bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 # Distinct gut metagenomics and metaproteomics signatures in
2 # prediabetics and treatment-naïve type 2 diabetics

3 Huanzi Zhong [1, 2, 3]†, Huahui Ren [1, 2, 3]†, Yan Lu [4]†, Chao Fang [1, 2, 3], Guixue Hou [1, 2],
4 Ziyi Yang [1, 2], Bing Chen [1, 2], Fangming Yang [1, 5], Yue Zhao [1, 2], Zhun Shi [1, 2], Baojin
5 Zhou [1, 2], Jiegen Wu [1], Hua Zou [1, 5], Jin Zi [1, 2], Jiayu Chen [2], Xiao Bao [2], Yihe Hu [4],
6 Yan Gao [4], Jun Zhang [4], Xun Xu [1, 2], Yong Hou [1, 2,], Huanming Yang [1, 6], Jian Wang [1, 6],
7 Siqi Liu [1, 2], Huijue Jia [1, 2], Lise Madsen [1, 3, 8], Susanne Brix [9], Fang Liu [4]*, Karsten
8 Kristiansen [1, 2, 3]*, Junhua Li [1, 2, 7]*

9 1. BGI-Shenzhen, Shenzhen 518083, China.
10 2. China National GeneBank, Shenzhen 518120, China.
11 3. Laboratory of Genomics and Molecular Biomedicine, Department of Biology,
12 University of Copenhagen, 2100 Copenhagen, Denmark.
13 4. Suzhou Centre for Disease Control and Prevention, Suzhou 215007, China
14 5. BGI Education Centre, University of Chinese Academy of Sciences
15 6. James D. Watson Institute of Genome Sciences, Hangzhou 310058, China.
16 7. School of Biology and Biological
17 Engineering, South China University of Technology, Guangzhou 510006, China.
18 8. Institute of Marine Research, P.O. Box 7800, 5020 Bergen, Norway.
19 9. Department of Biotechnology and Biomedicine, Technical University of
20 Denmark, Soltofts Plads, 2800 Kgs. Lyngby, Denmark.

21 * Correspondence should be addressed to Junhua Li, lijunhua@genomics.cn; Karsten
22 Kristiansen, (kk@bio.ku.dk), or Fang Liu (13306135766@163.com)

23 † Equal contributor
24
25
26
27

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

28 ## Abstract (254 words)

29 ### Background

30 The gut microbiota plays important roles in modulating host metabolism. Previous
31 studies have demonstrated differences in the gut microbiome of T2D and prediabetic
32 individuals compared to healthy individuals, with distinct disease-related microbial
33 profiles being reported in groups of different age and ethnicity. However,
34 confounding factors such as anti-diabetic medication hamper identification of the gut
35 microbial changes in disease development.

36 ### Method

37 We used a combination of in-depth metagenomics and metaproteomics analyses of
38 faecal samples from treatment-naïve type 2 diabetic (TN-T2D, n=77), pre-diabetic
39 (Pre-DM, n=80), and normal glucose tolerant (NGT, n=97) individuals to investigate
40 compositional and functional changes of the gut microbiota and the faecal content of
41 microbial and host proteins in Pre-DM and treatment-naïve T2D individuals to
42 elucidate possible host-microbial interplays characterising different disease stages.

43 ### Findings

44 We observed distinct differences characterizing the gut microbiota of these three
45 groups and validated several key features in an independent TN-T2D cohort. We also
46 demonstrated that the content of several human antimicrobial peptides and pancreatic
47 enzymes differed in faecal samples between three groups, such as reduced faecal level
48 of antimicrobial peptides and pancreatic enzymes in TN-T2D.

49 ### Interpretation

50 Our findings suggest a complex, disease stage-dependent interplay between the gut
51 microbiota and the host and emphasize the value of metaproteomics to gain further
52 insight into interplays between the gut microbiota and the host.

53

54 ### Funding

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

55 National Key Research and Development Program of China, No. 2017YFC0909703,
56 Shenzhen Municipal Government of China, No. JCYJ20170817145809215, and
57 National Natural Science Foundation of China, No. 31601073.

58 **Keywords**

59 Metagenomics, metaproteomics, prediabetes, treatment-naïve type 2 diabetes

60

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

61 ## Introduction (6,070 words for the main text)

62 Type 2 diabetes mellitus (T2D) is a chronic heterogeneous disorder associated
63 with hyperglycaemia and low grade inflammation [1,2]. The prevalence has increased
64 dramatically in Westernized countries, and also in China, where 11.6% and 36% of
65 Chinese adults suffer from diabetes and prediabetes (Pre-DM), respectively [3]. Due
66 to complications and comorbidities related to the development of T2D,
67 comprehensive characterization of phenotypic, metabolic and molecular changes of
68 the host and the gut microbiota in pre-DM and T2D compared to NGT is needed to
69 enable early identification of prediabetic individuals at high risk of T2D development.
70 Cross-sectional metagenomic studies have linked alterations in the gut microbiome to
71 T2D and prediabetes [4–7]. However, a few recent intervention studies have reported
72 profound impact of antidiabetic drugs on the human gut microbiome, such as
73 metformin, acarbose and glucagon-like peptide-1 (GLP-1) based therapies [8–13],
74 emphasizing the importance of controlling for medication in studies on association
75 between the microbiota and T2D. Moreover, distinct disease-related microbial profiles
76 have been reported in different age and ethnic groups [4–7], making it difficult to
77 identify the microbes possibly involved in disease development. Thus, detailed
78 information on the gut microbial species associated with T2D onset and progression is
79 still limited. Whereas information from metagenomics is limited to identification of
80 the presence of genes, taxa, and their inferred functional capacity, introduction of
81 additional omics approaches including metabolomics, metatranscriptomics, and
82 metaproteomics have increased our knowledge of microbial activity in health and
83 disease [14–17]. For instance, recent metatranscriptomics studies on inflammatory
84 bowel disease and cirrhosis cohorts have revealed considerable discrepancies between
85 data obtained from metagenomics vs metatranscriptomics analyses [17,18]. As
86 metaproteomics enables identification of microbial and human proteins
87 simultaneously in faecal samples [14,19,20], such an approach offers a potential for
88 deciphering both active microbial functions and host-microbiota interactions.
89

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

90 In the present study, we examined 254 stool samples collected from a Chinese cohort
91 combining shotgun metagenomics and metaproteomics analyses. We characterized
92 substantial differences between NGT, Pre-DM and TN-T2D individuals. Of note,
93 consistent aberrations in Pre-DM and TN-T2D individuals included lower abundances
94 of *Clostridiales* species and higher abundances of *Megasphaera elsdenii* compared to
95 NGT individuals. Several robust microbial compositional changes were detected at
96 both the DNA and protein levels, such as an enrichment of *E. coli* in Pre-DM
97 individuals and an increased abundance of *Bacteroides spp.* in TN-T2D patients.
98 Several Pre-DM-specific features were furthermore uncovered, including a reduced
99 capacity for processes involved in energy metabolism and bacterial growth, and an
100 enrichment of *Prevotella* proteins as detected by metaproteomics. Thus, our findings
101 revealed distinct characteristics of the intestinal ecosystem in the Pre-DM stage. Of
102 note, proteomics analyses of the faecal samples revealed that the levels of a number of
103 human proteins including several antimicrobial peptides (AMPs) differed in faecal
104 samples from NGT, Pre-DM, and TN-T2D individuals, suggesting that specific
105 differences in the host response amongst groups might also influence the composition
106 of the gut microbiota, or vice versa. In conclusion, our study provides a basis for
107 further analyses integrating faecal metagenomics and metaproteomics which may lead
108 to a better understanding of mechanisms underlying the development of Pre-DM and
109 T2D.
110

111 ## Materials and Methods

112 **Suzhou T2D study population**

113 The study population recruited from community residents from Suzhou, comprised 97
114 Chinese adults with normal glucose tolerance (NGT), 80 prediabetes patients
115 (Pre-DM) and 77 newly diagnosed, treatment naïve type 2 diabetes patients (TN-T2D).
116 All TN-T2D patients and Pre-DM individuals were screened and newly diagnosed
117 according to the 2011 WHO criteria via well-trained staffs from the Suzhou Centre for
118 Disease Prevention and Control (CDC), as described in detail in a recent published

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

119 lipidomic study based on this cohort [21]. All enrolled 254 individuals have reported
120 with no anti-diabetic treatments; thus, none have had taken insulin, or any oral or
121 injectable anti-diabetic medication before. Stool samples for metagenomics were
122 self-collected in 2ml faecal containers and immediately stored at -80°C and
123 transported to the laboratory on dry ice. DNA was extracted as previously described
124 [4]. A summary of sample information is presented in **Table S1.** In addition, shotgun
125 metagenomic datasets of stools from 94 anti-diabetic medication TN-T2D patients
126 from Shanghai [9], a city near to Suzhou, were used for validation purpose.
127

128 **Method for Metagenomics**

129 **1. Generation of BGISEQ-500 based faecal metagenome data set**

130 In this study, we performed DNA library construction and the combinatorial
131 probe-anchor synthesis (cPAS)-based BGISEQ-500 sequencing for metagenomics
132 (single-end; read length of 100bp) and applied the same quality control workflow to
133 filter the low-quality reads in accordance with the recently published metagenomic
134 study using this new platform [22]. The remaining high-quality reads were then
135 aligned to hg19 to remove human reads [23]. Metagenomic data statistics is provided
136 in **Table S2.**
137

138 **2. Profiling of metagenomic samples and biodiversity analysis**

139 High-quality non-human reads were aligned to the 9.9M integrated gene catalogue
140 (IGC) by SOAP2 using the criterion of identity ≥ 90% [23]. Sequence-based gene
141 abundance profiling was performed as previously described. The relative abundances
142 of phyla, genera, species and KOs were calculated by the sum of the relative
143 abundance of their annotated genes. The alpha diversity (within-sample diversity) was
144 quantified by the Shannon index using the relative abundance profiles at gene, genus
145 and KO levels as described [23]. The beta diversity (between-sample diversity) was
146 calculated using Bray-Curtis dissimilarity (R version 3.3.2, vegan package 2.4-4).
147

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

148 **3. Metagenome-wide association analysis (MWAS)**

149 MWAS was performed on the Suzhou T2D cohort as previously described [4] . Using
150 non-parametric Kruskal-Wallis test (R version 3.3.2 stats package), we identified
151 266,015 genes showing significant different abundances between the NGT, Pre-DM
152 and TN-T2D groups ($P < 0.05$). After clustering, a total of 126 MLGs (≥100 genes)
153 were generated from these genes. The relative abundance of each MLG was summed
154 using the relative abundance values of all genes from this MLG. The taxonomic
155 annotation of each MLG was determined if more than 50% of genes in this MLG
156 could be assigned to a certain taxon according to their IGC annotation. The genes of
157 85 unclassified MLGs were further annotated using a reference sequence database
158 including 1520 high-quality genomes cultivated from healthy Chinese individuals
159 [24], resulted in the taxonomic annotations of 11 additional MLGs (See detailed
160 information in **Table S5**).

161

162 **Method for Metaproteomics**

163 **1. Sample preparation and LC-MS/MS analysis**

164 Faecal samples from 84 individuals from NGT, Pre-DM, and TN-T2D individuals
165 were used for metaproteome analysis using isobaric tags for relative and absolute
166 quantitation (iTRAQ)–coupled-liquid chromatography tandem mass spectrometry
167 (LC-MS/MS) (**Figure S1**). Each group consisted of 28 randomly selected individual
168 samples with matched age, sex and BMI by propensity score matching (R version
169 3.3.2, MatchIt package 2.4-21) [25] (**Table S3**). Faecal samples were processed using
170 the filter-aided sample preparation (FASP) protocol [26]. Briefly, 100mg frozen faeces
171 from each individual were suspended in 500μl lysis buffer (4% SDS, 100mM
172 dithiothreitol, 100mM Tris-HCL (pH=7.8) with freshly added protease inhibitors
173 (cOmplete™, EDTA-free Protease Inhibitor Cocktail, Roche Applied Science). The
174 samples were incubated for 5 min at 100 °C, followed by sonication to decrease the
175 viscosity. The protein supernatants were collected after centrifugation at 30,000g at
176 4 °C for 30 min and then quantified using a 2D-quant kit (Sigma). For each diagnostic

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

177 group, protein extracts in equal amounts from four individuals were pooled, and the
178 selected 28 samples were thus aliquoted into 7 mixtures. A reference sample was
179 created by pooling equal amounts of protein from each of 84 individual sample and 28
180 samples from self-reported T2D patients. Each mixture containing 100μg proteins
181 was loaded onto a 10 kDa cut-off spin column (Vivacon 500, Sartorius AG,
182 Goettingen, Germany). The lysate was adjusted to 8M urea by centrifuging to remove
183 SDS and low-molecular-weight material. After reduction by dithiothreitol (DTT) and
184 alkylation by iodoacetamide (IAM), 8M urea was added and centrifuged to remove
185 any remaining reagent such as IAM. The urea buffer was then replaced with 0.5M
186 triethylammonium bicarbonate (TEAB) and the sample was washed with 0.5M TEAB
187 5 times. Trypsin (Promega, Madison, WI, USA) was added to digest the protein at a
188 protein: trypsin ratio of 50:1 and the mixtures were incubated for 18 hours at 37 °C.
189 The resulting peptides were eluted twice with 100μl 0.5M TEAB by centrifuging at
190 12,000 g for 30 min and vacuum-dried. The peptide mixture samples were then
191 dissolved in 0.5M TEAB and labelled with 8-plex iTRAQ reagents according to the
192 manufacturer's protocol (AB Sciex, USA). For each diagnostic group, 7 mixtures
193 were labelled with tags from I113 to I119. To perform the iTRAQ quantitation
194 throughout the whole experiment, we labelled the reference sample by tag 121 in each
195 iTRAQ run. Thus, three independent 8-plex iTRAQ runs were conducted.
196 Subsequently, labelled peptides were separated on a LC-20AB HPLC system
197 (Shimadzu, Kyoto, Japan) with an Ultremex SCX column (Phenomenon, Torrance,
198 CA) and collected into 20 fractions. Each fraction was analysed via a NanoLC system
199 coupled with a Q Exactive mass spectrometry (Thermo Fisher Scientific, San Jose,
200 CA) as described previously [27].

