Building comprehensive MS-friendly databases for proteomic analysis of bacterial species of unknown genetic background

Machado, K.C.T.[1], Fortuin, S.[3], Tomazella, G.G.[4,5], Fonseca, A.F.[1], Warren, R.[3], Wiker, H.G.[4], de Souza S.J.[1,2] and de Souza, G.A.[1,*]

[1]Bioinformatics Multidisciplinary Environment, UFRN, Brazil.
[2]The Brain Institute, UFRN, Brazil.
[3]Stellenbosch University, Cape Town, South Africa.
[4]The Gade Institute, University of Bergen, Norway.
[5]Institute for Bioinformatics and Biotechnology, Natal, Brazil.
*Corresponding author: gadsouza@gmail.com

ABSTRACT

In proteogenomic analysis of prokaryotes of unknown genetic background, merging different gene annotation from genomic data of all strains for a given species is a valuable strategy to help the characterization of the sample. It is also relevant for identification of important amino acid polymorphisms and validation of coding regions. We designed a bioinformatic tool which constructs fasta databases including conserved and unique sequences of strains of a given species. By using mass spectrometry data collected from 8 clinical strains from *Mycobacterium tuberculosis*, we checked protein identification performance of three sequence databases, one including all proteins from 65 sequenced strains; one built using our tool using the same 65 strains; and one using the assembly of model strain H37Rv. Finally, we built databases for 10 species with complete sequenced genomes and monitored features which are relevant for probabilistic-based protein identification by proteomics. We observed that as expected increase in database complexity correlates with pangenomic complexity. However *Mycobacterium tuberculosis* and *Bortedella pertusis* generated very complex databases even having low pangenomic complexity or no pangenome at all respectively. This indicate that differences in gene annotation is higher than average between strains of those species.

INTRODUCTION

While genomic annotation approaches undertook considerable advances in the past years (1-3), correct gene prediction is still challenging even for prokaryotes. Particularly the correct assignment of coding open reading frames (ORFs), the presence and classification of pseudogenes, and very often observed is also differences in translational starting site (TSS) choices (4-6). Using amino acid sequence information, usually collected through bottom-up mass spectrometry (MS)-based proteomic approaches, has become an accurate manner to further validate and confirm proposed gene models, the so called Proteogenomics approach (for recent reviews see (7, 8)).

In proteogenomics, researchers will often construct customized protein sequence databases which will be inspected against peptide sequence data collected by MS (9). Database customization is often achieved using two different strategies: if the genomic information of the prokaryote under study is known, a 6-frame translation is routinely employed (10, 11); if genomic information is unknown, an alternative is to construct a database using *ab initio* gene predictions from related strains of the same species, taking into consideration variations caused by SNPs, indels, divergent TSS choice, etc (12, 13). These approaches are not mutually exclusive, as gene annotation from related strains can be used to further optimize 6-frame translation approaches (14).

However, peptide identification in MS-based proteomics is often performed through probabilistic calculations between the observed peptide fragmentation pattern and theoretical data from a sequence database. Furthermore, database size is one of many parameters which can alter the search space and consequently the probabilistic calculations performed during peptide identification (15). Therefore, building databases using either 6-frame translations or sequence merging approaches will result in larger databases which, at some point, might become detrimental for the proteomic analysis (16). Particularly for approaches merging annotation from related strains, database size becomes a more evident issue as the amount of complete sequenced genomes available increases. Such issue might as well contribute differently depending on the species under study, for

example, we observed in previous work that building such databases for *Mycobacterium tuberculosis* (Mtb) complex is not incremented heavily as new strains are considered (17), since genetic similarity between strains is high and suggest clonal expansion, and there is no reports of examples of horizontal gene transfer (18, 19). But this will surely not be the case for species where a complex pangenome is observed, i.e, the genome of all strains of a given species contains only a set of genes from a larger pool of accessory genes for that species (20, 21). In addition, this wil be relevant as proteomic analysis will eventually become more routinely used to investigate bacterial communities (in similar manner as metagenomic analysis). Proper identification of peptides originated from more than one species will require merging of more databases, as already illustrated (22). It is then critical to address the impact of such approach in the size of customized databases used in MS-based proteomics, as well as in the performance for peptide identification.

We developed a computational tool which can merge genomic information from different strains, and create fasta databases following “MS-friendly” rules (13, 23). The tool defines unique annotated proteins as well as protein homologues across strains, adding all unique sequence information in the final database on a non-redundant manner. To test the approach, we performed a proteomic identification of 8 clinical strains of Mtb, using three databases: the routinely used laboratory strain H37Rv, and a concatenated database of all 65 available strains with or without processing by our tool. Our data shows an increase number of ORFs identifications when compared to single database search against the gold-standard strain, and similar performance when compared to a concatenated database without processing. Finally, to quantify database parameters, we then generated databases for ten different bacterial species with complete sequenced genomes, using increasing number of strains per round, and quantified the number of entries and unique peptide sequences for each species. As expected, species with large pangenomes such as *Escherichia coli* showed the larger database size increase per number of strains used, and database size in general correlated with pangenome size. The exceptions were Mtb, which has the 2nd smaller pangenome

size, yet it had the 4th in database size increment rate, and *Bortedella pertussis*, which has no pangenome described and it was the 7th in database increase, with a increment rate similar than *Staphylococcus aereus* which has a core genome corresponding to 34% of the total genetic pool of the species.

METHODS

Species selection

Ten bacterial species were selected for database construction: *Acinetobacter baumannii* (89)*, Bordetella pertusis* (348)*, Campylobacter jejuni* (114)*, Chlamydia trachomatis* (82)*, Escherichia coli* (425)*, Listeria monocytogenes* (148)*, Mycobacterium tuberculosis* (65)*, Pseudomonas aeruginosa* (105) and *Staphylococcus aereus* (194). *Burkholderia mallei* and *Burkholderia pseudomallei* (109) are considered genetically very similar (24, 25) and were analyzed together. Number in parenthesis represents strains with complete genome sequenced according to Genbank (26). Protein sequences were obtained using the assembly file available at ftp://ftp.ncbi.nih.gov/genomes/genbank/bacteria/).

Script design and availability

The tool was designed in PERL (version 5.24.1) and is present as two modules: rand.pl) provides the sequence alignment and creates the outputs with unique entries and homologues; create_bd.pl process the homologues output, and create the final database and the log file. In house BLAST installation (version 2.7.1 or later) is required. All script outputs are saved as txt files in the data folder. Scripts are made available at https://github.com/karlactm/Proteogenomics.

