On the impact of the pangenome and annotation discrepancies while building protein sequence databases for bacteria proteogenomics

Machado, K.C.T.[1], Fortuin, S.[2], Tomazella, G.G.[1,3,4], Fonseca, A.F.[1], Warren, R.[2], Wiker, H.G.[3], de Souza S.J.[1,5] and de Souza, G.A.[1,6,*]

[1]Bioinformatics Multidisciplinary Environment, Universidade Federal do Rio Grande do Norte, 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, South Africa.

[3]The Gade Research Group for Infection and Immunity, Department of Clinical Science, University of Bergen, Norway.

[4]The Institute of Bioinformatics and Biotechnology, Natal-RN, Brazil.

[5]The Brain Institute, Universidade Federal do Rio Grande do Norte, Brazil.

[6]Dept. of Biochemistry, Universidade Federal do Rio Grande do Norte, Brazil

***Correspondence**:

Gustavo de Souza
gadsouza@gmail.com

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

Running title: Multi-strain proteomics sequence databases

Abstract

In proteomics, peptide information within mass spectrometry data from a specific organism sample is routinely challenged against a protein sequence database that best represent such organism. However, if the species/strain in the sample is unknown or poorly genetically 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 ten 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 *Bortedella pertusis* generated very complex databases even having low pangenomic complexity or no pangenome at all. This suggests that discrepancies in gene annotation is higher than average between strains of those species. We further tested database performance by using mass spectrometry data from eight clinical strains from *Mycobacterium 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.

59 **Introduction**

60 While genome annotation approaches undertook considerable advances in the past years(Bocs
61 et al., 2003;Davidsen et al., 2010;Vallenet et al., 2013), correct gene prediction is still challenging
62 even for prokaryotes. Particularly, the correct assignment of open reading frames (ORFs), the
63 presence and classification of pseudogenes anddifferences in translational starting site (TSS)(de
64 Souza et al., 2008;de Souza et al., 2009;Cuklina et al., 2016) are the main sources of variability.
65 Consequently, for any given strain or species, its predicted protein sequence information will
66 normally contain inaccuracies that will interfere with peptide identification in mass spectrometry
67 (MS)-based proteomics. The use of accurate protein sequence databases is a key step in most
68 proteomic approaches, and this is particularly critical for studies involving samples containing strains
69 with little to none genetic information or samples containing multiple strains (metaproteomics)
70 (Tanca et al., 2013).

71 In such cases when the establishment of a gold-standard annotation that can represent the
72 sample under investigation is difficult, a viable alternative is to construct customized protein
73 sequence databases which are then inspected against peptide sequence data collected by MS
74 (Nesvizhskii, 2014). This has been largely employed in proteogenomics, where "novel" sequences
75 are inserted in the customized database and, if identified, are further used to validate and confirm
76 proposed gene models and other genetic polymorphisms (for recent reviews see (Renuse et al.,
77 2011;Ruggles et al., 2017)). Database customization is often achieved using two different strategies:
78 i) if the genomic information of the prokaryote under study is known, a 6-frame translation is
79 possible (Fermin et al., 2006;Baerenfaller et al., 2008); ii) if genomic information is unknown, an
80 alternative is to construct a database by merging *ab initio* gene predictions from related strains of the
81 same species, taking into consideration variations caused by SNPs, indels, divergent TSS choice,
82 among others (de Souza et al., 2010;Omasits et al., 2017). These approaches are not mutually

83 exclusive, as gene annotation from related strains can be used to further optimize 6-frame translation
84 approaches (Castellana et al., 2008).

85 However, peptide identification in MS-based proteomics is often performed through
86 probabilistic calculations between the observed peptide fragmentation pattern and theoretical data
87 from a sequence database. Therefore, database size will: i) alter the search space and consequently
88 the probabilistic calculations performed during peptide identification (Nesvizhskii, 2010); ii) make
89 protein inference more difficult, especially in multistrain databases (Nesvizhskii and Aebersold,
90 2005); and iii) demand more computational power due to the handling of larger files. Therefore,
91 building databases using either 6-frame translations or sequence merging approaches which are larger
92 than regular can become, at some point, detrimental for the proteomic analysis (Blakeley et al.,
93 2012).

94 Such issue might as well contribute differently depending on the species under study, how
95 much genomic information is available (number of strains with genome sequenced) and genomic
96 features particular to it. For example, with few sequenced strains at the time, we previously observed
97 that the size of such merged database for *Mycobacterium tuberculosis* (Mtb) complex was not
98 heavily incremented as more strains were considered (de Souza et al., 2011). Each new strain added
99 to the database contributed with only few genetic polymorphisms when compared to the reference
100 H37Rv strain. However, the current 65 Mtb strains with complete genome sequenced demonstrated
101 the existence of a pangenome in this species, i.e., the genome of all strains of a given species contains
102 only a set of genes from a larger pool of accessory genes for that species (Rasko et al.,
103 2008;McInerney et al., 2017). It is then critical to address the impact of such approach in the size of
104 customized databases used in MS-based proteomics, as well as in the performance for peptide
105 identification.

106 Many reports have demonstrated that database size could be controlled by creating protein
107 entries where identical peptides between homologue proteins are reported once and only unique
108 polymorphic tryptic peptides are inserted when merging sequences (Schandorff et al., 2007;Omasits
109 et al., 2017;Liao et al., 2018). It is unclear if genomic features such the presence of a pangenome
110 could impact database customization of specific species. We then decided to test this by creating
111 protein sequence databases for ten species using strains with complete genome sequences and a
112 modified version of MSMSpdbb (de Souza et al., 2010). Database parameters such as the number of
113 entries and unique peptide sequences for each species using increasing number of strains per round
114 were then quantified. As expected, species with large pangenomes such as *Escherichia coli* showed
115 the larger increase in database size per number of strains used, and database size in general correlated
116 with pangenome size. Finally we performed proteomic identification and computational performance
117 oftwo different datasets from: i) eight clinical strains of Mtb using a database with 65 complete strain
118 genomes; and ii) two *S. aureus* datasets (Depke et al., 2015;Hoegl et al., 2018) using a database with
119 194 complete strain genomes.
120

121 **Experimental Procedures**

122 **Species selection**

123 Ten bacterial species were selected for database construction: *Acinetobacter baumannii* (89),
124 *Bordetella pertusis* (348), *Campylobacter jejuni* (114), *Chlamydia trachomatis* (82), *Escherichia coli*
125 (425), *Listeria monocytogenes* (148), *Mycobacterium tuberculosis* (65), *Pseudomonas aeruginosa*
126 (105) and *Staphylococcus aereus* (194). *Burkholderia mallei* and *Burkholderia pseudomallei* (109)
127 are considered genetically very similar (Godoy et al., 2003;Song et al., 2010) and were analyzed
128 together. Number in parenthesis represents strains with complete genome sequenced according to

129 Genbank (Benson et al., 2013) in August 2017. Protein sequences were obtained using the assembly
130 file available at ftp://ftp.ncbi.nih.gov/genomes/genbank/bacteria/.
131
132 **Script design and availability**
133 We redesigned MSMSpdbb (de Souza et al., 2010) in PERL (version 5.24.1) and is present as
134 two modules: all_fasta.pl provides the sequence alignment and creates the outputs with unique entries
135 and homologues; create_bd.pl process the homologues output, and create the final database and the
136 log file. In house BLAST installation (version 2.7.1 or later) is required. All script outputs are saved
137 as txt files in the data folder. Rand.pl is also available to reproduce the experimental design described
138 below.
139
140 **Data processing**
141 Even though MSMSpdbb is already published, since we redesigned how homology is defined
142 (using a Bidirectional Best Hit rather than a Cluster BLAST approach), we provide below further
143 information about the method. To assign gene homology, our script initially gathers the protein
144 sequence data from all strains and then performs pair wise alignments using BLASTP (Altschul et
145 al., 1990) through a Bidirectional Best Hit (BHH) method (Overbeek et al., 1999). The script
146 performs in a manner that the strain with most number of protein entries is initially selected as the
147 query dataset and consequently aligned to the remaining strains sequence datasets (subjects). Two
148 sequences from different strains will be defined as homologues if: i) they are the best hit possible for
149 all alignments performed; ii) sequence identity is equal or higher than 50%; iii) sequence similarity is
150 equal or higher than 70%; and iv) sequence coverage is equal or higher than 70%. These have to be
151 correlated on both directions (BHH) of the alignment.

