genomes in the 14 previously described phylogroups.

To further validate GRUMPS' ability to accurately identify and remove low-quality genomes from species level datasets, we also compared the results obtained by GRUMPS for both *E. faecium* and *M. abscessus* to recent studies[27,31] related to the population structure of these two species. For *M. abscessus*, the results obtained by both the "regular" and "strict" cleaning modes of GRUMPS revealed the known population structure of this species with each of the known subspecies groups visible and distinctly separate from each other. For *E. faecium*, the results obtained by the "regular" cleaning mode of GRUMPS revealed the known population structure of the species with all three recognized clades visible. However, the results obtained by the "strict" cleaning mode of GRUMPS removed all clade B strains from the dataset. While this could be viewed as overcleaning, recent studies on clade B strains of *E. faecium* indicate that these strains should be reclassified as members of the species *E. lactis*. Our investigation based on the genetic similarity between clade B strains and type strains for both *E. faecium* and *E. lactis* also support this reclassification. Thus, the results obtained by the "strict" mode of GRUMPS for *E. faecium* reflect a more accurate population structure for the species.

Benchmarking results showed that GRUMPS is a computationally efficient tool able to process an *E. coli* dataset containing more than 23,000 genomes in under 30 minutes with local computational resources. While processor clock speed and RAM speed contribute to the overall walltime of any GRUMPS method, the major computational bottlenecks of GRUMPS are the size of the input data (which increases the read and write time in addition to the increases the walltime of the cleaning process) and the disk speed of the drive where the dataset is being read and written to.

In a nutshell, GRUMPS is impactful for three main reasons: the ability to accurately identify and remove low-quality genomes from species level datasets, the ability to isolate uncleaned species level datasets from genus level datasets, and the ability to rapidly uncover the population structure of any bacterial species. GRUMPS provides a methodology to assist researchers in obtaining high-quality

datasets for comparative genomic studies even when utilizing big data. In addition, GRUMPS and its helper scripts provides an easy-to-use platform for creating high quality figures and investigating the population structure of any bacterial species.

Methods

Test datasets

To conduct the analysis, the prokaryotes summary table was downloaded from NCBI on April 6, 2020; the total number of entries in GenBank on this date was 248,003. The table was then filtered to only include entries with 'Bacteria' in the 'Organism Groups' column resulting in a set of 243,260 possible genomes. To reduce the impact of low-quality genomes, we exclude genomes with more than 500 contigs which resulted in a reduced dataset of 231,608 genomes. To reduce the impact of uncultured genomes, we removed any genome with ‘uncultured’ in the '#Organism Name' column which resulted in a reduced final set of 218,099 potential genomes.

To generate an approximate abundance of species within this reduced dataset, the '#Organism Name' column was isolated and each entry was split by white space and only entries with one or more white spaces were retained. Next, we joined the first two splits with a single white space and appended them to a new list. We then trimmed any leading or trailing ‘ characters to improve the accuracy of counting unique genus/species strings and counted the frequency of each unique genus/species string; the number of unique genus/species strings was 13,802. To reduce the impact of poorly annotated genomes, we removed any genus/species strings that contained either ‘bacterium’ or ‘sp.’ which resulted in a set of 11,913 unique genus/species strings.

In order to demonstrate the scalability of GRUMPS, we assessed the population structure of at least 50% of the final reduced set of bacterial genomes by including the most numerous species until there were at least 109,050 genomes in the test set. This resulted in a test set containing 31 species of bacteria and 110,220 genomes. It should be noted that based on the results of our previous work and in accordance with the literature, we considered genomes belonging to the genus *Shigella* to be within the species *Escherichia coli* and counted all *Shigella* labeled genomes and *E. coli* labeled genomes as the same species. We then used the GCAs of all 110,220 genomes in the test dataset to download the genomic .fna file using BatchEntrez.

Plasmid Dataset

Using the metadata provided by NCBI, we identified a set of 877 complete *E. coli* genomes that contained plasmid sequences and we created a duplicate copy for each genomic file that excluded all plasmid sequences. Mash sketches (k=21 and s=10,000) were generated using all 877 genomes: a set that included the plasmid sequences and a set with the plasmid sequences removed. Mash distances were obtained for both sets and converted into a distance matrix using the helper script provided in GRUMPS. We then utilized both ‘regular’ and ‘strict’ cleaning modes of GRUMPS on each dataset. Both cleaning modes were compared by identifying any genomes that were present in the plasmid included dataset but missing from the plasmid removed dataset.