Data processing

Basically, to assign gene homology, our script initially gathers the protein sequence data from all strains and then performs pair wise alignments using BLASTP (27) through a Bidirectional Best Hit (BHH) method (28). The script performs in a manner that the strain with most number of protein entries is initially selected as the query dataset and consequently aligned to the remaining strains

sequence datasets (subjects). Two sequences from different strains will be defined as homologues if: i) they are the best hit possible for all alignments performed; ii) sequence identity is equal or higher than 50%; iii) sequence similarity is equal or higher than 70%; and iv) sequence coverage is equal or higher than 70%. These have to be correlated on both directions (BHH) of the alignment.

When homologues are defined between query and subject strains, our script will proceed by indexing homologue IDs for later characterizations of polymorphisms. It will then define “unique” protein sequences from the query dataset (i.e., a sequence without a defined homologue in any of the subject strains), and add it to the partial version of the final database output (see Figure 1). In parallel, all entries from the subject strains which were properly aligned to a query sequence will be removed from the respective strains datasets. A new round of alignment will be performed after excluding the query strain and selecting one of the previous subject datasets as a new query strain. This will be performed successively until all strains are used as query. At this point, the script will had then defined all “unique” protein entries from all datasets which did not aligned to any other annotated sequence under the selected parameters.

Regarding all indexed homologues, our script then defines the longest version of the protein as reference (not necessarily from same strain used as query for the BHH step above). The reference sequence is copied integrally into the partial version of the final output. The script performs an *in silico* trypsin digest of the protein sequences, and compare amino acid composition of all generated tryptic peptides. Peptides shorter than 7 or longer than 35 amino acids are excluded. Whenever different tryptic peptide sequences are observed in the non-reference dataset (due to differences of TSS choices, SAAs or indels which result in frame shifts), they are added into the final output under an artificially created protein entry (13). Amino acid changes result from poorly annotated sequences, where X or U are present in the sequence as an amino acid, are not considered. Finally, a log file is created, reporting and classifying all differences observed, describing the amino acid changes and also in which strains those were observed.

Database analysis

To investigate how each dataset contributes to the final database size, for each species we constructed databases using 5, 10, 15, 30 and 65 randomly selected strains. This was performed a total of three times to decrease the chances that a randomly selected strain with very unique features might interfere with the final result. Each database had its MS search space size measured regarding number of entries and number of unique tryptic peptide sequences.

Mycobacterial Cell Culturing

Stock cultures of Mtb strains TB179, TB861, TB1430, TB1593, TB1659, TB1841, TB1945 and TB2666 were inoculated into mycobacterial growth indicator tubes and incubated until positive growth was detected using the Bactec 460 TB system (BD Biosciences). Approximately 0.2 mL was inoculated onto Löwenstein-Jensen medium and incubated over 6 weeks with weekly aeration until colony formation. Colonies were transferred into 20 mL of supplemented 7H9 Middlebrook medium (BD Biosciences) containing 0.2% (v/v) glycerol (Merck Laboratories), 0.1% Tween 80 (Merck Laboratories), and 10% dextrose, catalase. Once the culture reached an A600 of 0.9, one mL was inoculated into 80 mL of supplemented 7H9 Middlebrook medium and incubated until an A600 between 0.6 and 0.7 was reached. All steps were performed at 37 °C until Mtb cells in mid-log growth phase were used for whole cell lysate protein extraction.

Preparation of Crude Mycobacterial Extracts

Mycobacterial cells were collected by centrifugation (10 min at 2500 x g) at 4°C and resuspended in 1 mL of cold lysis buffer containing 10 mMTris-HCl, pH 7.4 (Merck Laboratories), 0.1% Tween 80 (Sigma-Aldrich), one tablet per 25 mL Complete protease inhibitor mixture (Roche Applied Science), and one tablet per 10 mL phosphatase inhibitor mixture (Roche Applied Science). Cells were transferred into 2 mL cryogenic tubes with O-rings, and the pellet was collected after centrifugation (5 min at 6000 x g) at 21 °C. An equal volume of 0.1 mm glass beads (Biospec Products Inc., Bartlesville, OK) was added to the pelleted cells. In

addition, 300 μL of cold lysis buffer including 10 μL of 2 units/ml RNase-free DNase I (New England Biolabs) was added, and the cell walls were lysed mechanically by bead beating for 20 s in aRibolyser (Bio101 Savant, Vista, CA) at a speed of 6.4. Thereafter the cells were cooled on ice for 1 min. The lysis procedure was repeated three times. The lysate was clarified by centrifugation (10,000 x g for5 min) at 21 °C, and the supernatant containing the whole cell lysate proteins was retained. Thereafter the lysate was filter-sterilized through a 0.22 μm-pore Acrodisc 25 mm PF syringe sterile filter (PallLife Sciences, Pall Corp., Ann Arbor, MI), quantified using the Coomassie Plus Assay Kit (Pierce), and stored at -80 °C until further analysis.

Gel Electrophoresis and In-gel Trypsin Digestion

Each whole cell lysates (60 μg) were mixed with electrophoretic sample buffer (NuPAGE kit, Invitrogen, CA) containing 100 mM DTT, and heated for 5 min at 95 °C prior to the electrophoretic run. Proteins were separated in triplicate using a NuPage 4–12% Bis-Tris Gel in 2-N-morpholine ethane sulfonic acid (MES) buffer (Invitrogen) at 200 V for approximately 40 min. Proteins were visualized using a Colloidal Coomassie Novex kit (Invitrogen). After staining, each triplicate was cut into 15 fractions, sliced into smaller pieces and submitted to in gel digestion with trypsin, as previously described (29). Briefly, the proteins in the gel pieces were reduced using 10 mM DTT for 1 h at 56 °C and alkylated with 55 mM iodoacetamide for 45 min at room temperature. The reduced and alkylated proteins were digested with 0.125 µg of trypsin (Sequence grade modified, Promega, WI, USA) for 16 h at 37 °C in 50 mM NH4HCO3, pH 8.0. The reaction was stopped by acidification with 2% trifluoracetic acid (Fluka, Buchs,Germany). The resulting peptide mixture was eluted from the gel slices, and further desalted using RP-C18 STAGE tips (30). The peptide mixture was dissolved in 0.1% formic acid 5% acetonitrile for nano-HPLC-MS analysis.

LC-MS/MS analysis

Peptides were separated by reversed-phase chromatography in an Acclaim PepMap 100 column (C18, 3 μm, 100 Å) (Dionex) capillary of 12 cm bed length and 100 μm inner diameter self-packed with ReproSilPur C18-AQ (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany), using a Dionex Ultimate 3000 nano-LC system (Dionex, Sunnyvale, CA) connected to a linear quadrupole ion trap-Orbitrap (LTQ-Orbitrap) mass spectrometer (Thermo Electron, Bremen, Germany) equipped with a nanoelectrosprayion source. Peptides were onto loaded to the column with a flow rate of 0.3 mL/min of 7–40% solvent B in 87 min and then 40–80% solvent B in 8 min. Solvent A was aqueous 2% ACN in 0.1% formic acid, and solvent B was aqueous 90% ACN in 0.1% formic acid.

