frontiers
in Microbiology

ORIGINAL RESEARCH
published: 20 June 2019
doi: 10.3389/fmicb.2019.01410

On the Impact of the Pangenome and Annotation Discrepancies While Building Protein Sequence Databases for Bacteria Proteogenomics

Karla C. T. Machado[1], *Suereta Fortuin*[2], *Gisele Guicardi Tomazella*[1,3,4], *Andre F. Fonseca*[1], *Robin Mark Warren*[2], *Harald G. Wiker*[3], *Sandro Jose de Souza*[1,5] and *Gustavo Antonio de Souza*[1,6]*

[1] Bioinformatics Multidisciplinary Environment, Universidade Federal do Rio Grande do Norte, Natal, Brazil, [2] DST/NRF Centre of Excellence for Biomedical Tuberculosis Research/SAMRC Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Stellenbosch, South Africa, [3] The Gade Research Group for Infection and Immunity, Department of Clinical Science, University of Bergen, Bergen, Norway, [4] The Institute of Bioinformatics and Biotechnology, Natal, Brazil, [5] The Brain Institute, Universidade Federal do Rio Grande do Norte, Natal, Brazil, [6] Department of Biochemistry, Federal University of Rio Grande do Norte (UFRN), Natal, Brazil

OPEN ACCESS

Edited by:
George Tsiamis,
University of Patras, Greece

Reviewed by:
Sergei Moshkovskii,
Institute of Biomedical Chemistry,
Russian Academy of Medical
Sciences (RAMS), Russia
Amit Kumar Yadav,
Translational Health Science
and Technology Institute, India
Sneha M. Pinto,
Yenepoya University, India

***Correspondence:**
Gustavo Antonio de Souza
gadsouza@gmail.com

Specialty section:
This article was submitted to
Systems Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 26 February 2019
Accepted: 05 June 2019
Published: 20 June 2019

Citation:
Machado KCT, Fortuin S,
Tomazella GG, Fonseca AF,
Warren RM, Wiker HG, de Souza SJ
and de Souza GA (2019) On
the Impact of the Pangenome
and Annotation Discrepancies While
Building Protein Sequence Databases
for Bacteria Proteogenomics.
Front. Microbiol. 10:1410.
doi: 10.3389/fmicb.2019.01410

In proteomics, peptide information within mass spectrometry (MS) data from a specific organism sample is routinely matched against a protein sequence database that best represent such organism. However, if the species/strain in the sample is unknown or genetically poorly characterized, it becomes challenging to determine a database which can represent such sample. Building customized protein sequence databases merging multiple strains for a given species has become a strategy to overcome such restrictions. However, as more genetic information is publicly available and interesting genetic features such as the existence of pan- and core genes within a species are revealed, we questioned how efficient such merging strategies are to report relevant information. To test this assumption, we constructed databases containing conserved and unique sequences for 10 different species. Features that are relevant for probabilistic-based protein identification by proteomics were then monitored. As expected, increase in database complexity correlates with pangenomic complexity. However, *Mycobacterium tuberculosis* and *Bordetella pertussis* generated very complex databases even having low pangenomic complexity. We further tested database performance by using MS data from eight clinical strains from *M. tuberculosis*, and from two published datasets from *Staphylococcus aureus*. We show that by using an approach where database size is controlled by removing repeated identical tryptic sequences across strains/species, computational time can be reduced drastically as database complexity increases.

Keywords: databases, proteomics, proteogenomics, mass spectrometry, pangenome

INTRODUCTION

In the past decade, the development of Next Generation Sequencing methods has made genome sequencing more affordable and consequently more accessible. However, while genome annotation approaches also undertook considerable advances in the past years (Bocs et al., 2003; Davidsen et al., 2010; Vallenet et al., 2013), correct gene prediction is still challenging even for prokaryotes. Particularly, the correct assignment of open reading frames (ORFs), the presence and classification of pseudogenes and differences in translational starting site (TSS) (de Souza et al., 2008, 2009; Cuklina et al., 2016) are the main sources of variability. Consequently, for any given strain or species, its predicted protein sequence information will normally contain inaccuracies that will interfere with peptide identification in mass spectrometry (MS)-based proteomics. The use of accurate protein sequence databases is a key step in most proteomic approaches, and this is particularly critical for studies involving samples containing strains with little to no genetic information or samples containing multiple strains (metaproteomics) (Tanca et al., 2013).

In such cases when the establishment of a gold-standard annotation that can represent the sample under investigation is difficult, a viable alternative is to construct customized protein sequence databases which are then inspected against peptide sequence data collected by MS (Nesvizhskii, 2014). This has been largely employed in proteogenomics, where "novel" sequences are inserted in the customized database and, if identified, are further used to validate and confirm proposed gene models and other genetic polymorphisms (for recent reviews see Renuse et al., 2011; Ruggles et al., 2017). Database customization is often achieved using two different strategies: (i) through a 6-frame translation of the genome of the strain (Fermin et al., 2006; Baerenfaller et al., 2008); (ii) or by constructing a database merging *ab initio* gene predictions from related strains of the same species, taking into consideration variations caused by SNPs, indels, divergent TSS choice, among others (de Souza et al., 2010; Omasits et al., 2017). These approaches are not mutually exclusive, as gene annotation from related strains can be used to further optimize 6-frame translation approaches (Castellana et al., 2008).

However, peptide identification in MS-based proteomics is often performed through probabilistic calculations between the observed peptide fragmentation pattern and theoretical MS/MS data from a sequence database. Therefore, database size will: (i) alter the search space and consequently the probabilistic calculations performed during peptide identification (Nesvizhskii, 2010); (ii) make protein inference more difficult, especially in multistrain databases (Nesvizhskii and Aebersold, 2005); and (iii) demand more computational power due to the handling of larger files. Therefore, building databases using either 6-frame translations or sequence merging approaches which are larger than regular can become, at some point, detrimental for the proteomic analysis (Blakeley et al., 2012).

Such issue might as well-contribute differently depending on the species under study, how much genomic information is available (number of strains with genome sequenced) and genomic features particular to it. When we first developed the MSMSpdbb approach (de Souza et al., 2010), the number of genomic information available was a fraction of the amount currently existent. For example, at the time there were only complete genomes sequenced for eight strains from the *Mycobacterium tuberculosis* (Mtb) complex (five from *M. tuberculosis* and three from *M. bovis*). For such dataset we observed that the size of a merged database Mtb complex was not heavily incremented as more strains were considered (de Souza et al., 2011). Each new strain added to the database contributed with only few genetic polymorphisms when compared to the reference H37Rv strain. However, the current 65 Mtb strains with complete genome sequenced demonstrated the existence of a pangenome in this species, 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 (Rasko et al., 2008; McInerney et al., 2017). We then wondered if the approach would still be valuable as 100s of genome information are available to a single species. 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.

Many reports have demonstrated that database size could be controlled by creating protein entries where identical peptides between homolog proteins are reported once and only unique polymorphic tryptic peptides are inserted when merging sequences (Schandorff et al., 2007; Omasits et al., 2017; Liao et al., 2018). It is unclear if genomic features such as the presence of a pangenome could impact database customization of specific species. We then decided to test this by creating protein sequence databases for 10 species using strains with complete genome sequences and a modified version of MSMSpdbb (de Souza et al., 2010). Database parameters such as the number of entries and unique peptide sequences for each species using increasing number of strains per round were then quantified. As expected, species with large pangenomes such as *Escherichia coli* showed the larger increase in database size per number of strains used, and database size in general correlated with pangenome size. Finally we performed proteomic identification and computational performance of two different datasets from: (i) eight clinical strains of Mtb using a database with 65 complete strain genomes; and (ii) two *S. aureus* datasets (Depke et al., 2015; Hoegl et al., 2018) using a database with 194 complete strain genomes.