201

202 2. Database searching and protein identification

203 For protein database searching, we used Mascot (Version 2.3) [28] as the search
204 engine with the following parameters: trypsin was used as default enzyme and up to
205 two missed cleavages were allowed. Carbamidomethyl (C), iTRAQ8plex (N-term)

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

206 and iTRAQ8plex (K) were chosen as fixed modifications, and Oxidation (M) was
207 chosen as variable modification. The peptide mass tolerance was set to 10 ppm and
208 the fragment mass tolerance to 0.03 Da.
209 A two-step search method was applied. The MS/MS spectra were first searched
210 against a collection of three protein sequence databases, including *Homo sapiens*
211 sequences retrieved from SwissProt (release 2014_11), and human gut microbial
212 protein sequences of IGC genes mapped by sequencing reads from our 254
213 metagenomic samples. The detailed search parameters are presented in **Table S4**. The
214 Mascot search yielded a set of scored peptide-spectrum matches (PSMs) and the
215 proteins were inferred from the PSMs. Subsequently, a target-decoy protein database
216 was created containing the above-mentioned proteins and the reversed sequences from
217 these proteins. A second round search based on the target-decoy database was
218 performed to control for false positives as described elsewhere [29]. The PSMs were
219 re-scored by Mascot Percolator [30] integrated into IQuant [31], and filtered at false
220 discovery rate (FDR) $\leq$ 0.01. To improve the confidence in identification, peptides
221 supported by $\geq$ 2 spectra were retained and protein identifications were thus inferred.
222

223 **3. Meta-protein Generation**

224 Due to the shared similarity of metagenomic protein reference sequences, a microbial
225 peptide hit is typically returned from several proteins within and between species. To
226 avoid inflating numbers and alleviate taxonomic ambiguities of identified microbial
227 proteins, several processes were performed to reduce data redundancy. We first
228 grouped the microbial proteins with at least one shared peptide to generate protein
229 clusters (**Figure S2**). Each cluster was then processed according to the maximum
230 parsimony principle. The minimum protein sets containing all peptides of each cluster
231 were selected and defined as the meta-protein representing the cluster (**Figure S2**).
232 Individual proteins which only contained unique peptides were also assigned as a
233 meta-protein. All redundant non-meta-protein sequences were thus omitted in
234 subsequent analyses.

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

235

236 **4. Protein Quantification**

237 Protein quantification was performed by IQuant [31] in the following three steps.
238 We first normalized the intensities of iTRAQ reporter ions for all spectra across the
239 eight iTRAQ-labelled samples (I113…I119, I121) using the formula (1) as follows:

240

241 $$\overline{s_{i-k}} = \frac{s_{i-k}}{median(s_{1-k}:s_{n-k})}, \text{ where k}=I113\ldots I119, I121 \quad (1)$$

242

243 Where $\overline{s_{i-k}}$ is the normalized relative intensity of spectrum *i* in the label *k*.
244

245 The reporter ion ratios were then determined using the formula (2):

246 $$\overline{r_{i-k}} = \frac{\overline{s_{i-k}}}{s_{i-121}}, \text{ where k} = I113 \ldots I119 \quad (2)$$

247 Where $\overline{r_{i-k}}$ is the ratio of relative intensity of spectrum *i* in the label *k*, with S_{i-121},
248 the relative intensity of the global QC labelled with 121 tags, as denominators.

249

250 For protein quantification, only unique peptides were taken into consideration. The
251 relative protein ratio was calculated using the mean relative intensity ratio of all
252 unique peptide spectra in each protein using the formula (3):

253 $$\overline{p_k} = mean(\overline{r_{1-k}}:\overline{r_{p-k}}), \text{where } k = I113 \ldots I119 \quad (3)$$

254 Where $\overline{p_k}$ is the protein ratio in label K and acts as an indication of the relative
255 proportions of that protein between the differently labelled samples.

256

257 **5. Protein annotation**

258 For microbial meta-proteins, taxonomic and functional annotations of identified
259 proteins were derived from the putative protein-coding IGC genes. As a result, we
260 linked 64.15% (8777 of 11,980) of the meta-proteins with annotation at the phylum or
261 lower taxonomical levels and 80.27% (10983 of 11,980) with KEGG Ontology (KO)

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

262 annotation. For human proteins, functional annotations were obtained from
263 UniProtKB/Swiss-Prot (release 2014_11).
264
265 **Statistical analyses of metagenomes and metaproteomes**
266 **MLG-based random forest classification**
267 Relative abundance data of all MLGs were subjected to random forest (RF) analysis
268 to perform five-fold cross validation (R 3.3.2, caret package 6.0-77) [32]. The
269 combinations of optimal MLGs markers maximising the discrimination accuracy
270 between each two groups were thus determined by RF using an embedded feature
271 selection strategy as previously reported [33]. The importance values of
272 model-selected MLGs were calculated using "mean decrease in accuracy" strategy.
273
274 **Spearman's rank coefficient correlation**
275 Spearman's rank coefficient correlation (SCC) analysis was used for correlations
276 between MLG profiles and phenotypic factors, and between number of meta-proteins
277 and metagenomic abundances at the genus level, and between the levels of proteins.
278 The significance cut-off for SCC was set at an FDR adjusted $P < 0.05$.
279
280 **Enrichment analysis of KEGG modules**
281 Differentially enriched KEGG modules were identified according to reporter Z-scores
282 [34]. Z-score for each KO was first calculated from Benjamín-Hochberg (BH)-adjusted P values
283 from Wilcoxon rank-sum tests of comparisons between each two groups. The aggregated Z-score
284 for each module was calculated using Z-scores of all individual KOs belonging to the
285 corresponding module. A module was considered significant at a |reporter Z-score | ≥
286 1.96.
287
288 **Other statistical analyses**
289 Kruskal–Wallis test was conducted to detect the differences in continuous phenotypic
290 factors, microbial diversity, richness and MLG relative abundances between

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

291 multi-groups. *Dunn's post hoc* tests followed by pairwise comparisons were applied to
292 explore the differential phenotypes and MLGs between each two groups (R version
293 3.3.2, PMCMR package 4.1). The *Dunn's post hoc* p-values were adjusted with
294 the Benjamini-Hochberg method among multiple pairwise comparisons. The
295 significance cut-off was set as a *Dunn's post hoc P* value less than 0.05. Wilcoxon
296 rank-sum test was performed for comparisons of MLG relative abundances between
297 published TN-T2D patients from Shanghai [9] and NGT or Pre-DM from the Suzhou
298 cohort in this study for validation purposes. The significance cut-off of Wilcoxon
299 rank-sum test was set as a *P* value less than 0.05. Detailed information on enrichment
300 of MLGs between groups is provided in **Table S5**.
301 Wilcoxon rank-sum test was conducted to detect differences in protein levels between
302 each two groups. The significance cut-off for proteins was set as a *P* value less than
303 0.05, and a fold change of protein levels > 1.2 or < 0.8. Chi-square test was conducted
304 to detect the distribution of differences in discrete phenotypic factors, such as sex and
305 treatment distribution between groups, and to identify differences in taxonomic and
306 functional assignments between metagenomic and metaproteomic datasets. The
307 significant cut-off was set as *P* value less than 0.05.
308

309 **Data availability**

310 Metagenomic sequencing data for 254 faecal samples can be accessed from China
311 Nucleotide Sequence Archive (CNSA) with the dataset identifier CNP0000175. The
312 mass spectrometry metaproteomics data have been deposited to the ProteomeXchange
313 Consortium via the PRIDE partner repository with the dataset identifier PXD013452
314 and 10.6019/PXD013452.
315

316 **Results**

317 **Experimental design**

318 The cohort consisted of 77 TN-T2D patients, 80 Pre-DM individuals and 97 NGT
319 individuals from Suzhou, China (**Methods**, **Table S1**). The three groups were

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

320 matched regarding body mass index (BMI) and sex ($P > 0.05$), but individuals with
321 TN-T2D (mean age 66 +/- 8 years) were on average 5 years older than individuals in
322 the two other groups (**Table S1**). Shotgun metagenomics was performed on faecal
323 samples from all participants, whereas metaproteomics profiling was performed on a
324 subgroup of 84 participants, including 28 age-, sex-, and BMI-matched individuals
325 from each group (**Figure 1**).
326
327 **Distinct metagenomics profiles in Chinese prediabetic and type 2 diabetic**
328 **individuals**
329 Shotgun metagenomic sequencing of the 254 stool DNA samples was performed
330 using the BGISEQ-500 platform and raw reads were filtered and aligned to the
331 integrated gene catalogue (IGC) of the human gut microbiome to generate gene,
332 taxonomic and functional profiles as previously described (**Methods, Table S2**). In
333 line with previous studies [4–6], no significant differences in microbial gene-based
334 richness, alpha-diversity, and beta-diversity were found between the NGT, Pre-DM,
335 and TN-T2D individuals (**Figure S3,** Kruskal-Wallis (KW) test, $P > 0.05$). Using a
336 metagenome-wide association approach [4], we identified 266,015
337 T2D-associated genes (KW test, $P < 0.05$) and clustered these genes into 126
338 metagenomic linkage groups (MLGs, ≥100 genes, **Table S5**).
339 We further applied the KW test to detect statistically significant differences in the
340 relative abundances of MLGs between individuals with NGT, Pre-DM, and TN-T2D.
341 Compared to NGT individuals, the abundances of MLGs from the *Clostridia* class,
342 such as *Butyrivibrio crossotus* (MLG-2076), *Dialister invisus* (MLG-3376) and
343 *Roseburia hominis* (MLG-14865 and MLG-14920) were significantly lower in
344 individuals with Pre-DM or TN-T2D (**Figure 2A, Table S5,** *Dunn's post hoc test*, $P <$
345 0.05), which is in agreement with previous findings in a Danish T2D cohort [6]. In
346 addition, we found that the abundance of the butyrate-producing *Faecalibacterium*
347 *prausnitzii* (MLG-4560) was lower in Pre-DM compared to both NGT and TN-T2D
348 individuals. On the contrary, MLGs annotated to *Escherichia coli* (MLG-7919 and
349 MLG-7840*)*, *Streptococcus salivarius* (MLG-6991 and MLG-7099), and *Eggerthella*

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

350 *sp*. (MLG-351) were highly enriched in Pre-DM compared to NGT individuals
351 (**Figure 2A,** P < 0.05). An increased abundance of *Streptococcus* operational
352 taxonomic units (OTUs) was also recently reported in a Danish prediabetic cohort [7].
353 Additionally, Pre-DM individuals also exhibited a significant enrichment in *E. coli*
354 abundance compared to TN-T2D individuals (**Figure 2A,** P < 0.05). Moreover, we
355 detected significantly lower abundances of *Akkermansia muciniphila* (MLG-2159)
356 and *Clostridium bartlettii* (MLG-7540) and higher abundances of *Bacteroides caccae*
357 (MLG-10234 and MLG-10325), *Bacteroides finegoldii* (MLG-10154 and
358 MLG-10159), and *Collinsella intestinalis* (MLG-10084) in TN-T2D patients
359 compared with NGT and Pre-DM individuals (**Figure 2A,** P < 0.05). Finally, the
360 abundance of *Megasphaera elsdenii* (MLG-1568) was significantly higher in both
361 TN-T2D and Pre-DM individuals than in NGT individuals (**Figure 2A,** P < 0.05), in
362 line with the positive correlation between the relative abundance of the genus
363 *Megasphaera* and T2D recently reported in a large cohort with about 7000 individuals
364 from South China [35]. Several key findings were further validated in faecal samples
365 of 94 treatment naïve T2D patients in Shanghai (Gu et al., 2017a) , such as a lower
366 abundance of *A. muciniphila* and *C. bartlettii* compared to NGT and Pre-DM
367 individuals, and a lower abundance of *E.coli* compared to Pre-DM individuals in this
368 study (**Figure 2A, Table S5,** *Wilcoxon* rank test, P < 0.05). A summary of gut
369 microbial taxa reported in previously published cross-sectional T2D or prediabetes
370 studies is presented in **Table S6.**
371 We next performed Spearman's rank correlation analysis to explore the associations
372 between host phenotypes and MLGs. *M. elsdenii* and four unannotated MLGs
373 enriched in TN-T2D individuals showed significantly positive correlations to
374 glycaemic indices, including homeostasis model assessment of insulin resistance
375 (HOMA-IR), fasting blood glucose (FBG), 2h post-load glucose (2h-PG), and HbA1c,
376 whereas MLGs enriched in NGT were negatively correlated with the abovementioned
377 indices (adjusted P < 0.05, **Figure S4A-B**). Very few MLGs showed significant
378 correlations with non-glycaemic indices, such as age, BMI and systolic blood pressure

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

379 (SBP) (**Figure S4**).