The mass spectrometer was operated in the data-dependent mode to automatically switch between Orbitrap-MS and LTQ-MS/MS acquisition. Survey full-scan MS spectra (from m/z 300 to 2000) were acquired in the Orbitrap with resolution of R= 60,000 at m/z 400 (after accumulation to a target of 1,000,000 charges in the LTQ). The method used allowed sequential isolation of the most intense ions (up to six, depending on signal intensity) for fragmentation on the linear ion trap using collisionally induced dissociation at a target value of 100,000 charges. For accurate mass measurements, the lock mass option was enabled in MS mode, and the polydimethylcyclosiloxane ions generated in the electrospray process from ambient air were used for internal recalibration during the analysis (31). Target ions already selected for MS/MS were dynamically excluded for 60 s. General mass spectrometry conditions were as follows: electrospray voltage, 1.5 kV; no sheath or auxiliary gas flow. Ion selection threshold was 500 counts for MS/MS, an activation Q-value of 0.25 and activation time of 30 ms were also applied for MS/MS.

Data analysis

Raw MS files of 15 fractions in triplicates from 8 cell lineages were analyzed by MaxQuant version 1.5.2.8. MS/MS spectra were searched by the Andromeda search engine (32) against either the H37Rv database, a database containing all sequences from 65 strains without any processing, or the database generated by

our tool using the same 65 strains. MaxQuant analysis included an initial search with a precursor mass tolerance of 20 ppm which were used for mass recalibration (33). In the main Andromeda search precursor mass and fragment mass were searched with initial mass tolerance of 6 ppm and 0.5 Da, respectively. The search included variable modifications of oxidation (Met), N-terminal acetylation (protein). Carbamidomethyl cysteine was added as a fixed modification. Minimal peptide length was set to 7 amino acids and a maximum of two miscleavages was allowed. The false discovery rate (FDR) was set to 0.01 for peptide and protein identifications. In the case of identified peptides that are shared between two proteins, these are combined and reported as one protein group. Protein and peptide datasets were filtered to eliminate the identifications from the reverse database, and common contaminants.

RESULTS

Our tool was designed to use gene assembly data from Genbank from selected bacterial strain genomes to construct protein sequence databases which considered all possible sequence variations (SAP, TSS choice, for example) but on a non-redundant manner (i.e., without extensive entry usage for very similar sequences). On average, the assemblies used on this work had at least 910 genes (*Chlamydia trachomatis*, average value to all strains) to 6092 genes (*Pseudomonas aeruginosa*). The analyzed species were selected not solely due to higher number of strains with complete genome sequences available. They were also selected considering the complexity of their pangenomes. Nine of the selected species have pangenomes which can be very diverse (for example, *Escherichia coli*, 2,459 core genes to approximately 26,000 pangenomic genes) or just a fraction of the core genome (such as *Chlamydia trachomatis* and Mtb, which have only one accessory gene to every five core genes) (20, 21, 34).

Figure 1 illustrates the tool workflow: briefly, a query strain is selected, and a pair wise comparison is performed with the remaining subject strains using a bidirectional approach. Unique entries from query strain (i.e., sequences which no other protein entries in subject strains had the minimal identity, similarity and

coverage percentage thresholds) are saved in the final database. Protein sequences across strains that share the required levels of identity and similarity are further processed. The longest version is used as a reference (regardless of which strain is originated from) and saved into the final database. Tryptic peptides between reference and non-reference sequences are compared. When a tryptic peptide from a non-reference sequence cannot be matched to any of the reference sequence (due to a different TSS choice or a SNP for example), a new protein entry is created in the final database. This new entry will contain all non-reference tryptic peptides that do not match the reference sequence.

Once this is finished, the query strain is excluded from the strain list, and all proteins from subject strains which were matched to a protein in the query strain (and therefore already processed to define polymorphisms) are excluded from the subject strains datasets. New pair wise comparisons will be performed using a new query strain. This process will be repeated until no strains are left to be used as query, or until no protein entries are left in yet-to-be selected query strains. A log file is also created, reporting all entries that were matched and classifying all tryptic peptides which were added to the final database, showing the type of event (TSS choice, SNP etc) and in case of SNP, the resulting amino acid substitution.

Since our group possessed unpublished MS data for clinical strains of Mtb, we optimized our tool using this data and databases created from 65 strains with complete genome for this species. The processed Mtb database created by the tool contained 15,996 protein entries, from those 4,506 are entries containing polymorphic peptides. The remaining 11,490 entries contain reference sequences from homologue comparisons and sequences which were unique to each strain. As the median value for a single Mtb strain is 4,056 annotated proteins, our data suggest that each of the 64 strains processed after the query strain incremented the database with approximately 114 unique sequences. Considering that Mtb genome sequences are highly similar between strains and its pangenome is relatively small, most of this increment should be a result of differences in genomic annotation. In comparison, a merely concatenated Mtb database using all

sequences from the 65 strains would create a database with approximately 265,000 protein entries.

Figure 2 illustrates a typical homologue comparison leading to the detection of polymorphic peptides which will be added on an artificial entry. The hypothetical protein AEM98552 (Rv0104 on the Tuberculist annotation (35)) was set as reference, and 9 other different variants were observed. Those variants are result of the different combinations of: two additional TSS choices at position 9 and 118; three SAPs at positions 402 (Tyr → His), 477 (Gly → Arg) and 493 (Leu → Val); a not-classified genetic modification that leads to alternate amino acid sequence after position 475 (yellow box, named as frameshift/indel); and finally, two polymorphisms leading to premature stop codons and generating different c-terminal tryptic peptides ending at position 378 and 426. For example, the variant from strain CAS/NITR204 (AGL25567) contains two SAPs, while the variant from strain CCDC5180 (AEJ48929) has a different TSS choice at position 9 of reference, and a stop codon happens prematurely, ending the protein at position 378. Unique tryptic peptides characterizing all such polymorphisms are added in the database as an additional entry.

