

1 **BiG-MAP: an automated pipeline to profile metabolic gene cluster abundance** 2 **and expression in microbiomes**

3

4 Victoria Pascal Andreu¹, Hannah E. Augustijn^{*1}, Koen van den Berg^{*1}, Justin J. J.
5 van der Hooff¹, Michael A. Fischbach^{2#}, Marnix H. Medema^{#1}

6

7 1. Bioinformatics Group, Wageningen University, Wageningen, the Netherlands

8 2. Department of Bioengineering and ChEM-H, Stanford University, Stanford, USA

9 *Contributed equally

10 #Corresponding authors: fischbach@fischbachgroup.org, marnix.medema@wur.nl

11

12 **Abstract**

13 Microbial gene clusters encoding the biosynthesis of primary and secondary
14 metabolites play key roles in shaping microbial ecosystems and driving microbiome-
15 associated phenotypes. Although effective approaches exist to evaluate the metabolic
16 potential of such bacteria through identification of metabolic gene clusters in their
17 genomes, no automated pipelines exist to profile the abundance and expression levels
18 of such gene clusters in microbiome samples to generate hypotheses about their
19 functional roles and to find associations with phenotypes of interest. Here, we describe
20 BiG-MAP, a bioinformatic tool to profile abundance and expression levels of gene
21 clusters across metagenomic and metatranscriptomic data and evaluate their
22 differential abundance and expression between different conditions. To illustrate its
23 usefulness, we analyzed 47 metagenomic samples from healthy and caries-
24 associated human oral microbiome samples and identified 58 gene clusters, including
25 unreported ones, that were significantly more abundant in either phenotype. Among
26 them, we found the *muc* operon, a gene cluster known to be associated to tooth decay.
27 Additionally, we found a putative reuterin biosynthetic gene cluster from a

28 *Streptococcus* strain to be enriched but not exclusively found in healthy samples;
29 metabolomic data from the same samples showed masses with fragmentation
30 patterns consistent with (poly)acrolein, which is known to spontaneously form from the
31 products of the reuterin pathway and has been previously shown to inhibit pathogenic
32 *Streptococcus mutans* strains. Thus, we show how BiG-MAP can be used to generate
33 new hypotheses on potential drivers of microbiome-associated phenotypes and
34 prioritize the experimental characterization of relevant gene clusters that may mediate
35 them.

36

37 **Importance**

38 Microbes play an increasingly recognized role in determining host-associated
39 phenotypes by producing small molecules that interact with other microorganisms or
40 host cells. The production of these molecules is often encoded in syntenic genomic
41 regions, also known as gene clusters. With the increasing numbers of (multi-)omics
42 datasets that can help understanding complex ecosystems at a much deeper level,
43 there is a need to create tools that can automate the process of analyzing these gene
44 clusters across omics datasets. The current study presents a new software tool called
45 BiG-MAP, which allows assessing gene cluster abundance and expression in
46 microbiome samples using metagenomic and metatranscriptomic data. In this
47 manuscript, we describe the tool and its functionalities, and how it has been validated
48 using a mock community. Finally, using an oral microbiome dataset, we show how it
49 can be used to generate hypotheses regarding the functional roles of gene clusters in
50 mediating host phenotypes.

51

52 **Running title:** BiG-MAP: profiling gene clusters across microbiomes

53 **Keywords: metabolic gene cluster, metagenomics, metatranscriptomics,**
54 **microbiome-associated phenotype**

55

56 **Introduction**

57 Bacteria can produce diverse sets of small molecules that interact with other microbes
58 or with their host. These metabolites include members of both primary and secondary
59 metabolism and cover a wide chemical diversity^{1,2}. These pathways and metabolites
60 are often specific to certain strains or species and help them to compete for space and
61 resources³, e.g. through antimicrobial, nutrient-scavenging or immunomodulatory
62 activities⁴. The genes that encode these pathways are often physically clustered and
63 are also known as Biosynthetic Gene Clusters (BGCs) or Metabolic Gene Clusters
64 (MGCs)^{5,6}—the latter being a broader definition that also includes catabolic pathways.
65 Several studies have indicated metabolites produced from such gene clusters to be
66 the major drivers of specific phenotypic traits; for instance, pseudomonads in the
67 rhizosphere of sugar beet plants were shown to produce the antifungal non-ribosomal
68 peptide (NRP) thanamycin, which protects plants from fungal infections⁷. Another
69 example from primary metabolism is trimethylamine, a diet derived-molecule that is
70 processed by bacteria harboring a gene cluster that includes both *CutC* and *CutD*, and
71 has been associated with an increased risk of suffering from cardiovascular disease⁸.
72 Therefore, mining genomes for BGCs or MGCs enables moving the field towards a
73 deeper understanding of function at the molecular level and determine the role a given
74 microbe plays in the ecosystem⁹.

75

76 Several tools have been developed to mine genomes for these gene clusters, like
77 antiSMASH¹⁰, gutSMASH

78 (<https://github.com/victoriapascal/gutsmash/tree/gutsmash/>) or DeepBGC¹¹. In
79 contrast to other tools for functional profiling of microbial communities, such as
80 HUMAnN2¹², MetaPath¹³, FMAP¹⁴ and Metatrans¹⁵, these do not depend on pathways
81 that are present in reference databases like KEGG¹⁶ or MetaCyc¹⁷, which only include
82 pathways for which most or all enzymatic steps have been elucidated. In fact, the
83 majority of gene clusters identified by antiSMASH and many gene clusters predicted
84 by gutSMASH encode pathways for which the catalytic steps, intermediates, and final
85 products are yet unknown. However, known pathways that are encoded by gene
86 clusters can also be reliably detected. The detection of complete gene clusters instead
87 of individual enzyme-coding genes likely decreases false positive detections of
88 enzymes that show sequence similarity to reference enzyme sequences but are part
89 of different functional contexts. For these reasons, identification of gene clusters of
90 known and unknown function provides a useful basis to look for functional
91 explanations of microbiome-associated phenotypes of interest. As phenotypes are
92 often triggered by metabolites at physiologically relevant concentrations, while
93 samples without the phenotype lack these metabolites or have them at lower
94 concentrations, assessing gene cluster abundance and expression levels across
95 samples is crucial to predict associations with the phenotype in question. Another
96 significant advantage of profiling the community by combining different omics data is
97 to prioritize the characterization of putative gene clusters that are highly abundant or
98 expressed in samples of interest and thus, help elucidating novel compounds and their
99 biosynthetic pathways.

100

101 Here, we present designed BiG-MAP (Biosynthetic Gene cluster Meta'omics
102 Abundance Profiler), which provides a streamlined and automated process to

103 determine BGC/MGC abundance and expression in bacterial communities by
104 mapping metagenomic and metatranscriptomic reads to gene cluster sequences from
105 reference genomes or metagenomic assemblies. BiG-MAP uses MinHash-based
106 redundancy filtering and groups BGCs into families with BiG-SCAPE¹⁸ to avoid
107 ambiguous mapping, and uses these to output and visualize profiles of MGC
108 abundance or expression levels across samples. Additionally, it calculates differential
109 abundance or expression using either parametric or nonparametric tests. We validate
110 the tool using simulated metagenomic data and show how MGC abundance and
111 expression levels are accurately recapitulated. Finally, to showcase its usefulness, we
112 applied BiG-MAP on a large publicly available metagenome dataset from the human
113 oral microbiome and describe how it successfully identified gene clusters related to
114 bacteria's specialized primary and secondary metabolism that are (potentially)
115 relevant for caries development. Among others, this collection includes the previously
116 reported *pdu* and cobalamin gene cluster involved in the reuterin synthesis and the
117 *muc* operon, gene clusters that were predicted by gutSMASH and antiSMASH,
118 respectively. Thus, BiG-MAP suggests new lines to explore further the onset and
119 development of oral cavities.