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

162 Regarding all indexed homologues, our script then defines the longest version of the protein
163 as reference (not necessarily from same strain used as query for the BHH step above). The reference
164 sequence is copied integrally into the partial version of the final output. The script performs an *in*
165 *silico* trypsin digest of the protein sequences, and compare amino acid composition of all generated
166 tryptic peptides. Peptides shorter than 7 or longer than 35 amino acids are excluded. Whenever
167 different tryptic peptide sequences are observed in the non-reference dataset (due to differences of
168 TSS choices, SAAs or indels which result in frame shifts), they are added into the final output under
169 an artificially created protein entry (Omasits et al., 2017). Amino acid changes result from poorly
170 annotated sequences, where X or U are present in the sequence as an amino acid, are not considered.
171 Finally, a log file is created, reporting and classifying all differences observed, describing the amino
172 acid changes and also in which strains those were observed.

173

174 **Database analysis – Experimental design and statistical rationale**

175 To investigate how each dataset contributes to the final database size, for each species we
176 constructed databases using 5, 10, 15, 30 and 65 randomly selected strains. This was performed a
177 total of three times to decrease the chances that a randomly selected strain with very unique features
178 might interfere with the final result. Each database had its MS search space size measured regarding
179 number of entries and number of unique tryptic peptide sequences. Pangenomic size calculations
180 were based on the panX tool (Ding et al., 2018).
181

182 **Mycobacterial Cell Culturing**

183 Stock cultures of Mtb strains TB179, TB861, TB1430, TB1593, TB1659, TB1841, TB1945
184 and TB2666 were inoculated into mycobacterial growth indicator tubes and incubated until positive
185 growth was detected using the Bactec 460 TB system (BD Biosciences). Approximately 0.2 mL was
186 inoculated onto Löwenstein-Jensen medium and incubated over 6 weeks with weekly aeration until
187 colony formation. Colonies were transferred into 20 mL of supplemented 7H9 Middlebrook medium
188 (BD Biosciences) containing 0.2% (v/v) glycerol (Merck Laboratories), 0.1% Tween 80 (Merck
189 Laboratories), and 10% dextrose, catalase. Once the culture reached an A600 of 0.9, one mL was
190 inoculated into 80 mL of supplemented 7H9 Middlebrook medium and incubated until an A600
191 between 0.6 and 0.7 was reached. All steps were performed at 37 °C until Mtb cells in mid-log
192 growth phase were used for whole cell lysate protein extraction.
193

194 **Preparation of Crude Mycobacterial Extracts**

195 Mycobacterial cells were collected by centrifugation (10 min at 2500 x g) at 4°C and
196 resuspended in 1 mL of cold lysis buffer containing 10 mMTris-HCl, pH 7.4 (Merck Laboratories),
197 0.1% Tween 80 (Sigma-Aldrich), one tablet per 25 mL Complete protease inhibitor mixture (Roche
198 Applied Science), and one tablet per 10 mL phosphatase inhibitor mixture (Roche Applied Science).

199 Cells were transferred into 2 mL cryogenic tubes with O-rings, and the pellet was collected after
200 centrifugation (5 min at 6000 x g) at 21 °C. An equal volume of 0.1 mm glass beads (Biospec
201 Products Inc., Bartlesville, OK) was added to the pelleted cells. In addition, 300 µL of cold lysis
202 buffer including 10 µL of 2 units/ml RNase-free DNase I (New England Biolabs) was added, and the
203 cell walls were lysed mechanically by bead beating for 20 s in a Ribolyser (Bio101 Savant, Vista,
204 CA) at a speed of 6.4. Thereafter the cells were cooled on ice for 1 min. The lysis procedure was
205 repeated three times. The lysate was clarified by centrifugation (10,000 x g for 5 min) at 21 °C, and
206 the supernatant containing the whole cell lysate proteins was retained. Thereafter the lysate was
207 filter-sterilized through a 0.22 µm-pore Acrodisc 25 mm PF syringe sterile filter (PallLife Sciences,
208 Pall Corp., Ann Arbor, MI), quantified using the Coomassie Plus Assay Kit (Pierce), and stored at -
209 80 °C until further analysis.
210

211 **Gel Electrophoresis and In-gel Trypsin Digestion**

212 Each whole cell lysate (60 µg) were mixed with electrophoretic sample buffer (NuPAGE kit,
213 Invitrogen, CA) containing 100 mM DTT, and heated for 5 min at 95 °C prior to the electrophoretic
214 run. Proteins were separated in triplicate using a NuPage 4–12% Bis-Tris Gel in 2-N-morpholine
215 ethane sulfonic acid (MES) buffer (Invitrogen) at 200 V for approximately 40 min. Proteins were
216 visualized using a Colloidal Coomassie Novex kit (Invitrogen). After staining, each triplicate was cut
217 into 15 fractions, sliced into smaller pieces and submitted to in gel digestion with trypsin, as
218 previously described (Tomazella et al., 2012). Briefly, the proteins in the gel pieces were reduced
219 using10 mM DTT for 1 h at 56 °C and alkylated with 55 mM iodoacetamide for 45 min at room
220 temperature. The reduced and alkylated proteins were digested with 0.125 µg of trypsin (Sequence
221 grade modified, Promega, WI, USA) for 16 h at 37 °C in 50 mM NH4HCO3, pH 8.0. The reaction
222 was stopped by acidification with 2% trifluoracetic acid (Fluka, Buchs, Germany). The resulting

223 peptide mixture was eluted from the gel slices, and further desalted using RP-C18 STAGE tips
224 (Rappsilber et al., 2003). The peptide mixture was dissolved in 0.1% formic acid 5% acetonitrile for
225 nano-HPLC-MS analysis.
226
227 **LC-MS/MS analysis**
228 Peptides were separated by reversed-phase chromatography in an Acclaim PepMap 100
229 column (C18, 3 µm, 100 Å) (Dionex) capillary of 12 cm bed length and 100 µm inner diameter self-
230 packedwith ReproSilPur C18-AQ (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany), using a
231 Dionex Ultimate 3000 nano-LC system connected to a linear quadrupole ion trap-Orbitrap (LTQ-
232 Orbitrap) mass spectrometer (Thermo Electron, Bremen, Germany) equipped with a nanoelectrospray
233 ion source. Peptides were onto loaded to the column with a flow rate of 0.3 mL/min of 7–40%
234 solvent B in 87 min and then 40–80% solvent B in 8 min. Solvent A was aqueous 2% ACN in 0.1%
235 formic acid, and solvent B was aqueous 90% ACN in 0.1% formic acid.
236 The mass spectrometer was operated in the data-dependent mode to automatically switch
237 between Orbitrap-MS and LTQ-MS/MS acquisition. Survey full-scan MS spectra (from m/z 300 to
238 2000) were acquired in the Orbitrap with resolution of R= 60,000 at m/z 400 (after accumulation to a
239 target of 1,000,000 charges in the LTQ). The method used allowed sequential isolation of the most
240 intense ions (up to six, depending on signal intensity) for fragmentation on the linear ion trap using
241 collisionally induced dissociation at a target value of100,000 charges. For accurate mass
242 measurements, the lock mass option was enabled in MS mode, and the polydimethylcyclosiloxane
243 ions generated in the electrospray process from ambient air were used for internal recalibration
244 during the analysis(Olsen et al., 2005). Target ions already selected for MS/MS were dynamically
245 excluded for 60 s. General mass spectrometry conditions were as follows: electrospray voltage, 1.5

246 kV; no sheath or auxiliary gas flow. Ion selection threshold was 500 counts for MS/MS, an activation
247 Q-value of 0.25 and activation time of 30 ms were also applied for MS/MS.
248
249 **MS Data analysis**
250 Raw MS files of all Mtb and *S. aureus* datasets were analyzed by MaxQuant version 1.5.2.8.
251 Peptides in MS/MS spectra were identified by the Andromeda search engine (Cox et al., 2011b). For
252 Mtb dataset, MS files were searched against a database containing all sequences from 65 strains
253 without any processing (263,683 entries), or the database generated by our tool using the same 65
254 strains (15,979 entries). The *S. aureus* dataset was searched against a database containing all
255 sequences from 194 strains without any processing (523,281 entries) or the MSMSpdbb generated
256 database (15,073 entries). All bacterial sequences were obtained from Genbank as written above
257 (August 2017). MaxQuant analysis included an initial search with a precursor mass tolerance of 20
258 ppm which were used for mass recalibration (Cox et al., 2011a). Trypsin without proline restriction
259 was used as enzyme specificity, with two allowed miscleavages and no unspecific cleavage allowed.
260 In the main Andromeda search precursor mass and fragment mass were searched with initial mass
261 tolerance of 6 ppm and 0.5 Da, respectively. The search included variable modifications of oxidation
262 of methionine and protein N-terminal acetylation. Carbamidomethyl cysteine was added as a fixed
263 modification. Minimal peptide length was set to 7 amino acids. The false discovery rate (FDR) was
264 set to 0.01 for peptide and protein identifications, and since the aim is to compare database
265 performance, no further score or PEP thresholds were used for individual MS spectra. In the case of
266 identified peptides that are shared between two proteins, these are combined and reported as one
267 protein group. Protein and peptide datasets were filtered to eliminate the identifications from the
268 reverse database and from common contaminants.
269