MATERIALS AND METHODS

Species Selection

Ten bacterial species were selected for database construction: *Acinetobacter baumannii* (89), *Bordetella pertussis* (348), *Campylobacter jejuni* (114), *Chlamydia trachomatis* (82), *Escherichia coli* (425), *Listeria monocytogenes* (148), *Mycobacterium tuberculosis* (65), *Pseudomonas aeruginosa* (105) and *Staphylococcus aureus* (194). *Burkholderia mallei* and *Burkholderia pseudomallei* (109) are considered genetically very similar (Godoy et al., 2003; Song et al., 2010) and were analyzed together. Number in parenthesis represents strains with complete genome sequenced according to GenBank (Benson et al., 2013) in August 2017. Protein sequences were obtained using the assembly file available at ftp://ftp.ncbi.nih.gov/genomes/genbank/bacteria/.

Script Design and Availability

We redesigned MSMSpdbb (de Souza et al., 2010) in PERL (version 5.24.1) and is present as two modules: all_fasta.pl provides the sequence alignment and creates the outputs with unique entries and homologs; pep_trip.pl process the homologs 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. Rand.pl and create_db.pl are also available to reproduce the experimental design described below (see section "Database Analysis – Experimental Design and Statistical Rationale"). For detailed information on how to use and how each script (all_fasta.pl and pep_trip.pl) performs, see **Supplementary Text**. A briefer description of the tool is given below.

Data Processing

Even though MSMSpdbb is already published, we changed how orthology is defined (using a Bidirectional Best Hit rather than a Cluster BLAST approach) because BHH is considered a more appropriate manner to define orthology. We provide below further information about the method. To assign gene homology, all_fasta.pl initially gathers the protein sequence data from all strains and then performs pair wise alignments using BLASTP (Altschul et al., 1990) through a Bidirectional Best Hit (BHH) method (Tatusov et al., 1997; Bork et al., 1998; Overbeek et al., 1999). 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 homologs 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 homologs are defined between query and subject strains, all_fasta.pl will proceed by indexing homolog IDs for later characterizations of polymorphisms. It will then define "unique" protein sequences from the query dataset (i.e., a sequence without a defined homolog 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.

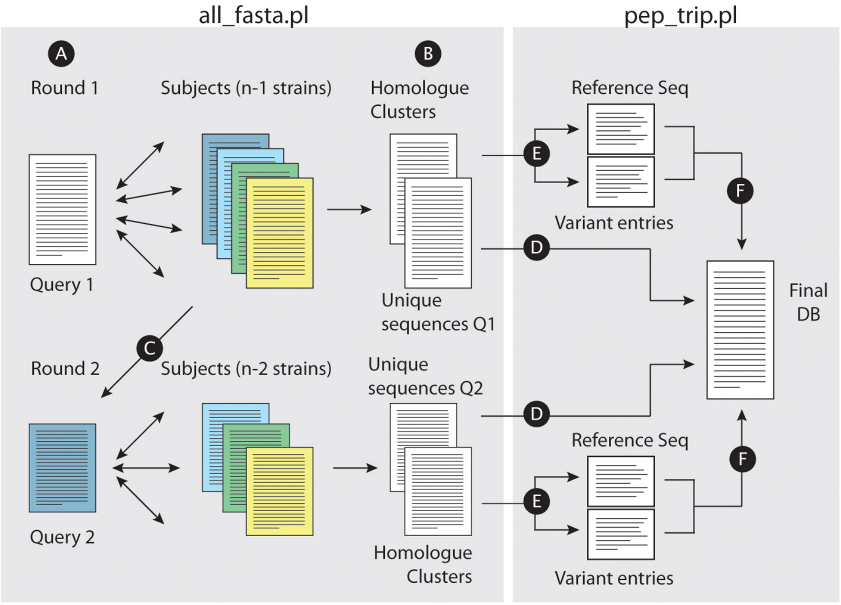

FIGURE 1 | Building concatenated databases from a dataset of fasta files. **(A)** For a given set of fasta files, the script all_fasta.pl selects a Query file (Query 1) and BLASTP is performed against the remaining files. A bidirectional approach is performed for homolog identification across all strains. **(B)** The script writes two files, one with uniquely annotated entries from Query 1 and other with homolog clusters. **(C)** Query 1 fasta file is removed from the dataset, and the remaining subject files will have entries removed (proteins which were already defined as a homolog of a Query 1 protein); a new fasta file is selected as Query (Query 2) for a new BLASTP round. Such rounds will be performed until all files are analyzed. When this is achieved, the script pep_trip.pl will build the database and write the log file reporting all tasks performed. First, uniquely annotated entries are copied "as is" into the final fasta database **(D)**. **(E)** For each homolog group, a reference sequence is chosen, and each polymorphic tryptic peptides from non-reference homologs are added into individual entries for each peptide ("Variant entries"); reference and variant entries are then copied into the final database **(F)**.

Regarding all indexed homologs, our script pep_trip.pl then defines the longest version of the protein as reference (not necessarily from same strain used as query for the BHH step above). This was done for two reasons: first, amino acid length is an easier parameter to track; and second, the longer version also provides the less redundant outcome for the remaining TSS variants that our script must consider. For example, if a shorter version was selected as reference, our script would need to have additional coding to compare and sort out any repeated sequences in the other possible TSS choices. 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, without allowing any miscleavages, and compares 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), each peptide is added into the final output under a created protein entry (Omasits et al., 2017). Amino acid changes as a 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. One must note that our approach only works for samples treated with trypsin, since variant peptides are added as tryptic peptide entries in the final database.

Database Analysis – Experimental Design and Statistical Rationale

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 with respect to number of entries and number of unique tryptic peptide sequences. Pangenomic size calculations were based on the panX tool (Ding et al., 2018).

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, 1 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 × *g*) at 4°C and resuspended in 1 mL of cold lysis buffer containing 10 mM Tris-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 × *g*) at 21°C. An equal volume of 0.1 mm glass beads (Biospec Products, Inc., Bartlesville, OK, United States) 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 a Ribolyser (Bio101 Savant, Vista, CA, United States) 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 × *g* for 5 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, United States), 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 lysate (60 μg) were mixed with electrophoretic sample buffer (NuPAGE kit, Invitrogen, Carlsbad, CA, United States) 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 (Tomazella et al., 2012). 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, Fitchburg, WI, United States) for 16 h at 37°C in 50 mM NH_4HCO_3, 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 (Rappsilber et al., 2003). 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 connected to a linear quadrupole ion trap-Orbitrap (LTQ-Orbitrap) mass spectrometer (Thermo Electron, Bremen, Germany) equipped with a nanoelectrospray ion 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 (Olsen et al., 2005). Target ions already selected for MS/MS were dynamically excluded for 60 s. General MS 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.