120

121 **Results and discussion**

122 **An approach to map metagenomics and metatranscriptomic reads to gene** 123 **clusters**

124 BiG-MAP maps shotgun sequencing reads onto gene clusters that have been either
125 predicted by antiSMASH¹⁹ or gutSMASH (manuscript in preparation). It is a Python-
126 based pipeline, which allows downloading datasets from SRA repository, aligning
127 metagenomic or metatranscriptomic reads to gene clusters detected in reference

128 genome collections or in a metagenomic assembly, providing normalized counts
129 across samples, performing differential analyses, and visualizing the results. The tool
130 requires three main inputs: (1) a gene cluster collection obtained from running any
131 “SMASH-based” algorithm, (2) the meta’omic dataset in FASTQ or FASTA format or,
132 alternatively, the Sequence Read Archive (SRA) accession numbers to download it,
133 and (3) a metadata file with sample information to segregate them into groups and
134 compare their gene cluster content.

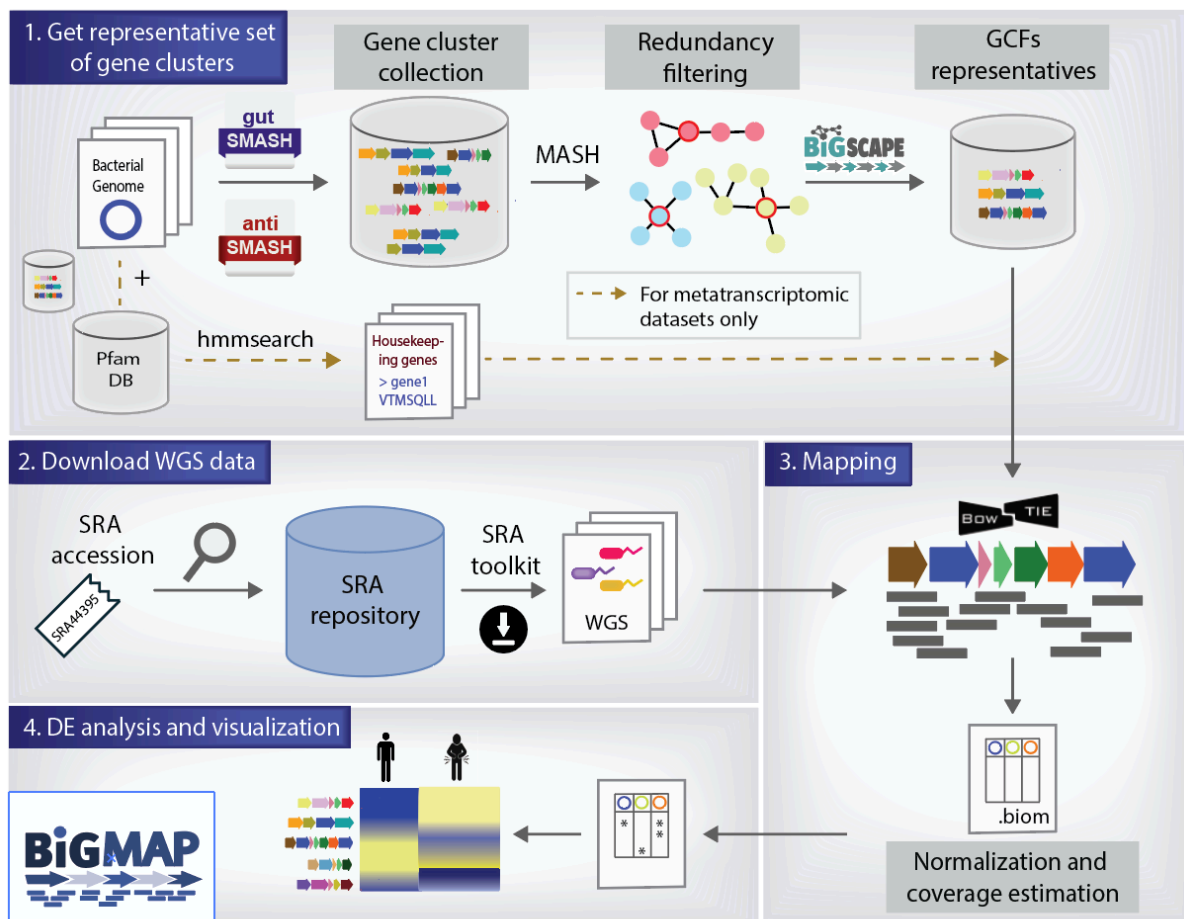
135

136 BiG-MAP is composed of four different modules (see Fig. 1): (1) BiG-MAP.family,
137 which performs redundancy filtering on the input collection of predicted gene clusters
138 and provides a set of representative gene clusters for the mapping process. (2) BiG-
139 MAP.download, which uses a list of SRA accession ids to download the shotgun data
140 if present in the SRA database (this step is optional). (3) BiG-MAP.map, which maps
141 reads from the metagenomic or metatranscriptomic samples onto the set of
142 representative gene clusters obtained from BiG-MAP.family. (4) BiG-MAP.analyse,
143 which normalizes the counts for sparsity and sequencing depth, performs differential
144 abundance/expression analysis and visualizes the output.

145 The BiG-MAP.family module performs a redundancy analysis on the gene cluster
146 collection to remove almost identical sequences, in order to reduce the computing time
147 and avoid ambiguous mapping. To achieve this, the protein sequences of the gene
148 clusters are used as input for MASH²⁰, a MinHash-based algorithm to estimate
149 sequence distance. Next, a representative gene cluster is selected using medoids
150 calculation. The resulting representatives are then clustered into Gene Cluster
151 Families (GCFs) using BiG-SCAPE¹⁸, an algorithm that uses three different distance
152 metrics to group MGCs into families based on sequence and architectural similarity.

153 This step helps to group more distantly related homologous gene clusters that likely
154 have the same chemical products but that are encoded in more distantly related
155 organisms. In such cases, BiG-MAP maps reads to the family representatives
156 separately, but also allows reporting combined abundance or expression levels per
157 family to find associations with phenotypes at a higher level. In order to set an
158 expression baseline when using metatranscriptomic data, BiG-MAP screens bacterial
159 genomes whose gene clusters have been included in the non-redundant
160 representative set of gene clusters for five house-keeping genes known to have stable
161 expression levels using HMMer (for details, see Methods section titled *BiG-*
162 *MAP.family: Creating a non-redundant MGC representative collection*). Next, the
163 reads are mapped to the representative gene clusters using the short-read aligner
164 Bowtie2²¹. The obtained raw read counts are then converted to RPKM (Reads Per
165 Kilobase Million) values, which are averaged over the GCF size (based on BiG-
166 SCAPE clustering). In the last module, RPKM values are then normalized using
167 Cumulative Sum Scaling²² (CSS) to account for sparsity. Moreover, for each aligned
168 gene cluster we assess its coverage to control for gene clusters that are only partially
169 mapped to by meta'omic reads. We report two coverage values in the intermediate
170 files; one for the whole gene cluster and the other considering only the core genes of
171 the BGC/MGC; showing both these numbers is often insightful in cases where borders
172 of gene clusters called by antiSMASH or gutSMASH are imprecise and reads may be
173 mapped to regions flanking the actual gene cluster. Subsequently, BiG-MAP detects
174 differentially abundant or expressed gene clusters by using either zero-inflated
175 gaussian distribution mixture models (ZIG-models) or using a Kruskal-Wallis model.
176 Finally, all the generated results are displayed into a plot that includes a heatmap for
177 the gene clusters abundance/expression values, a bar plot for the log fold change, the

178 coverage values and finally another heatmap for the housekeeping gene expression
179 values when analyzing metatranscriptomes (see Suppl. Fig. S2). The output folders
180 contain different intermediate and final results as for instance the BiG-SCAPE results,
181 the resulting bedgraphs, the raw and normalized RPKM counts for each sample (in
182 BIOM format²³) and after applying the fitZIG and Kruskal Wallis tests in tab-separated
183 tables and mapping coverage values for each gene cluster and sample. Altogether,
184 this tool presents a streamlined method to functionally profile meta'omics data by
185 mapping reads to known or putative gene clusters.