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; non-reference peptides are now inserted as new entries 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. Nine of the selected species have pangenomes that 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) (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 a pair wise comparison is performed with the remaining subject strains. Unique entries from the query strain (i.e., sequences with no significant hit to any other protein entries in subject strains) are then 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 then used as a reference (regardless of which strain it 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 do not

294 match any of the reference sequence (due to a different TSS choice or a SNP for example), a new
295 protein entry is created in the final database. This new entry will contain all non-reference tryptic
296 peptides that do not match the reference sequence. This process will be repeated until no strains are
297 left to be used as query, or until no protein entries are left in yet-to-be selected query strains. A log
298 file is also created, reporting all entries that were matched and classifying all tryptic peptides which
299 were added to the final database, showing the type of event (TSS choice, SNP etc) and in case of
300 SNP, the resulting amino acid substitution.

301 Figure 2 illustrates a typical homologue comparison leading to the detection of polymorphic
302 peptides, which will be added to the database as an artificial entry. The hypothetical protein
303 AEM98552 (Rv0104 on the Tuberculist annotation (Lew et al., 2011)) was set as reference, and 9
304 other different variants were observed. Those variants are the result of the different combinations
305 including: two additional TSS choices at position 9 and 118; three SAPs at positions 402 (Tyr →
306 His), 477 (Gly→Arg) and 493 (Leu→ Val); a not-classified genetic modification that leads to
307 alternate amino acid sequence after position 475 (yellow box, named as frameshift/indel); and finally,
308 two polymorphisms leading to premature stop codons and generating different c-terminal tryptic
309 peptides ending at position 378 and 426. For example, the variant from strain CAS/NITR204
310 (AGL25567) contains two SAPs, while the variant from strain CCDC5180 (AEJ48929) has a
311 different TSS choice at position 9 of reference, and a premature stop codon, truncating the protein at
312 position 378. Unique tryptic peptides characterizing all such polymorphisms are added in the
313 database as an additional entry.

314

315 **Database size increase ratio per species**

316 Databases were created for ten species to measure if the rate of increase in the database size
317 could be impacted by unique features of each genome, such as the size of a pangenome. For a fair
318 comparison among species, we built databases using a maximum of 65 strains, since at the time Mtb

319 had only 65 strains with complete sequenced genomes available. To assist the visualization of how
320 the database size is incremented as more strains are used, we also created databases using 5, 10, 20
321 and 30 randomly selected strains of each species. And to avoid outlier behavior in casea strain with
322 very unique annotation being randomly selected, every database was created three times in total. The
323 rand.pl script provided with the tool performs all such steps automatically, if the data needs to be
324 reproduced. We then counted total number of protein entries and tryptic peptides in each database.

325 Supplementary Table 1 shows all values measured for all tests including: number of proteins
326 (Supplementary Figure S1), number of annotated proteins (excluding artificial entries created to
327 accommodate polymorphisms) (Figure 3 and Table I), and number of peptides with or without
328 miscleavages allowed (Supplementary Figure S2). Figure 3 shows graphically the increase in number
329 of annotated proteins in all species, and Table I describe number of annotated proteins in the database
330 using 65 strains, rate of database size increase, number of annotated proteins added per strain on
331 average, and size of pangenome for each species. As expected, the increase rate in the number of
332 annotated proteins in the final database correlated with the size of the gene pool in the pangenome. *E.*
333 *coli,* which has the largest genetic pool described, had indeed the larger database increase rate of all
334 species, at 4.6x increase compared to the average number of genes per strain. The three exceptions
335 for this were: *B. pseudomallei/mallei* with the 2nd increase rate (3.53x) even though the pangenome
336 size of *B. pseudomallei* is the 4th most complex (no data is available regarding *B. mallei* pangenome);
337 Mtb, which has the 2nd less complex pangenome size, representing only 21% of total genetic pool,
338 yet it has 4th database increase rate (2.83x), surpassing *P. aeruginosa* which has a larger genome size
339 and a bigger genetic pool; and *B. pertussis*, with no pangenome described so far and yet with a
340 database increase rate (2.46x) which surpassed three species with characterized pangenomes.

341

342 **Proteomic performance in diverse MS datasets**

343 When we first developed the MSMSpdbb approach (de Souza et al., 2010), the number of
344 genomic information available was a fraction of the amount currently in existence. For example, the
345 analysis validation performed on an Mtb dataset used a database with only 8 strains (5 from *M.*
346 *tuberculosis* and 3 from *M. bovis*). We then wondered if the approach would still be valuable as
347 hundreds of genome information is available to a single species. This was tested by performing
348 proteomic analysis in two independent datasets.

349 The MSMSpdbb processed database created by the tool contained 15,996 (Mtb), 15,073 (*S.*
350 *aureus*) protein entries. In comparison, merely concatenated databases using all sequences from the
351 available strains for each species created a database with 263,683 protein entries for Mtb and 523,281
352 protein entries for *S. aureus*. From now on, all MSMSpdbb processed databases are called DB1, and
353 simply concatenated databases are named DB2 for simplicity. The MS datasets were then challenged
354 for peptide identification using the respective DB1 and DB2 databases for each species. It is
355 important to note that all theoretical unique tryptic peptides present in DB2 are also present in DB1
356 (data not show). This means that MSMSpdbb did not lose peptide sequences during its database
357 construction, and that the differences observed below regarding number of identified peptides is not
358 due to missing information in DB1. The complete results folders from each MaxQuant search
359 including protein and peptide lists with relevant identification parameters are provided (see Data
360 availability section).

361 In all datasets, when comparing the results from DB1 and DB2 searches, the number of
362 peptides identified was very similar in both searches (Figure 4). For example, the Mtb dataset
363 identified 32,986 unique peptides in both searches, while 515 peptides were identified only in the
364 DB1 search and 1,116 peptides only in the DB2 search result (Figure 4A). When all peptides are
365 considered, 95.28% of the peptides were identified in both searches. For *S. aureus,* DB1 and DB2
366 searches were in agreement for 92% of the identified peptides in all searches combined. We observed

367 that the majority of such database exclusive identifications are distributed in the lower range of the
368 spectra scoring. For all datasets, the median MaxQuant score for peptides identified in only DB1 and
369 DB2 were in the range of 52 to 75, while the median scores of peptides identified both searches was
370 from 128-137 (see histograms Figure 4). If a MaxQuant posterior error probability (PEP) score lower
371 than 0.01 is also considered together with regular MS2 score, the numbers of exclusive peptides drop
372 by half in DB1 and DB2 for all datasets considered (data not show).

373 Multiple TSS polymorphisms for the same protein were also observed even for the same
374 strains. Figure S3 shows the position of three peptides; all predicted as possible N-terminal peptides
375 of the protein lipase lipV. MIIDLHVQR (Score 93.4, PEP 0.0019, Figure S3B) represents the most
376 5' prediction; LTIHGVTEHGR (Score 85.9, PEP 0.0005, Figure S3C) starts at position 19 of
377 reference protein and HGVTEHGR (Score 100, PEP 0.0077, Figure S3D) starts at position 22 of
378 reference. The longer form is the most abundant in almost all samples, except S1430 where only
379 LTIHGVTEHGR was identified and S2666 where the N-terminal HGVTEHGR is slightly
380 overrepresented than MIIDLHVQR. Finally, the DB1 results also included 5 polymorphic peptides
381 from the artificial entries, which could not be found in the DB2 database. Closer inspection of these
382 peptides revealed that they all contain missing cleavage sites. For example, the identification of
383 peptide GEGAAVVVIKEGSIGGGTVFVYSGR from gene PpsC includes two polymorphic peptides
384 that were concatenated in DB1, GEGAAVVVIK corresponding to positions 274-283 in the reference
385 protein, and EGSIGGGTVFVYSGR corresponding to positions 564-578. Clearly, the artificial
386 entries in this case contain two concatenated polymorphic peptides which are not adjacent in the
387 reference protein. The fact that we accept miscleavages as a parameter during the search allows a
388 clearly false-positive identification for artificial entries. The spectra and identification score for this
389 concatenated peptide is given in Supplementary Figure S7, and as expected, the identified spectra has
390 poor score and sequence coverage. Supplementary Table 2 lists all five false-positive peptides.