Precision, Recall, and Accuracy Analyses

We identified the three most numerous species (*E. coli*, *Salmonella enterica*, and *Staphylococcus aureus*) in our test data to create multiple sets of confusion data to assess the accuracy, sensitivity, and specificity of GRUMPS. Using the sketches generated in the initial test datasets, pairwise Mash distances for all 45,558 genomes in these three species were obtained. These distances were then

converted into a distance matrix using the provided helper script. To create 10 confusion datasets for each species, we randomly selected genomes totaling 10% of the total size of the species under investigation from the other two species, 5% from each. This resulted in a set of 30 confusion distance matrices that were then cleaned using both 'regular' and 'strict' cleaning modes implemented in GRUMPS resulting in 60 total sets. To calculate accuracy, specificity, and sensitivity of both cleaning modes per species, we considered the output of each cleaning method from the original test data of that species to be the ground truth, meaning any genome retained by the respective cleaning method in the original test data analysis were considered true positives and any genome removed by the respective cleaning method in the original test data analysis were considered true negatives. For each of the 60 sets, true positives were genomes that were present in both the original set and the test set, true negatives were genomes that were absent in both the original set and the test set, false positives were genomes that were present in the test set but absent in the original set, and false negatives were genomes that were removed in the test set but present in the original set.

Benchmarking

To benchmark the speed of GRUMPS' "summary", "regular", and "strict" modes, we produced a set of distance matrices with genomes ranging from 500 to 23,532 with a step size of 500 from the *E. coli* dataset used in the validation section. For each mode, we ran GRUMPS 10 times and recorded the wallclock time which was then averaged to obtain the wallclock time for each cleaning mode. For both "regular" and "strict" cleaning modes, we utilized the following different sets of optional features to assess the computational increase associated with the optional features: with heatmap with sigma, with heatmap without sigma, without heatmap with sigma, and without heatmap without sigma. Benchmarking was performed using a workstation with the following configuration: Ubuntu 22.04.1, 2.1 GHz Intel Xeon Gold 6230, 200GB of Micron DDR4 2933 RAM, and a 1TB Micron 2200S NVMe.

A MacBook Pro (Mid 2015 model) with 16GB of 1600MHz DDR RAM, a 2.8GHz Quad-Core Intel Core i7 processor (i7-4980HQ), and a 1TB Apple SATA SSD SM1024G took approximately 55 minutes to run GRUMPS in 'regular' mode with both the heatmap and sigma filtering options enabled on the full *E. coli* dataset containing 23,532 genomes.

GRUMPS Workflow

An overview of the typical GRUMPS workflow is shown in Figure 2a. In summary, researchers need to identify and download the genomes they intend to analyze before users run Mash (or another comparison method which has definitive species-level cutoffs such as ANI) on the genome set. Based on extensive testing, we generally recommend that genomes are sketched first using a k-mer size of 21 and 10,000 sketches before Mash distances are calculated to increase both performance and reproducibility; however, it is possible that these settings will not work for every bacterial genome that has or will be sequenced and different settings might be required depending on the dataset of interest. Once Mash distances are obtained, the provided helper script `distmat_converter` can be utilized to create a NxN symmetrical distance matrix (where N is the number of genomes) required by `grumps`. If a comparison method other than Mash is utilized, such as ANI, the helper script `distmat_converter` should be run with '-m ani'. When `distmat_converter` is run with '-m ani', users can either convert their values to Mash distances with '-c yes' or invert their values with '-i yes'. Only one of these options can be used at a time and '-c yes' is prioritized over '-i yes'. Converting to Mash distances is accomplished by the following: 1-(ANI/100). Inverting is accomplished in a similar fashion: 100-ANI. We highly recommend users to convert their values to Mash distances as GRUMPS was designed to work with values in this range. If users choose to invert instead of convert, users will need to manually specify their cutoff to an appropriate value. A cutoff of 5 would correspond to a species boundary of 95% ANI value. Once users have a distance matrix, one of the available cleaning methods described in detail

below can be leveraged to clean the dataset. Finally, the supplied R script `r_grumps` can be used to produce graphical representations of the dataset. The available cleaning modes in GRUMPS and their function are summarized in Figure 2b.