186 GCF (Gene Cluster Family), WGS (Whole Genome Sequencing), DE (Differential Expression)

187 **Figure 1. BiG-MAP workflow.** BiG-MAP is composed of four different modules: (1)
188 BiG-MAP.family returns a representative set of non-redundant gene clusters based on
189 sequence similarity, given a set of predicted gene clusters by either gutSMASH or
190 antiSMASH. This module also looks for the protein sequences of 5 housekeeping

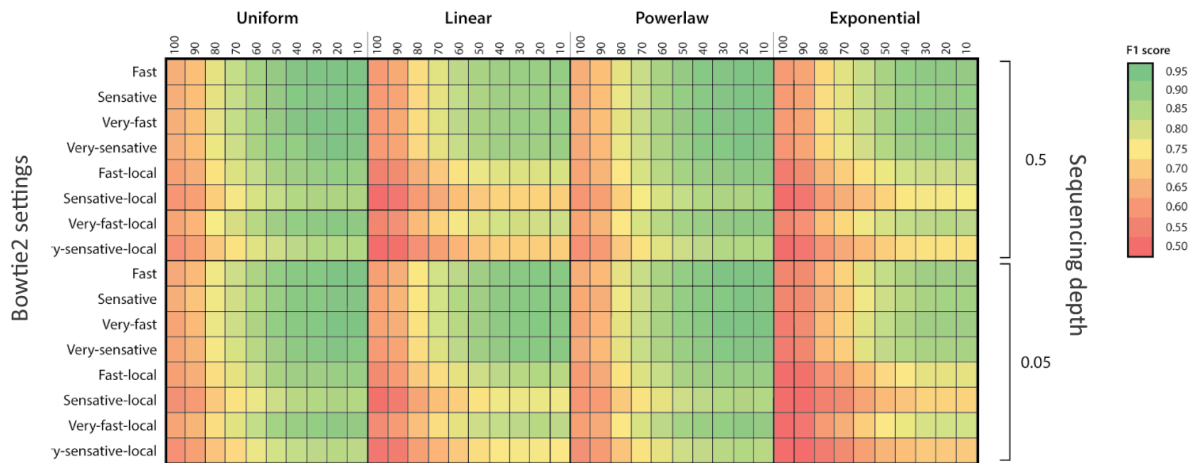
191 genes from the bacteria encoding the representative gene clusters when reads from
192 metatranscriptomic sequences are going to be used. (2) BiG-MAP.download
193 downloads a set of metagenomes/metatranscriptomes based on their SRA
194 accessions. (3) BiG-MAP.map aligns omics reads to the representative set of gene
195 clusters using Bowtie and (4) BiG-MAP.analyse computes normalized read counts,
196 performs differential abundance/expression analysis of gene clusters across different
197 conditions, and visualizes the results (see Suppl. Fig S1 and S2 as an example).

198

199 **Assessing and validating BiG-MAP performance using simulated data**

200 In order to evaluate the overall performance of BiG-MAP and in particular, all the
201 default parameters chosen as defaults, such as the Bowtie alignment mode and the
202 MASH similarity score cut-off, we designed a mock microbial community for
203 metagenome simulation. From the Culturable Genome Reference (CGR) genome
204 collection²⁴, we randomly chose 101 CGR genomes to simulate metagenome reads
205 from and to use as input for gutSMASH. To assess the impact of different sequencing
206 depths (coverage of 0.5x and 0.05x) and community structure (uniform, linear, power-
207 law and exponential), we simulated eight different metagenomic libraries. Since the
208 gene cluster content and their abundance levels in simulated data is known (ground
209 truth), this allowed us to assess the recall and precision of the BiG-MAP assignments
210 using MASH dissimilarity scores ranging from 10-100 and the eight different alignment
211 modes available in Bowtie across the eight different simulated data libraries. From
212 these results we computed the F1-score or harmonic mean of precision and recall (see
213 Fig. 2), which showed that the community structure slightly affects BiG-MAP results.
214 Moreover, since the highest F1 scores were obtained when using MASH score cut-off
215 (similarity) of 0.8 and using “fast” alignment mode (end-to-end), we set these

216 parameters as defaults. Still, the user is able to change them as desired by indicating
217 it with the appropriate flag when running BiG-MAP.
218



219
220 **Figure 2. BiG-MAP validation using simulated metagenomes.** F1 score heatmap
221 using simulated metagenomes constructed to assess the best MASH dissimilarity cut-
222 off across four different microbial community structures, two different sequencing
223 depth values and eight different Bowtie alignment modes.

224
225 **Analysis of the oral microbiome: revealing the presence of gene clusters**
226 **associated with health and disease**

227 The oral cavity is a natural habitat for many bacteria that reside in or on the gingival
228 sulcus, tongue, teeth and cheeks, among other surfaces. These bacteria take part in
229 important processes such as initial digestion of food, but are also associated with
230 several oral diseases such as caries²⁵ and periodontitis²⁶. It is known that these
231 bacteria can organize themselves to form biofilms, which can play a causal role in the
232 development of these diseases²⁷. There are different functional and metabolic
233 pathway alterations that have been associated with the onset of disease via the
234 production of small molecules^{28,29,30,31}. For instance, tetramic acid produced by the

235 caries-associated bacterium *Streptococcus mutans* has been linked to tooth decay³².
236 For this reason, in order to functionally profile these oral communities and acquire
237 further insights into the MGCs that might be involved, we studied a dataset of 47 oral
238 microbiome samples³⁰ for which paired metagenomics and metabolomics data have
239 been acquired and further analyzed using BiG-MAP (see *Methods Assessing the pdu*
240 *operon abundance by surveying different oral metagenomic samples* and *Evaluating*
241 *the presence of the muc operon in caries-associated metagenomes* sections).
242
243 To evaluate possible molecular mechanisms underpinning caries formation, we first
244 analyzed the available MS/MS data together with the metabolite feature abundance
245 table using Pathway Activity Level Scoring (PALS)³³, which uses molecular families
246 obtained using molecular networking³⁴ to group similar metabolites, and PLAGES³⁵ to
247 find differentially expressed metabolite groups between two conditions. PALS showed
248 a very consistent and strong differential abundance between healthy and caries
249 volunteers of a number of features in a metabolite group that we could annotate with
250 polymer-like structures based on their C₃H₄O mass differences. With MASST
251 searches³⁶ across all public data present in GNPS-MassIVE, we could confirm the
252 occurrence of these differential features in various microbial, human, and
253 environmental-related public datasets (see *Methods* and *Supplementary Methods* for
254 further information on the metabolomics data analysis). Based on the above
255 information, we concluded that these polymer-like structures might well represent
256 molecules called polyacroleins (metabolite identification level 3 - annotated compound
257 class), which are known to spontaneously form from a component of the antimicrobial
258 set of molecules called reuterin³⁷, and which have a matching mass difference
259 between different polymer lengths. The formation of (poly)acrolein has been shown to

260 contribute strongly to the antimicrobial activity of reuterin³⁷. Reuterin is produced by
261 lactobacilli from a genomic island containing a *pdu-like* operon together with a
262 cobalamin biosynthetic gene cluster³⁸. Of note, acrolein is an ubiquitous compound
263 that can be found in the human body for various reasons, such as the endogenous
264 production of it, the ingestion of different food sources or due to exposure to different
265 environmental conditions³⁹. There are various known routes that can converge into
266 the formation of acrolein, as it can be formed spontaneously from glycerol and 3-
267 hydroxypropionaldehyde³⁷. Furthermore, glycerol metabolism from gut bacteria has
268 also been found to produce this molecule⁴⁰. Typically, the acrolein polymerization
269 occurs under alkaline conditions⁴¹, thus, it is more likely to accumulate in saliva from
270 healthy samples, as caries typically acidifies the oral cavity. Indeed, our results show
271 that the possible polyacroleins are more abundant in samples of healthy volunteers.
272 Interestingly, the presence of acrolein has been linked to inhibition of *Streptococcus*
273 *mutans*, a well-known cariogenic bacteria^{42,43}.