Our MS dataset was then challenged for peptide/protein identification using three databases: one was the assembly of the strain H37Rv, which is the laboratory strain more extensively annotated (35) and validated by proteogenomics (4, 36), and therefore is the most used database for strains of unknown background. The other two databases are the concatenated entries of all 65 Mtb strains with complete genome sequenced, with or without being processed by our tool. Those databases from now on will be denominated DB1 and DB2, respectively, for simplicity. By using solely the reference strain H37Rv, 2,801 proteins were identified in the clinical strains of Mtb. Using DB1 in the same dataset identified 3,139 proteins, an increment of 338 proteins (12%). From those, 143 proteins were annotated in multiple strains but H37Rv, while 195 where annotated in a single strain.

This data clearly shows the importance of creating multiple annotation strategies for proper characterization of clinical strain of unknown genomic

background. In addition to increasing genomic coverage at protein identification level, such databases also allow the characterization of polymorphisms present across all strains of a gives species. In the Mtb dataset, we identified 1,131 peptides characterizing 502 TSS choices, 376 SAPs and 253 peptides containing “uncharacterized” sequences, such as the region in yellow boxes in Figure 2. We classified those as uncharacterized because, considering our tool does not compare genomic information of the strains yet, we can only assume that such changes in primary structure result from indels which could be short and lead to a frameshift, or are large enough to create a new peptide sequence and stop codon. All polymorphic peptides and their identification features are shown in Supplementary Table I.

Multiple TSS polymorphisms for the same protein were also observed even for the same strains. Figure 3 shows the position of three peptides, all predicted as possible N-terminal peptides of the protein lipase lipV. MIIDLHVQR (Score 93.4, PEP 0.0019, spectrum in Figure S1) represents the further upstream prediction, LTIHGVTEHGR (Score 85.9, PEP 0.0005, Figure S2) starts at position 19 of reference protein and HGVTEHGR (Score 113.2, PEP 0.0077, Figure S3) starts at position 22 of reference. The longer form is the most abundant in practically all samples, except S1430 where only LTIHGVTEHGR was identified and S2666 where the N-terminal HGVTEHGR is slightly overrepresented than MIIDLHVQR. Finally, the DB1 results also included 5 polymorphic peptides from the artificial entries which could not be found in the log file created in parallel to the database. Closer inspection of these peptides revealed that they all contain missing cleavage sites. For example, the identification of peptide GEGAAVVVIKEGSIGGGTVFVYSGR from gene PpsC includes two polymorphic peptides which were concatenated in the database, GEGAAVVVIK which correspond to positions 274-283 in the reference protein, and EGSIGGGTVFVYSGR which correspond to positions 564-578. Clearly, the artificial entries in this case contains two concatenated polymorphic peptides which are not adjacent in the reference protein. The fact that we accept miscleavages as a parameter during the search allows a clearly false-positive identification for artificial

entries. The spectra and identification score for this concatenated peptide is given in Supplementary Figure S4, and as expected, the identified spectra has poor score and sequence coverage. Supplementary Table II lists all 5 peptides.

When comparing the results from DB1 and DB2 searches, the number of peptides identified was very similar: the DB1 search resulted in the identification of 33,501 tryptic peptides, while DB2 resulted in 34,102 peptides identified. Therefore, our processed database identified 601 peptides less than DB2, a mere difference of 1.75%. This difference is very close to the 1% variation expected to from the probabilistic nature of the identification approaches used by MaxQuant (selected as a 1% allowed false-discovery rate), which means both performances should be considered very similar. In addition, a close inspection of those peptides indicated that a good number of the peptides identified in the DB2 search were peptides which should had been added on DB1 as non-reference peptides, but they weren't due to the parameters selected in our processing. For example, we did not consider possible miscleavages when comparing reference/non-reference homologues, the *in silico* trypsin digest generated peptides at 100% enzyme efficiency. Some of the identification was also non-reference peptides with more than 35 amino acids, which we excluded in our processing.

Finally, we further created databases from another 9 species, in order to measure if the rate of increase in the database size could be impacted by unique features of each genome, such as size of a pangenome. For a fair comparison among species, we built databases using a maximum of 65 strains, since Mtb had only 65 strains with complete sequenced genomes available. To assist the visualization of how the database size is incremented as more strains are used, we also created databases using 5, 10, 20 and 30 randomly selected strains of each species. And to avoid outlier behavior due to a very unique strain annotation being randomly selected, we created every database three times in total. We then counted total number of protein entries and tryptic peptides in each database.

Supplementary Table 3 shows all values measured for all tests for: number of proteins (Supplementary Figure 5), number of annotated proteins (excluding artificial entries created to accommodate polymorphisms) (Figure 4 and Table I),

number of peptides with or without miscleavages allowed (Supplementary Figure 6). Figure 4 shows graphically the increase in number of annotated proteins in all species, and Table I describe number of annotated proteins in the database using 65 strains, rate of database size increase, number of annotated proteins added per strain on average, and size of pangenome for each species. As expected, the increase rate in the number of annotated proteins in the final database correlated with the size of the gene pool in the pangenome. *E. coli* which has the largest genetic pool described had indeed the larger database increase rate of all species, at 4.6x compared to the average number of genes per strain. The three exceptions for this were: *B. pseudomallei/mallei* with the 2nd increase rate (3.53x) even though the pangenome size of *B. pseudomallei* is the 4th most complex (no data is available regarding *B. mallei* pangenome); Mtb, which has the 2nd less complex pangenome size, representing only 21% of total genetic pool, yet it has 4th database increase rate (2.83x), surpassing *P. aeruginosa* which has a larger genome size and a bigger genetic pool; and *B. pertussis*, with no pangenome described so far and yet with a database increase rate (2.46x) which surpassed three species with characterized pangenomes.

DISCUSSION

Constructing customized databases is a key step in proteogenomics (9), or for the characterization of samples with unknown genetic background such as bacterial clinical strains (17, 29). Such databases allow the validation of coding regions and confirmation of sequence polymorphisms normally omitted from reference databases. This is a critical characterization, considering that even genomes with two decades of considerable investigation still have gene annotation issues (37). This is true even for prokaryotes (4, 5), with their simpler genomic structure and higher coding density.

We had previously developed a tool named MSMSpdbb which creates MS-friendly databases for prokaryote species using strain annotations (12), however at the time we used database rules which are nowadays incompatible to peptide search engines. Further motivation to create an up-to-date tool came with the

publication of the tool iPtgxDB from Omasitis and colleagues (13), which designs similar MS-friendly databases using either 6-frame translation or different annotations datasets for a single genome sequence. Therefore, iPtgxDB designs databases for single strains, while like MSMSpdbb we focused in building databases for a given species. While iPtgxDB is better designed for proper proteogenomics research, we believe the strength in our approach is the analysis of samples with unknown genetic background, which in the future includes the characterization of complex microbiome samples.