380 To assess the discriminative power of MLGs in T2D and identify key MLGs
381 differentiating individuals with respect to different disease stages, we applied a
382 feature selection approach and constructed Random Forest (RF) classification models
383 comparing the groups **(Methods)**. Remarkably, the RF models provided high
384 performances regarding classification of samples from the two different disease stages,
385 with area under the ROC curve (AUC) values from 0.90 to 0.94 (**Figure 2B**). Apart
386 from taxonomically unclassified MLGs, the most discriminatory MLG for separating
387 TN-T2D and NGT was *A. muciniphila.* Moreover, MLGs annotated to *F. prausnitzii*
388 and *E. coli* both showed to be important in separating Pre-DM samples from TN-T2D
389 and NGT samples (**Figure 2C**), indicating the unique microbial signatures of lower
390 abundance of *F. prausnitzii* and higher abundance of *E. coli* in Pre-DM individuals.
391 We also validated the predictive power of the RF models between TN-T2D and other
392 two groups, which showed an accuracy of 76. 6% (72 of 94 patients) for disease
393 prediction in a previously described TN-T2D cohort from Shanghai (**Table S7**) [9].

394 We next performed KEGG enrichment analyses to examine possible differential
395 patterns of microbial functional potentials in NGT, Pre-DM and TN-T2D individuals
396 (**Table S8**). Interestingly, we observed a significant enrichment in modules
397 comprising several sugar phosphotransferase systems (PTS), ATP-binding cassette
398 transporters (ABC transporters) of amino acids, and bacterial secretion systems in the
399 gut microbiota of Pre-DM compared to NGT individuals (reporter score $\geq$ 1.96,
400 **Figure 2D**). Likewise, in line with previous findings in several Chinese cohorts with
401 metabolic diseases, such as atherosclerotic cardiovascular disease (ACVD), obesity
402 and T2D [36], a similar enrichment was found in TN-T2D patients compared with
403 NGT individuals (**Figure 2D**). The abundances of the transport system for microcin C,
404 a peptide-nucleotide antibiotic produced by *Enterobacteria* [37], and the transport
405 system for autoinducer-2 (AI-2), a quorum sensing signalling molecule reported in
406 Proteobacteria [38], were also significant higher in Pre-DM than in NGT individuals
407 (**Figure 2D**). Except for enrichment of type II-IV secretion and AI-2 transport systems

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

408 in Pre-DM vs TN-T2D, we found no other KEGG modules for PTS and ABC
409 transporters to differ significantly in abundance between Pre-DM and TN-T2D
410 individuals (**Figure 2D**). However, Pre-DM individuals displayed a significant
411 reduction with respect to several energy and nucleotide metabolism modules
412 compared to both NGT and TN-T2D individuals, including modules of V-type ATPase,
413 pyruvate: ferredoxin oxidoreductase, and bacterial ribosomal proteins **(Figure 2D)**.
414 Taken together, these results indicate the possible involvement of substantial
415 compositional and functional disease-related gut microbial changes in the pre-diabetic
416 stage.
417
418 **Gut metaproteomics simultaneously identifies faecal levels of microbial and**
419 **human proteins**
420 To gain further insights into functional changes in the gut microbiota associated with
421 T2D, we conducted metaproteomic analyses using iTRAQ (isobaric peptide tags for
422 relative and absolute quantification) and LC-MS/MS-based protocols on 84 samples,
423 with 28 samples derived from each of the three diagnostic groups **(Methods, Figure**
424 **S1**). Using the strict parameters of 2 peptide-spectrum matches (PSMs) per protein, <
425 10 ppm mass error and 1% PSM-level FDR (**Methods**), we identified a total of
426 145,014 high quality PSMs corresponding to 15,670 proteins, including 15,245
427 (97.29%) microbial proteins and 425 (2.71%) human proteins (**Table S9**). As reported
428 [14,19,20], one microbial peptide often exhibits matches to multiple proteins with
429 high sequence similarity, resulting in difficulties in identifying the microbial origin of
430 individual peptides. To alleviate ambiguities, we applied a maximum parsimony
431 principle reported in recent studies [14]' [39] and generated 11,980 non-redundant
432 meta-proteins (78.58% of microbial proteins) containing at least one unique microbial
433 peptide. The relative intensities of these unique peptides were further used for
434 meta-protein quantification (**Methods, Table S9**). The number of identified
435 meta-proteins ranged between 5,067 in the Pre-DM samples to 8,134 in the TN-T2D
436 samples **(Table S9)**. Venn diagrams showed that only 2782 meta-proteins (34.2%-54.9%
437 of the total number of meta-proteins per group) were shared among the three groups

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

438 (**Figure S5A**), indicating differential microbial expression patterns at the protein level
439 among the groups. Taxonomic annotations indicated a higher percentage of unique
440 Proteobacteria meta-proteins in Pre-DM individuals, compared to the other groups
441 (Chi-square test, $P < 0.05$, **Figure S5B**), whereas no difference in the distributions of
442 the uniquely detected meta-proteins associated with a wide range of functions was
443 found between the three groups **(Figure S5C)**.
444

445 **Concordance and discordance of microbiota features between metagenomes and**
446 **metaproteomes**

447 Based on annotated microbial features, we next investigated the consistency as well as
448 the divergence of microbial composition and function at the DNA and protein level**.**
449 At the phylum level, more than 90% genes and meta-proteins were consistently
450 assigned to three major phyla, namely Firmicutes, Bacteroidetes and Proteobacteria
451 (**Figure 3A**). Despite the overall consistency, we found a significantly higher
452 percentage of the annotated proteins to be assigned to Bacteroidetes (41%) compared
453 to the percentage of genes annotated to Bacteroidetes (25%) (Chi-square test, $P < 0.05$**,**
454 **Figure 3A**), suggesting that Bacteroidetes might display an overall higher protein
455 production than the other phyla across the 84 samples. At the genus level, the
456 composition of the metaproteomes was biased towards a limited number of genera.
457 Among 212 common metagenomically-identified genera detected in at least 10% of
458 the 84 samples, only 81 genera (38.21%) could be detected based on metaproteomics
459 (**Table S10**). Spearman's rank correlation analysis was subsequently performed to
460 determine the relationship between the number of meta-proteins and the abundances
461 at the genus level based on metagenomics. The more abundant a given genus was
462 based on metagenomics analysis, the more of the identified meta-proteins were
463 assigned to this genus (Spearman's correlation coefficient (SCC) = 0.726, P =
464 5.21E-08, **Figure 3B, Table S9**), with *Bacteroides* (n=1664), *Prevotella* (n=818) and
465 *Faecalibacterium* (n=719) harbouring most assigned meta-proteins. For a few genera,
466 such as *Anaerotruncus* (n=9), *Paraprevotella* (n=9) and *Enterococcus* (n=7), we were

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

467 only able to identify less than 10 meta-proteins although their median metagenomic
468 abundances were greater than 1E-04 **(Table S10)**.
469 Comparing KEGG functional categories based on metagenomics and metaproteomics
470 data, we observed large differences in the relative contribution of individual
471 categories between the two datasets (Chi-square test, $P < 0.05$, **Figure 3C**), in
472 accordance with several previous studies [14,19,20]. For instance, as determined by
473 metaproteomics, 24% and 18% of the proteins were assigned to carbohydrate
474 metabolism and translation categories, whereas the corresponding metagenomic
475 percentages of the two categories were only 11% and 4%, respectively (**Figure 3C).**
476 We found that 1508 meta-proteins, accounting for 12.59% of all identified
477 meta-proteins, could be assigned to 10 KEGG orthologues (KO). The top KOs
478 harboured 360 proteins annotated as Ca-activated chloride channel homologues
479 (K07114), whereas the remaining KOs comprised proteins representing abundant
480 house-keeping proteins such as elongation factors, large subunit ribosomal proteins
481 (K02355, K02358 and K02395), chaperones (K04077 and K04043), and
482 glyceraldehyde 3-phosphate dehydrogenase (K00134) as well as flagellin proteins
483 (K02406) **(Table S11, Figure S6A)**.
484 Aiming to link the microbial protein patterns to metagenomic microbial abundances,
485 we next conducted a fold-change analysis of meta-proteins. In agreement with our
486 metagenomic findings **(Figure 2A)**, the Proteobacteria meta-proteins (mainly from
487 *Escherichia*, *Citrobacter* and *Enterobacter*) exhibited enrichment in the Pre-DM
488 group, whereas *Bacteroides* meta-proteins were enriched in TN-T2D individuals
489 (**Figure 3D, Table S12,** $P < 0.05$ and fold change of protein intensities > 1.2).
490 Surprisingly, *Prevotella* meta-proteins were selectively enriched in Pre-DM
491 individuals (**Figure 3D),** although no *Prevotella* annotated metagenomic MLGs
492 exhibited significantly higher abundance. At the functional level, we observed that the
493 level of meta-proteins involved in carbohydrate metabolism tended to be lower in
494 NGT compared to Pre-DM and TN-T2D individuals, including those involved in the
495 metabolism of succinate (**Figure 3E, Figure S6B, Table S11**).

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

496

497 **Functional characteristics of faecal excreted human proteins in T2D**

498 Among the 425 detected human proteins, we identified 218 human proteins that were
499 shared among the NGT, Pre-DM, and TN-T2D groups, accounting for 59.6% to 85.2%
500 of the identified human proteins in each group **(Figure S7A).** We next annotated the
501 human proteins with Gene Ontology (GO) terms to obtain insight into the functional
502 characteristics of the human proteins excreted in faeces (**Table S13**). Among the
503 identified proteins, 181 (42.59%) had previously been identified in faecal samples by
504 metaproteomics, indicative of their general presence (**Table S14**) [14,19,20]. These
505 included several intestinal mucin proteins, such as MUC-1, MUC-2, MUC-4, MUC5B,
506 MUC12 and MUC-13 as well as members of annexins (ANXA1- ANXA7, a family of
507 calcium-binding proteins) (**Table S14**). We identified 233 of the faecal human
508 proteins to have tissue-specific annotation, amongst which 151 proteins (64.81%)
509 were reported to exhibit high expression in the digestive system, and the remaining
510 proteins were annotated to be highly expressed in blood or other tissues such as
511 epidermis **(Table S13).** Of interest, 18 of the human proteins were annotated as AMPs
512 [40] (**Table S13**). Several human proteins involved in glucose metabolism, including
513 the sodium/glucose cotransporter 1, were detected in faecal samples of TN-T2D
514 patients only (**Figure S6B**). Inhibitors of this protein have been proposed for
515 antidiabetic treatment [26]. Additionally, the TMAO-producing enzyme, dimethylaniline
516 monooxygenase [N-oxide-forming] 3 (FMO3) was also identified exclusively in the
517 TN-T2D group (**Table S13**). On the other hand, we found that ras
518 GTPase-activating-like protein (IQGAP1) and unconventional myosin-Ic (MYO1C)
519 were uniquely identified in the NGT group (**Figure S7B**). Loss of IQGAP1 and
520 MYO1C has been related to impairment of insulin signalling [43–45], but whether
521 their presence in faeces has functional implications remains to be established.

522

523 Forty-nine of the human proteins present in faeces were found to differ significantly
524 in intensity between at least two of the groups **(Figure 4A, Table S15)**. We found

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

525 significantly higher levels of four AMPs, including defensin-5, neutrophil defensin-1,
526 lysozyme c, as well as secreted phospholipase A2, all with important roles in the
527 defence against bacteria [46–48], in faecal samples from NGT individuals than in
528 samples from TN-T2D individuals (**Figure 4A**). We also found higher levels of
529 mucin-5AC samples from NGT compared to TN-T2D individuals, suggesting
530 possible effects on the mucus barrier in TN-T2D. Interestingly, the level of the
531 antimicrobial cathepsin G, reported to inhibit the growth of several organisms from
532 the Proteobacteria phylum [49], was higher in samples from Pre-DM than NGT and
533 TN-T2D, and this was coupled to lower levels of alpha-1-antichymotrypsin and
534 alpha-1-antitrypsin, both known inhibitors of cathepsin G [50] (**Figure 4A**),
535 suggesting that Pre-DM individuals have initiated strategies to activate a defence
536 system against the enhanced relative abundances of *E. coli*. By contrast, we found that
537 several proteins within the immunoglobulin superfamily were present at lower levels
538 in samples from Pre-DM compared to NGT or TN-T2D (**Figure 4A)**. Individuals with
539 Pre-DM also exhibited lower levels of galectin-3, a lectin with
540 beta-galactoside-binding ability. Galectin-3 has been reported to bind
541 lipopolysaccharides (LPS) from *E. coli* and play a role as a negative regulator of
542 LPS-mediated inflammation [51]. In addition, galectin-3 was also reported to improve
543 epithelial intercellular contact via desmoglein-2 stabilization [52]. Taken together,
544 these finding indicate that the gut ecosystem in Pre-DM individuals exhibits trait
545 compatible with the upregulation of defence systems against an increased abundance
546 of Proteobacteria simultaneously with the downregulation of factors capable of
547 reducing the impact of the inflammation-inducing activity of LPS. We also found that
548 several digestive enzymes differed in levels in faeces from NGT, Pre-DM, and
549 TN-T2D individuals. Thus, we found lower levels of proteases (trypsin and
550 chymotrypsin and their precursors) and lipases, and higher amylase (AMY1) levels in
551 TN-T2D (**Figure 4A**). It is also interesting to note that the level of dipeptidyl
552 peptidase 4 (DDP4), known to inhibit insulin secretion via its action on GLP-1, was
553 lower in individuals with Pre-DM than in TN-T2D individuals. A network analysis