274

275 Based on these findings, we were motivated to look for the presence of the *pdu* operon
276 in the metagenomics samples, in order to identify candidate MGCs that might be
277 involved in acrolein formation. To this end, we ran gutSMASH on the 1,440 genomes
278 from the Human Microbiome Oral Database (HMOD, <http://www.homd.org/>) available
279 in April 2020. Interestingly, gutSMASH identified a *pdu-like* operon in the genome of
280 *Streptococcus sp. F0442* that also includes a cobalamin (vitamin B12) biosynthetic
281 region and is architecturally similar (cumulative Blast bit score of 13,271) to the
282 *Lactobacillus reuteri* one (see Fig. 3A). Therefore, to assess the abundance of the
283 predicted gene clusters in the oral microbiome we used our gutSMASH run, which
284 predicted 3,352 gene clusters, as input for the BiG-MAP.family module, to filter out

285 redundant MGCs. Next, the reads of the 47 oral metagenomes (24 healthy and 23
286 caries-related) were mapped onto the 1,544 representative gene clusters using BiG-
287 MAP.map and the counts were further normalized and parsed with BiG-MAP.analyse.
288 We found that 56 gene clusters predicted by gutSMASH were significantly differentially
289 abundant between caries-related and healthy samples when using Kruskal Wallis.
290 Despite the fact that the *pdu* operon was not among these, we could see that it was
291 still somewhat more abundant in healthy samples (mean: 5.30 RPKM counts/sample)
292 when compared to the diseased group (mean: 4.16 RPKM counts/sample). Motivated
293 by this, we sought to assess its presence in a larger oral microbiome dataset by using
294 48 paired publicly available paired-end metagenome samples, which also included
295 metagenomes from samples suffering from periodontitis and plaque formation, all
296 considered as disease-related samples. These were used in combination with the
297 already analyzed ones, making a total of 96 samples; 33 caries-related, 34 healthy,
298 10 periodontitis-related and 19 involved in plaque development and all were used as
299 input for BiG-MAP (see Methods section titled *Assessing the pdu operon abundance*
300 *by surveying different oral metagenomic samples*). From this run, we found 164 gene
301 clusters differentially abundant between groups (using Kruskal Wallis test), and the
302 *pdu* operon was among them. While healthy samples on average have 5.15 RPKM
303 counts/sample mapping to this gene cluster, diseased ones have 3.05 (p-value=
304 0.0004 using Kruskal Wallis). We also evaluated the coverage of the read mapping
305 within the expanded metagenomic datasets and found that within healthy samples,
306 not all samples contain this gene cluster. For instance, from 34 healthy samples in the
307 extended dataset, we could find 15 of them that appear not to have the *Streptococcus*
308 *sp. F0442 pdu* operon (coverage below 0.5), while the rest had fairly high coverage
309 scores with a mean coverage value of 0.79 (selecting the samples with coverage

310 values of at least 0.5), implying the presence of this operon or a close homologue of
311 it (see Fig. 3B). Overall, this MGC constitutes a potential source for polyacrolein
312 production, and the hypothesis that it could be involved in inhibition of *Streptococcus*
313 *mutans* strains in non-acidic conditions is intriguing. As, logically, expression of the
314 MGC would be required for conferring a metabolic and potentially disease-suppressive
315 phenotype, metatranscriptomics analysis of samples where putative polyacrolein
316 accumulation is observed could be an interesting follow-up analysis in the future to
317 test the hypothesis of the involvement of this MGC in its production. Additionally, more
318 detailed chemical analysis of the putative polyacroleins is required to confirm their
319 structural identity. Nonetheless, this analysis illustrates how BiG-MAP analysis,
320 especially when combined with complimentary omics data such as metabolomics, can
321 generate concrete and relevant hypotheses about microbiome-associated phenotypes
322 that can be tested in the laboratory.

A

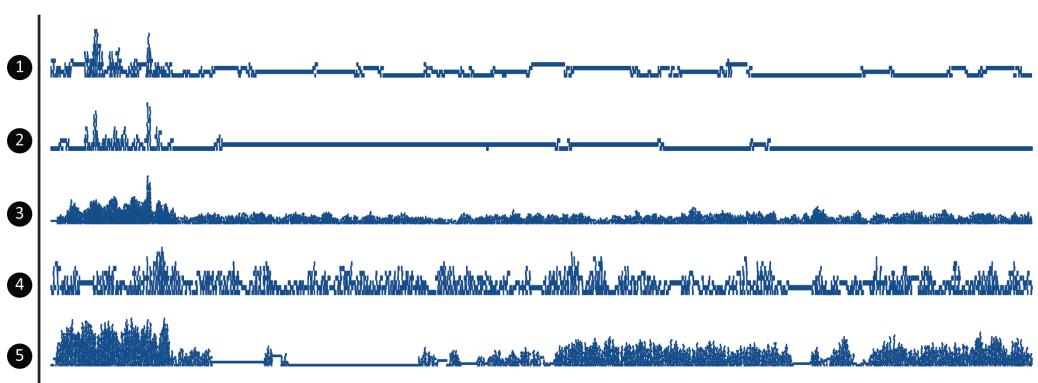
Streptococcus sp. F0442 | KB373314.1 | gutSMASH-predicted *pdu* operon



Lactobacillus reuteri JCM 1112 | AP007281.1 | *pdu* operon



B



○ *pdu* metabolic genes ● Cobalamin biosynthetic genes

323

324 **Figure 3. Detection of a *pdu* / cobalamin operon in healthy oral metagenomes.**

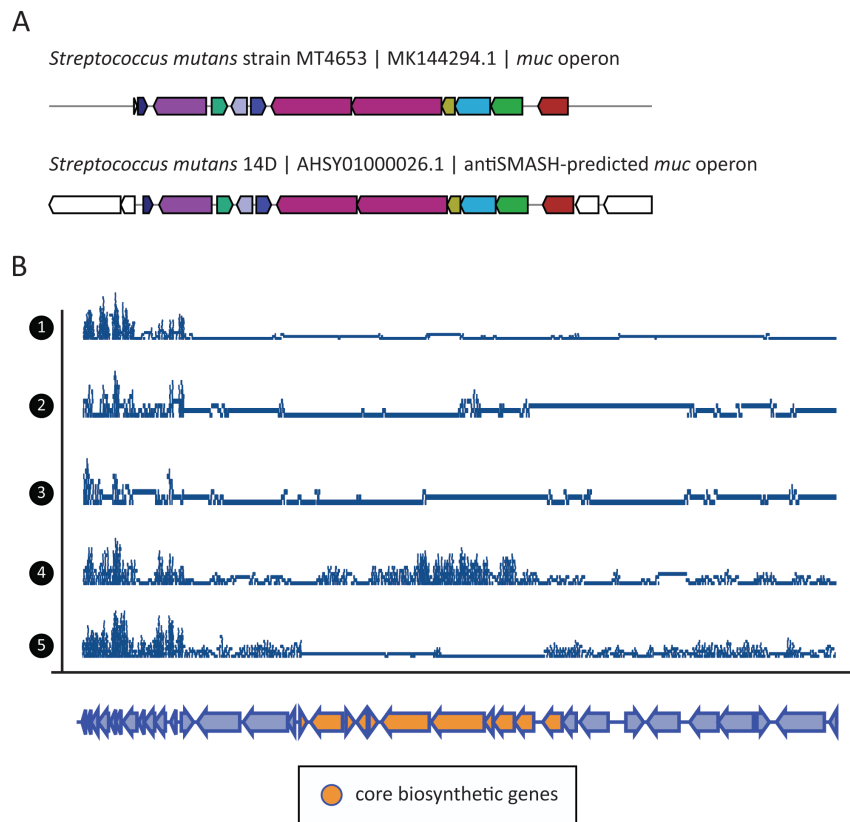
325 (A) MultiGeneBlast comparison between the *pdu* operon found in *Streptococcus sp.*
326 *F0442* by gutSMASH and the characterized one from *Lactobacillus reuteri*
327 (AP007281). (B) Read coverage of five randomly chosen healthy metagenomes along
328 the gutSMASH-predicted *pdu* gene cluster. The coverage graphs, which were plotted
329 using the Sushi R package (version 3.5.1)⁴⁴, show that some samples (3 and 4)
330 contain reads that cover the whole gene cluster, while in other samples, reads hardly
331 cover the cluster (1 and 2) or only part of it (5).