Another critical difference in how we designed the database is related to the polymorphic peptides: all peptides were concatenated into a single artificial amino acid sequence, which is added into the database under a single protein entry. The alternative would be to create a single protein entry to each polymorphic peptide being considered, which would increase the number of fasta headers to be processed by the search engine. While, to our knowledge, it has not been tested if there is any clear advantage to use any of the mentioned structures, it is possible that by concatenating polymorphic peptides and keeping number of entries to a minimum will result is lesser computational time for the analysis, since the search engine will have less entries to read/write during any I/O process. While such differences might be negligible for databases being created with the current level of genomic information, this might become more relevant as more genomic data is created or as metaproteomic databases are attempted. The downside of concatenating sequences is the identification of incorrect, miscleaved peptides containing two fully tryptic peptides which are not in tandem in the protein primary structure. However, with the data we have collected, we observed only a small number of such events compared to the total number of polymorphic peptides detected (0.44%) in clinical samples of *Mtb*.

As such set of test samples were analyzed using three different databases, some observations became clear: first, performance (i.e., number of protein and peptide hits) was superior using DB1 or DB2 compared to using a single annotation from a model strain by at least 12% in number of proteins identified. In addition to identification coverage, the observations of multiple polymorphisms also

justify the use of multiple strain databases. Surprisingly we even identified multiple N-terminal predictions for the same protein, in some cases observed in the same strain. While normally the identification of a protein N-terminal peptide in proteogenomics is used to validate and confirm the TSS of the protein, our data suggest that excluding the remaining TSS choices from the database might be undesirable, considering they might be later identified when additional strains are analyzed by MS. Overall, DB1 and DB2 performance was very similar, and differences were negligible even though DB2 did perform slightly better. However, while it was not our concern to benchmark computational performance for this work, the searches for DB1 took half of the computational time than DB2 searches. As more genomic information is used, or as more complex the databases become to incorporate multiple species for metaproteomics, it is not absurd to predict that processing entries as we did for DB1 will be clearly advantageous in order to save computational power.

When databases for ten species were created, surprisingly Mtb and *B. pertussis* databases size increased at rates higher than expected based on the low complexity of their core and pangenomic genes. Previous data from Mtb suggested that database size increment was not heavy, when at the time only five Mtb strains and three strains of *M. bovis* were considered (17). Our data here however shows that Mtb database size increment was one of the most prominent, arguably the most relevant if its small pangenome is considered. This suggests that as additional strains were sequenced, variations in gene annotation performed were more acute then observed for the other species that were investigated here. Something similar could be assumed about *B. pertusis* databases, which has no pangenome and database increase was higher than three species with characterized pangenomes. However, *C. trachomatis, L. monocytogenes* and *C. jejuni* have smaller genomes than compared to *B. pertussis*, all also having higher density of coding regions (38-40). More nucleotide regions in the genome marked as non-coding will offer more opportunity for different strategies to generate conflicting gene annotation data, which could explain *B. pertussis* database size increase in our analysis. The database size increase for *Burkholderia pseudomallei*

and *Burkholderia mallei* was not taken into consideration by us on this analysis, because while we decided to merge both species as one based on genotyping data (25), large scale rearrangements in *B. mallei* (41) might be interfering with our analysis.

Finally, another motivation to this work has been that, up to date, little has been done with relative good protein identification coverage regarding bacterial metaproteomics, with one rare recent exception (22). In general, the databases employed in those researches are built through simply concatenating all annotated proteins for the species under investigation. Even though we had no access to the database used in the reference above, it is simple to estimate that the database used probably had close to one million entries, taking into account an average of 2,000 proteins for all 462 genomes described in the Human Oral Microbiome repository (42). As spectra identification happens at peptide level in bottom-up proteomics and most peptides in such databases will have several identical copies in many homologue proteins, such concatenated databases are largely composed by peptides which aren't unique enough to characterize which species are in the sample. This will result in very large databases with redundant information that consequently demand a large computational capacity during the analysis. The strategy described here to streamline annotations from different strains of the same bacterial species can be also used for metaproteomic databases, aiming to reduce their redundancy and optimize species identification using unique peptide sequences.

REFERENCES

1. Bocs, S., Cruveiller, S., Vallenet, D., Nuel, G., and Medigue, C. (2003) AMIGene: Annotation of MIcrobial Genes. *Nucleic acids research* 31, 3723-3726

2. Davidsen, T., Beck, E., Ganapathy, A., Montgomery, R., Zafar, N., Yang, Q., Madupu, R., Goetz, P., Galinsky, K., White, O., and Sutton, G. (2010) The comprehensive microbial resource. *Nucleic acids research* 38, D340-345

3. Vallenet, D., Belda, E., Calteau, A., Cruveiller, S., Engelen, S., Lajus, A., Le Fevre, F., Longin, C., Mornico, D., Roche, D., Rouy, Z., Salvignol, G., Scarpelli, C., Thil Smith, A. A., Weiman, M., and Medigue, C. (2013) MicroScope--an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic acids research* 41, D636-647