391

392 **Database size impact in computational performance**

393 As shown before, database size reduction from DB2 to DB1 was approximately 16x to Mtb
394 and 35x to *S. aureus*. By inspecting MaxQuant log folder we calculated the amount of minutes used
395 for each step in its proteomics pipeline. For Mtb, the time spent for the whole pipeline was reduced
396 by half, from 853 minutes for DB2 to 434 minutes to DB1 (Figure 5). Main differences were
397 observed for steps related to the peptide search, i.e., the 'main search' (5x) and the 'second peptide
398 search' (4x) steps, and during the creation of results and other final files named 'writing tables'
399 (3.5x). For *S. aureus* the differences were far more evident, as the complete pipeline took 7,553
400 minutes for DB2 and 2,147 minutes for DB1. Time reduction in peptide search steps were similar to
401 Mtb at 4.5x and 4x respectively. But in addition to those, other steps which are related to read and
402 finish the main and the second search results had improvements of 42 and 19.6 times respectively.
403 Steps related to assignment and assembly of protein groups had improvements of 42 times, and
404 'writing tables' had a time improvement of 10.4 times.

405

406 **Discussion**

407 Database size is a known bottleneck in proteomics, and its implications had been already
408 discussed in proteogenomics and metaproteomics research (Jagtap et al., 2013;Heyer et al., 2017).
409 Therefore, controlling database information is of key importance. A manner to achieve this is to
410 construct customized databases where identical sequences/tryptic peptides from homologue proteins
411 are not exhaustively present in different entries within the database. We had already applied this for
412 the characterization of samples with unknown genetic background such as bacterial clinical strains
413 (de Souza et al., 2011;Tomazella et al., 2012). Such databases allow the validation of coding regions
414 and confirmation of sequence polymorphisms normally omitted from reference databases. This is a
415 critical characterization, considering that even genomes with two decades of considerable

416 investigation still have gene annotation issues (Abascal et al., 2018). It is true even for prokaryotes
417 (de Souza et al., 2008;de Souza et al., 2009), with their simpler genomic structure and higher coding
418 density. However, as the amount of genomic information exponentially increases with time, we
419 wished to further evaluate the impact of such approach with the current amount of available genomic
420 data.

421 First, we were curious to investigate how different species with diverse genetic features (such
422 as the presence of a pangenome) would impact database size and structure. As expected, database
423 size increase correlated with the pangenome complexity of each species. Surprisingly Mtb and *B.*
424 *pertussis* databases size increased at rates higher than expected based on the low complexity of their
425 core and pangenomic genes. Previous data from Mtb suggested that database size increment was not
426 heavy, when at the time only five Mtb strains and three strains of *M. bovis* were considered (de Souza
427 et al., 2011). Our data here however shows that Mtb database size increment was one of the most
428 prominent, arguably the most relevant if its small pangenome is considered. This suggests that as
429 additional strains were sequenced, variations in gene annotation performed were more acute then
430 observed for the other nine species. Something similar could be assumed about *B. pertusis* databases,
431 which has no pangenome and database increase was higher than three species with characterized
432 pangenomes. However, *C. trachomatis, L. monocytogenes* and *C. jejuni* have smaller genomes than
433 compared to *B. pertussis*, all also having higher density of coding regions (Parkhill et al.,
434 2000;Parkhill et al., 2003;Thomson et al., 2008). More nucleotide regions in the genome marked as
435 non-coding will offer more opportunity for different strategies to generate conflicting gene
436 annotation data, which could explain *B. pertussis* database size increase in our analysis. The database
437 size increase for *Burkholderia pseudomallei* and *Burkholderia mallei* was not taken into
438 consideration by us on this analysis, because while we decided to merge both species as one based on

439 genotyping data (Godoy et al., 2003), large scale rearrangements in *B. mallei* (Losada et al., 2010)
440 might be interfering with our approach.

441 We then selected two MS datasets from Mtb and from *S. aureus* to identify peptides using a
442 routine probabilistic-based approach, but now using the complete set of available genomes (194
443 strains) for *S. aureus*. Overall, DB1 and DB2 performance was very similar, and differences were
444 negligible and mostly could be ignored to their low scoring nature. A minority of valid identifications
445 exclusive to DB2 were N-terminal peptides containing miscleavages that were not selected by
446 MSMSpdbb (i.e., the peptides without miscleavages were shorter than 7 amino acids and not present
447 in the selected reference protein) and therefore were absent in DB1 due to the parameters used for
448 DB1 construction. It is worth mentioning that surprisingly, we have identified multiple N-terminal
449 predictions for the same protein, in some cases observed in the same strain (Supplementary Figure 3).
450 While normally the identification of a protein N-terminal peptide in proteogenomics is used to
451 validate and confirm the TSS of the protein, our data suggest that excluding the remaining TSS
452 choices from the database might be undesirable, considering they might be later identified when
453 additional strains are analyzed by MS.

454 The advantage of using a concatenated database is evident when computational performance
455 is measured. The main steps in Andromeda/MaxQuant pipeline which involve Fasta reading ('main'
456 and 'second peptide' searches) were similarly improved in both datasets tested. In *S. aureus* dataset,
457 where DB1 concatenation is larger when compared to DB2 (from approximately 500,000 entries to
458 15,000 entries), time processing improvements were observed in additional stages of the pipeline.
459 Particularly steps related to handle of peptide search engine results and assembly of the protein
460 groups to be reported as final protein identifications. The step which creates the final output files
461 ('Writing tables'), even though do not directly use the Fasta file, was also significantly improved.
462 This makes sense, as MaxQuant will report all entries from the database that share identical peptides.

463 Entries from a protein with high sequence identity present in all 194 strains for *S. aureus*, for
464 example, will have all 194 entries recorded in the output if that protein is identified. For DB1, since
465 those 194 highly identical sequences are concatenated into a single reference entry, the output
466 reporting is vastly simplified.

467 As mass spectrometry evolves and as detection sensitivity and peptide identification coverage
468 improves, one of the last bottlenecks in metaproteomics research is database design. Building
469 databases for metaproteomics is often achieved by two approaches: either by collecting metagenomic
470 or metatranscriptomic data from the sample; or by concatenating protein sequences from a public
471 repository such as Uniprot (for recent reviews see (Heyer et al., 2017;Starr et al., 2018)). Regarding
472 the metagenomic approach, even with the advances of next-generation sequencing methods, reads
473 and contigs assembly in complex samples is still challenging (Heyer et al., 2017). On the other hand,
474 concatenating protein sequences from multiple organisms from a public repository results in very
475 large databases, making the analysis time costly and susceptive to high false-positive rates (Tanca et
476 al., 2013). False-positive issues had been mostly solved by database size restrictions using a multi-
477 step peptide search or multiple peptide searches using smaller datasets of refined sequences on each
478 search and further analysis to combine results (Jagtap et al., 2013;Muth et al., 2016;Zhang et al.,
479 2016;Liao et al., 2018;Muth et al., 2018). While those are successful methods, they are mostly
480 provided as complete pipelines. We believe that providing solutions solely at Fasta file level offers
481 more freedom to researchers, as they can use whichever peptide search engines or pipelines they
482 prefer. We predict that controlling database size such as done by MSMSpdbb without losing peptide
483 identification coverage and drastically reducing computational processing time will be of key
484 importance for metaproteomics.