332

333 Another example of a gene cluster that has been found relevant in the oral cavity is
334 the *muc* operon, which has been shown to be responsible for the production of tetramic
335 acid, which is known to inhibit the colonization of commensal bacteria in the oral cavity.
336 This gene cluster encodes a hybrid between a polyketide synthase and nonribosomal
337 peptide synthetase (PKS/NRPS)³². In order to further test this association and assess
338 the abundance of the *muc* operon in the oral cavity, a collection of 170 *Streptococcus*
339 *mutans* genomes collected from Tang *et al.*³² and Liu *et al.*⁴⁵ was run through
340 antiSMASH¹⁰, which predicted a total of 1,849 BGCs. After obtaining 41 representative
341 gene clusters with BiG-MAP.family module, reads from the 47 oral microbiome
342 metagenomes were mapped onto the predicted gene clusters and further processed
343 using BiG-MAP.map and BiG-MAP.analyse subsequently. From the results, two gene
344 clusters were found to be significantly differentially abundant between healthy and
345 disease samples when using the fitZIG model: an NRPS from *Streptococcus mutans*
346 N29 and the *muc* operon from *Streptococcus mutans* 14D. The *muc* operon from this
347 strain shows high similarity to the one characterized by Tang *et al.*³² (Cumulative Blast
348 bit score of 9,056) (see Fig. 4A). However, the mean read core coverage in both

349 groups is low; 0.283 in healthy and 0.372 in caries-associated samples, which imply
350 the presence of some of the *muc* operon genes but not the complete gene cluster (see
351 Fig. 4B). Nonetheless, within both groups we see that some samples have reads
352 mapping to the complete gene cluster, with coverages values close to 1. When filtering
353 out samples with coverage values < 0.5 ; leaving only 6 samples in each group, the
354 mean coverage rises to 0.803 in healthy and 0.991 in disease. This is because there
355 are nine healthy samples that have a core coverage value of 0 and also five disease
356 samples that do not have reads mapping to the core genes of the *muc* operon.
357 Interestingly, depending at which stage you check which group is more enriched with
358 this gene cluster—either before or after normalization and depending on which
359 differential abundance test you apply—one group or the other seems to have higher
360 counts. The average abundance of raw RPKM counts in healthy is 16854.08
361 compared to 12815.69 in disease samples. After being normalized, healthy samples
362 have on average 11.22 RPKM counts/sample, slightly lower than the disease group
363 that has 11.28 RPKM counts/sample. When using the two available differential
364 abundance testing methods, we see that when applying the fitZIG model the difference
365 in abundance between healthy and disease samples is significant (more abundant in
366 disease) but not when testing it with Kruskal-Wallis. This is illustrated in the fitZIG BiG-
367 MAP output heatmap (Suppl. Fig. S3), which shows that despite the *muc* operon is
368 significantly more abundant in disease samples, the abundance of this gene cluster
369 across all samples is generally very similar. Therefore, despite finding this operon
370 being more abundant in caries-prone samples when applying the fitZIG model,
371 suggesting that indeed the *muc* operon plays a role in the caries development, the oral
372 microbiota from healthy donors seem to also harbor this PKS/NRPS. Hence, the
373 microbiota from healthy samples may have a mechanism to counteract the inhibiting

374 effect of tetramic acid, or there might be a difference in expression of the gene cluster
375 between healthy and diseased subjects.



376

377

378 **Figure 4. Detection of the *muc* operon in a subset of caries-associated samples.**

379 (A) MultiGeneBlast comparison between the *muc* operon characterized from

380 *Streptococcus mutans* strain MT4653.1 and the antiSMASH predicted one from

381 *Streptococcus mutans* 14D. (B) Read coverage of five random chosen caries-related

382 metagenomes along the antiSMASH predicted *muc* gene cluster. The coverage

383 graphs, which were plotted using the Sushi R package (version 3.5.1)⁴⁴, show that

384 despite the fact that the *muc* operon is generally not very highly covered by reads from

385 the randomly picked examples, some seem to truly contain for this operon, such as

386 sample 4, where the core biosynthetic genes look to be abundant at sufficient levels.

387 Full data pertaining all samples can be found in Fig. S3.

388

389 In addition, we also assessed the presence of the *muc* operon in the extended dataset
390 that includes 96 metagenomic samples in total (see *Evaluating the presence of the*
391 *muc operon in caries-associated metagenomes* Methods section). However, neither
392 the *muc* nor the other BGCs predicted from the *Streptococcus* genomes were
393 significantly more abundant in either group. This could be explained because within
394 the 96 samples there are not only healthy or caries-associated metagenomes but also
395 metagenomes from patients suffering from periodontitis and samples from a study that
396 observes how a biofilm evolves over time; therefore, it might be that the community
397 structure of all these samples differ quite a lot in terms of BGC content but also
398 regarding the presence of *Streptococcus mutans*. All in all, our results suggest that
399 the abundance of the *muc* operon is not very predictive for a healthy or disease state
400 of the microbiome by itself, and other factors likely play (more) important roles.

401

402 **Conclusions**

403 Overall, combining different omics datasets is a very useful approach to understand
404 which microbes are doing what and poses a promising avenue to better understand
405 complex biological processes. Here, we presented BiG-MAP, a command-line tool that
406 it is able to profile the abundance and expression of a collection of gene clusters
407 across metagenomic and metatranscriptomic data. Each of the steps in the BiG-MAP
408 pipeline is robust, as demonstrated using simulated metagenomes. Indeed, BiG-MAP
409 can discover interesting and relevant potential associations between genomic regions
410 and phenotypes, which can guide experimental efforts to test MGC function. It is worth
411 noting the usefulness of the gene cluster mapping coverage values, since they allow
412 the user to discern between the real presence of predicted gene clusters of interest
413 and spurious read mapping. Also, the associations that can be found using BiG-MAP

414 strongly depend on the WGS data sequencing depth and sample size, as for instance
415 in the examples described in our study, we found both gene clusters (*pdu-like* operon
416 and *muc*) only significant in either dataset (reduced or extended one). Moreover, from
417 the BiG-MAP output folders, which include raw and processed results, it is possible to
418 extract valuable information, such as the differences within groups, distribution of
419 reads across a gene cluster, raw and normalized RPKM counts, etc. Overall, we
420 believe BiG-MAP will help researchers solving biologically complex questions by
421 integrative multi-omics approaches, to obtain deeper insights into the relationships
422 between microbial metabolic capacities and microbiome-associated phenotypes.

423

424 **Methods**

425 **Code availability**

426 BiG-MAP is implemented in Python 3 as a command line package. It consists of four
427 modules: `BiG-MAP.download`, `BiG-MAP.family`, `BiG-MAP.map`, and `BiG-`
428 `MAP.analyse`. The code is available at: <https://github.com/medema-group/BiG-MAP>
429 together with documentation on how to install BiG-MAP and its dependencies and a
430 short tutorial on how to run it.