4. de Souza, G. A., Malen, H., Softeland, T., Saelensminde, G., Prasad, S., Jonassen, I., and Wiker, H. G. (2008) High accuracy mass spectrometry analysis as a tool to verify and improve gene annotation using Mycobacterium tuberculosis as an example. *BMC genomics* 9, 316
5. de Souza, G. A., Softeland, T., Koehler, C. J., Thiede, B., and Wiker, H. G. (2009) Validating divergent ORF annotation of the Mycobacterium leprae genome through a full translation data set and peptide identification by tandem mass spectrometry. *Proteomics* 9, 3233-3243
6. Cuklina, J., Hahn, J., Imakaev, M., Omasits, U., Forstner, K. U., Ljubimov, N., Goebel, M., Pessi, G., Fischer, H. M., Ahrens, C. H., Gelfand, M. S., and Evguenieva-Hackenberg, E. (2016) Genome-wide transcription start site mapping of Bradyrhizobium japonicum grown free-living or in symbiosis - a rich resource to identify new transcripts, proteins and to study gene regulation. *BMC genomics* 17, 302
7. Renuse, S., Chaerkady, R., and Pandey, A. (2011) Proteogenomics. *Proteomics* 11, 620-630
8. Ruggles, K. V., Krug, K., Wang, X., Clauser, K. R., Wang, J., Payne, S. H., Fenyo, D., Zhang, B., and Mani, D. R. (2017) Methods, Tools and Current Perspectives in Proteogenomics. *Molecular & cellular proteomics : MCP* 16, 959-981
9. Nesvizhskii, A. I. (2014) Proteogenomics: concepts, applications and computational strategies. *Nature methods* 11, 1114-1125
10. (!!! INVALID CITATION !!!)
11. Baerenfaller, K., Grossmann, J., Grobei, M. A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., Zimmermann, P., Grossniklaus, U., Gruissem, W., and Baginsky, S. (2008) Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science* 320, 938-941
12. de Souza, G. A., Arntzen, M. O., and Wiker, H. G. (2010) MSMSpdbb: providing protein databases of closely related organisms to improve proteomic characterization of prokaryotic microbes. *Bioinformatics* 26, 698-699
13. Omasits, U., Varadarajan, A. R., Schmid, M., Goetze, S., Melidis, D., Bourqui, M., Nikolayeva, O., Quebatte, M., Patrignani, A., Dehio, C., Frey, J. E., Robinson, M. D., Wollscheid, B., and Ahrens, C. H. (2017) An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics. *Genome research* 27, 2083-2095
14. Castellana, N. E., Payne, S. H., Shen, Z., Stanke, M., Bafna, V., and Briggs, S. P. (2008) Discovery and revision of Arabidopsis genes by proteogenomics. *Proceedings of the National Academy of Sciences of the United States of America* 105, 21034-21038
15. Nesvizhskii, A. I. (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of proteomics* 73, 2092-2123
16. Blakeley, P., Overton, I. M., and Hubbard, S. J. (2012) Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *Journal of proteome research* 11, 5221-5234
17. de Souza, G. A., Arntzen, M. O., Fortuin, S., Schurch, A. C., Malen, H., McEvoy, C. R., van Soolingen, D., Thiede, B., Warren, R. M., and Wiker, H. G. (2011) Proteogenomic analysis of polymorphisms and gene annotation divergences in prokaryotes using a clustered mass spectrometry-friendly database. *Molecular & cellular proteomics : MCP* 10, M110 002527
18. Namouchi, A., Didelot, X., Schock, U., Gicquel, B., and Rocha, E. P. (2012) After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection. *Genome research* 22, 721-734
19. Schurch, A. C., Kremer, K., Warren, R. M., Hung, N. V., Zhao, Y., Wan, K., Boeree, M. J., Siezen, R. J., Smith, N. H., and van Soolingen, D. (2011) Mutations in the regulatory network underlie the recent clonal expansion of a dominant subclone of the Mycobacterium tuberculosis

Beijing genotype. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* 11, 587-597
20. Rasko, D. A., Rosovitz, M. J., Myers, G. S., Mongodin, E. F., Fricke, W. F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N. R., Chaudhuri, R., Henderson, I. R., Sperandio, V., and Ravel, J. (2008) The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *Journal of bacteriology* 190, 6881-6893
21. McInerney, J. O., McNally, A., and O'Connell, M. J. (2017) Why prokaryotes have pangenomes. *Nature microbiology* 2, 17040
22. Grassl, N., Kulak, N. A., Pichler, G., Geyer, P. E., Jung, J., Schubert, S., Sinitcyn, P., Cox, J., and Mann, M. (2016) Ultra-deep and quantitative saliva proteome reveals dynamics of the oral microbiome. *Genome medicine* 8, 44
23. Schandorff, S., Olsen, J. V., Bunkenborg, J., Blagoev, B., Zhang, Y., Andersen, J. S., and Mann, M. (2007) A mass spectrometry-friendly database for cSNP identification. *Nature methods* 4, 465-466
24. Song, H., Hwang, J., Yi, H., Ulrich, R. L., Yu, Y., Nierman, W. C., and Kim, H. S. (2010) The early stage of bacterial genome-reductive evolution in the host. *PLoS pathogens* 6, e1000922
25. Godoy, D., Randle, G., Simpson, A. J., Aanensen, D. M., Pitt, T. L., Kinoshita, R., and Spratt, B. G. (2003) Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, Burkholderia pseudomallei and Burkholderia mallei. *Journal of clinical microbiology* 41, 2068-2079
26. Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2013) GenBank. *Nucleic acids research* 41, D36-42
27. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *Journal of molecular biology* 215, 403-410
28. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. (1999) The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America* 96, 2896-2901
29. Tomazella, G. G., Risberg, K., Mylvaganam, H., Lindemann, P. C., Thiede, B., de Souza, G. A., and Wiker, H. G. (2012) Proteomic analysis of a multi-resistant clinical Escherichia coli isolate of unknown genomic background. *J Proteomics* 75, 1830-1837
30. Rappsilber, J., Ishihama, Y., and Mann, M. (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem* 75, 663-670
31. Olsen, J. V., de Godoy, L. M., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S., and Mann, M. (2005) Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol Cell Proteomics* 4, 2010-2021
32. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* 10, 1794-1805
33. Cox, J., Michalski, A., and Mann, M. (2011) Software lock mass by two-dimensional minimization of peptide mass errors. *J Am Soc Mass Spectrom* 22, 1373-1380
34. Ding, W., Baumdicker, F., and Neher, R. A. (2018) panX: pan-genome analysis and exploration. *Nucleic acids research* 46, e5
35. Lew, J. M., Kapopoulou, A., Jones, L. M., and Cole, S. T. (2011) TubercuList--10 years after. *Tuberculosis* 91, 1-7
36. Kelkar, D. S., Kumar, D., Kumar, P., Balakrishnan, L., Muthusamy, B., Yadav, A. K., Shrivastava, P., Marimuthu, A., Anand, S., Sundaram, H., Kingsbury, R., Harsha, H. C., Nair, B., Prasad, T. S., Chauhan, D. S., Katoch, K., Katoch, V. M., Kumar, P., Chaerkady, R., Ramachandran, S.,