MS Data Analysis

In addition to in-house generated Mtb data, we have also analyzed *S. aureus* MS files found in ProteomeXchange as PXD006483 (Hoegl et al., 2018) and PXD000702 (Depke et al., 2015). Raw MS files of all Mtb and *S. aureus* datasets were analyzed by MaxQuant version 1.5.2.8. Peptides in MS/MS spectra were identified by the Andromeda search engine (Cox et al., 2011b). For Mtb dataset, MS files were searched against a database containing all sequences from 65 strains without any processing (263,683 entries), or the database generated by our tool using the same 65 strains (15,979 entries). The *S. aureus* dataset was searched against a database containing all sequences from 194 strains without any processing (523,281 entries) or the MSMSpdbb generated database (15,073 entries). All bacterial sequences were obtained from GenBank as written above (August 2017). MaxQuant analysis included an initial search with a precursor mass tolerance of 20 ppm which were used for mass recalibration (Cox et al., 2011a). Trypsin without proline restriction was used as enzyme specificity, with two allowed miscleavages and no unspecific cleavage allowed. 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 of methionine and protein N-terminal acetylation. Carbamidomethyl cysteine was added as a fixed modification. Minimal peptide length was set to seven amino acids. The false discovery rate (FDR) was set to 0.01 for peptide and protein identifications, and since the aim is to compare database performance, no further score or PEP thresholds were used for individual MS spectra. 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 from common contaminants.

To test if time performance was biased by the MaxQuant environment, we have also submitted both MS datasets for peptide identification against reduced and concatenated databases using Comet (Eng et al., 2013), OMSSA (Geer et al., 2004), and X!Tandem (Craig and Beavis, 2004). All searches were performed separately using SearchGUI (Vaudel et al., 2011), using the same search parameters as for the MaxQuant searches. Data regarding analysis time were retrieved from the SearchGUI log file generated during each search.

RESULTS

Implementation and Database Format

MSMSpdbb was originally 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) on a non-redundant manner (i.e., without extensive entry usage for very similar sequences). At the current version we modified it to perform BBH BLAST, which is an approach more used to define orthologous sequences, rather than Cluster BLAST. Database formatting was also modified; each non-reference peptides containing polymorphisms are now inserted as a new entry in the database.

On average, the assemblies used on this work had a number of genes ranging from 910 in *Chlamydia trachomatis* (average value to all strains) to 6092 in *Pseudomonas aeruginosa*. The analyzed species were selected not solely due to a higher number of strains with complete genome sequences available but also considering the complexity of their pangenomes. All selected species have pangenomes that can be very diverse (for example, *E. 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) (Rasko et al., 2008; McInerney et al., 2017; Ding et al., 2018).

Figure 1 illustrates the approach workflow: briefly, a query strain is selected, and pair wise comparisons are performed with the remaining subject strains (**Figure 1A**). All_fasta.pl will generate two files, one containing unique entries from the query strain, which did not find an homolog in any of the subject strains (i.e., sequences with no BHH significant hit to any other protein entries in subject strains), and another file with all homolog protein sequence clusters from the strains that shared the required levels of identity and similarity in BHH manner (**Figure 1B**). All_fasta.pl will then

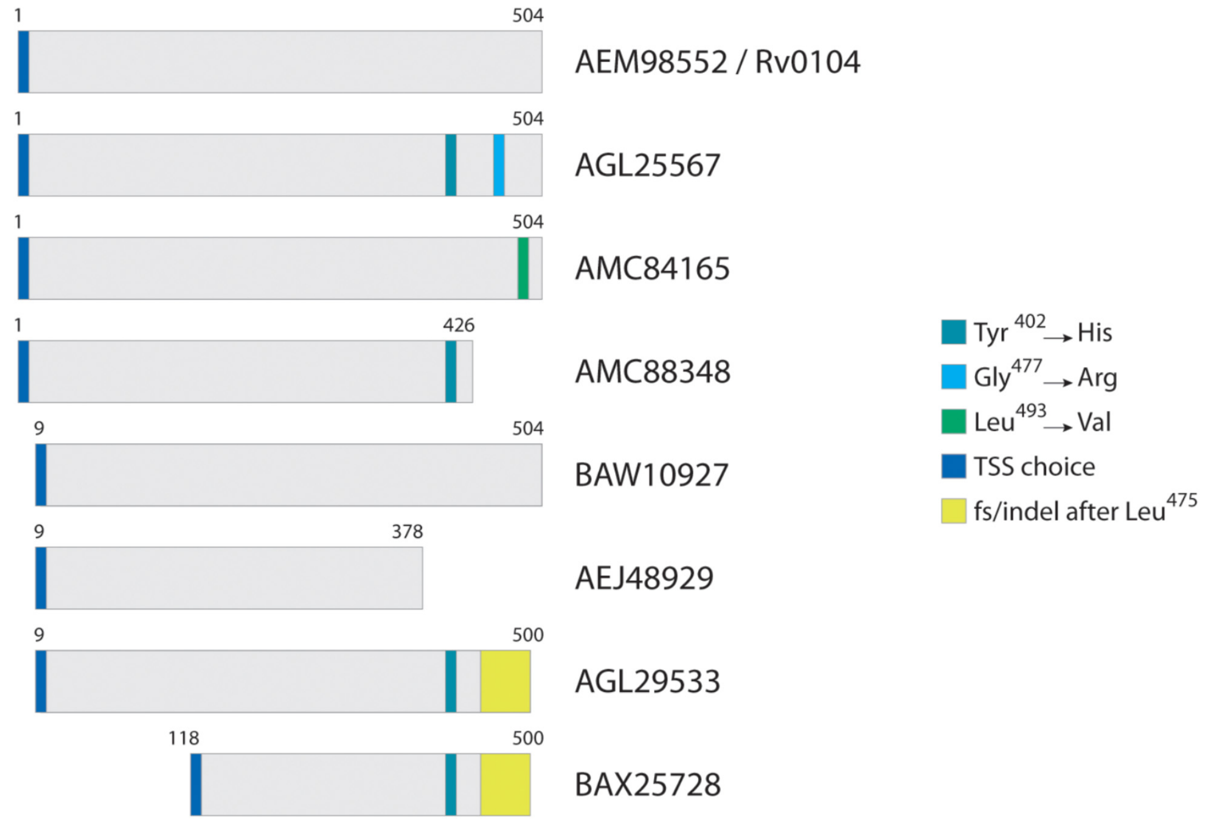

FIGURE 2 | Typical homolog comparison. For protein Rv0104, in this case selected as reference, three SAPs were identified across all Mtb strains, in addition to two TSS choices, two premature stop codons, and a complete amino acid sequence change (region in yellow) most probably from an indel.

remove Query 1 strain from the dataset, and reduce the fasta files from Subject strains by removing proteins that were already identified as homologs to a Query 1 protein. One of the subject strains will be selected as Query 2, and a new round of BHH alignment will be performed using the remaining subject strains (**Figure 1C**). This will be repeated until all subject strains are used as Query. At this point, the script pep_trip.pl will create the final database by first copying all uniquely annotated sequences (**Figure 1D**). For each homolog cluster, the script will select the longest homolog to be used as a reference sequence (regardless of which strain it is originated from), and save it into the final database. The remaining homologs will have their tryptic peptides compared to the reference sequence. Each tryptic peptide from a non-reference sequence that does not match any of the reference sequence (due to a different TSS choice or a SNP for example), will be added in the database as a new entry. Pep_trip.pl will finish the database (**Figure 1F**) and also create a log file reporting all entries that were inserted into database, their origin, how they were clustered and compared, 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. Detailed information of this process can be found in the **Supplementary Text**.