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

554 revealed significant correlations between 20 human proteins showing significant
555 differences in levels in two-pairwise comparisons between NGT, Pre-DM and
556 TN-T2D individuals (**Figure 4B**). For instance, we identified a negative correlation
557 between the defensin-5 and TN-T2D-enriched peptidyl-prolyl cis-trans isomerase B
558 (PPIB) (**Figure 4B,** SCC, adjusted $P < 0.05$), the latter previously reported to be
559 associated with islet dysfunction [53].
560 Aiming to investigate possible host-microbial protein interactions in the human gut,
561 we next investigate the possible correlation between the discriminatory bacterial and
562 human proteins. Interestingly, we found significantly negative correlations between
563 several Pre-DM-enriched *E. coli* proteins and human proteins involved in innate
564 immune responses (HV304, HV305) and adhesion (CEAM6, CEAM7), whereas
565 positive correlations were found between *E. coli* proteins and cathepsin G,
566 Cytochrome c (CYC) and trypsin−1 (TRY1) (**Figure 4C,** adjusted $P < 0.05$).
567 Conversely, NGT-enriched proteins from *F. prausnitzii* showed positive correlations
568 with several NGT-enriched digestive enzymes from the exocrine pancreas, such as
569 chymotrypsin-like elastase family member 3A (CEL3A), chymotrypsinogen B2
570 (CTRB2) and carboxypeptidases (CBPA1 and CBPB1).
571

572 Discussion

573 Our comparative study using metagenomics and metaproteomics in normal glucose
574 tolerant, pre-diabetics and treatment naïve T2D individuals provides important novel
575 findings with regard to disease-stage specifications at the gut bacterial and host level.
576 A substantial number of Pre-DM associated features were revealed at both the
577 metagenomics and metaproteomics level. Of specific note are the significantly higher
578 abundance of Proteobacteria species (dominated by *E. coli*) and the lower levels of
579 host proteins which potentially are involved in Proteobacteria-specific responses in
580 Pre-DM, such as galectin-3 and proteins within the immunoglobulin superfamily.
581 Furthermore, significantly higher levels of *Prevotella* proteins were uniquely detected
582 in Pre-DM individuals although the abundance of *Prevotella* was not significantly

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

583 enriched in this group based on metagenomics data. *Prevotella copri* has previously
584 been shown to produce branched-chain amino acids (BCAA), reported to correlate
585 with BCAA blood levels and insulin resistance [54]. However, in the present study
586 only two enzymes related to the synthesis of BCAAs were detected among the
587 identified *Prevotella* proteins with no differences in levels between the three groups.
588
589 Only a modest number of relatively highly abundant faecal proteins were identified in
590 the current study. This reflects the current methodological challenges in microbial
591 protein extraction, identification, and annotation as reported previously [55,56], as
592 well as the detection limitations of MS-based proteomics [57]. For instance, we
593 identified less than 50 proteins from each of several taxa with median abundances in
594 the 0.1 % ranges based on metagenomics data (such as NGT-enriched *Dialister*,
595 *Butyrivibrio* and *Haemophilus*). Nevertheless, metaproteomics provides a valuable
596 addition to not only estimating expression of microbial proteins, but also to delineate
597 host-microbial protein interactions in different disease stages. In this regard, we
598 identified higher levels of several host-derived AMPs in NGT individuals compared
599 to TN-T2D and Pre-DM individuals, suggesting a possible stronger host defence
600 against invading (disease-related) microbes in NGT individuals. By contrast,
601 significant negative associations were found between Pre-DM-enriched *E. coli*
602 proteins and several human proteins, including AMPs, adhesion molecules and
603 galectin-3, all involved in intestinal barrier function. It is also worth to note the
604 significant changes in levels and types of digestive enzymes identified in the faecal
605 samples, where TN-T2D showed enhanced alpha-amylase (AMY1) levels, as
606 compared to pancreatic-derived lipases and proteases. However, the level of
607 pancreatic alpha-amylase (AMYP) was lower in Pre-DM compared to the two other
608 groups. A metaproteomics study has reported lower faecal AMYP levels in type 1
609 diabetes (T1D) patients compared to their healthy relatives [10], whereas no difference
610 in levels of AMY1 was reported between T1D and controls, suggesting different
611 amylase responses might be present in Pre-DM, TN-T2D and T1D patients based on

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

612 metaproteomics data. Differences in levels of secreted digestive enzymes from the
613 exocrine pancreas in NGT, Pre-DM and T2D have to our notice not been addressed
614 previously, although it may be of major importance in relation to the metabolic state
615 in T2D.
616 Together, our findings suggest that unique and nonlinear changes of the intestinal
617 ecosystem might exist in Pre-DM individuals before transition to T2D. Further
618 large-scale, longitudinal follow-up studies are needed to delineate how microbial
619 functions changes from prediabetes to diabetes and to address the nature of
620 interactions between the gut microbiota and the host in the transitional phases leading
621 to overt T2D.
622

623 **Acknowledgements**
624 We thank Prof. Yan Ren for helpful discussion on designing the metaproteomic
625 experiments. We thank Dr Cong Lin and Dr Zhe Zhang for helpful discussion and
626 suggestions on developing the manuscript. We gratefully acknowledge colleagues at
627 BGI for DNA extraction, library preparation and shotgun sequencing experiments,
628 and helpful discussions.
629 **Funding Sources**
630 This work was supported by grants from National Key Research and Development
631 Program of China (No. 2017YFC0909703), Shenzhen Municipal Government of
632 China (No. JCYJ20170817145809215) and National Natural Science Foundation of
633 China (No. 31601073).
634

635 **Declarations of interests**

636 The authors declare no competing interests.

637
638 **Author contributions**
639 J.L. and H.Z. designed and coordinated the study. F.L. and J.Z. oversaw the blood and
640 faecal sample collection. Y.L, B.C., J.C., X. B., Y.H. and Y.G. participated in sample

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

641 collection and provided phenotypic information. G.H., B.Z, J.Z. and S.L. carried out
642 the metaproteomic experiments. H.Z., H.R., F.Y., Z.S, and H.Zou. performed the
643 bioinformatic analyses of metagenomic data. H.Z., H.R., C.F., B.Z, G.H., Y.Z. and
644 J.W. performed the bioinformatic analyses of metaproteomic data. H.Z. and H.R
645 performed integrative analyses of metagenomic and metaproteomic data. Y.Z.
646 performed revision of the figures. H.Z. interpreted together with J.L., S.B. and K.K.
647 the data and wrote the first version of the manuscript. J.L., K.K., S.B., and L.M.
648 performed revision of the manuscript. H.Z., H.R., C.F., G.H., F.Y., Z.Y., Y.Z., Z.S.,
649 J.W, L.M., S.B., K.K. and J.L. participated in discussions. All authors contributed to
650 the revision of the manuscript. All authors read and approved the final manuscript.
651
652

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

653 **Reference**

654 [1] Stumvoll M, Goldstein BJ, Van Haeften TW. Type 2 diabetes: Principles of
655 pathogenesis and therapy. Lancet, vol. 365, 2005, p. 1333–46.
656 doi:10.1016/S0140-6736(05)61032-X.
657 [2] Pickup JC. Inflammation and Activated Innate Immunity in the Pathogenesis of
658 Type 2 Diabletes. Diabetes Care 2004;27:813–23.
659 doi:10.2337/diacare.27.3.813.
660 [3] Wang L, Gao P, Zhang M, Huang Z, Zhang D, Deng Q, et al. Prevalence and
661 ethnic pattern of diabetes and prediabetes in China in 2013. JAMA - J Am Med
662 Assoc 2017;317:2515–23. doi:10.1001/jama.2017.7596.
663 [4] Wang J, Qin J, Li Y, Cai Z, Li S, Zhu J, et al. A metagenome-wide association
664 study of gut microbiota in type 2 diabetes. Nature 2012;490:55–60.
665 doi:10.1038/nature11450.
666 [5] Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B,
667 et al. Gut metagenome in European women with normal, impaired and diabetic
668 glucose control. Nature 2013. doi:10.1038/nature12198.
669 [6] Forslund K, Hildebrand F, Nielsen T, Falony G, Le Chatelier E, Sunagawa S,
670 et al. Disentangling type 2 diabetes and metformin treatment signatures in the
671 human gut microbiota. Nature 2015;528:262–6. doi:10.1038/nature15766.
672 [7] Allin KH, Tremaroli V, Caesar R, Jensen BAH, Damgaard MTF, Bahl MI, et al.
673 Aberrant intestinal microbiota in individuals with prediabetes. Diabetologia
674 2018;61:810–20. doi:10.1007/s00125-018-4550-1.
675 [8] Wu H, Esteve E, Tremaroli V, Khan MT, Caesar R, Mannerås-Holm L, et al.
676 Metformin alters the gut microbiome of individuals with treatment-naive type 2
677 diabetes, contributing to the therapeutic effects of the drug. Nat Med
678 2017;23:850–8. doi:10.1038/nm.4345.
679 [9] Gu Y, Wang X, Li J, Zhang Y, Zhong H, Liu R, et al. Analyses of gut
680 microbiota and plasma bile acids enable stratification of patients for
681 antidiabetic treatment. Nat Commun 2017;8:1785.

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

682 doi:10.1038/s41467-017-01682-2.
683 [10] Zhao L, Chen Y, Xia F, Abudukerimu B, Zhang W, Guo Y, et al. A
684 glucagon-like peptide-1 receptor agonist lowers weight by modulating the
685 structure of gut microbiota. Front Endocrinol (Lausanne) 2018.
686 doi:10.3389/fendo.2018.00233.
687 [11] Moreira G V., Azevedo FF, Ribeiro LM, Santos A, Guadagnini D, Gama P, et
688 al. Liraglutide modulates gut microbiota and reduces NAFLD in obese mice. J
689 Nutr Biochem 2018. doi:10.1016/j.jnutbio.2018.07.009.
690 [12] Olivares M, Neyrinck AM, Pötgens SA, Beaumont M, Salazar N, Cani PD, et
691 al. The DPP-4 inhibitor vildagliptin impacts the gut microbiota and prevents
692 disruption of intestinal homeostasis induced by a Western diet in mice.
693 Diabetologia 2018. doi:10.1007/s00125-018-4647-6.
694 [13] Liao X, Song L, Zeng B, Liu B, Qiu Y, Qu H, et al. Alteration of gut
695 microbiota induced by DPP-4i treatment improves glucose homeostasis.
696 EBioMedicine 2019. doi:10.1016/j.ebiom.2019.03.057.
697 [14] Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, et al.
698 Integrated multi-omics of the human gut microbiome in a case study of familial
699 type 1 diabetes. Nat Microbiol 2016;2. doi:10.1038/nmicrobiol.2016.180.
700 [15] Abu-Ali GS, Mehta RS, Lloyd-Price J, Mallick H, Branck T, Ivey KL, et al.
701 Metatranscriptome of human faecal microbial communities in a cohort of adult
702 men. Nat Microbiol 2018;3:356–66. doi:10.1038/s41564-017-0084-4.
703 [16] Liu R, Hong J, Xu X, Feng Q, Zhang D, Gu Y, et al. Gut microbiome and
704 serum metabolome alterations in obesity and after weight-loss intervention. Nat
705 Med 2017;23:859–68. doi:10.1038/nm.4358.
706 [17] Schirmer M, Franzosa EA, Lloyd-Price J, McIver LJ, Schwager R, Poon TW,
707 et al. Dynamics of metatranscription in the inflammatory bowel disease gut
708 microbiome. Nat Microbiol 2018;3:337–46. doi:10.1038/s41564-017-0089-z.
709 [18] Bajaj JS, Thacker LR, Fagan A, White MB, Gavis EA, Hylemon PB, et al. Gut
710 microbial RNA and DNA analysis predicts hospitalizations in cirrhosis. JCI

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

711 Insight 2018;3:1–12. doi:10.1172/jci.insight.98019.
712 [19] Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, Halfvarson J,
713 et al. Shotgun metaproteomics of the human distal gut microbiota. ISME J
714 2009;3:179–89. doi:10.1038/ismej.2008.108.
715 [20] Young JC, Pan C, Adams RM, Brooks B, Banfield JF, Morowitz MJ, et al.
716 Metaproteomics reveals functional shifts in microbial and human proteins
717 during a preterm infant gut colonization case. Proteomics 2015;15:3463–73.
718 doi:10.1002/pmic.201400563.
719 [21] Zhong H, Fang C, Fan Y, Lu Y, Wen B, Ren H, et al. Lipidomic profiling
720 reveals distinct differences in plasma lipid composition in healthy, prediabetic,
721 and type 2 diabetic individuals. Gigascience 2017;6.
722 doi:10.1093/gigascience/gix036.
723 [22] Fang C, Zhong H, Lin Y, Chen B, Han M, Ren H, et al. Assessment of the
724 cPAS-based BGISEQ-500 platform for metagenomic sequencing. Gigascience
725 2018;7:1–8. doi:10.1093/gigascience/gix133.
726 [23] Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog
727 of reference genes in the human gut microbiome. Nat Biotechnol
728 2014;32:834–41. doi:10.1038/nbt.2942.
729 [24] Zou Y, Xue W, Luo G, Deng Z, Qin P, Guo R, et al. 1,520 reference genomes
730 from cultivated human gut bacteria enable functional microbiome analyses. Nat
731 Biotechnol 2019. doi:10.1038/s41587-018-0008-8.
732 [25] Austin PC. An introduction to propensity score methods for reducing the
733 effects of confounding in observational studies. Multivariate Behav Res
734 2011;46:399–424. doi:10.1080/00273171.2011.568786.
735 [26] Wiśniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample
736 preparation method for proteome analysis. Nat Methods 2009;6:359–62.
737 doi:10.1038/nmeth.1322.
738 [27] Guo J, Ren Y, Hou G, Wen B, Xian F, Chen Z, et al. A Comprehensive
739 Investigation toward the Indicative Proteins of Bladder Cancer in Urine: From