431

432 **BiG-MAP.download: Data collection**

433 This module allows to retrieve sequencing data present in the SRA database using
434 the SRA toolkit (<https://github.com/ncbi/sra-tools>). To initially develop, test and
435 validate this, we used an IBD cohort that contains metagenomic and
436 metatranscriptomic data from 78 individuals, 21 suffering from UC, 46 individuals with
437 CD, and 11 healthy samples⁴⁶. These samples were retrieved using the SRA

438 accession IDs under BioProject PRJNA389280 tool (see Suppl. Fig S1 and S2
439 generated from this dataset).

440

441 **BiG-MAP.family: Creating a non-redundant MGC representative collection**

442 The family module uses as input a directory that contains the gene cluster prediction
443 outputted by the antiSMASH⁴⁷ or gutSMASH algorithms
444 (<https://github.com/victoriapascal/gutsmash>). The predicted gene clusters are then
445 subjected to a redundancy filtering step based on their mutual sequence similarity. For
446 that, the protein sequences of the gene clusters are extracted and used as input for
447 MASH²⁰ sketch, which creates sketches from the raw sequences. The sketches are
448 then used to calculate the distances between sequences using MASH dist. The
449 resulting tab-delimited file with the pairwise distance comparisons is used to group
450 together gene clusters with above a 0.8 default similarity cut-off (see Figure 2). Next,
451 to pick the best representative of each group, medoids are computed (see formula
452 below). For this, a distance matrix is created comparing all distances between pairs of
453 gene clusters; the one with minimal cumulative distance value is picked as
454 representative of that group. Additionally, the selected gene clusters are subjected to
455 another round of clustering using BiG-SCAPE¹⁸, to group gene clusters into GCFs at
456 a 0.3 similarity cut-off (default value), from which a random representative is picked.

457

$$458 \quad x_{medoid} = \underset{y \in \{x_1, x_2, \dots, x_n\}}{\operatorname{argmin}} \sum_{i=1}^n d(y, x_i)$$

459

460 If metatranscriptomes will be used in the BiG-MAP.map module, an additional step is
461 performed to set an expression baseline. For this, the protein sequences of the

462 genomes whose gene clusters form the non-redundant representative gene cluster
463 collection are scanned using hmmsearch (hmmsearch version 3.1b2) for five
464 housekeeping-coding proteins: DNA gyrase A (PF00521), DNA gyrase B (PF00204),
465 Recombinase A (PF00154), DNA directed RNA polymerase A (PF01000), and DNA
466 directed RNA polymerase B (PF00562). The selection of these Pfam domains was
467 based on the findings by Rocha *et al.*⁴⁸ that these housekeeping genes show highly
468 stable expression across samples. Next, the gathered protein sequences are also
469 used as queries in the mapping module to align metatranscriptomic reads to gene
470 clusters.

471

472 **BiG-MAP.map: mapping reads to a non-redundant gene cluster collection**

473 This module relies on Bowtie2²¹ (version 2.3.4.3) to align reads to a given sequence.
474 From the reference gene cluster sequences selected by the medoid calculation,
475 Bowtie index files are created. Next, Bowtie2 aligns reads to these index files that by
476 default uses the fast alignment mode. The resulting alignment is stored in SAM format
477 and converted to BAM format to later be parsed by SAMtools⁴⁹ (version 1.9). The
478 alignments are then sorted by leftmost coordinates, the aligned reads are counted and
479 corrected by GCF and gene cluster size consecutively. Later, the corrected raw counts
480 are converted to TPM counts (Transcripts Per Kilobase Million) and consecutively to
481 RPKM (Reads Per Kilobase Million) counts to account for sequencing depth.

482

483 Another functionality that was added in this module was to compute the read coverage
484 of each gene cluster using the coordinates in the sorted BAM files. To do so, the sorted
485 alignment files are converted to bedgraphs using BEDtools⁵⁰ (v2.28.0), that allow to
486 estimate the number of covered bases of each cluster (*coverage*) by subtracting the

487 number of non-covered bases (*ncb*) to the length of each cluster (*cl*) as indicated in
488 the formula below.

$$489 \quad \text{coverage} = \frac{cl - ncb}{cl}$$

490

491 The same procedure is followed to compute the RPKM counts and the coverage of
492 the core genes within a gene cluster, which strictly considers the core metabolic genes
493 within each gene cluster. This information is taken from the antiSMASH/gutSMASH
494 (or any other “SMASH” related algorithm) Genbank output files that flag the key coding
495 genes that are needed for the synthesis of a given molecule. Once the core genes are
496 identified, the alignment information concerning them is retrieved using SAMtools.
497 Next, in the same manner as RPKM are computed for the whole gene clusters, reads
498 aligned to the core region are pulled out, counted and corrected to finally get the RPKM
499 counts. To perform the coverage calculation, the locations of the core genes are
500 extracted from the bedgraph to evaluate the coverage score using the aforementioned
501 formula.

502

503 **BiG-MAP.analyse: Normalization of RPKM counts and finding differentially** 504 **expressed/abundant MGCs**

505 In order to account for sparse high-throughput sequencing RPKM are normalized
506 using Cumulative Sum Scaling (CSS) from the R Bioconductor package
507 MetagenomeSeq²². BiG-MAP offers two different statistics to account for differentially
508 abundant/expressed gene clusters, the parametric zero inflated gaussian distribution
509 mixture model (ZIG-models) that assumes normal distribution of values or the non-
510 parametric Kruskal-Wallis test. Relatively small changes in gene cluster
511 abundance/expression are expected thus, ZIG-model values are adjusted with log2

512 fold-change that ultimately helps fitting the model to a log-normal distribution.
513 Alternatively, Kruskal-Wallis can be run on the normalized RPKM counts, which allows
514 to assess whether the distribution of ranks for one group significantly differs from the
515 distribution of ranks for the other group. Additionally, FDR correction is applied to
516 correct for multiple hypothesis testing. Finally, heatmaps are produced to visualize the
517 results using the Seaborn python package (<https://github.com/mwaskom/seaborn>).

518

519 **Testing BiG-MAP performance using a mock community**

520 To test BiG-MAP performance, 101 bacterial genomes were randomly chosen from
521 the CGR collection²⁴. Thus, the gutSMASH-predicted MGCs from each genome were
522 used as ground truth (<https://github.com/victoriapascal/gutsmash>, version 0.8, github
523 commit stamp: 569e860). Next, paired-end reads were generated with a mean read
524 length of 100 bp from the 101 CGR bacterial genomes using Grinder v0.5.3⁵¹. Two
525 different read coverage thresholds were used (0.5x and 0.05x) in combination with
526 four different community structures: uniform, linear, power-law and exponential. Both
527 the MGCs and the simulated reads were used as input for BiG-MAP, which was run
528 ranging the MASH similarity thresholds between 10-100% in intervals of 10% along
529 the eight different Bowtie2 alignment modes. From each individual run, true positive,
530 false positive and false negatives rates were calculated to evaluate the precision and
531 recall, which was ultimately used to compute the harmonic mean of precision and
532 recall, also known as the F1-score. The results were plotted in a heatmap using the
533 ComplexHeatmap package in R⁵².

534

535 **Assessing the *pdu* operon abundance by surveying different oral metagenomic**
536 **samples.**

537 To find possible leads on metabolic perturbances between healthy and caries-related
538 samples, the processed mass spectra (MGF format) and metabolomics feature tables
539 from Aleti, G. *et al.*³⁰ were downloaded from GNPS-MassIVE³⁴ accession ID
540 MSV000081832 to perform re-analysis. Feature-based Molecular Networks⁵³ were
541 run using GNPS release version 21
542 ([https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ef4f64542ab24a7fb0802ceacbcf](https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ef4f64542ab24a7fb0802ceacbcfa071)
543 [a071](https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ef4f64542ab24a7fb0802ceacbcfa071),
544 [https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9c95754d1fdc42b4a43b16919c3](https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9c95754d1fdc42b4a43b16919c398ecd)
545 [98ecd](https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9c95754d1fdc42b4a43b16919c398ecd)). The resulting molecular family information together with the metabolite feature
546 tables and sample information (metadata) were loaded into PALS
547 (<https://pals.glasgowcompbio.org/app/>)³³, to identify metabolite families differing in
548 activity between healthy and caries-related samples. From the results, three out of
549 seven candidate metabolites in one differentially expressed molecular family showing
550 clear different abundance patterns between healthy and caries samples were further
551 examined using GNPS MASST (<https://masst.ucsd.edu>)³⁶, the ChemCalc MF finder⁵⁴,
552 and PubChem⁵⁵, leading to the putative annotation of polyacrolein-related metabolites
553 in healthy samples, which may be produced from a *pdu-like* operon that requires the
554 presence of the cobalamin biosynthetic genes (see Supplementary material for further
555 information).