Figure 2 illustrates a typical homolog comparison leading to the detection of polymorphic peptides. The hypothetical protein AEM98552 [Rv0104 on the Tuberculist annotation (Lew et al., 2011)] was set as reference, and nine other different variants were observed. Those variants are the result of the different combinations including: 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 premature stop codon, truncating the protein at position 378. Each unique tryptic peptide characterizing all such polymorphisms are added in the database as an additional entry.

Database Size Increase Ratio per Species

Databases were created for 10 species to measure if the rate of increase in the database size could be impacted by unique features of each genome, such as the size of a

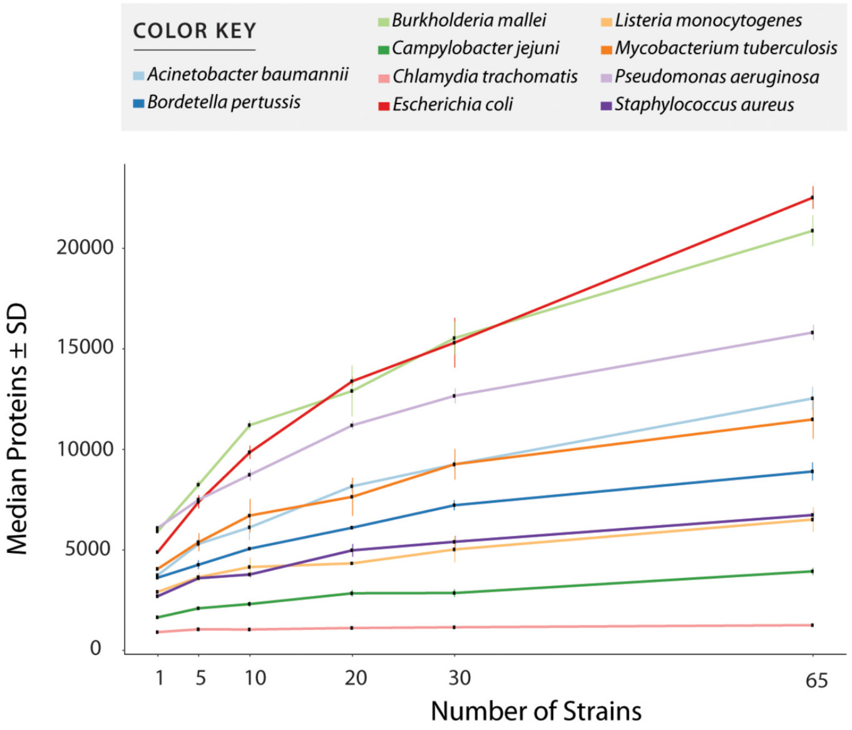

FIGURE 3 | Number of entries per database after redundancy removal. The graph shows data when only considering reference and uniquely annotated sequences in all strains, i.e., customized entries containing polymorphic peptides where not counted. Values in x axis show the number of strains used for database creation. All values plotted in the graph are given in **Supplementary Table S1**.

TABLE 1 | Database size increase and pangenome size in 10 selected species.

	1 strain	Final DB Ref	Gain per strain	DB increase	Pangenome?	Pangenome size (%/100)*
E. coli	4893	22525	271	4.60	Y	0.91
B. pseudomallei/mallei	5917	20876	230	3.53	Y	0.70**
A. baumanii	3746	12536	135	3.35	Y	0.79
M. tuberculosis	4056	11490	114	2.83	Y	0.21
P. aeruginosa	6092	15814	150	2.60	Y	0.81
S. aureus	2697	6735	62	2.50	Y	0.67
B. pertussis	3619	8902	81	2.46	Y	0.15
C. jejuni	1644	3934	35	2.39	Y	0.62
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) in March 2019. **Pangenome size for B. pseudomallei.*

pangenome. For a fair comparison among species, we built databases using a maximum of 65 strains, since at the time 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 in case a strain with very unique annotation being randomly selected, every database was created three times in total. The rand.pl script provided with the tool performs all such steps automatically, if the data needs to be reproduced (replacing all_fasta.pl described in section "Materials and Methods"). We then counted total number of protein entries and tryptic peptides in each database.

Supplementary Table S1 shows all values measured for all tests including number of annotated proteins (excluding all entries created to accommodate polymorphisms) (**Figure 3** and **Table 1**), and number of peptides with or without miscleavages allowed (**Supplementary Figure S1**). **Figure 3** shows graphically the increase in number of annotated proteins in all species, and **Table 1** 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 increase 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 3rd 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 the smaller pangenome but with a database increase rate (2.46x) which surpassed three species with larger pangenomes.

Proteomic Performance in Diverse MS Datasets

When we first developed the MSMSpdbb approach (de Souza et al., 2010), the number of genomic information available was a fraction of the amount currently in existence. For example, the analysis validation performed on an Mtb dataset used a database with only eight strains (five from *M. tuberculosis* and three from *M. bovis*). We then wondered if the approach would still be valuable as 100s of genome information are available to a single species. This was tested by performing proteomic analysis in two independent datasets.

The MSMSpdbb processed databases created by the tool contained 15,996 (Mtb), and 15,073 (*S. aureus*) protein entries. In comparison, merely concatenated databases using all sequences from the available strains for each species created a database with 263,683 protein entries for Mtb and 523,281 protein entries for *S. aureus*. From now on, all MSMSpdbb processed databases are called DB1, and simply concatenated databases are named DB2 for simplicity. The MS datasets were then challenged for peptide identification using the respective DB1 and DB2 databases for each species. It is important to note that all theoretical unique tryptic peptides present in DB2 are also present in DB1 (data not show), except for variant peptides in DB2 that are shorter than 7 amino acids or longer than 35 amino acids, which were not considered by the pep_trip.pl script which created DB1. The complete results folders from each MaxQuant search including protein and peptide lists with relevant identification parameters are provided (see section "Data Availability").

In all datasets, when comparing the results from DB1 and DB2 searches, the number of peptides identified was very similar in both searches (**Figure 4**). For example, the Mtb dataset identified 32,986 unique peptides in both searches, while 515 peptides were identified only in the DB1 search and 1,116 peptides only in the DB2 search result (**Figure 4A**). When all peptides are considered, 95.28% of the peptides were identified in both searches. For *S. aureus,* DB1 and DB2 searches were in agreement for 92% of the identified peptides in all searches combined. We observed that the majority of such database exclusive identifications are distributed in the lower range of the spectra scoring. For all datasets, the median MaxQuant score for peptides identified in only DB1 and DB2 were in the range of 52 to 75, while the median scores of peptides identified both searches was from 128 to 137 (see histograms **Figure 4**). If a MaxQuant posterior error probability (PEP) score lower than 0.01 is also considered as a parameter for valid MS2 identification, the numbers of exclusive peptides drop even further. For Mtb, DB1 exclusive peptides were reduced from 515 to 315 peptides, 0.95% of the total number of peptides identified. For DB2, the difference was even larger, from 1,116 to 424 peptides (1.29% of the total peptides identified). From these remaining 424 peptides with good PEP scores, 323 are peptides which are not present in DB1 because they did not meet requirements used by our script, i.e., they are not reference peptides, they contain polymorphisms (i.e., they should had been added in DB1 as an additional entry), but they are fully tryptic peptides with less than seven amino acids. The same was observed for *S. aureus* dataset: DB1 exclusive peptides dropped from 1,500 to 1,075 (now 2.7% of total) and DB2 dropped from 2,916 to 1,738 (now 4.8% of total). While 4.8% variation is still a high difference, from those 1,738 peptides, 1,121 are fully tryptic peptides with less than seven amino acids not present in DB1.