Dash, D., and Pandey, A. (2011) Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry. *Molecular & cellular proteomics : MCP* 10, M111 011627
37. Abascal, F., Juan, D., Jungreis, I., Martinez, L., Rigau, M., Rodriguez, J. M., Vazquez, J., and Tress, M. L. (2018) Loose ends: almost one in five human genes still have unresolved coding status. *Nucleic acids research*
38. Thomson, N. R., Holden, M. T., Carder, C., Lennard, N., Lockey, S. J., Marsh, P., Skipp, P., O'Connor, C. D., Goodhead, I., Norbertzcak, H., Harris, B., Ormond, D., Rance, R., Quail, M. A., Parkhill, J., Stephens, R. S., and Clarke, I. N. (2008) Chlamydia trachomatis: genome sequence analysis of lymphogranuloma venereum isolates. *Genome research* 18, 161-171
39. Parkhill, J., Wren, B. W., Mungall, K., Ketley, J. M., Churcher, C., Basham, D., Chillingworth, T., Davies, R. M., Feltwell, T., Holroyd, S., Jagels, K., Karlyshev, A. V., Moule, S., Pallen, M. J., Penn, C. W., Quail, M. A., Rajandream, M. A., Rutherford, K. M., van Vliet, A. H., Whitehead, S., and Barrell, B. G. (2000) The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences. *Nature* 403, 665-668
40. Parkhill, J., Sebaihia, M., Preston, A., Murphy, L. D., Thomson, N., Harris, D. E., Holden, M. T., Churcher, C. M., Bentley, S. D., Mungall, K. L., Cerdeno-Tarraga, A. M., Temple, L., James, K., Harris, B., Quail, M. A., Achtman, M., Atkin, R., Baker, S., Basham, D., Bason, N., Cherevach, I., Chillingworth, T., Collins, M., Cronin, A., Davis, P., Doggett, J., Feltwell, T., Goble, A., Hamlin, N., Hauser, H., Holroyd, S., Jagels, K., Leather, S., Moule, S., Norberczak, H., O'Neil, S., Ormond, D., Price, C., Rabbinowitsch, E., Rutter, S., Sanders, M., Saunders, D., Seeger, K., Sharp, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Unwin, L., Whitehead, S., Barrell, B. G., and Maskell, D. J. (2003) Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. *Nature genetics* 35, 32-40
41. Losada, L., Ronning, C. M., DeShazer, D., Woods, D., Fedorova, N., Kim, H. S., Shabalina, S. A., Pearson, T. R., Brinkac, L., Tan, P., Nandi, T., Crabtree, J., Badger, J., Beckstrom-Sternberg, S., Saqib, M., Schutzer, S. E., Keim, P., and Nierman, W. C. (2010) Continuing evolution of Burkholderia mallei through genome reduction and large-scale rearrangements. *Genome biology and evolution* 2, 102-116
42. Chen, T., Yu, W. H., Izard, J., Baranova, O. V., Lakshmanan, A., and Dewhirst, F. E. (2010) The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database : the journal of biological databases and curation* 2010, baq013

FIGURE LEGENDS

Figure 1 – **Building concatenated databases from related strains.** (A) For a given set of fasta files, a Query file is selected and BLASTP is performed against the remaining files. A bidirectional approach is performed. (B) Uniquely annotated entries in the Query are defined and moved to the final fasta database (D), while entries where homologues were found in the subject files are indexed for further processing. (C) For each homologue group, a reference sequence is chosen, and polymorphic peptides are added into a customized entry, both being copied into the

final database. (E) The Query file is removed from the dataset, the remaining subject files will have entries removed (proteins which were already defined as a homologue of a Query protein), and a new fasta file is selected as Query for a new BLASTP. Such rounds will be performed until all files are analyzed.

Figure 2 – Typical homologue comparison. For protein Rv0104, in this case selected as reference, three SAPs were identified across all Mtb strains, in addition to two TSS choices, two premature stop codons, and a complete amino acid sequence change (region in yellow) most probably from an indel.

Figure 3 – Multiple TSS identified in protein lipV. Three possible TSS choices were identified in clinical Mtb strains. The most upstream prediction (MIIDLHNQR) is the most abundant variant observed in all strains, except strain S1430 which has only the variant with TSS at position 19 of the reference sequence, and strain S2666 which TSS at position 22 is predominant. Two possible TSS choice variants were also observed at high levels for strain S1593.

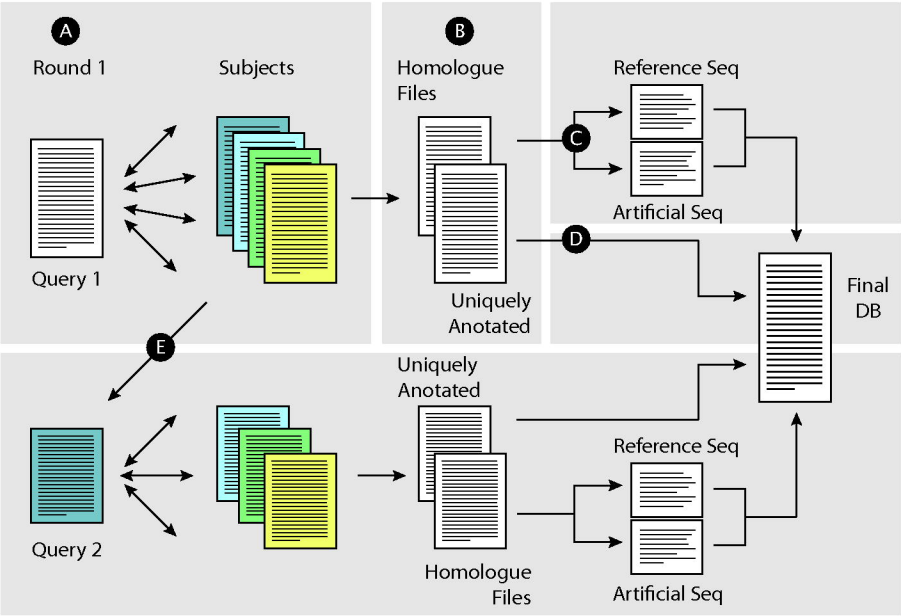
Figure 4 – Number of entries per database after redundancy removal. The graph shows data when only considering reference and uniquely annotated sequences in all strains, i.e., customized entries containing polymorphic peptides where not counted. Values in x axis show the number of strains used for database creation. All values plotted in the graph are given in Supplementary Table III.

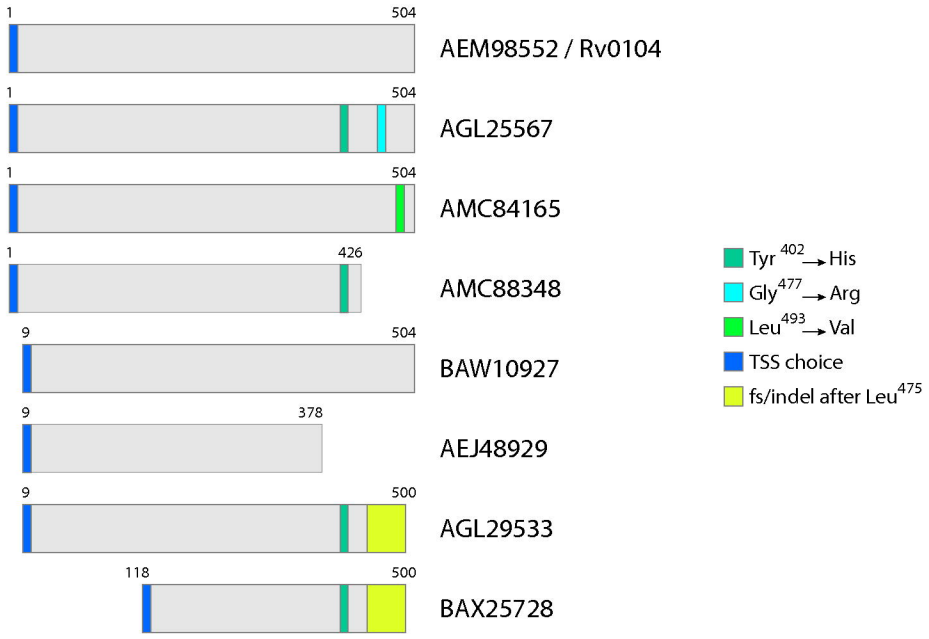
Table I - Database size increase and pangenome size in ten selected species