Cleaning Options

Summary Mode

The "summary" mode is intended to be used to quickly obtain a statistical overview of the dataset. Users are encouraged to use this mode first on any species level dataset before using any of the cleaning methods provided by GRUMPS to get an initial understanding of the distribution of genetic similarity values for the dataset. The summary mode by default provides a histogram of the distances and three summary matrices (genome level summary, summary of the means, and the overall summary). Both the mean and overall matrices provide the following: the count of genomes in the dataset, the mean and standard deviation of all distances in the input distance matrix, the minimum and maximum of the input distance matrix, and a range of percentiles for all distances in the input distance matrix. The genome summary matrix provides similar information as the mean and overall matrices on a per genome basis with additional percentiles to manually assess and evaluate the quality of each genome. The intended use of the overall matrix is to assess the distribution of distances through the whole dataset while the genome summary matrix is intended to quickly identify the presence of genomes that could be outliers within the input dataset.

Regular Mode

The "regular" mode is the intended default mode for cleaning datasets utilizing GRUMPS. In the regular mode, the specified cutoff (default 0.05) is increased by 2% to create an outlier distance value that is outside the specified cutoff. To enable rapid outlier detection, an artificial set of outliers is

created with a total number of members equal to 5% of the total dataset size in order to minimize the bias created by this set of artificial outliers. The distance for each artificial outlier between all artificial outliers and genomes in the dataset is equal to the outlier distance value. These outlier distance values are appended to the input distance matrix. The modified distance matrix is then converted into a presence/absence matrix to facilitate a non-biased k-means partitioning, where presence represents distances equal or below the specified cutoff and absence represents distances above the specified cutoff. The resultant presence/absence matrix is then partitioned using k-means clustering (k = 2) to bifurcate the dataset into inliers and outliers. The outliers are determined by the cluster that contains the artificial outliers. It is important to note that for some datasets regular mode with the default cutoff may not sufficiently clean the dataset which can be observed in the clustered heatmap output by this mode. In these cases, users can either re-run regular mode and modify the cutoff used to clean or clean the initial dataset using strict mode. For re-running regular mode with a modified cutoff, we recommend using summary mode to determine the overall mean and standard deviation of all distances in the initial dataset and add these two values together to obtain a cutoff better suited for the initial dataset due to the variation of similarity values between different species of bacteria.

Strict Mode

The strict mode is intended to be used with datasets that are highly similar and have an overall mean distance much lower than the default cutoff of 0.05. This mode is practically identical to running in regular mode with the default cutoff. However, the strict mode includes a final cleaning step that automatically scales with the provided data. As GRUMPS is designed around the assumption that bacterial species dataset follows an approximately normal distribution of sequence similarity values, strict mode utilizes the 68-95-99.7 rule, also known as the three-sigma rule, to automatically identify outliers using the statistical distribution of distances in the initial dataset. In GRUMPS, strict mode is

implemented by finding the overall mean of the initial dataset and adding three times the overall standard deviation of the dataset to obtain a new cutoff (empirical rule of 99.7%). This new cutoff is then used to remove any genome that has an overall average distance to the entire dataset higher than the statistically derived cutoff.

Sigma Mode

The sigma mode is intended to be utilized in combination with the "regular" or "strict" modes of GRUMPS, though it can be applied to any dataset without needing to run either "regular" or "strict" mode. This standalone mode is provided to give researchers flexibility when using GRUMPS to clean their dataset and can be run after any of the cleaning modes provided by GRUMPS. Note that if GRUMPS is run in "regular" or "strict" mode without providing the command line argument "-s no", GRUMPS will automatically perform the sigma cleaning step. The sigma mode utilizes a similar principle to the strict mode but uses the distribution of distances for each genome as a filtering criterion as opposed to the means of each genome. For each genome, the behavior of the left and right tails of the distribution is assessed by calculating the percentiles 0.03% and 99.7% to assess the left and right tail behavior respectively. For each side, the mean and standard deviation are calculated using the respective percentile values and a cutoff is determined for each side by summing the mean and three times the standard deviation for the respective side. Genomes that have left or right tail values that are above the respective cutoff are removed.

Target Mode

The target mode is intended to be utilized by researchers who have either a single genome or a set of genomes (for instance, whole genome sequences of type strains) or a set of reference genomes and wants to use them as references to clean the given dataset. In the target mode, a set of reference genomes in the dataset of investigation are used to remove genomes from the dataset that have a

distance larger than the specified cutoff to the set of reference genome(s). Due to using only a select number of genomes during the filtering step for identifying outlier genomes, this mode is less reliable than the other cleaning methods (regular mode, strict mode) provided by GRUMPS. However, we implemented this mode to enable researchers to generate datasets focused on a collection of genomes they wish to investigate such as type strain sequences for the species under investigation.