We also had validated if a false-positive bias was present in identified peptides from proteins that are annotated/present in fewer strains, compared to peptides present in most strains of both Mtb and *S. aureus*. For this, we divided the identified peptides according to the number of strains containing those peptides, and visualized their score distribution. As a control, we observed the score distribution of the peptides identified in decoys (reversed) entries. For Mtb, we divided peptides as those present in 33 strains or more; in 13 to 32 strains; or those annotated in 12 strains or less. For *S. aureus*, groups were those annotated in 100 strains or more, 21 to 99 strains, or 20 strains or less. **Supplementary Figure S2** shows the histograms for all groups. Score distribution was similar to all identified peptides, regardless if they are commonly observed between strains or rarer. As a control, we compared the score distribution of those peptides to the score of decoy (reverse) peptides, i.e., true false-positives. As expected, the score distribution of peptides belonging to decoy identifications was very low.

Interestingly, multiple TSS polymorphisms for the same protein were also observed even for the same strains. **Supplementary Figure S3** 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) represents the most 5′ prediction; LTIHGVTEHGR (Score 85.9, PEP 0.0005) starts at position 19 of reference protein and HGVTEHGR (Score 100, PEP 0.0077) starts at position 22 of reference. The longer form is the most abundant in almost all samples, except S1430 where only LTIHGVTEHGR was identified and S2666 where the N-terminal HGVTEHGR is slightly overrepresented than MIIDLHVQR.

Database Size Impact in Computational Performance

As shown before, database size reduction from DB2 to DB1 was approximately 16x to Mtb and 35x to *S. aureus*. By inspecting MaxQuant log folder we calculated the amount of minutes used for each step in its proteomics pipeline. For Mtb, the time spent for the whole pipeline was reduced by half, from 853 min for DB2

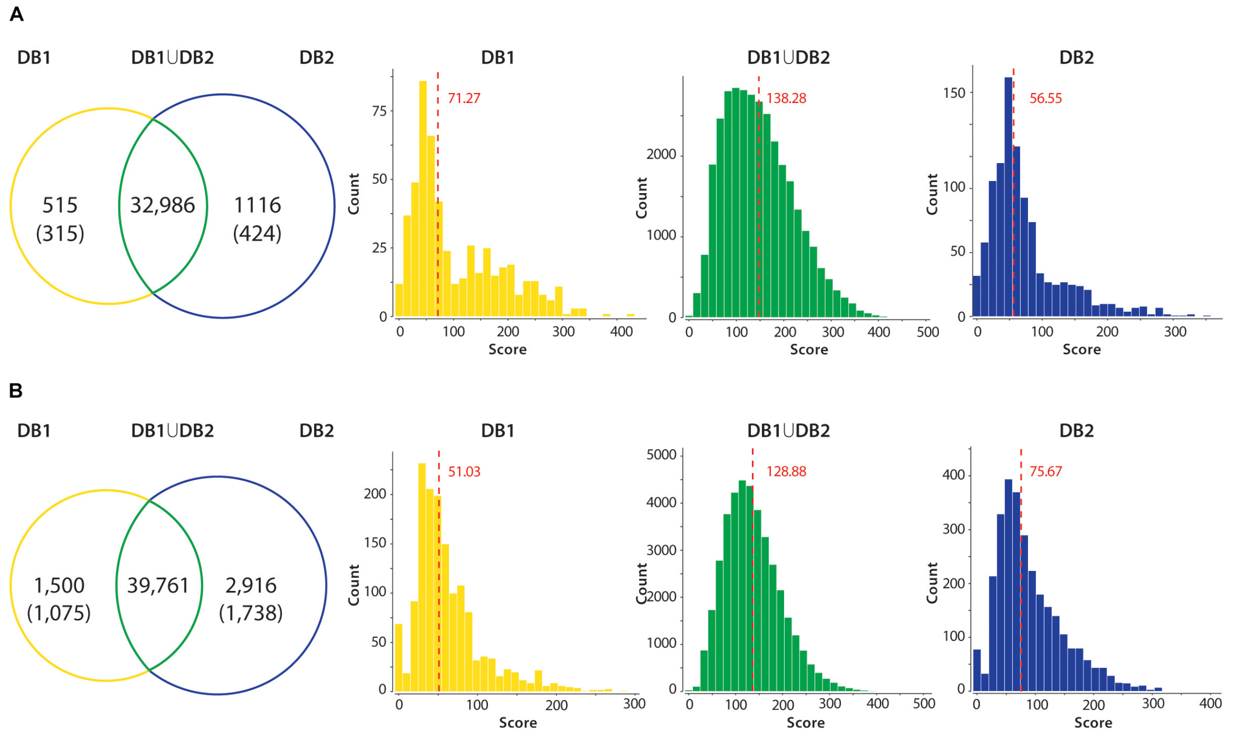

FIGURE 4 | Peptide identification performance in reduced (DB1) vs. concatenated (DB2) databases. All unique tryptic peptides identified in DB1 and DB2 in the Mtb **(A)** and *S. aureus* **(B)** datasets are compared. Venn diagrams show that majority of tryptic peptides were identified in both searches. Unique peptides identified in each dataset had their PEP score inspected. Values in parenthesis shown number of peptides with valid scoring (PEP < 0.01). Peptide MS2 scores of each group were plotted in a histogram, and a median was calculated. Median of the score distribution for peptides exclusive to DB1 or DB2 ranged from 51.03 to 75.67, while median scores of peptides identified on both searches were 138.28 for Mtb and 128.88 for *S. aureus*.