485

486 **Data availability**

487 Scripts are fully available at https://github.com/karlactm/Proteogenomics. All Mtb mass spectrometry
488 files, and MaxQuant search parameters and results are available at ProteomeXchange with the
489 accession number PXD011080. *S. aureus* mass spectrometry files are found in ProteomeXchange as
490 PXD006483 (Hoegl et al., 2018) and PXD000702 (Depke et al., 2015).
491

492 **Acknowledgements**

493 This work was supported by: grant 23038.004629/2014-19 to SJS and GAS from the Coordenação de
494 Aperfeiçoamento de Pessoal do Ensino Superior; grants 305233/2015-7 to SJS and 406630/2016-0 to
495 GAS from Conselho Nacional de DesenvolvimentoCientífico e Tecnológico (CNPq); and grants
496 175141 (HWG) and 183418 (HGW/RW) from the Norwegian Research Council. The manuscript has
497 been released as a Pre-Print at bioRxiv with doi: https://doi.org/10.1101/378117.
498

499 **Authors' contribution**

500 KCTM designed the tool and performed bioinformatic analysis; SF and GGT generated Mtb samples
501 and proteomics data; RW, HGW, SJS and GAS supervised work and data analysis. GAS provided
502 *S.aureus* proteomic analysis. KCTM and GAS designed experiments and analysis.
503

504 **Disclosure declaration**

505 All authors declare that there are no conflicts of interest associated to this research.
506

507 **References**

508 Abascal, F., Juan, D., Jungreis, I., Martinez, L., Rigau, M., Rodriguez, J.M., Vazquez, J., and Tress, M.L. (2018).
509 Loose ends: almost one in five human genes still have unresolved coding status. *Nucleic Acids Res*.
510 Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J*
511 *Mol Biol* 215**,** 403-410.
512 Baerenfaller, K., Grossmann, J., Grobei, M.A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., Zimmermann, P.,
513 Grossniklaus, U., Gruissem, W., and Baginsky, S. (2008). Genome-scale proteomics reveals
514 Arabidopsis thaliana gene models and proteome dynamics. *Science* 320**,** 938-941.

515 Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2013).
516 GenBank. *Nucleic Acids Res* 41**,** D36-42.
517 Blakeley, P., Overton, I.M., and Hubbard, S.J. (2012). Addressing statistical biases in nucleotide-derived
518 protein databases for proteogenomic search strategies. *J Proteome Res* 11**,** 5221-5234.
519 Bocs, S., Cruveiller, S., Vallenet, D., Nuel, G., and Medigue, C. (2003). AMIGene: Annotation of MIcrobial
520 Genes. *Nucleic Acids Res* 31**,** 3723-3726.
521 Castellana, N.E., Payne, S.H., Shen, Z., Stanke, M., Bafna, V., and Briggs, S.P. (2008). Discovery and revision of
522 Arabidopsis genes by proteogenomics. *Proc Natl Acad Sci U S A* 105**,** 21034-21038.
523 Cox, J., Michalski, A., and Mann, M. (2011a). Software lock mass by two-dimensional minimization of peptide
524 mass errors. *J Am Soc Mass Spectrom* 22**,** 1373-1380.
525 Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., and Mann, M. (2011b). Andromeda: a
526 peptide search engine integrated into the MaxQuant environment. *J Proteome Res* 10**,** 1794-1805.
527 Cuklina, J., Hahn, J., Imakaev, M., Omasits, U., Forstner, K.U., Ljubimov, N., Goebel, M., Pessi, G., Fischer,
528 H.M., Ahrens, C.H., Gelfand, M.S., and Evguenieva-Hackenberg, E. (2016). Genome-wide
529 transcription start site mapping of Bradyrhizobium japonicum grown free-living or in symbiosis - a
530 rich resource to identify new transcripts, proteins and to study gene regulation. *BMC Genomics* 17**,**
531 302.
532 Davidsen, T., Beck, E., Ganapathy, A., Montgomery, R., Zafar, N., Yang, Q., Madupu, R., Goetz, P., Galinsky, K.,
533 White, O., and Sutton, G. (2010). The comprehensive microbial resource. *Nucleic Acids Res* 38**,** D340-
534 345.
535 De Souza, G.A., Arntzen, M.O., Fortuin, S., Schurch, A.C., Malen, H., Mcevoy, C.R., Van Soolingen, D., Thiede,
536 B., Warren, R.M., and Wiker, H.G. (2011). Proteogenomic analysis of polymorphisms and gene
537 annotation divergences in prokaryotes using a clustered mass spectrometry-friendly database. *Mol*
538 *Cell Proteomics* 10**,** M110 002527.
539 De Souza, G.A., Arntzen, M.O., and Wiker, H.G. (2010). MSMSpdbb: providing protein databases of closely
540 related organisms to improve proteomic characterization of prokaryotic microbes. *Bioinformatics* 26**,**
541 698-699.
542 De Souza, G.A., Malen, H., Softeland, T., Saelensminde, G., Prasad, S., Jonassen, I., and Wiker, H.G. (2008).
543 High accuracy mass spectrometry analysis as a tool to verify and improve gene annotation using
544 Mycobacterium tuberculosis as an example. *BMC Genomics* 9**,** 316.
545 De Souza, G.A., Softeland, T., Koehler, C.J., Thiede, B., and Wiker, H.G. (2009). Validating divergent ORF
546 annotation of the Mycobacterium leprae genome through a full translation data set and peptide
547 identification by tandem mass spectrometry. *Proteomics* 9**,** 3233-3243.
548 Depke, M., Michalik, S., Rabe, A., Surmann, K., Brinkmann, L., Jehmlich, N., Bernhardt, J., Hecker, M.,
549 Wollscheid, B., Sun, Z., Moritz, R.L., Volker, U., and Schmidt, F. (2015). A peptide resource for the
550 analysis of Staphylococcus aureus in host-pathogen interaction studies. *Proteomics* 15**,** 3648-3661.
551 Ding, W., Baumdicker, F., and Neher, R.A. (2018). panX: pan-genome analysis and exploration. *Nucleic Acids*
552 *Res* 46**,** e5.
553 Fermin, D., Allen, B.B., Blackwell, T.W., Menon, R., Adamski, M., Xu, Y., Ulintz, P., Omenn, G.S., and States,
554 D.J. (2006). Novel gene and gene model detection using a whole genome open reading frame
555 analysis in proteomics. *Genome Biol* 7**,** R35.
556 Godoy, D., Randle, G., Simpson, A.J., Aanensen, D.M., Pitt, T.L., Kinoshita, R., and Spratt, B.G. (2003).
557 Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis
558 and glanders, Burkholderia pseudomallei and Burkholderia mallei. *J Clin Microbiol* 41**,** 2068-2079.
559 Heyer, R., Schallert, K., Zoun, R., Becher, B., Saake, G., and Benndorf, D. (2017). Challenges and perspectives
560 of metaproteomic data analysis. *J Biotechnol* 261**,** 24-36.
561 Hoegl, A., Nodwell, M.B., Kirsch, V.C., Bach, N.C., Pfanzelt, M., Stahl, M., Schneider, S., and Sieber, S.A.
562 (2018). Mining the cellular inventory of pyridoxal phosphate-dependent enzymes with functionalized
563 cofactor mimics. *Nat Chem* 10**,** 1234-1245.