Clique Mode

The "clique" mode is intended to be utilized by researchers who want to investigate a dataset that either is known to contain multiple species, such as an entire genus, or potentially contains more than one species. In this mode the distance matrix is loaded into a network graph using the python package NetworkX, where genome IDs are assigned to the labels of nodes and pairwise distances between genomes set to edge weights. This network graph is then trimmed such that any weighted edge between nodes that is above the cutoff are removed resulting in a collection of sub-graphs. Each sub-graph contains a collection of nodes that contain at least one weighted edge to another node within the subgraph equal or less than the specified cutoff. For each sub-graph, a sub-distance matrix containing all nodes of the sub-graph parsed out of the initial distance matrix is written out. Genome identifiers without any weighted edges equal to or lower than the specified cutoff are saved in a separate file. These genomes with no edges connected are typically very low-quality, more likely deposited with misclassified taxonomic labels, or could be members of a species with only a single member in the initial dataset. These genomes should be further investigated before assuming the reason for their exclusion from any of the resultant sub-graphs. The resultant sub-graphs can then be cleaned using one of cleaning methods in GRUMPS to obtain cleaned species datasets. We recommend that each sub-graph should be summarized with the summary mode of GRUMPS before cleaning.

Small Mode

The “small” mode is intended to be utilized on small datasets when “regular” or “strict” mode fails to partition the data via k-means clustering where k=2. As a feature of both “regular” and “strict” modes, a warning is raised when these modes fail to partition a small dataset. For most species level datasets this should not be an issue, but some datasets obtained using the “clique” mode might be small enough to cause “regular” or “strict” mode to fail. This mode calculates the average distance of each genome to all other genomes in the dataset excluding self-comparisons. For example, a genome in a dataset of three genomes might have a distance of 0.055 to all genomes in the dataset (true average distance is 0.055) but the average distance of this genome would be 0.037 if the self-comparison is included. Genomes that have an average distance above the cut-off are removed from the dataset.

Remover Mode

The “remover mode” is intended to assist with manually removing genomes from the dataset. While the remover mode is intended to be used in parallel with “summary” mode, it can also be useful to users that need to remove a set of genomes from the dataset for any reason. The genomes to remove should be provided in a text file as input where each line has a genome identifier in the provided dataset.

Author contributions

Conceptualization: K.Z.A.

Data curation: K.Z.A.

Formal Analysis: K.Z.A.

Funding acquisition: S.J.

Investigation: K.Z.A. Z.U. I.N. M.R. S.J.

Methodology: K.Z.A.

Software: K.Z.A.

Supervision: S.J. Z.U.

Validation: K.Z.A. Z.U.

Visualization: K.Z.A. Z.U.

Writing- original draft: K.Z.A. Z.U.

Writing – review & editing: K.Z.A. Z.U. I.N. M.R. S.J.

Competing interesting

Authors declare no competing interests.

Supplementary Information

Supplementary Figure 1 Comparison of *Escherichia coli* population structure from Abram et al. 2021 and GRUMPS "regular" mode. Panel **a** contains a heatmap of pairwise Mash distance values for a set of 10,667 *E. coli* genomes published in Abram et al. 2021. Panel **b** contains a heatmap of pairwise Mash distance values for a set of 23,287 *E. coli* genomes produced by GRUMPS with "regular" cleaning mode and "sigma" filtering. The color bars at the top of both heatmaps identify the phylogroups as predicted in Abram et al. 2021. The colors in the heatmap are based on the pairwise Mash distances. Shades of teal represent similarity between genomes, with the darkest teal corresponding to identical genomes reporting a Mash distance of 0. Shades of brown represent low genetic similarity per Mash distance, with the darkest brown indicating a maximum distance of ~ 0.039. Genomes of relative median genetic similarity have the lightest color. While the ordering of the phylogroups differs between panel **a** and **b**, the classification remains the same.

References

1. Zielezinski, A., Vinga, S., Almeida, J. & Karlowski, W. M. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology* **18**, 186 (2017).
2. Brockhurst, M. A. *et al.* The Ecology and Evolution of Pangenomes. *Current Biology* **29**, R1094–R1103 (2019).
3. Joseph, S. J. & Read, T. D. Bacterial population genomics and infectious disease diagnostics. *Trends in Biotechnology* **28**, 611–618 (2010).
4. Ciufo, S. *et al.* Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *International Journal of Systematic and Evolutionary Microbiology* **68**, 2386–2392 (2018).
5. Deng, X., Phillippy, A. M., Li, Z., Salzberg, S. L. & Zhang, W. Probing the pan-genome of Listeria monocytogenes: new insights into intraspecific niche expansion and genomic diversification. *BMC Genomics* **11**, 500 (2010).
6. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome”. *PNAS* **102**, 13950–13955 (2005).
7. Vernikos, G., Medini, D., Riley, D. R. & Tettelin, H. Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* **23**, 148–154 (2015).
8. Udaondo, Z., Molina, L., Segura, A., Duque, E. & Ramos, J. L. Analysis of the core genome and pangenome of *P seudomonas putida*: The *Pseudomonas putida* pangenome. *Environmental Microbiology* **18**, 3268–3283 (2016).
9. Udaondo, Z., Duque, E. & Ramos, J.-L. The pangenome of the genus Clostridium. *Environmental Microbiology* **19**, 2588–2603 (2017).