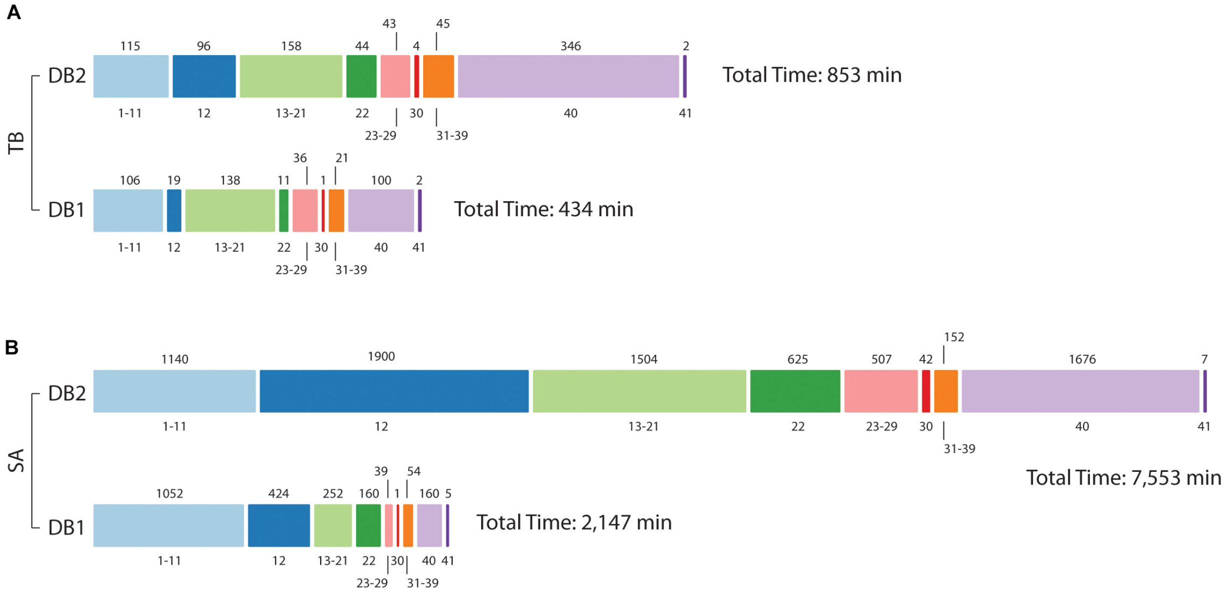

FIGURE 5 | Computational demand in reduced (DB1) vs. concatenated (DB2) databases. Processing times for Mtb **(A)** and *S. aureus* **(B)** datasets. The bar is divided into MaxQuant stage number, as described in Neuhauser et al. (2013) (value at bottom) and time in minutes is at the top of the bar. For Mtb, analysis was hasten mainly at stage 12 (Main search), 22 (Second peptide search), and 40 (Writing tables); in addition to those, for *S. aureus* improvements were also observed for post-search steps 13–21 and 23–29, and step 30 ("Protein groups assembly").

to 434 min to DB1 (**Figure 5**). Main differences were observed for steps related to the peptide search, i.e., the 'main search' (5x) and the 'second peptide search' (4x) steps, and during the creation of results and other final files named 'writing tables' (3.5x). For *S. aureus* the differences were far more evident, as the complete pipeline took 7,553 min for DB2 and 2,147 min for DB1. Time reduction in peptide search steps were similar to Mtb at 4.5x and 4x respectively. But in addition to those, other steps which are related to read and finish the main and the second search results had improvements of 42 and 19.6 times respectively. Steps related to assignment and assembly of protein groups had improvements of 42 times, and 'writing tables' had a time improvement of 10.4 times.

Arguably, such differences could be specific to the MaxQuant environment. Therefore, Mtb dataset and databases were additionally tested by three different peptide search engines: X!Tandem (Craig and Beavis, 2004), OMSSA (Geer et al., 2004), and Comet (Eng et al., 2013). All searches were performed independently using SearchGUI environment (Vaudel et al., 2011). Those searches are comparable to the "Main search" step in MaxQuant, which represents the task performed by Andromeda peptide search engine (Cox et al., 2011b). All processing times are shown in **Supplementary Figure S4**. Peptide searches using the concatenated DB2 database, searches took 174 min in X!Tandem, 1,455 min in OMSSA and 2,708 min in Comet. For the reduced DB1 database, the searches took 58 min in X!Tandem, 322 min in OMSSA and 184 min in Comet, representing analyses time reduction of three times, 4.51x and 14.7x respectively.

DISCUSSION

Database size is a known bottleneck in proteomics, and its implications had been already discussed in proteogenomics and metaproteomics research (Jagtap et al., 2013; Heyer et al., 2017). Therefore, controlling database information is of key importance. A manner to achieve this is to construct customized databases where identical sequences/tryptic peptides from homolog proteins are not exhaustively present in different entries within the database. We had already applied this for the characterization of samples with unknown genetic background such as bacterial clinical strains (de Souza et al., 2011; Tomazella et al., 2012). 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 (Abascal et al., 2018). It is true even for prokaryotes (de Souza et al., 2008, 2009), with their simpler genomic structure and higher coding density. However, as the amount of genomic information exponentially increases with time, we wished to further evaluate the impact of such approach with the current amount of available genomic data.

First, we were curious to investigate how different species with diverse genetic features (such as the presence of a pangenome) would impact database size and structure. As expected, database size increase correlated with the pangenome complexity of each species. 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 (de Souza et al., 2011). 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. Something similar happened to *B. pertusis* databases, which has the smallest pangenome and database increase was higher than three species with bigger 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 (Parkhill et al., 2000, 2003; Thomson et al., 2008). 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 *B. pseudomallei* and *B. 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 (Godoy et al., 2003), large scale rearrangements in *B. mallei* (Losada et al., 2010) might be interfering with our approach.

We then selected two MS datasets from Mtb and from *S. aureus* to identify peptides using a routine probabilistic-based approach, but now using the complete set of available genomes (194 strains) for *S. aureus*. Overall, DB1 and DB2 performance was very similar, and differences were negligible (highest difference was 2.5% for DB1 exclusive peptide identifications in *S. Aureus*), as most could be possessing low MS2 scores and low PEP values. Valid identifications which are exclusive to DB2 (323 and 1121 peptides in Mtb and *S. aureus,* respectively) were peptides containing miscleavages that were not selected by MSMSpdbb (i.e., the peptides without miscleavages were shorter than seven amino acids and not present in the selected reference protein) and therefore were absent in DB1 due to the parameters used for DB1 construction. We also investigated if "rarer" identified peptides, i.e., those containing polymorphism or from proteins present in a minority of the strains of the species, could have an identification bias. MS2 score distribution was very similar, regardless if peptide belonged to a commonly observed variant or not, and distribution had higher medians than peptides identified from decoy entries, i.e., distinguishable false-positives. Therefore we conclude no bias exists toward common or rare peptides. It is worth mentioning that surprisingly, we have identified multiple N-terminal predictions for the same protein, in some cases observed in the same strain (**Supplementary Figure S3**). 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.

The advantage of using a concatenated database is evident when computational performance is measured. The main steps in Andromeda/MaxQuant pipeline which involve Fasta reading ('main' and 'second peptide' searches) were similarly improved in both datasets tested. Similar observations were obtained regardless if other peptide search engines were used. In *S. aureus* dataset, where DB1 concatenation is larger when compared to DB2 (from approximately 500,000 entries to 15,000 entries), time processing improvements were observed in additional stages of the MaxQuant pipeline. Particularly steps related to handle of peptide search engine results and assembly of the protein groups to be reported as final protein identifications. The step which creates the final output files ('Writing tables'), even though do not directly use the Fasta file, was also significantly improved. This makes sense, as MaxQuant will report all entries from the database that share identical peptides. Entries from a protein with high sequence identity present in all 194 strains for *S. aureus*, for example, will have all 194 entries recorded in the output if that protein is identified. For DB1, since those 194 highly identical sequences are concatenated into a single reference entry, the output reporting is vastly simplified.

As MS evolves and as detection sensitivity and peptide identification coverage improves, one of the last bottlenecks in metaproteomics research is database design. Building databases for metaproteomics is often achieved by two approaches: either by collecting metagenomic or metatranscriptomic data from the sample; or by concatenating protein sequences from a public repository such as UniProt (for recent reviews see Heyer et al., 2017; Starr et al., 2018). Regarding the metagenomic approach, even with the advances of next-generation sequencing methods, proper genomic screening of a complex sample is still costly, and assembly of short reads and contigs in complex samples is challenging due to sequence similarity (Heyer et al., 2017). On the other hand, concatenating protein sequences from multiple organisms from a public repository results in very large databases, making the analysis time costly and susceptive to high false-positive rates (Tanca et al., 2013). False-positive issues had been mostly solved by database size restrictions using a multi-step peptide search or multiple peptide searches using smaller datasets of refined sequences on each search and further analysis to combine results (Jagtap et al., 2013; Muth et al., 2016, 2018; Zhang et al., 2016; Liao et al., 2018). While those are successful methods, they are mostly provided as complete pipelines. We believe that providing solutions solely at Fasta file level offers more freedom to researchers, as they can use whichever peptide search engines or pipelines they prefer. We predict that controlling database size such as done by MSMSpdbb without losing peptide identification coverage and drastically reducing computational processing time will be of key importance for metaproteomics.