564 Jagtap, P., Goslinga, J., Kooren, J.A., Mcgowan, T., Wroblewski, M.S., Seymour, S.L., and Griffin, T.J. (2013). A
565 two-step database search method improves sensitivity in peptide sequence matches for
566 metaproteomics and proteogenomics studies. *Proteomics* 13**,** 1352-1357.
567 Lew, J.M., Kapopoulou, A., Jones, L.M., and Cole, S.T. (2011). TubercuList--10 years after. *Tuberculosis (Edinb)*
568 91**,** 1-7.
569 Liao, B., Ning, Z., Cheng, K., Zhang, X., Li, L., Mayne, J., and Figeys, D. (2018). iMetaLab 1.0: a web platform for
570 metaproteomics data analysis. *Bioinformatics* 34**,** 3954-3956.
571 Losada, L., Ronning, C.M., Deshazer, D., Woods, D., Fedorova, N., Kim, H.S., Shabalina, S.A., Pearson, T.R.,
572 Brinkac, L., Tan, P., Nandi, T., Crabtree, J., Badger, J., Beckstrom-Sternberg, S., Saqib, M., Schutzer,
573 S.E., Keim, P., and Nierman, W.C. (2010). Continuing evolution of Burkholderia mallei through
574 genome reduction and large-scale rearrangements. *Genome Biol Evol* 2**,** 102-116.
575 Mcinerney, J.O., Mcnally, A., and O'connell, M.J. (2017). Why prokaryotes have pangenomes. *Nat Microbiol*
576 2**,** 17040.
577 Muth, T., Kohrs, F., Heyer, R., Benndorf, D., Rapp, E., Reichl, U., Martens, L., and Renard, B.Y. (2018). MPA
578 Portable: A Stand-Alone Software Package for Analyzing Metaproteome Samples on the Go. *Anal*
579 *Chem* 90**,** 685-689.
580 Muth, T., Renard, B.Y., and Martens, L. (2016). Metaproteomic data analysis at a glance: advances in
581 computational microbial community proteomics. *Expert Rev Proteomics* 13**,** 757-769.
582 Nesvizhskii, A.I. (2010). A survey of computational methods and error rate estimation procedures for peptide
583 and protein identification in shotgun proteomics. *J Proteomics* 73**,** 2092-2123.
584 Nesvizhskii, A.I. (2014). Proteogenomics: concepts, applications and computational strategies. *Nat Methods*
585 11**,** 1114-1125.
586 Nesvizhskii, A.I., and Aebersold, R. (2005). Interpretation of shotgun proteomic data: the protein inference
587 problem. *Mol Cell Proteomics* 4**,** 1419-1440.
588 Neuhauser, N., Nagaraj, N., Mchardy, P., Zanivan, S., Scheltema, R., Cox, J., and Mann, M. (2013). High
589 performance computational analysis of large-scale proteome data sets to assess incremental
590 contribution to coverage of the human genome. *J Proteome Res* 12**,** 2858-2868.
591 Olsen, J.V., De Godoy, L.M., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S.,
592 and Mann, M. (2005). Parts per million mass accuracy on an Orbitrap mass spectrometer via lock
593 mass injection into a C-trap. *Mol Cell Proteomics* 4**,** 2010-2021.
594 Omasits, U., Varadarajan, A.R., Schmid, M., Goetze, S., Melidis, D., Bourqui, M., Nikolayeva, O., Quebatte, M.,
595 Patrignani, A., Dehio, C., Frey, J.E., Robinson, M.D., Wollscheid, B., and Ahrens, C.H. (2017). An
596 integrative strategy to identify the entire protein coding potential of prokaryotic genomes by
597 proteogenomics. *Genome Res* 27**,** 2083-2095.
598 Overbeek, R., Fonstein, M., D'souza, M., Pusch, G.D., and Maltsev, N. (1999). The use of gene clusters to infer
599 functional coupling. *Proc Natl Acad Sci U S A* 96**,** 2896-2901.
600 Parkhill, J., Sebaihia, M., Preston, A., Murphy, L.D., Thomson, N., Harris, D.E., Holden, M.T., Churcher, C.M.,
601 Bentley, S.D., Mungall, K.L., Cerdeno-Tarraga, A.M., Temple, L., James, K., Harris, B., Quail, M.A.,
602 Achtman, M., Atkin, R., Baker, S., Basham, D., Bason, N., Cherevach, I., Chillingworth, T., Collins, M.,
603 Cronin, A., Davis, P., Doggett, J., Feltwell, T., Goble, A., Hamlin, N., Hauser, H., Holroyd, S., Jagels, K.,
604 Leather, S., Moule, S., Norberczak, H., O'neil, S., Ormond, D., Price, C., Rabbinowitsch, E., Rutter, S.,
605 Sanders, M., Saunders, D., Seeger, K., Sharp, S., Simmonds, M., Skelton, J., Squares, R., Squares, S.,
606 Stevens, K., Unwin, L., Whitehead, S., Barrell, B.G., and Maskell, D.J. (2003). Comparative analysis of
607 the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella
608 bronchiseptica. *Nat Genet* 35**,** 32-40.
609 Parkhill, J., Wren, B.W., Mungall, K., Ketley, J.M., Churcher, C., Basham, D., Chillingworth, T., Davies, R.M.,
610 Feltwell, T., Holroyd, S., Jagels, K., Karlyshev, A.V., Moule, S., Pallen, M.J., Penn, C.W., Quail, M.A.,
611 Rajandream, M.A., Rutherford, K.M., Van Vliet, A.H., Whitehead, S., and Barrell, B.G. (2000). The
612 genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable
613 sequences. *Nature* 403**,** 665-668.

614 Rappsilber, J., Ishihama, Y., and Mann, M. (2003). Stop and go extraction tips for matrix-assisted laser
615 desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem*
616 75**,** 663-670.
617 Rasko, D.A., Rosovitz, M.J., Myers, G.S., Mongodin, E.F., Fricke, W.F., Gajer, P., Crabtree, J., Sebaihia, M.,
618 Thomson, N.R., Chaudhuri, R., Henderson, I.R., Sperandio, V., and Ravel, J. (2008). The pangenome
619 structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic
620 isolates. *J Bacteriol* 190**,** 6881-6893.
621 Renuse, S., Chaerkady, R., and Pandey, A. (2011). Proteogenomics. *Proteomics* 11**,** 620-630.
622 Ruggles, K.V., Krug, K., Wang, X., Clauser, K.R., Wang, J., Payne, S.H., Fenyo, D., Zhang, B., and Mani, D.R.
623 (2017). Methods, Tools and Current Perspectives in Proteogenomics. *Mol Cell Proteomics* 16**,** 959-
624 981.
625 Schandorff, S., Olsen, J.V., Bunkenborg, J., Blagoev, B., Zhang, Y., Andersen, J.S., and Mann, M. (2007). A mass
626 spectrometry-friendly database for cSNP identification. *Nat Methods* 4**,** 465-466.
627 Song, H., Hwang, J., Yi, H., Ulrich, R.L., Yu, Y., Nierman, W.C., and Kim, H.S. (2010). The early stage of bacterial
628 genome-reductive evolution in the host. *PLoS Pathog* 6**,** e1000922.
629 Starr, A.E., Deeke, S.A., Li, L., Zhang, X., Daoud, R., Ryan, J., Ning, Z., Cheng, K., Nguyen, L.V.H., Abou-Samra,
630 E., Lavallee-Adam, M., and Figeys, D. (2018). Proteomic and Metaproteomic Approaches to
631 Understand Host-Microbe Interactions. *Anal Chem* 90**,** 86-109.
632 Tanca, A., Palomba, A., Deligios, M., Cubeddu, T., Fraumene, C., Biosa, G., Pagnozzi, D., Addis, M.F., and
633 Uzzau, S. (2013). Evaluating the impact of different sequence databases on metaproteome analysis:
634 insights from a lab-assembled microbial mixture. *PLoS One* 8**,** e82981.
635 Thomson, N.R., Holden, M.T., Carder, C., Lennard, N., Lockey, S.J., Marsh, P., Skipp, P., O'connor, C.D.,
636 Goodhead, I., Norbertzcak, H., Harris, B., Ormond, D., Rance, R., Quail, M.A., Parkhill, J., Stephens,
637 R.S., and Clarke, I.N. (2008). Chlamydia trachomatis: genome sequence analysis of lymphogranuloma
638 venereum isolates. *Genome Res* 18**,** 161-171.
639 Tomazella, G.G., Risberg, K., Mylvaganam, H., Lindemann, P.C., Thiede, B., De Souza, G.A., and Wiker, H.G.
640 (2012). Proteomic analysis of a multi-resistant clinical Escherichia coli isolate of unknown genomic
641 background. *J Proteomics* 75**,** 1830-1837.
642 Vallenet, D., Belda, E., Calteau, A., Cruveiller, S., Engelen, S., Lajus, A., Le Fevre, F., Longin, C., Mornico, D.,
643 Roche, D., Rouy, Z., Salvignol, G., Scarpelli, C., Thil Smith, A.A., Weiman, M., and Medigue, C. (2013).
644 MicroScope--an integrated microbial resource for the curation and comparative analysis of genomic
645 and metabolic data. *Nucleic Acids Res* 41**,** D636-647.
646 Zhang, X., Ning, Z., Mayne, J., Moore, J.I., Li, J., Butcher, J., Deeke, S.A., Chen, R., Chiang, C.K., Wen, M., Mack,
647 D., Stintzi, A., and Figeys, D. (2016). MetaPro-IQ: a universal metaproteomic approach to studying
648 human and mouse gut microbiota. *Microbiome* 4**,** 31.

649

650

651

652

653

654

655

656

657

658

659

660 **Tables**

661 **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*						

662

663

664 **Figure Legends**

665 **Figure 1 – Building concatenated databases from related strains.** (A) For a given set of fasta
666 files, a Query file is selected and BLASTP is performed against the remaining files. A bidirectional
667 approach is performed. (B) Uniquely annotated entries in the Query are defined and moved to the
668 final fasta database (D), while entries where homologues were found in the subject files are indexed
669 for further processing. (C) For each homologue group, a reference sequence is chosen, and
670 polymorphic peptides are added into a customized entry ("Artificial Seq"), both being copied into the
671 final database. (E) The Query file is removed from the dataset, the remaining subject files will have
672 entries removed (proteins which were already defined as a homologue of a Query protein), and a new
673 fasta file is selected as Query for a new BLASTP round. Such rounds will be performed until all files
674 are analyzed.