556

557 For the analysis of the *pdu* operon and its presence in the oral microbiome, 1,440 oral
558 bacteria genomes were downloaded from the HOMD collection
559 ([http://www.homd.org/?name=GenomeList&link=GenomeList&type=all oral](http://www.homd.org/?name=GenomeList&link=GenomeList&type=all%20oral)). Next,
560 these genomes were used as input for gutSMASH (version 0.8). The comparison
561 between the two *pdu* operons from *Lactobacillus reuteri* (AP007281) and

562 *Streptococcus sp. F0442* (GCA_000314795.2) was done using MultiGeneBlast⁵⁶.
563 Next, all predicted gene clusters were used as input for the BiG-MAP family module.
564 At the same time, the oral metagenomics datasets were downloaded using the BiG-
565 MAP.download module by providing the SRA accession IDs associated to the
566 PRJNA478018, PRJNA396840, and PRJNA398963 BioProject IDs. Once the
567 metagenomes were downloaded, BiG-MAP.map was run using the output of the family
568 module and the metagenomic reads in fastq format. Finally, the RPKM counts were
569 normalized, processed and visualized using BiG-MAP.analyse.

570

571 **Evaluating the presence of the *muc* operon in caries-associated metagenomes**

572 AntiSMASH was used to predict BGCs from a total of 170 *Streptococcus mutans*
573 genomes reported in Tang *et al*³² and Liu *et al*⁴⁵. Within the predicted BGCs, the *muc*
574 operon was found and compared to the *muc* operon characterized by Hao *et al*.⁵⁷
575 using MultiGeneBlast⁵⁶. The predicted BGCs were then used as input for the BiG-
576 MAP.family module. Both, the representative BGCs and metagenomic reads were
577 then used as input in the subsequent BiG-MAP.map mapping module using the
578 metagenomes from the following three BioProjects: PRJNA478018, PRJNA396840,
579 and PRJNA398963. Finally, the raw mapping counts were normalized and further
580 processed and visualized using BiG-MAP.analyse.

581

582 **Data availability**

583 The supporting information for this article can be found in the Supplementary material
584 and in the Zenodo repository (<https://zenodo.org/>) with the following DOI:
585 10.5281/zenodo.4320501. The metabolomics data used for reanalysis is available
586 from GNPS-MassIVE accession ID MSV000081832.

587

588 **Acknowledgements**

589 We thank Daria Zuzanna Świgoń, Arno Hagenbeek, Sarah van den Broek, Jeanine
590 Boot and Robert Koetsier for preliminary results on the *pdu* operon, which provided us
591 the lead to further explore these datasets. We also acknowledge the guidance
592 provided by Rens Holmer in the early stage of this study and Dr Madeleine Ernst for
593 her help in locating the relevant files of the relevant metabolomics data files from the
594 Aleti *et al.* study.

595

596 **Funding information**

597 This work was supported by the Chan- Zuckerberg Biohub (M.A.F.), and the U.S.
598 Defense Advanced Research Projects Agency's Living Foundries program award
599 HR0011-15- C-0084 (M.A.F. and V.P.A.) and an ASDI eScience grant
600 (ASDI.2017.030) from the Netherlands eScience Center (J.J.J.v.d.H. and M.H.M.).

601

602 **Conflicts of interest**

603 MHM is a co-founder of Design Pharmaceuticals and a member of the scientific
604 advisory board of Hexagon Bio. M.A.F. is a co-founder and director of Federation Bio,
605 a co-founder of Revolution Medicines, and a member of the scientific advisory board
606 of NGM Biopharmaceuticals.

607

608 **References**

- 609 1. Donia, M. S. & Fischbach, M. A. Small molecules from the human microbiota.
610 *Science*. **349**, 395–406 (2015).
- 611 2. Berendsen, R. L., Pieterse, C. M. J. & Bakker, P. A. H. M. The rhizosphere

- 612 microbiome and plant health. *Trends Plant Sci.* **17**, 478–486 (2012).
- 613 3. Hibbing, M. E., Fuqua, C., Parsek, M. R. & Peterson, S. B. Bacterial
614 competition: surviving and thriving in the microbial jungle. *Nat Rev Microbiol* **8**,
615 15–25 (2010).
- 616 4. Joseph M. Pickard, Melody Y. Zeng, R. C. Gut Microbiota: Role in Pathogen
617 Colonization, Immune Responses and Inflammatory Disease. *Immunol Rev.*
618 **279**, 70–89 (2017).
- 619 5. Garsin, D. A. Ethanolamine Utilization in Bacterial Pathogens: Roles and
620 Regulation. *Nat Rev Microbiol* **8**, 290–295 (2010).
- 621 6. Tracanna, V., de Jong, A., Medema, M. H. & Kuipers, O. P. Mining prokaryotes
622 for antimicrobial compounds: From diversity to function. *FEMS Microbiol. Rev.*
623 **41**, 417–429 (2017).
- 624 7. Mendes, R. *et al.* Deciphering the Rhizosphere Microbiome for Disease-
625 Suppressive Bacteria. *Science.* **911**, 1–5 (2011).
- 626 8. Brial, F., Le, A., Marc, L., Dumas, E. & Gauguier, D. Implication of gut
627 microbiota metabolites in cardiovascular and metabolic diseases. *Cell. Mol.*
628 *Life Sci.* **75**, 3977–3990 (2018).
- 629 9. Fischbach, M. A. Microbiome: Focus on Causation and Mechanism. *Cell* **174**,
630 785–790 (2018).
- 631 10. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome
632 mining pipeline. *Nucleic Acids Res* **47**, 81–87 (2019).
- 633 11. Hannigan, G. D. *et al.* A deep learning genome-mining strategy for biosynthetic
634 gene cluster prediction. *Nucleic Acids Res.* **47**, 110–123 (2019).
- 635 12. Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and
636 metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).

- 637 13. Liu, B. & Pop, M. MetaPath : identifying differentially abundant metabolic
638 pathways in metagenomic datasets. *BMC Proc.* **5**, 1–12 (2011).
- 639 14. Kim, J., Kim, M. S., Koh, A. Y., Xie, Y. & Zhan, X. FMAP : Functional Mapping
640 and Analysis Pipeline for metagenomics and metatranscriptomics studies.
641 *BMC Bioinformatics* **17**, 1–8 (2016).
- 642 15. Martinez, X. *et al.* MetaTrans: an open-source pipeline for
643 metatranscriptomics. *Sci. Rep.* **6**, 26447 (2016).
- 644 16. Kanehisa, M. KEGG for integration and interpretation of large-scale molecular
645 data sets. *Nucleic Acids Res.* **40**, 109–114 (2012).
- 646 17. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes.
647 *Nucleic Acids Res.* **46**, 633–639 (2018).
- 648 18. Navarro-muñoz, J. C. *et al.* A computational framework to explore large-scale
649 biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2019).
- 650 19. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome
651 mining pipeline. *Nucleic Acids Res* **47**, 81–87 (2019).
- 652 20. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation
653 using MinHash. *Genome Biol.* **17**, 1–14 (2016).
- 654 21. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2.
655 *Nat. Methods* **9**, 357–360 (2012).
- 656 22. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance
657 analysis for microbial marker-gene surveys. *Nat. Methods* **10**, 1–6 (2013).
- 658 23. Mcdonald, D. *et al.* The Biological Observation Matrix (BIOM) format or: how I
659 learned to stop worrying and love the ome-ome. *Gigascience* **1**, 1–6 (2012).
- 660 24. Zou, Y. *et al.* 1,520 reference genomes from cultivated human gut bacteria
661 enable functional microbiome analyses. *Nat. Biotechnol.* **37**, 179–187 (2019).