DATA AVAILABILITY

The scripts are fully available at https://github.com/karlactm/Proteogenomics. All the Mtb MS files, and MaxQuant search parameters and results are available at the ProteomeXchange with the accession number PXD011080. The MS files of *S. aureus* are available in the ProteomeXchange as PXD006483 (Hoegl et al., 2018) and PXD000702 (Depke et al., 2015).

AUTHOR CONTRIBUTIONS

KM designed the tool and carried out bioinformatic analysis. SF and GT generated the Mtb samples and proteomics data. AF assisted with optimization of script design and BHH analysis. RW, HW, SS, and GS supervised the work and data analysis. GS provided the proteomic analysis of *S. aureus*. KM and GS designed the experiments and carried out analysis.

FUNDING

This work was supported by the grant 23038.004629/2014-19 to SS and GS from the Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior; grants 305233/2015-7 to SS and 406630/2016-0 to GS from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); and the grants 175141 (HW) and 183418 (HW/RW) from the Norwegian Research Council. This research was also supported by the High Performance Computing Center at the UFRN (NPAD/UFRN). This manuscript has been released as a preprint at bioRxiv with the doi: 10.1101/378117 (Machado et al., 2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01410/full#supplementary-material

FIGURE S1 | Number of tryptic peptides per MSMSpdbb database. In **(A)** only fully tryptic peptides are considered, while in **(B)** peptides with up to two miscleavages are also counted, a standard search parameter choice in proteomics. Values in x axis show the number of strains used in the database. All values plotted in the graph are given in **Supplementary Table S1**.

FIGURE S2 | Score distribution in subpopulations of identified peptides. All identified peptides in *S. aureus* **(A)** and Mtb **(B)** datasets were divided according to their occurrence across all strains. For Mtb they were divided as peptides annotated in more than 33 strains, annotated in 13 to 32 strains, or in 12 strains or less. For *S. aureus* the values for similar groups is 100, 21 to 99 and 20 strains or less, respectively. Score medians in all valid identifications ranged from 121.5 to 138.2, while decoy identification had median scores of 59.2 in Mtb and 51.3 in *S. aureus* (red lines).

FIGURE S3 | Multiple TSS identified in protptionn lipV. **(A)** 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 (ITIHGVTEHGR) of the reference sequence, and strain S2666 which TSS at position 22 (HGVTEHGR) is predominant. Two possible TSS choice variants were also observed at high levels for strain S1593. **(B–D)** MS2 spectra for all three peptides mentioned above.

FIGURE S4 | Analysis time for peptide search engines other than Andromeda/MaxQuant. Bars shows, in minutes, the time spent for peptide identification using X!Tandem, OMSSA or Comet, using either the reduced DB1 or the concatenated DB2 databases.

TABLE S1 | Database size increase and pangenome size in 10 selected species.

REFERENCES

Abascal, F., Juan, D., Jungreis, I., Martinez, L., Rigau, M., Rodriguez, J. M., et al. (2018). Loose ends: almost one in five human genes still have unresolved coding status. *Nucleic Acids Res.* 46, 7070–7084. doi: 10.1093/nar/gky587

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1006/jmbi.1990.9999

Baerenfaller, K., Grossmann, J., Grobei, M. A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., et al. (2008). Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* 320, 938–941. doi: 10.1126/science.1157956

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. (2013). GenBank. *Nucleic Acids Res.* 41, D36–D42. doi: 10.1093/nar/gks1195

Blakeley, P., Overton, I. M., and Hubbard, S. J. (2012). Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J. Proteome Res.* 11, 5221–5234. doi: 10.1021/pr300411q

Bocs, S., Cruveiller, S., Vallenet, D., Nuel, G., and Medigue, C. (2003). AMIGene: annotation of MIcrobial genes. *Nucleic Acids Res.* 31, 3723–3726. doi: 10.1093/nar/gkg590

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. (1998). Predicting function: from genes to genomes and back. *J. Mol. Biol.* 283, 707–725. doi: 10.1006/jmbi.1998.2144

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. *Proc. Natl. Acad. Sci. U.S.A.* 105, 21034–21038. doi: 10.1073/pnas.0811066106

Cox, J., Michalski, A., and Mann, M. (2011a). Software lock mass by two-dimensional minimization of peptide mass errors. *J. Am. Soc. Mass Spectrom.* 22, 1373–1380. doi: 10.1007/s13361-011-0142-8

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011b). Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* 10, 1794–1805. doi: 10.1021/pr101065j

Craig, R., and Beavis, R. C. (2004). TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20, 1466–1467. doi: 10.1093/bioinformatics/bth092

Cuklina, J., Hahn, J., Imakaev, M., Omasits, U., Forstner, K. U., Ljubimov, N., et al. (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. doi: 10.1186/s12864-016-2602-9

Davidsen, T., Beck, E., Ganapathy, A., Montgomery, R., Zafar, N., Yang, Q., et al. (2010). The comprehensive microbial resource. *Nucleic Acids Res.* 38, D340–D345. doi: 10.1093/nar/gkp912

de Souza, G. A., Arntzen, M. O., Fortuin, S., Schurch, A. C., Malen, H., Mcevoy, C. R., et al. (2011). Proteogenomic analysis of polymorphisms and gene annotation divergences in prokaryotes using a clustered mass spectrometry-friendly database. *Mol. Cell. Proteom.* 10:M110002527. doi: 10.1074/mcp.M110.002527

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. doi: 10.1093/bioinformatics/btq004

de Souza, G. A., Malen, H., Softeland, T., Saelensminde, G., Prasad, S., Jonassen, I., et al. (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. doi: 10.1186/1471-2164-9-316

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. doi: 10.1002/pmic.200800955

Depke, M., Michalik, S., Rabe, A., Surmann, K., Brinkmann, L., Jehmlich, N., et al. (2015). A peptide resource for the analysis of Staphylococcus aureus in host-pathogen interaction studies. *Proteomics* 15, 3648–3661. doi: 10.1002/pmic.201500091

Ding, W., Baumdicker, F., and Neher, R. A. (2018). panX: pan-genome analysis and exploration. *Nucleic Acids Res.* 46:e5. doi: 10.1093/nar/gkx977

Eng, J. K., Jahan, T. A., and Hoopmann, M. R. (2013). Comet: an open-source MS/MS sequence database search tool. *Proteomics* 13, 22–24. doi: 10.1002/pmic.201200439

Fermin, D., Allen, B. B., Blackwell, T. W., Menon, R., Adamski, M., Xu, Y., et al. (2006). Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol.* 7:R35.

Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., et al. (2004). Open mass spectrometry search algorithm. *J. Proteome Res.* 3, 958–964.