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

740 Surveying Cell Secretomes to Verifying Urine Proteins. J Proteome Res
741 2016;15:2164–77. doi:10.1021/acs.jproteome.6b00106.
742 [28] Pappin DJC, Creasy DM, Cottrell JS. Probability-based Protein Identification
743 by Searching Sequence Databases Using Mass Spectrometry Data.
744 Electrophoresis 1999;20:3551–67.
745 [29] Elias JE, Gygi SP. Target-Decoy Search Strategy for Mass Spectrometry-Based
746 Proteomics. Proteome Bioinforma 2010:55–71.
747 doi:10.1007/978-1-60761-444-9_5.
748 [30] Brosch M, Yu L, Hubbard T, Choudhary J. Accurate and sensitive peptide
749 identification with mascot percolator. J Proteome Res 2009;8:3176–81.
750 doi:10.1021/pr800982s.
751 [31] Wen B, Zhou R, Feng Q, Wang Q, Wang J, Liu S. IQuant: An automated
752 pipeline for quantitative proteomics based upon isobaric tags. Proteomics
753 2014;14:2280–5. doi:10.1002/pmic.201300361.
754 [32] Max K, Kuhn M. Building Predictive Models in R Using the caret Package. J
755 Stat Softw 2008. doi:10.1053/j.sodo.2009.03.002.
756 [33] Tett A, Pasolli E, Farina S, Truong DT, Asnicar F, Zolfo M, et al. Unexplored
757 diversity and strain-level structure of the skin microbiome associated with
758 psoriasis. Npj Biofilms Microbiomes 2017;3. doi:10.1038/s41522-017-0022-5.
759 [34] Patil KR, Nielsen J. Uncovering transcriptional regulation of metabolism by
760 using metabolic network topology. Proc Natl Acad Sci 2005;102:2685–9.
761 doi:10.1073/pnas.0406811102.
762 [35] He Y, Wu W, Zheng HM, Li P, McDonald D, Sheng HF, et al. Regional
763 variation limits applications of healthy gut microbiome reference ranges and
764 disease models. Nat Med 2018. doi:10.1038/s41591-018-0164-x.
765 [36] Jie Z, Xia H, Zhong SL, Feng Q, Li S, Liang S, et al. The gut microbiome in
766 atherosclerotic cardiovascular disease. Nat Commun 2017;8:845.
767 doi:10.1038/s41467-017-00900-1.
768 [37] Rebuffat S. Microcins in action: amazing defence strategies of Enterobacteria.

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

769 Biochem Soc Trans 2012;40:1456–62. doi:10.1042/BST20120183.
770 [38] Pereira CS, Thompson JA, Xavier KB. AI-2-mediated signalling in bacteria.
771 FEMS Microbiol Rev 2013;37:156–81.
772 doi:10.1111/j.1574-6976.2012.00345.x.
773 [39] Muth T, Behne A, Heyer R, Kohrs F, Benndorf D, Hoffmann M, et al. The
774 MetaProteomeAnalyzer: A powerful open-source software suite for
775 metaproteomics data analysis and interpretation. J Proteome Res
776 2015;14:1557–65. doi:10.1021/pr501246w.
777 [40] Wang G, Li X, Wang Z. APD3: The antimicrobial peptide database as a tool
778 for research and education. Nucleic Acids Res 2016;44:D1087–93.
779 doi:10.1093/nar/gkv1278.
780 [41] Song P, Onishi A, Koepsell H, Vallon V. Sodium glucose cotransporter SGLT1
781 as a therapeutic target in diabetes mellitus. Expert Opin Ther Targets
782 2016;20:1109–25. doi:10.1517/14728222.2016.1168808.
783 [42] Van De Laar FA, Lucassen PL, Akkermans RP, Van De Lisdonk EH, Rutten
784 GE, Van Weel C. ??-Glucosidase inhibitors for patients with type 2 diabetes:
785 Results from a Cochrane systematic review and meta-analysis. Diabetes Care
786 2005;28:154–63. doi:10.2337/diacare.28.1.154.
787 [43] Rittmeyer EN, Daniel S, Hsu S-C, Osman MA. A dual role for IQGAP1 in
788 regulating exocytosis. J Cell Sci 2008;121:391–403. doi:10.1242/jcs.016881.
789 [44] Chawla B, Hedman AC, Sayedyahossein S, Erdemir HH, Li Z, Sacks DB.
790 Absence of IQGAP1 protein leads to insulin resistance. J Biol Chem
791 2017;292:3273–89. doi:10.1074/jbc.M116.752642.
792 [45] Yip MF, Ramm G, Larance M, Hoehn KL, Wagner MC, Guilhaus M, et al.
793 CaMKII-Mediated Phosphorylation of the Myosin Motor Myo1c Is Required
794 for Insulin-Stimulated GLUT4 Translocation in Adipocytes. Cell Metab
795 2008;8:384–98. doi:10.1016/j.cmet.2008.09.011.
796 [46] Wiesner J, Vilcinskas A. Antimicrobial peptides: The ancient arm of the human
797 immune system. Virulence 2010;1:440–64. doi:10.4161/viru.1.5.12983.

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

798 [47] Vidarsson G, Dekkers G, Rispens T. IgG subclasses and allotypes: From
799 structure to effector functions. Front Immunol 2014;5.
800 doi:10.3389/fimmu.2014.00520.
801 [48] Nevalainen TJ, Graham GG, Scott KF. Antibacterial actions of secreted
802 phospholipases A2. Review. Biochim Biophys Acta - Mol Cell Biol Lipids
803 2008;1781:1–9. doi:10.1016/j.bbalip.2007.12.001.
804 [49] MacIvor DM, Shapiro SD, Pham CT, Belaaouaj A, Abraham SN, Ley TJ.
805 Normal neutrophil function in cathepsin G-deficient mice. Blood
806 1999;94:4282–93.
807 [50] Duranton J, Adam C, Bieth JG. Kinetic mechanism of the inhibition of
808 cathepsin G by α1- antichymotrypsin and α1-proteinase inhibitor. Biochemistry
809 1998;37:11239–45. doi:10.1021/bi980223q.
810 [51] Li Y, Komai-Koma M, Gilchrist DS, Hsu DK, Liu F-T, Springall T, et al.
811 Galectin-3 Is a Negative Regulator of Lipopolysaccharide-Mediated
812 Inflammation. J Immunol 2008;181:2781–9.
813 doi:10.4049/jimmunol.181.4.2781.
814 [52] Jiang K, Rankin CR, Nava P, Sumagin R, Kamekura R, Stowell SR, et al.
815 Galectin-3 regulates desmoglein-2 and intestinal epithelial intercellular
816 adhesion. J Biol Chem 2014;289:10510–7. doi:10.1074/jbc.M113.538538.
817 [53] Lu H, Yang Y, Allister EM, Wijesekara N, Wheeler MB. The Identification of
818 Potential Factors Associated with the Development of Type 2 Diabetes. Mol
819 Cell Proteomics 2008;7:1434–51. doi:10.1074/mcp.M700478-MCP200.
820 [54] Pedersen HK, Gudmundsdottir V, Nielsen HB, Hyotylainen T, Nielsen T,
821 Jensen BAH, et al. Human gut microbes impact host serum metabolome and
822 insulin sensitivity. Nature 2016;535:376–81. doi:10.1038/nature18646.
823 [55] Wilmes P, Heintz-Buschart A, Bond PL. A decade of metaproteomics: Where
824 we stand and what the future holds. Proteomics 2015;15:3409–17.
825 doi:10.1002/pmic.201500183.
826 [56] Heyer R, Schallert K, Zoun R, Becher B, Saake G, Benndorf D. Challenges and

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

827 perspectives of metaproteomic data analysis. J Biotechnol 2017;261:24–36.
828 doi:10.1016/j.jbiotec.2017.06.1201.
829 [57] Schubert OT, Röst HL, Collins BC, Rosenberger G, Aebersold R. Quantitative
830 proteomics: Challenges and opportunities in basic and applied research. Nat
831 Protoc 2017;12:1289–94. doi:10.1038/nprot.2017.040.
832
833

bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

834 # Figure Legends

835 **Figure 1. Experimental overview.**

836 254 participants were recruited from the Suzhou cohort and diagnosed as treatment
837 naive T2D patients (TN-T2D, n=77, red), prediabetic individuals (Pre-DM, n=80,
838 blue) or individuals with normal glucose tolerance (NGT, n=97, green). Each
839 participant provided two stool samples. One set of stool samples was used for
840 metagenomic shotgun sequencing, followed by IGC-based taxonomic and functional
841 analyses. The other set of stool samples, comprising a total of 84 samples with 28
842 age-, BMI- and sex-matched participants from each group, was selected for
843 metaproteomic analyses using isobaric tags for relative and absolute
844 quantitation (iTRAQ)–coupled-liquid chromatography tandem mass spectrometry
845 (iTRAQ-LC-MS/MS) to provide information on the microbial and host proteins
846 present in stool samples.

847 A total of 11, 980 meta-proteins and 425 human proteins were identified in this study.
848 Microbial gene and protein profiling were used to determine alterations in the
849 abundance of microbial taxa and functions, and human protein profiling was used to
850 identify alterations in the abundance of human proteins in faecal samples from NGT,
851 Pre-DM and TN-T2D individuals.

852

853 **Figure 2. Determination of alterations in the abundance of MLGs and functional**
854 **modules.**

855 **(A)** Heatmap of statistically significant annotated MLGs discriminating between
856 TN-T2D, Pre-DM and NGT based on Z-scores. Red, MLGs enriched in high glucose
857 groups, blue, MLGs enriched in low glucose groups. *, indicates MLGs significantly
858 differed between any two groups in the Suzhou cohort; *Dunn's post hoc* test, $P < 0.05$.
859 #, indicates significant MLGs replicated in the treatment naïve T2D patients from
860 Shanghai (Gu et al., 2017a) compared with Pre-DM and NGT in the Suzhou cohort;
861 Wilcoxon rank-sum test, $P < 0.05$ (See **Table S5** for full list).

862 **(B)** Performance of cross-validated random forest (RF) classification models using

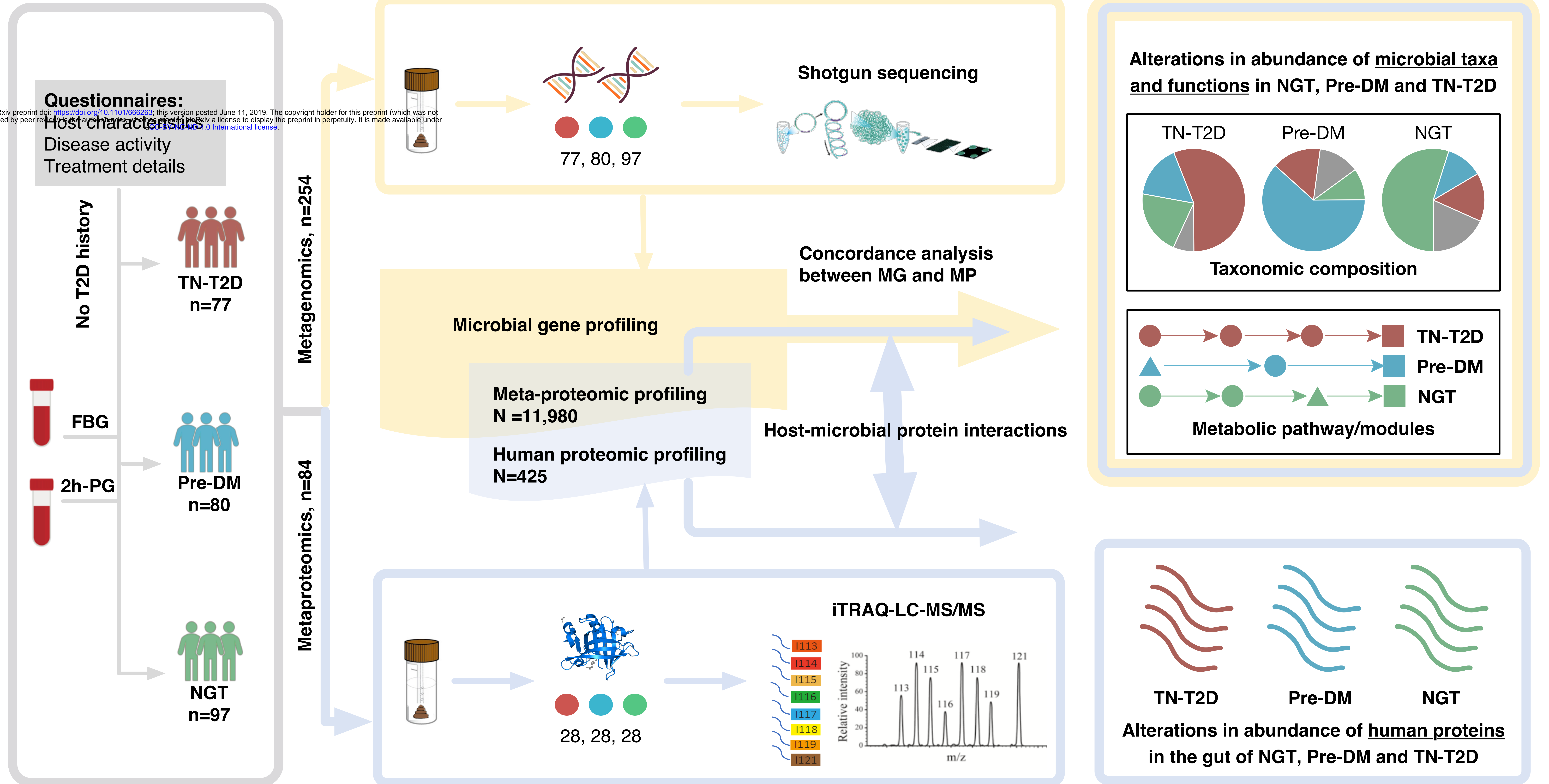
bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