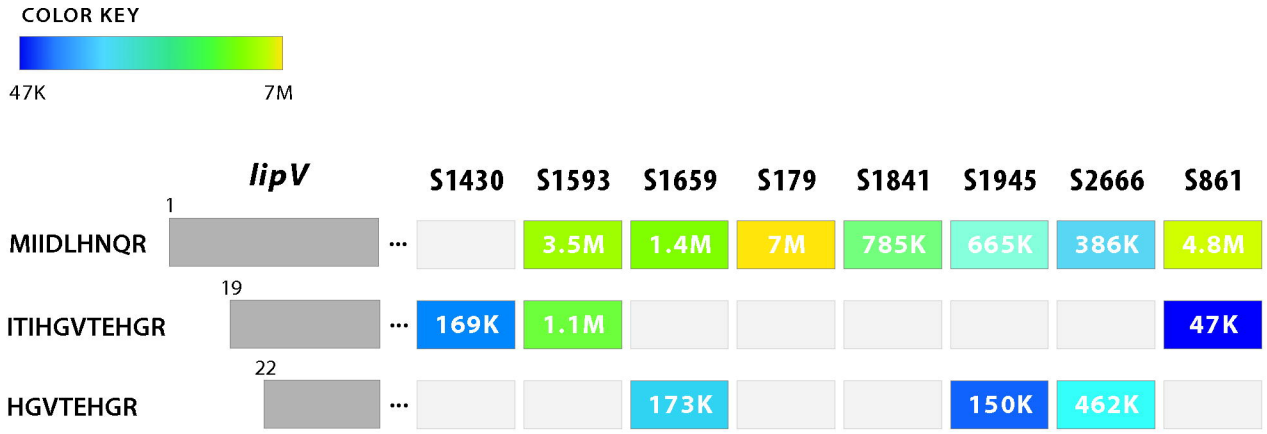
	1 strain	Final DB Ref	gain per strain	DB increase	pangenome?	Pangenome size (%/100)*
E. coli	4893	22525	271	4.60	Y	0.92
B. pseudomallei/mallei	5917	20876	230	3.53	Y	0.71**
A. baumanii	3746	12536	135	3.35	Y	0.83
M. tuberculosis	4056	11490	114	2.83	Y	0.21
P. aeruginosa	6092	15814	150	2.60	Y	0.79
S. aureus	2697	6735	62	2.50	Y	0.66
B. pertussis	3619	8902	81	2.46	N	0
C. jejuni	1644	3934	35	2.39	Y	0.6
L. monocytogenes	2913	6511	55	2.24	Y	0.58
C. trachomatis	910	1255	5	1.38	Y	0.19

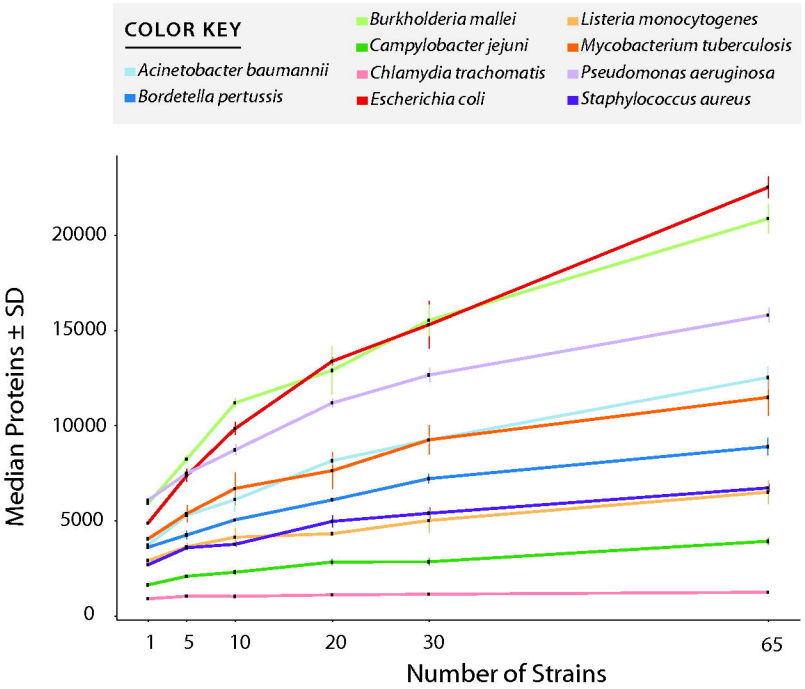
* As described in panX database (Ding et al, 2018).

** Pangenome size for *B. pseudomallei*


A
Round 1
Subjects
Query 1
B
Homologue
Files
Uniquely
Anotated
C
Reference Seq
Artificial Seq
D
Final
DB
E
Query 2
Uniquely
Anotated
Homologue
Files
Reference Seq
Artificial Seq


1
504
AEM98552 / Rv0104
1
504
AGL25567
1
504
AMC84165
1
426
AMC88348
9
504
BAW10927
9
378
AEJ48929
9
500
AGL29533
118
500
BAX25728
Tyr 402 → His
Gly477 → Arg
Leu493 → Val
TSS choice
fs/indel after Leu475


COLOR KEY
47K
7M
lipV
S1430
S1593
S1659
S179
S1841
S1945
S2666
S861
1
MIIDLHNQR
...
3.5M
1.4M
7M
785K
665K
386K
4.8M
19
ITIHGVTEHGR
...
169K
1.1M
47K
22
HGVTEHGR
...
173K
150K
462K


COLOR KEY
Acinetobacter baumannii
Bordetella pertussis
Burkholderia mallei
Campylobacter jejuni
Chlamydia trachomatis
Escherichia coli
Listeria monocytogenes
Mycobacterium tuberculosis
Pseudomonas aeruginosa
Staphylococcus aureus
Median Proteins ± SD
0
5000
10000
15000
20000
1
5
10
20
30
65
Number of Strains