Godoy, D., Randle, G., Simpson, A. J., Aanensen, D. M., Pitt, T. L., Kinoshita, R., et al. (2003). Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*. *J. Clin. Microbiol.* 41, 2068–2079. doi: 10.1128/jcm.41.5.2068-2079.2003

Heyer, R., Schallert, K., Zoun, R., Becher, B., Saake, G., and Benndorf, D. (2017). Challenges and perspectives of metaproteomic data analysis. *J. Biotechnol.* 261, 24–36. doi: 10.1016/j.jbiotec.2017.06.1201

Hoegl, A., Nodwell, M. B., Kirsch, V. C., Bach, N. C., Pfanzelt, M., Stahl, M., et al. (2018). Mining the cellular inventory of pyridoxal phosphate-dependent enzymes with functionalized cofactor mimics. *Nat. Chem.* 10, 1234–1245. doi: 10.1038/s41557-018-0144-2

Jagtap, P., Goslinga, J., Kooren, J. A., Mcgowan, T., Wroblewski, M. S., Seymour, S. L., et al. (2013). A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics* 13, 1352–1357. doi: 10.1002/pmic.201200352

Lew, J. M., Kapopoulou, A., Jones, L. M., and Cole, S. T. (2011). TubercuList–10 years after. *Tuberculosis* 91, 1–7. doi: 10.1016/j.tube.2010.09.008

Liao, B., Ning, Z., Cheng, K., Zhang, X., Li, L., Mayne, J., et al. (2018). iMetaLab 1.0: a web platform for metaproteomics data analysis. *Bioinformatics* 34, 3954–3956. doi: 10.1093/bioinformatics/bty466

Losada, L., Ronning, C. M., Deshazer, D., Woods, D., Fedorova, N., Kim, H. S., et al. (2010). Continuing evolution of Burkholderia mallei through genome reduction and large-scale rearrangements. *Genome Biol. Evol.* 2, 102–116. doi: 10.1093/gbe/evq003

Machado, K. C., Fortuin, S., Tomazella, G. G., Fonseca, A. F., Warren, R., Wiker, H. G., et al. (2019). On the impact of the pangenome and annotation discrepancies while building protein sequence databases for bacteria proteogenomics. *bioRxiv* 378117.

McInerney, J. O., Mcnally, A., and O'connell, M. J. (2017). Why prokaryotes have pangenomes. *Nat. Microbiol.* 2:17040.

Muth, T., Kohrs, F., Heyer, R., Benndorf, D., Rapp, E., Reichl, U., et al. (2018). MPA portable: a stand-alone software package for analyzing metaproteome samples on the go. *Anal. Chem.* 90, 685–689. doi: 10.1021/acs.analchem.7b03544

Muth, T., Renard, B. Y., and Martens, L. (2016). Metaproteomic data analysis at a glance: advances in computational microbial community proteomics. *Expert Rev. Proteom.* 13, 757–769. doi: 10.1080/14789450.2016.1209418

Nesvizhskii, A. I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteom.* 73, 2092–2123. doi: 10.1016/j.jprot.2010.08.009

Nesvizhskii, A. I. (2014). Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* 11, 1114–1125. doi: 10.1038/nmeth.3144

Nesvizhskii, A. I., and Aebersold, R. (2005). Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteom.* 4, 1419–1440. doi: 10.1074/mcp.r500012-mcp200

Neuhauser, N., Nagaraj, N., Mchardy, P., Zanivan, S., Scheltema, R., Cox, J., et al. (2013). High performance computational analysis of large-scale proteome data sets to assess incremental contribution to coverage of the human genome. *J. Proteome Res.* 12, 2858–2868. doi: 10.1021/pr400181q

Olsen, J. V., De Godoy, L. M., Li, G., Macek, B., Mortensen, P., Pesch, R., et al. (2005). Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteom.* 4, 2010–2021. doi: 10.1074/mcp.t500030-mcp200

Omasits, U., Varadarajan, A. R., Schmid, M., Goetze, S., Melidis, D., Bourqui, M., et al. (2017). An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics. *Genome Res.* 27, 2083–2095. doi: 10.1101/gr.218255.116

Overbeek, R., Fonstein, M., D'souza, M., Pusch, G. D., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U.S.A.* 96, 2896–2901. doi: 10.1073/pnas.96.6.2896

Parkhill, J., Sebaihia, M., Preston, A., Murphy, L. D., Thomson, N., and Harris, D. E. (2003). Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat. Genet.* 35, 32–40.

Parkhill, J., Wren, B. W., Mungall, K., Ketley, J. M., Churcher, C., Basham, D., et al. (2000). The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* 403, 665–668. doi: 10.1038/35001088

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. doi: 10.1021/ac026117i

Rasko, D. A., Rosovitz, M. J., Myers, G. S., Mongodin, E. F., Fricke, W. F., Gajer, P., et al. (2008). The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J. Bacteriol.* 190, 6881–6893. doi: 10.1128/JB.00619-08

Renuse, S., Chaerkady, R., and Pandey, A. (2011). Proteogenomics. *Proteomics* 11, 620–630. doi: 10.1002/pmic.201000615

Ruggles, K. V., Krug, K., Wang, X., Clauser, K. R., Wang, J., Payne, S. H., et al. (2017). Methods, tools and current perspectives in proteogenomics. *Mol. Cell. Proteom.* 16, 959–981. doi: 10.1074/mcp.MR117.000024

Schandorff, S., Olsen, J. V., Bunkenborg, J., Blagoev, B., Zhang, Y., Andersen, J. S., et al. (2007). A mass spectrometry-friendly database for cSNP identification. *Nat. Methods* 4, 465–466. doi: 10.1038/nmeth0607-465

Song, H., Hwang, J., Yi, H., Ulrich, R. L., Yu, Y., Nierman, W. C., et al. (2010). The early stage of bacterial genome-reductive evolution in the host. *PLoS Pathog.* 6:e1000922. doi: 10.1371/journal.ppat.1000922

Starr, A. E., Deeke, S. A., Li, L., Zhang, X., Daoud, R., Ryan, J., et al. (2018). Proteomic and metaproteomic approaches to understand host-microbe interactions. *Anal. Chem.* 90, 86–109. doi: 10.1021/acs.analchem.7b04340

Tanca, A., Palomba, A., Deligios, M., Cubeddu, T., Fraumene, C., Biosa, G., et al. (2013). Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. *PLoS One* 8:e82981. doi: 10.1371/journal.pone.0082981

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A genomic perspective on protein families. *Science* 278, 631–637. doi: 10.1126/science.278.5338.631

Thomson, N. R., Holden, M. T., Carder, C., Lennard, N., Lockey, S. J., Marsh, P., et al. (2008). *Chlamydia trachomatis*: genome sequence analysis of lymphogranuloma venereum isolates. *Genome Res.* 18, 161–171. doi: 10.1101/gr.7020108

Tomazella, G. G., Risberg, K., Mylvaganam, H., Lindemann, P. C., Thiede, B., De Souza, G. A., et al. (2012). Proteomic analysis of a multi-resistant clinical *Escherichia coli* isolate of unknown genomic background. *J. Proteom.* 75, 1830–1837. doi: 10.1016/j.jprot.2011.12.024

Vallenet, D., Belda, E., Calteau, A., Cruveiller, S., Engelen, S., Lajus, A., et al. (2013). MicroScope–an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Res.* 41, D636–D647. doi: 10.1093/bib/bbx113

Vaudel, M., Barsnes, H., Berven, F. S., Sickmann, A., and Martens, L. (2011). SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* 11, 996–999. doi: 10.1002/pmic.201000595

Zhang, X., Ning, Z., Mayne, J., Moore, J. I., Li, J., Butcher, J., et al. (2016). MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* 4:31. doi: 10.1186/s40168-016-0176-z

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.