863 relative abundance profiles of gut microbial MLGs, assessed by the area under the
864 ROC curve (AUC), 95% confidence intervals (CI). Orange, AUC for the RF model
865 classifying NGT (n=97) and Pre-DM (n=80). Grey, AUC for the RF model classifying
866 NGT (n=97) and TN-T2D (n=77). Blue, AUC for the RF model classifying Pre-DM
867 (n=80) and TN-T2D (n=77). The best cut-off points are marked on the ROC curves.
868 **(C)** Bar plot showing the 10 most discriminating MLGs in the RF models for
869 distinguishing between NGT, Pre-DM and TN-T2D. The bar lengths indicate the
870 importance of the selected MLGs, and colours represent enrichment in NGT (green),
871 Pre-DM (blue) and TN-T2D (red).
872 **(D)** Differential enrichment of KEGG modules comparing TN-T2D, Pre-DM and
873 NGT. Dashed lines indicate a reporter score of 1.96, corresponding to 95% confidence
874 in a normal distribution.
875
876 **Figure 3. Concordance and discordance of gut microbiome features in**
877 **metagenomes and metaproteomes.**
878 **(A)** Taxonomic distribution at the phylum level. Inner circle, metagenomes; Outer
879 circle, metaproteomes.
880 **(B)** Spearman's rank correlation between the median relative abundances of genera in
881 metagenomes of 84 samples selected for metaproteomics and the number of identified
882 meta-proteins assigned to the same genus. **(C)** Functional distribution at KEGG level
883 2. Inner circle, metagenomes; Outer circle, metaproteomes.
884 **(D-E)** Enrichment analysis of differentially expressed meta-proteins at taxonomic (d)
885 and functional levels (e) comparing NGT, Pre-DM and TN-T2D individuals. The
886 number of meta-proteins that exhibited significant differences in levels in each
887 pairwise comparison is shown. Colours represent enrichment in NGT (green),
888 Pre-DM (blue) and TN-T2D (red). Significant enrichment is defined as $P < 0.05$
889 (Wilcoxon rank-sum test) with a fold change of mean intensities > 1.2 in pairwise
890 comparisons.
891

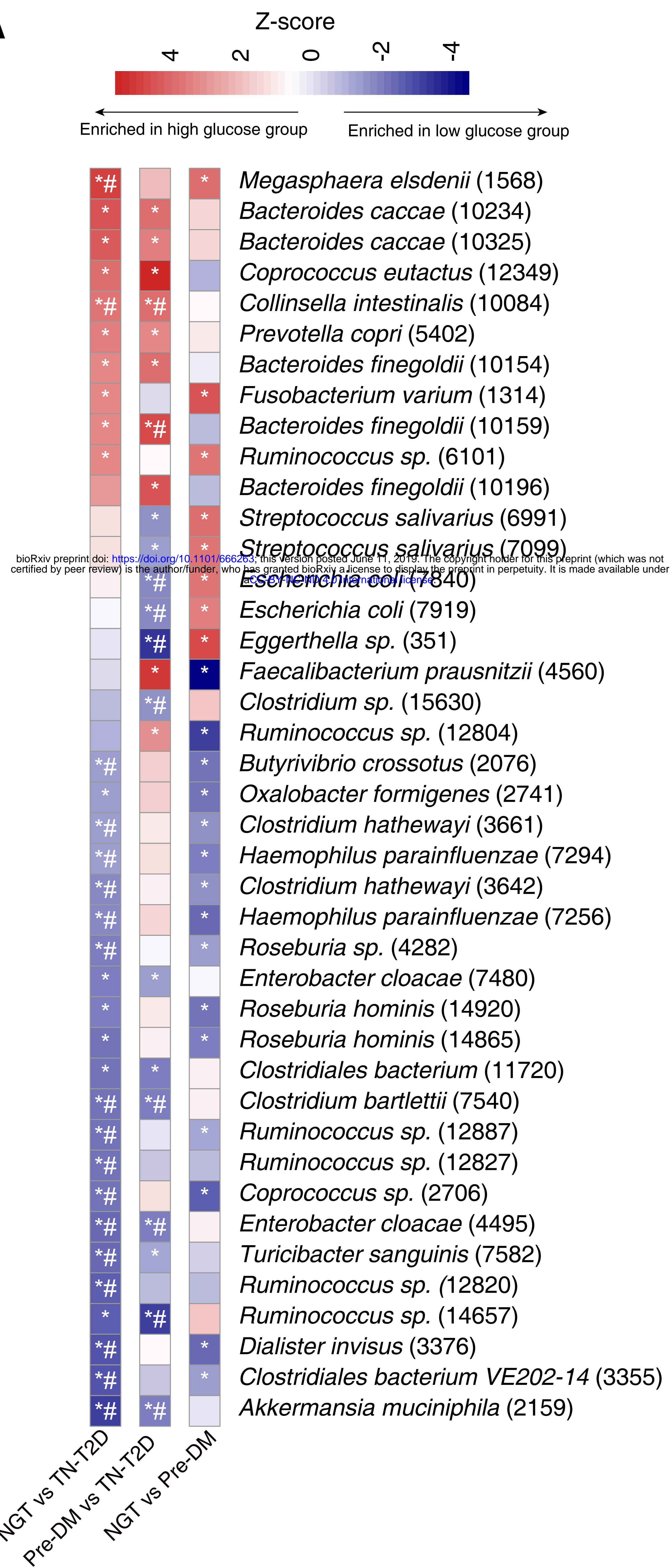
bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

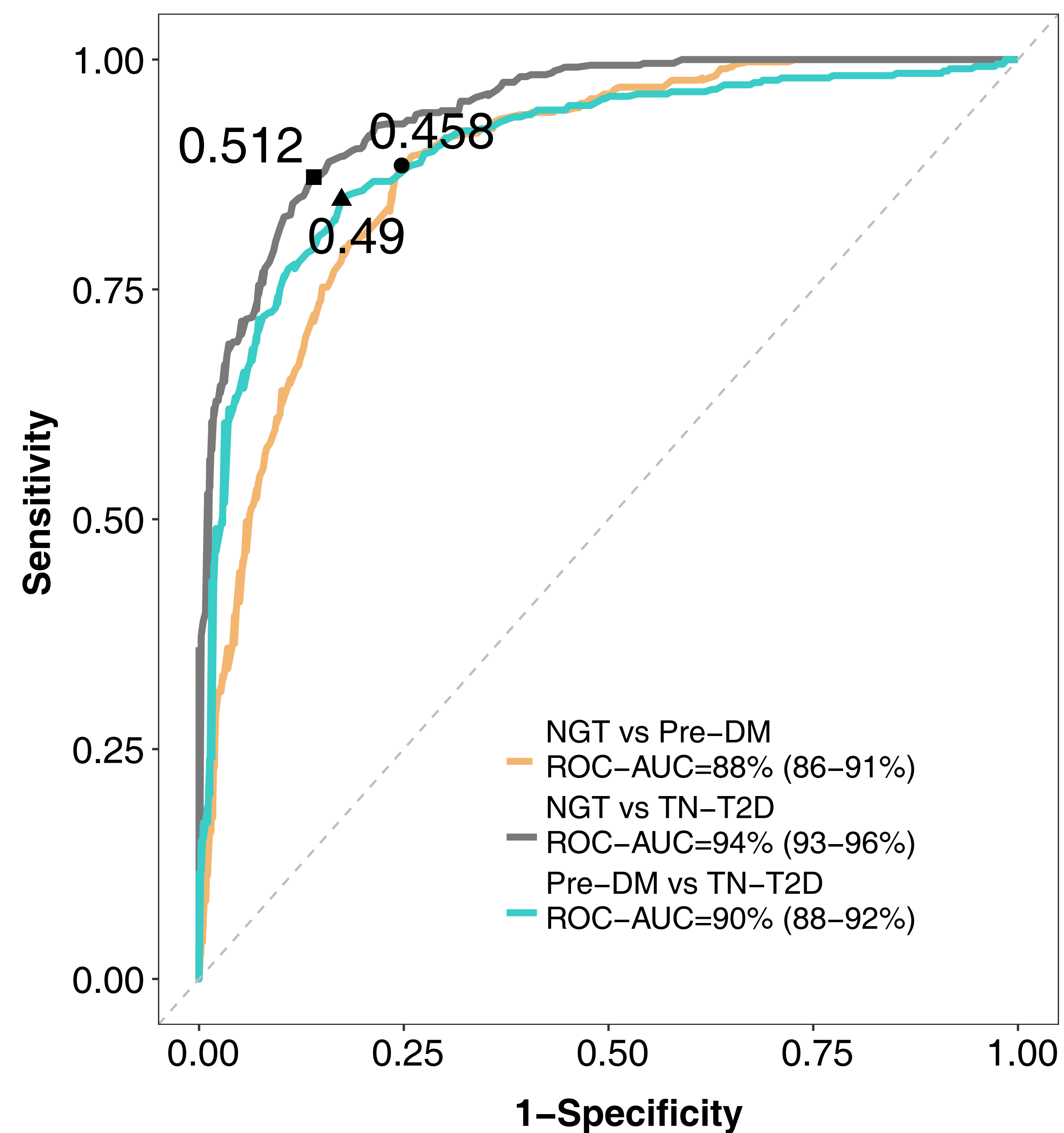
892 **Figure 4. Characterisation of human proteins in faecal samples from Chinese**
893 **NGT, Pre-DM, and TN-T2D individuals.**
894 **(A)** Heatmap showing levels of 49 discriminatory human proteins as fold change
895 between each two groups. *, $P < 0.05$ and fold change of protein levels > 1.2 or < 0.8.
896 **(B)** Protein-protein interaction network based on 20 discriminatory human proteins in
897 at least two pair-wise comparisons. The group signatures indicate human proteins
898 with significantly higher or lower levels in this group compared to others. Orange
899 indicates higher protein levels and blue indicates lower protein levels.
900 **(C)** Protein-protein interactions based on discriminatory meta-proteins in pair-wise
901 comparisons and discriminatory human proteins. Only discriminatory meta-proteins
902 annotated to the corresponding taxon of the MLGs were selected for the analysis.
903 The circles indicate human proteins and diamonds indicate meta-proteins. Detailed
904 information on the numbered meta-proteins is presented in **Table S12**. Colours
905 represent protein enrichment in NGT (green), Pre-DM (blue) and TN-T2D (red). Pink
906 line indicates positive correlation and grey line indicates negative correlation
907 (Spearman's rank correlations, adjusted $P < 0.05$).

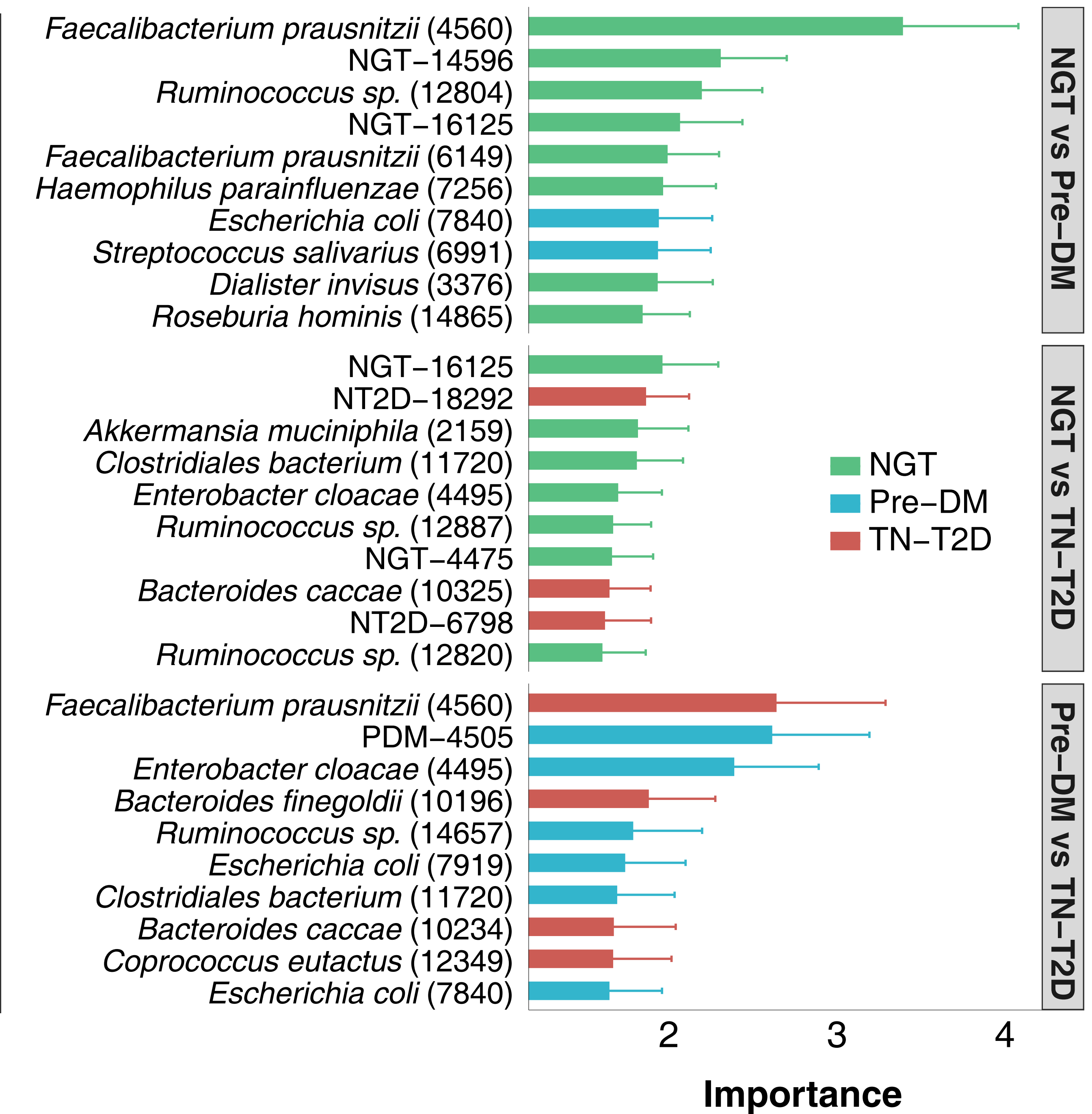
bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