10. Abram, K. Z., Jun, S.-R. & Udaondo, Z. Pseudomonas aeruginosa Pangenome: Core and Accessory Genes of a Highly Resourceful Opportunistic Pathogen. *Adv Exp Med Biol* **1386**, 3–28 (2022).
11. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 1–8 (2018).
12. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology* **17**, 132 (2016).
13. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
14. Manchanda, N. *et al.* GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations. *BMC Genomics* **21**, 193 (2020).
15. Abram, K. *et al.* Mash-based analyses of Escherichia coli genomes reveal 14 distinct phylogroups. *Commun Biol* **4**, 1–12 (2021).
16. Templeton, A. R. Scope and Basic Premises of Population Genetics. in *Population Genetics and Microevolutionary Theory* 1–18 (John Wiley & Sons, Ltd, 2006). doi:https://doi.org/10.1002/0470047356.ch1.
17. González-Ittig, R. E., Carletto-Körber, F. P. M., Vera, N. S., Jiménez, M. G. & Cornejo, L. S. Population genetic structure and demographic history of Streptococcus mutans (Bacteria: Streptococcaceae). *Biological Journal of the Linnean Society* **120**, 705–716 (2017).
18. Breitwieser, F. P., Pertea, M., Zimin, A. V. & Salzberg, S. L. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res.* **29**, 954–960 (2019).

19. Goig, G. A., Blanco, S., Garcia-Basteiro, A. L. & Comas, I. Contaminant DNA in bacterial sequencing experiments is a major source of false genetic variability. *BMC Biology* **18**, 24 (2020).
20. Weiss, S. *et al.* Tracking down the sources of experimental contamination in microbiome studies. *Genome Biology* **15**, 564 (2014).
21. Konstantinidis, K. T., Ramette, A. & Tiedje, J. M. The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**, 1929–1940 (2006).
22. Goris, J. *et al.* DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* **57**, 81–91 (2007).
23. Richter, M. & Rosselló-Móra, R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A* **106**, 19126–19131 (2009).
24. Konstantinidis, K. T. & Tiedje, J. M. Towards a Genome-Based Taxonomy for Prokaryotes. *Journal of Bacteriology* **187**, 6258–6264 (2005).
25. Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences* **102**, 2567–2572 (2005).
26. Kato-Maeda, M. *et al.* Comparing genomes within the species Mycobacterium tuberculosis. *Genome Res* **11**, 547–554 (2001).
27. Udaondo, Z., Abram, K., Kothari, A. & Jun, S.-R. Top-Down Genomic Surveillance Approach to Investigate the Genomic Epidemiology and Antibiotic Resistance Patterns of Enterococcus faecium Detected in Cancer Patients in Arkansas. 2022.11.23.22282607 Preprint at https://doi.org/10.1101/2022.11.23.22282607 (2022).
28. Belloso Daza, M. V., Cortimiglia, C., Bassi, D. & Cocconcelli, P. S. 2022. Genome-based studies indicate that the Enterococcus faecium Clade B strains belong to Enterococcus lactis species and

lack of the hospital infection associated markers. *International Journal of Systematic and Evolutionary Microbiology* **71**, 004948.

29. Belloso Daza, M. V. *et al.* Genomic Insights of Enterococcus faecium UC7251, a Multi-Drug Resistant Strain From Ready-to-Eat Food, Highlight the Risk of Antimicrobial Resistance in the Food Chain. *Frontiers in Microbiology* **13**, (2022).
30. Shridhar, P. B. *et al.* Whole genome sequence analyses-based assessment of virulence potential and antimicrobial susceptibilities and resistance of Enterococcus faecium strains isolated from commercial swine and cattle probiotic products. *Journal of Animal Science* **100**, skac030 (2022).
31. Matsumoto, Y. *et al*. Comprehensive subspecies identification of 175 nontuberculous mycobacteria species based on 7547 genomic profiles. *Emerging Microbes & Infections* **8**, 1043–1053 (2019).