675

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

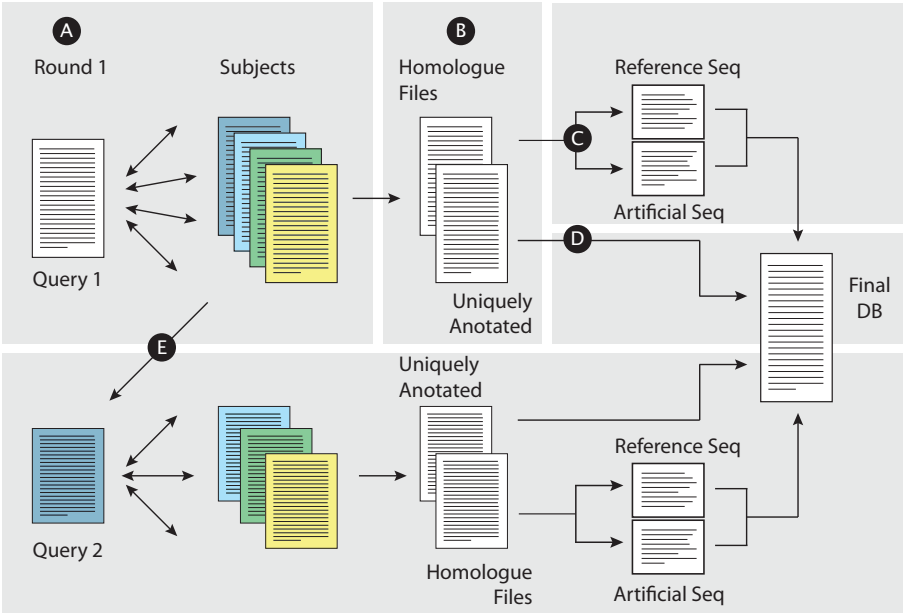
679

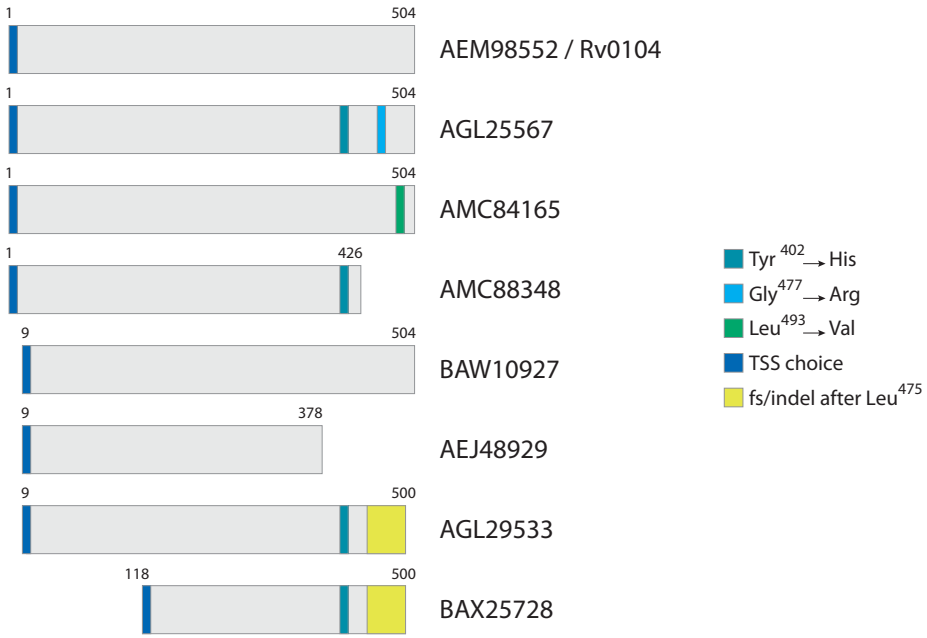
680 **Figure 3 – Number of entries per database after redundancy removal**. The graph shows data
681 when only considering reference and uniquely annotated sequences in all strains, i.e., customized
682 entries containing polymorphic peptides where not counted. Values in x axis show the number of
683 strains used for database creation. All values plotted in the graph are given in Supplementary Table 1.

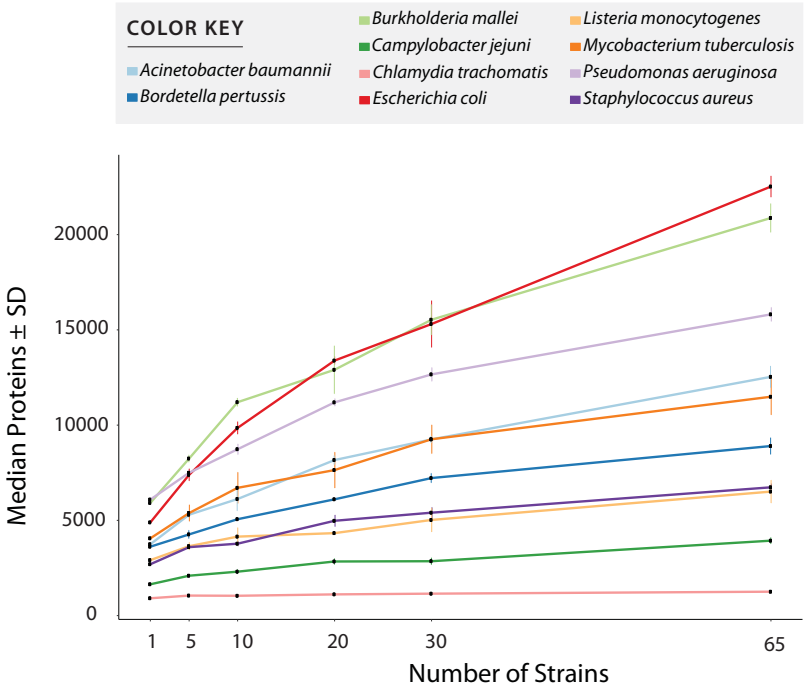
684

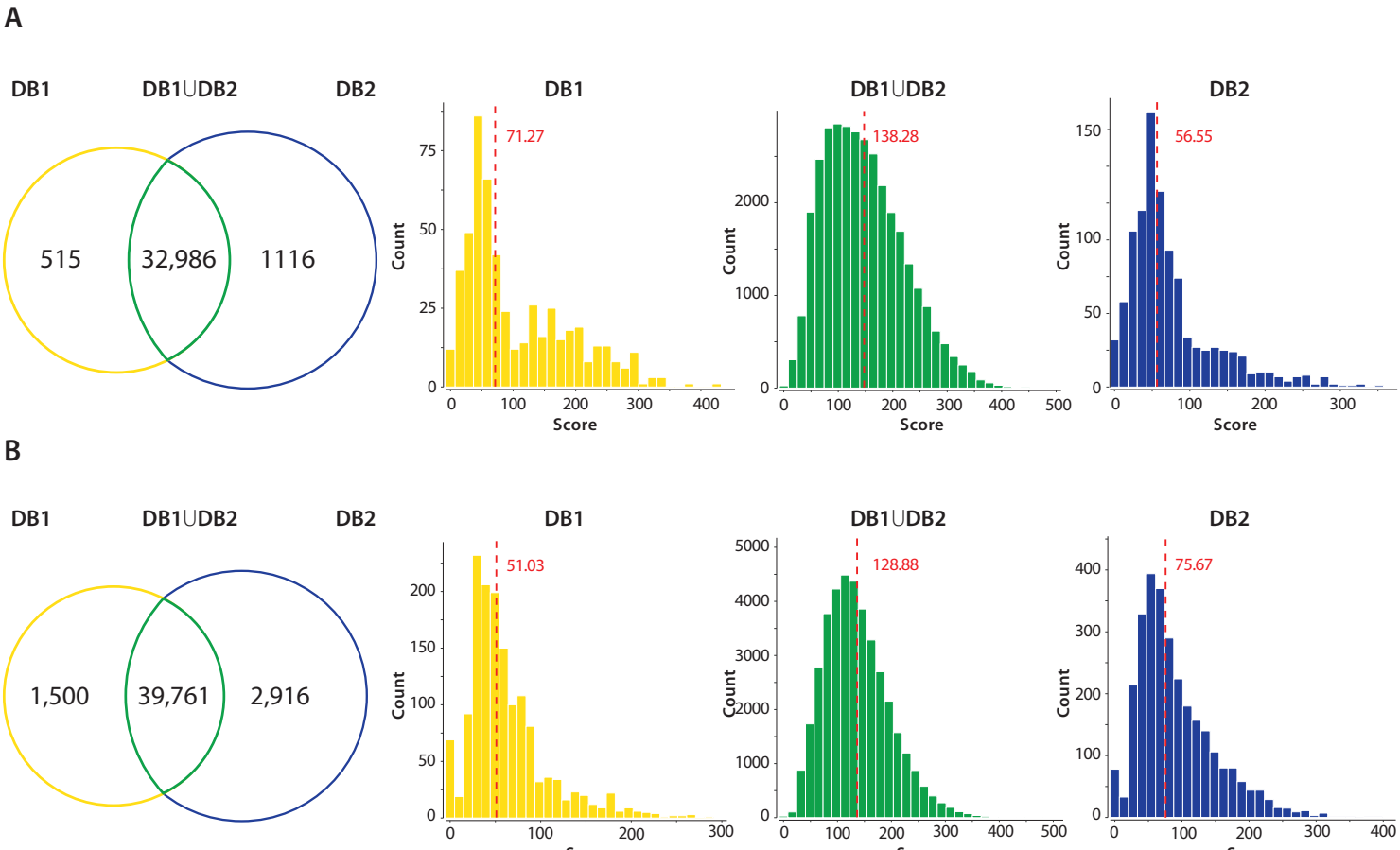
685 **Figure 4 – DB1 vs. DB2 peptide identification performance.** All unique tryptic peptides identified
686 in DB1 and DB2 in the Mtb (A) and *S. aureus* (B) datasets are compared. Venn diagrams show that
687 majority of tryptic peptides were identified in both searches. Peptide scores of each group were
688 plotted in a histogram, and a median was calculated. Median of the score distribution for peptides