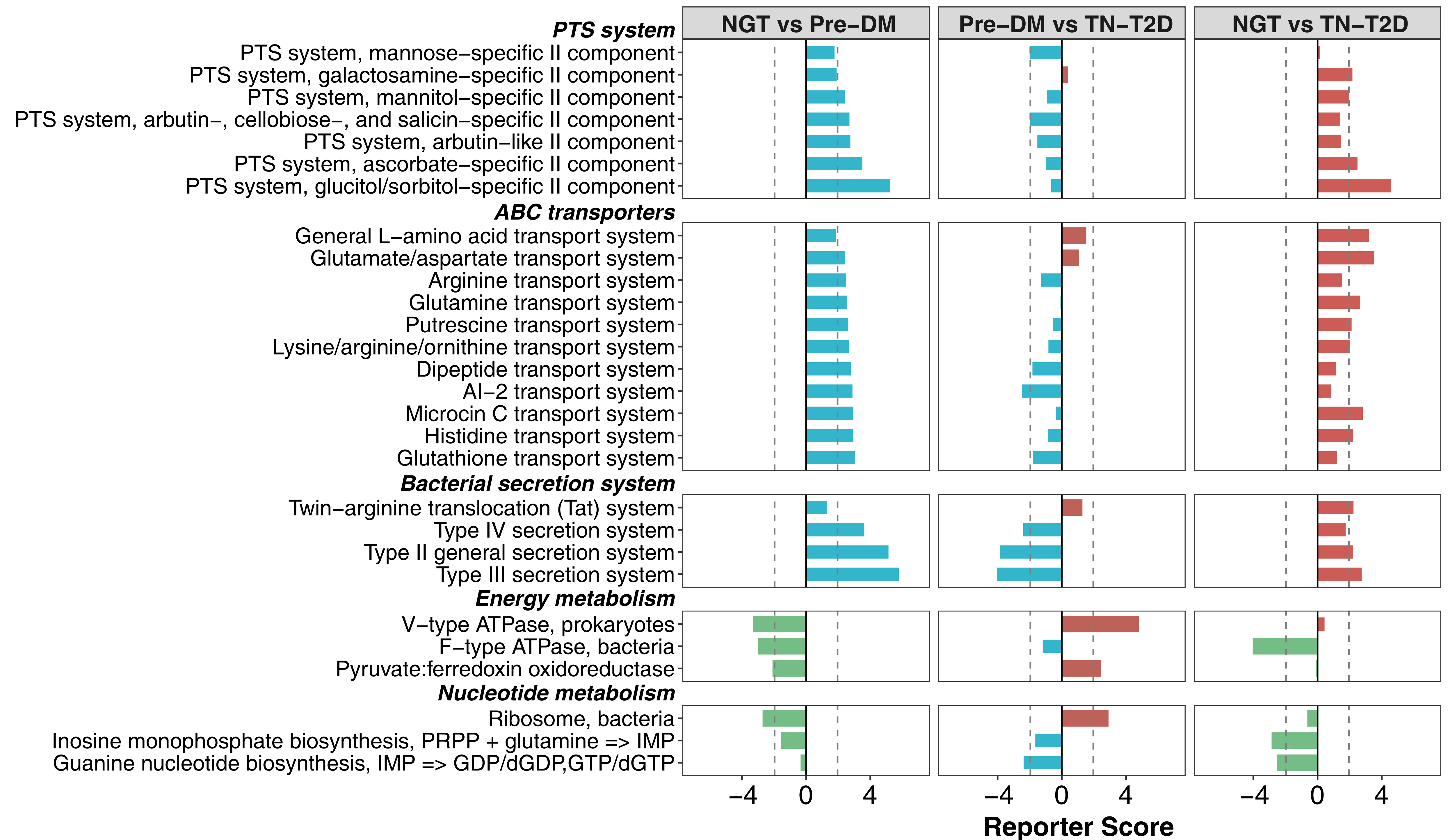


bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

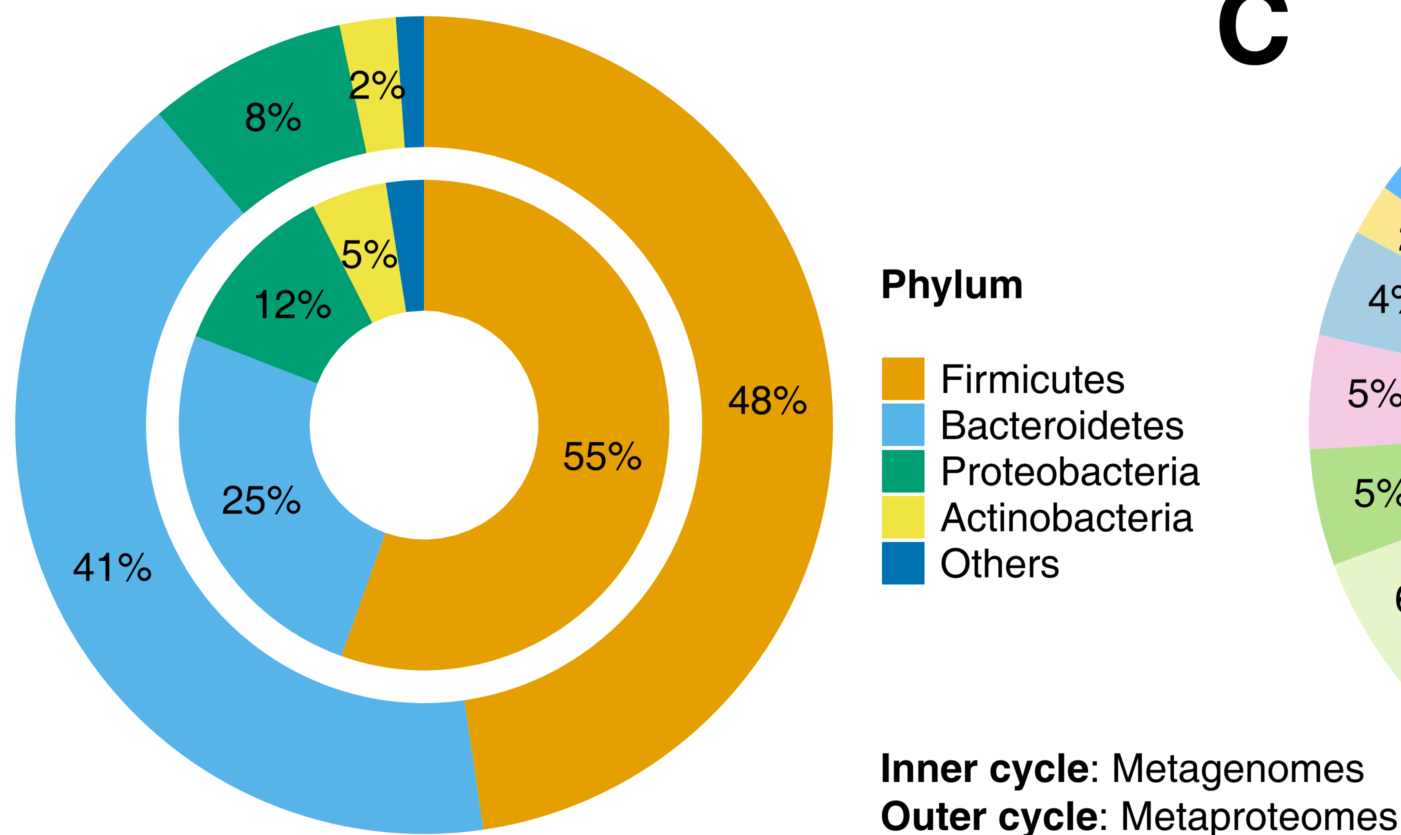


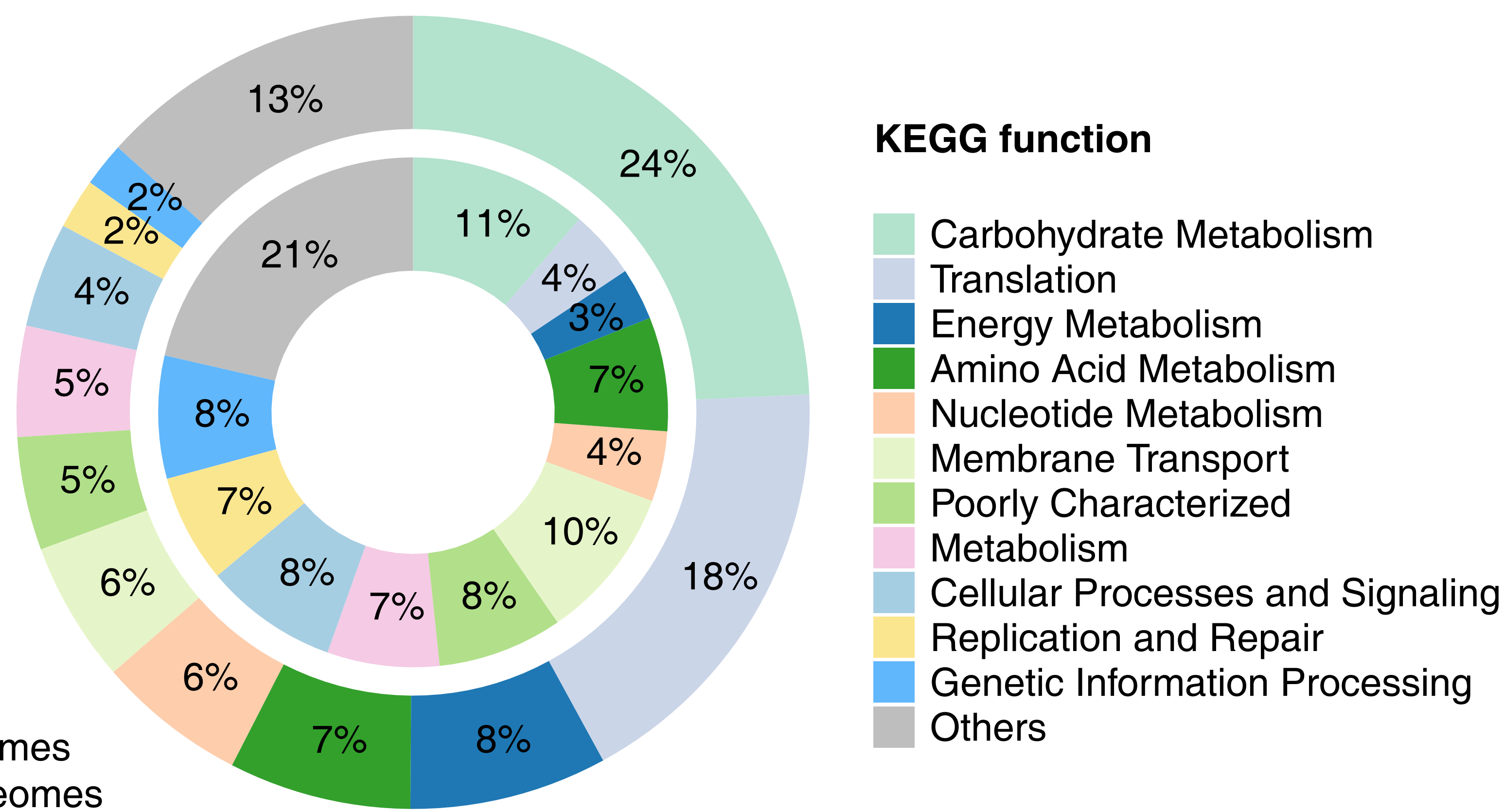


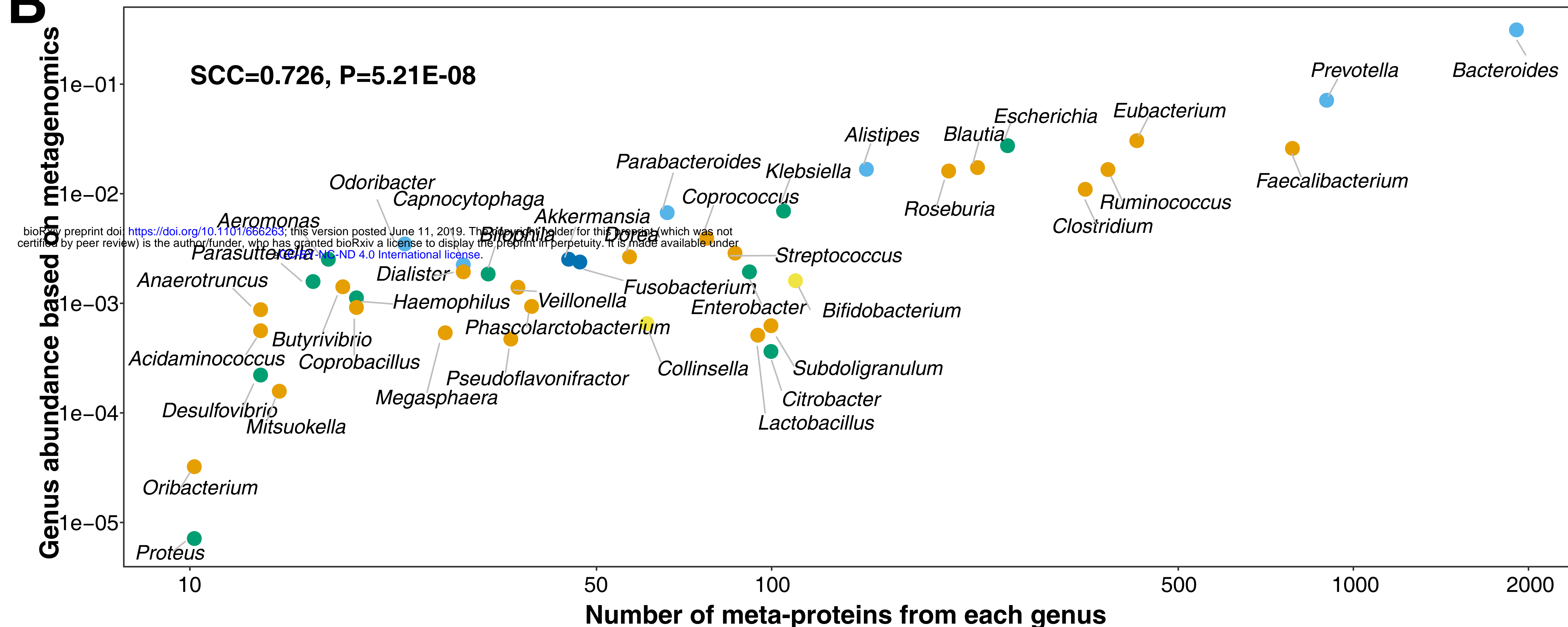


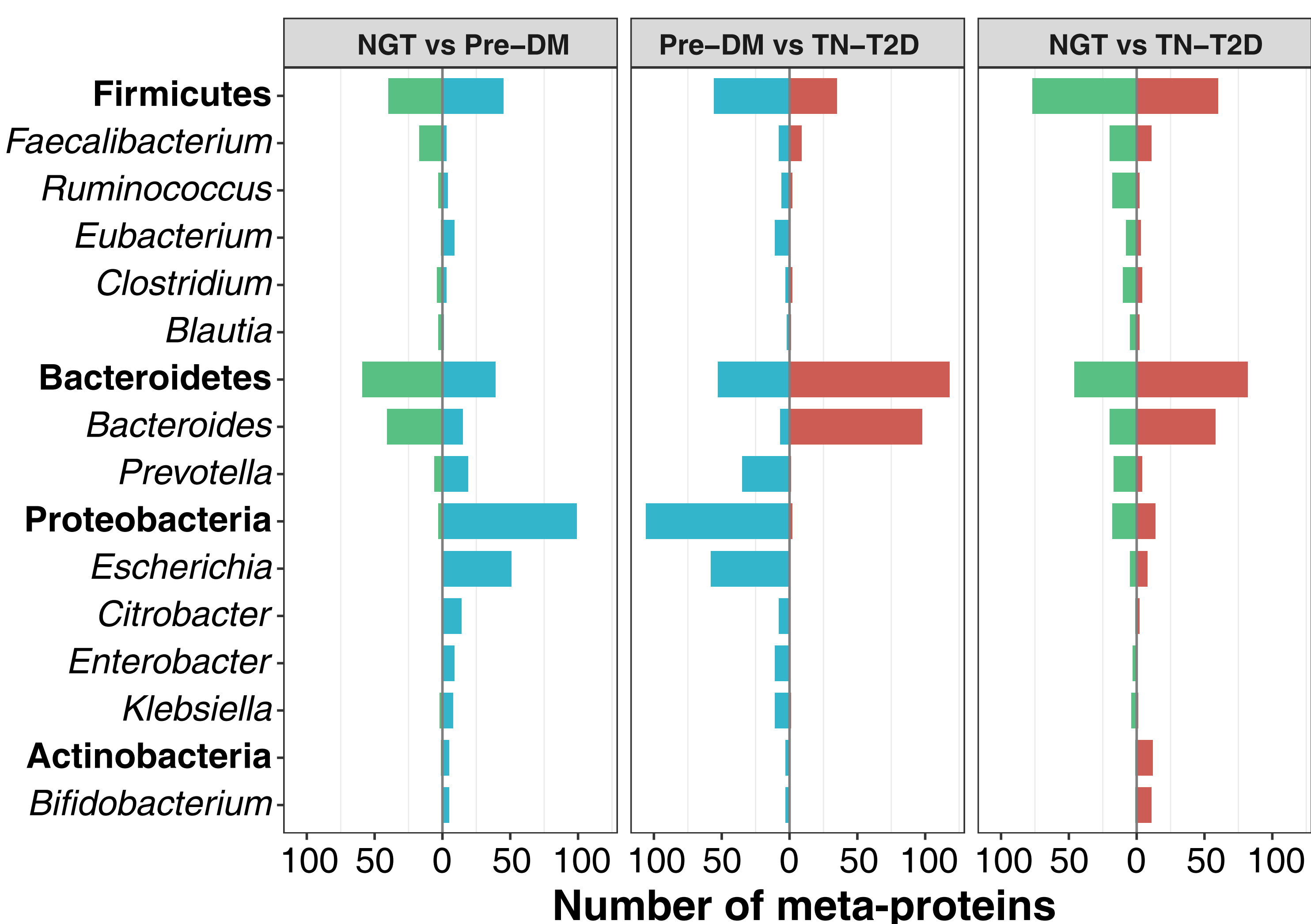