- 662 25. Deo, P. N. & Deshmukh, R. Oral microbiome : Unveiling the fundamentals. *J.*
663 *Oral Maxillofac. Pathol.* **23**, 122–128 (2019).
- 664 26. Graves, D. T., Corrêa, J. D. & Silva, T. A. The Oral Microbiota Is Modified by
665 Systemic Diseases. *J. Dent. Res.* **98**, 148 –156 (2019).
- 666 27. Dewhirst, F. E. *et al.* The Human Oral Microbiome. *J. Bacteriol.* **192**, 5002–
667 5017 (2010).
- 668 28. Garcia, S. S. *et al.* Targeting of *Streptococcus mutans* Biofilms by a Novel
669 Small Molecule Prevents Dental Caries and Preserves the Oral Microbiome.
670 *Dent. Res.* **96**, 807–814 (2017).
- 671 29. Edlund, A., Garg, N., Mohimani, H., Gurevich, A. & He, X. Metabolic
672 Fingerprints from the Human Oral Microbiome Reveal a Vast Knowledge Gap
673 of Secreted Small Peptidic Molecules. *mSystems* **2**, 1–16 (2017).
- 674 30. Aleti, G. *et al.* Identification of the Bacterial Biosynthetic Gene Clusters of the
675 Oral Microbiome Illuminates the Unexplored Social Language of Bacteria
676 during Health and Disease. *MBio* **10**, 1–19 (2019).
- 677 31. Sugimoto, Y., Camacho, F. R., Wang, S., Chankhamjon, P. & Odabas, A. A
678 metagenomic strategy for harnessing the chemical repertoire of the human
679 microbiome. *Science*. **366**, 1–17 (2019).
- 680 32. Tang, X. *et al.* Cariogenic *Streptococcus mutans* produces tetramic acid strain-
681 specific antibiotics that impair commensal colonization Cariogenic
682 *Streptococcus mutans* produces tetramic acid strain-specific antibiotics that
683 impair commensal colonization. *ACS Infect. Dis.* **6**, 563–571 (2020).
- 684 33. Mcluskey, K. *et al.* Decomposing metabolite set activity levels with PALS.
685 *bioRxiv* 1–12 (2020).
- 686 34. Wang, M. Perspective Sharing and community curation of mass spectrometry

- 687 data with Global Natural Products Social Molecular Networking. *Nat.*
688 *Biotechnol.* **34**, 828–837 (2016).
- 689 35. Tomfohr, J., Lu, J. & Kepler, T. B. Pathway level analysis of gene expression
690 using singular value decomposition. *BMC Bioinformatics* **11**, 1–11 (2005).
- 691 36. Wang, M. *et al.* Mass spectrometry searches using MASST. *Nat. Biotechnol.*
692 **38**, 23–26 (2020).
- 693 37. Engels, C. *et al.* Acrolein contributes strongly to antimicrobial and heterocyclic
694 amine transformation activities of reuterin. *Mol Nutr Food Res.* **6**, 1–13 (2016).
- 695 38. Orita, H. M. *et al.* Comparative Genome Analysis of *Lactobacillus reuteri* and
696 *Lactobacillus fermentum* Reveal a Genomic Island for Reuterin and Cobalamin
697 Production. *DNA Res.* **15**, 151–161 (2008).
- 698 39. Stevens, J. F. & Maier, C. S. Acrolein: Sources, metabolism, and biomolecular
699 interactions relevant to human health and disease. *Mol Nutr Food Res.* **52**, 7–
700 25 (2008).
- 701 40. Zhang, J., Sturla, S., Lacroix, C. & Schwab, C. Gut Microbial Glycerol
702 Metabolism as an Endogenous Acrolein. *MBio* **9**, 1–6 (2018).
- 703 41. Shlomo Margel and Erika Wiesel. Acrolein polymerization: Monodisperse,
704 homo, and hybrid microspheres, synthesis, mechanism, and reactions. *J.*
705 *Polym. Sci. Polym. Chem. Ed.* **22**, 145–158 (1984).
- 706 42. Nikawa, H. *et al.* *Lactobacillus reuteri* in bovine milk fermented decreases the
707 oral carriage of mutans streptococci. *Int. J. Food Microbiol.* **95**, 219–223
708 (2004).
- 709 43. Mi-Sun Kang. Inhibitory effect of *Lactobacillus reuteri* on periodontopathic and
710 cariogenic bacteria. *J. Microbiol.* **49**, 193–199 (2011).
- 711 44. Phanstiel, D. H., Boyle, A. P., Araya, C. L. & Snyder, M. P. Sushi . R : flexible ,

- 712 quantitative and integrative genomic visualizations for publication-quality multi-
713 panel figures. *Bioinformatics* **30**, 2808–2810 (2014).
- 714 45. Liu, L., Hao, T., Xie, Z., Horsman, G. P. & Chen, Y. Genome mining unveils
715 widespread natural product biosynthetic capacity in human oral microbe
716 *Streptococcus mutans*. *Sci. Rep.* **6**, 1–10 (2016).
- 717 46. Schirmer, M. *et al.* Dynamics of metatranscription in the inflammatory bowel
718 disease gut microbiome. *Nat. Microbiol.* **3**, 337–346 (2018).
- 719 47. Medema, M. H. *et al.* AntiSMASH: Rapid identification, annotation and analysis
720 of secondary metabolite biosynthesis gene clusters in bacterial and fungal
721 genome sequences. *Nucleic Acids Res.* **39**, 339–346 (2011).
- 722 48. Pacheco, L. G. C. Bacterial reference genes for gene expression studies by
723 RT-qPCR : survey and analysis. *Antonie Van Leeuwenhoek* **108**, 685–693
724 (2015).
- 725 49. Li, H. *et al.* The Sequence Alignment / Map format and SAMtools.
726 *Bioinformatics* **25**, 2078–2079 (2009).
- 727 50. Quinlan, A. R. & Hall, I. M. BEDTools : a flexible suite of utilities for comparing
728 genomic features. *Bioinformatics* **26**, 841–842 (2010).
- 729 51. Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P. & Tyson, G. W. Grinder : a
730 versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.* **40**, 1–
731 8 (2012).
- 732 52. Gu, Z., Eils, R. & Schlesner, M. Genome analysis Complex heatmaps reveal
733 patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**,
734 2847–2849 (2016).
- 735 53. Nothias, L. *et al.* Feature-based molecular networking in the GNPS analysis
736 environment. *Nat. Methods* **17**, 905–908 (2020).

- 737 54. Patiny, L. & Borel, A. ChemCalc: A Building Block for Tomorrow ' s Chemical
738 Infrastructure. *J. Chem. Inf. Model.* **53**, 1223–1228 (2012).
- 739 55. Kim, S. *et al.* PubChem 2019 update : improved access to chemical data.
740 *Nucleic Acids Res.* **47**, 1102–1109 (2019).
- 741 56. Medema, M. H., Takano, E. & Breitling, R. Detecting sequence homology at
742 the gene cluster level with multigeneblast. *Mol. Biol. Evol.* **30**, 1218–1223
743 (2013).
- 744 57. Hao, T. *et al.* An anaerobic bacterium host system for heterologous expression
745 of natural product biosynthetic gene clusters. *Nat. Commun.* **10**, 1–13 (2019).
746