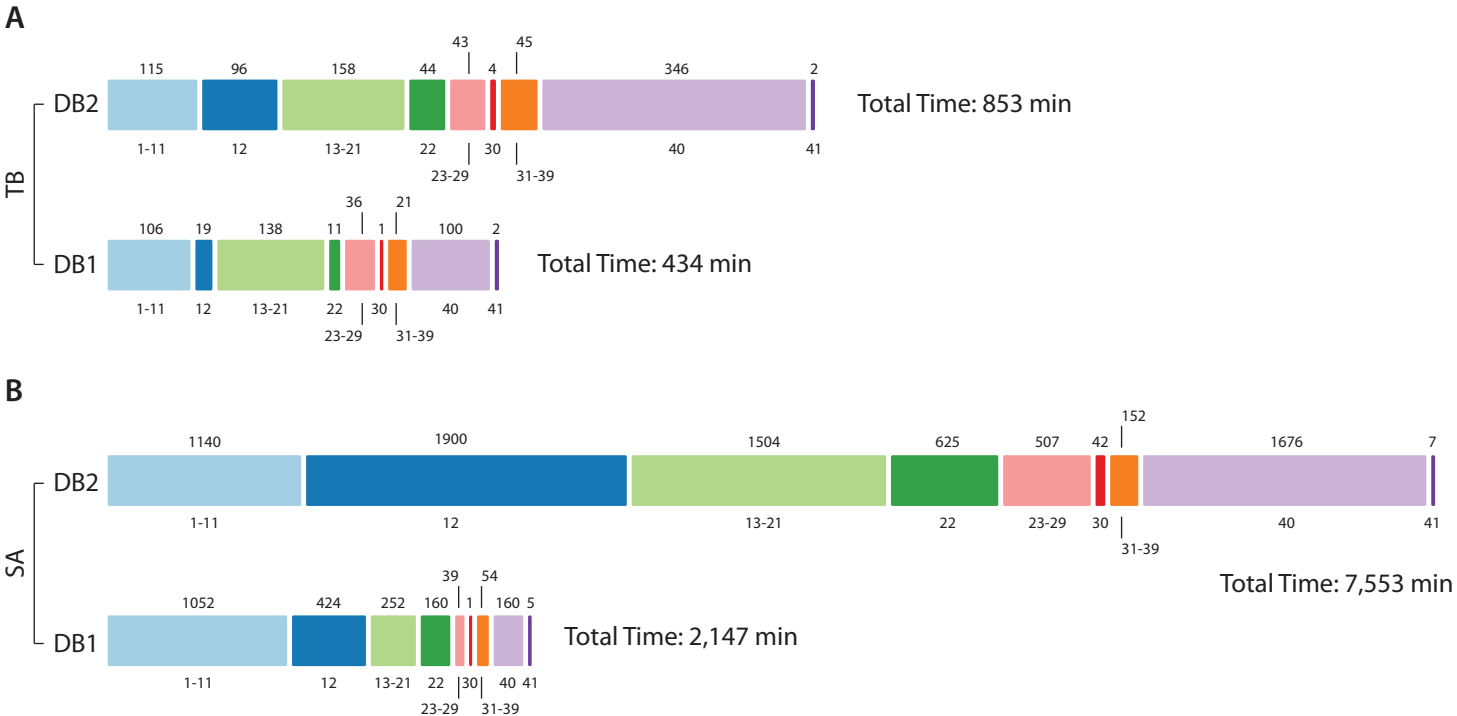
689 exclusive to DB1 or DB2 ranged from 51.03 to 75.67, while median scores of peptides identified on
690 both searches were 138.28 for Mtb and 128.88 for *S. aureus*.
691
692 **Figure 5 – DB1 vs DB2 computational demand.** Processing times for Mtb (A) and *S. aureus* (B)
693 datasets. The bar is divided into MaxQuant stage number, as described in (Neuhauser et al., 2013)
694 (value at bottom) and time in minutes is at the top of the bar. For Mtb, analysis was hasten mainly at
695 stage 12 (Main search), 22 (Second peptide search) and 40 (Writing tables); in addition to those, for
696 *S. aureus* improvements were also observed for post-search steps 13-21 and 23-29, and step 30
697 ("Protein groups assembly").
698
699 **Figure S1 – Total number of entries per MSMSpdbbdatabase**. Values in x axis show the number
700 of strains used for database creation. All values plotted in the graph are given in Supplementary
701 Table 1.
702
703 **Figure S2 – Number of tryptic peptides per MSMSpdbbdatabase**. In (A) only fully tryptic
704 peptides are considered, while in (B) peptides with up to two miscleavages are also counted, a
705 standard search parameter choice in proteomics. Values in x axis show the number of strains used in
706 the database. All values plotted in the graph are given in Supplementary Table 1.
707
708 **Figure S3 - Multiple TSS identified in protein lipV.** (A) Three possible TSS choices were
709 identified in clinical Mtb strains. The most upstream prediction (MIIDLHNQR) is the most abundant
710 variant observed in all strains, except strain S1430 which has only the variant with TSS at position 19
711 (ITIHGVTEHGR) of the reference sequence, and strain S2666 which TSS at position 22
712 (HGVTEHGR) is predominant. Two possible TSS choice variants were also observed at high levels
713 for strain S1593. (B-D) MS2 spectra for all three peptides mentioned above.
714
715 **Figure S4 – False-positive identification in DB1.** MS2spectrum of identified peptide
716 GEGAAVVVIKEGSIGGGTVFVYSGR, a concatenated sequence of peptide GEGAAVVVIK (pos
717 274-283 on PpsC protein) and peptide EGSIGGGTVFVYSGR (pos 564-578 on same protein). Note
718 the low scoring of the peptide identification. All five false-positives identified due to concatenation
719 of non-neighboring peptides are given in Supplementary Table 2.


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
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


A
DB1
DB1∪DB2
DB2
515
32,986
1116
DB1
71.27
DB1∪DB2
138.28
DB2
56.55
Count
Score
B
DB1
DB1∪DB2
DB2
1,500
39,761
2,916
DB1
51.03
DB1∪DB2
128.88
DB2
75.67
Count
Score


A
TB
DB2
115
96
158
44
43
4
45
346
2
1-11
12
13-21
22
23-29
30
31-39
40
41
Total Time: 853 min
DB1
106
19
138
11
36
1
21
100
2
1-11
12
13-21
22
23-29
30
31-39
40
41
Total Time: 434 min
B
SA
DB2
1140
1900
1504
625
507
42
152
1676
7
1-11
12
13-21
22
23-29
30
31-39
40
41
Total Time: 7,553 min
DB1
1052
424
252
160
39
1
54
160
5
1-11
12
13-21
22
23-29
30
31-39
40
41
Total Time: 2,147 min