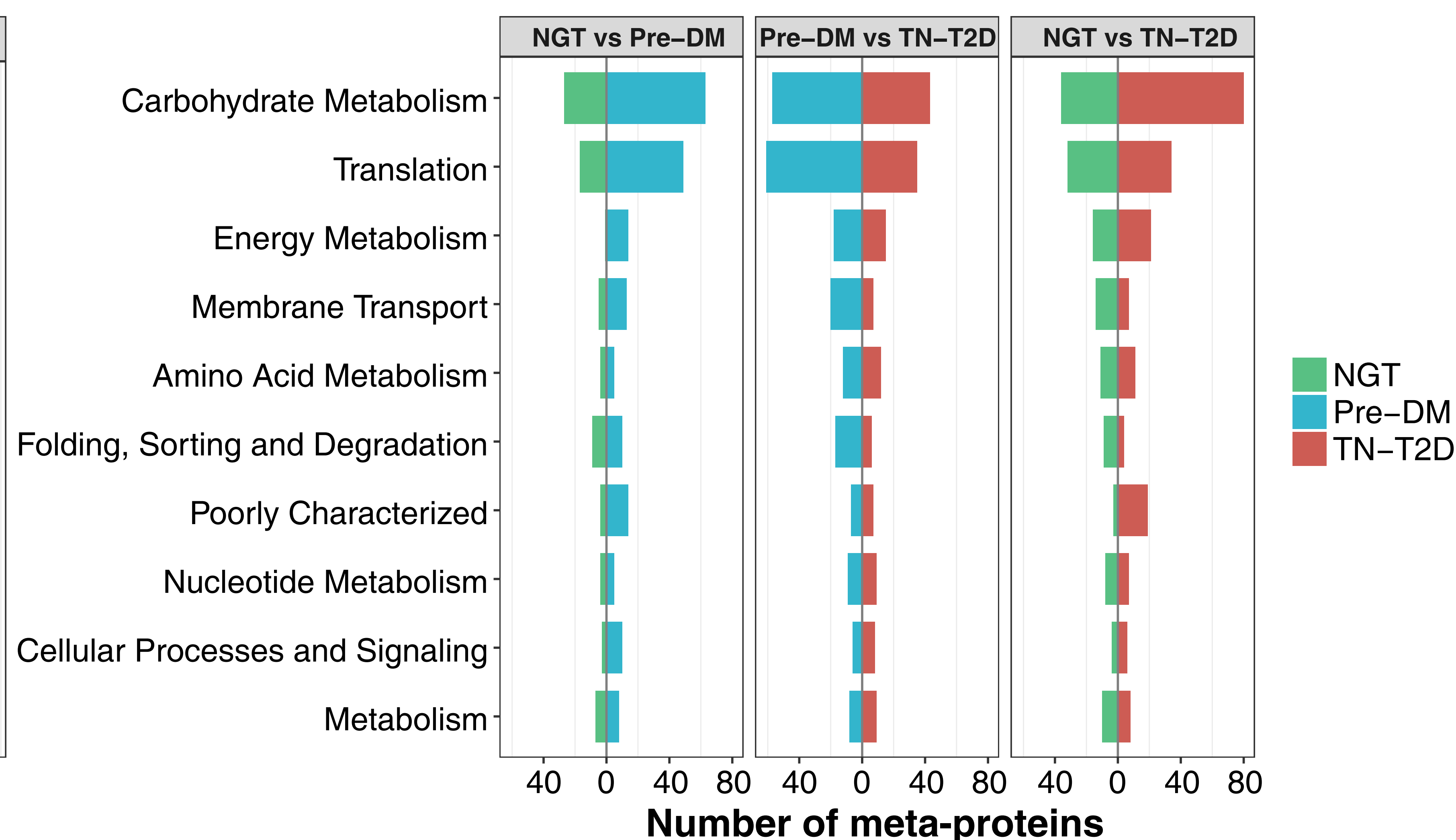
bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

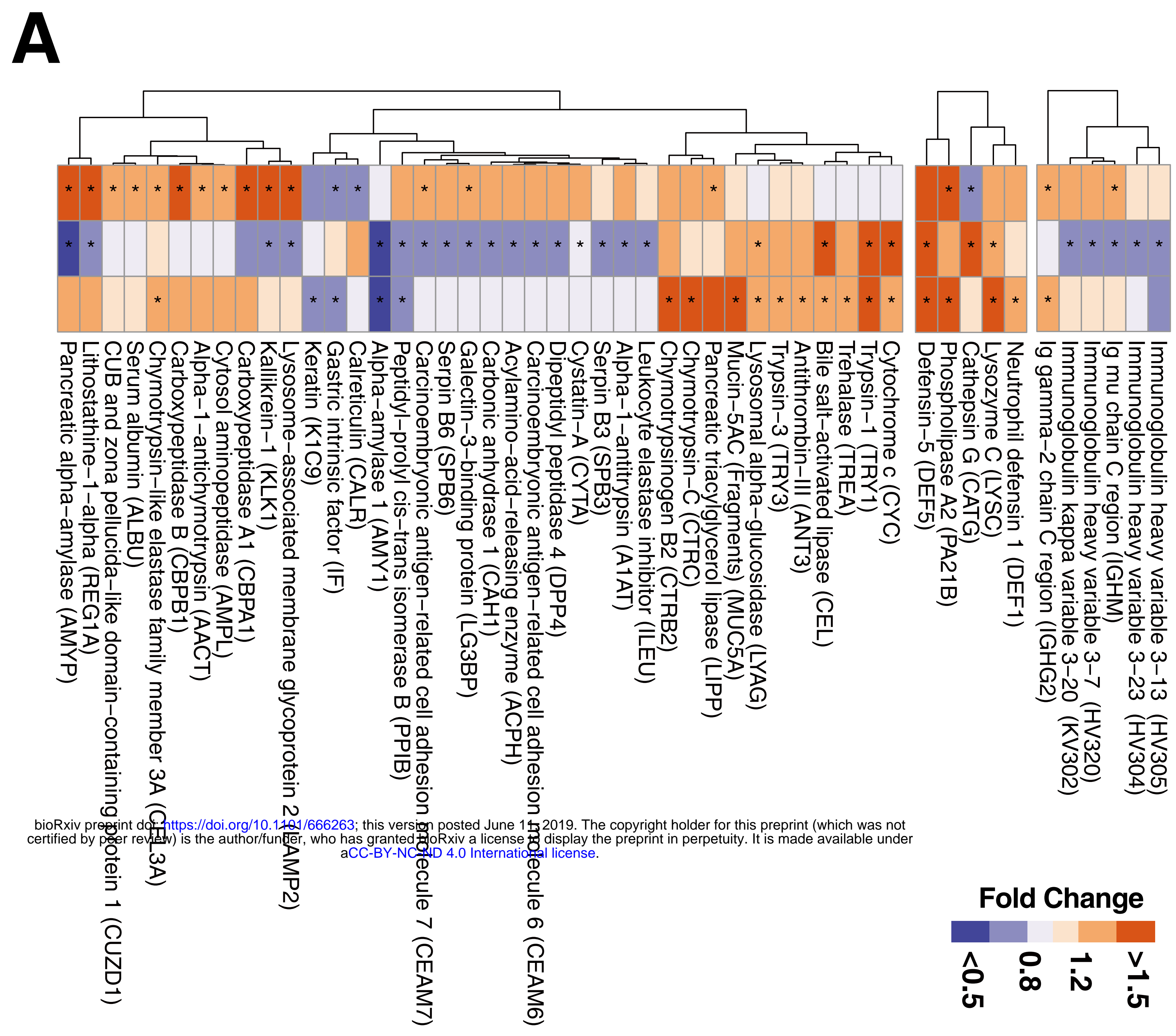












bioRxiv preprint doi: https://doi.org/10.1101/666263; this version posted June 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.