A Profile Hidden Markov Model to investigate the distribution and frequency of LanB-encoding lantibiotic modification genes in the human oral and gut microbiome

Calum J Walsh [1, 2] , Caitriona M Guinane [1] , Paul W O' Toole [2, 3] , Paul D Cotter [Corresp. 1, 3]

[1] Teagasc Food Research Centre, Moorepark, Co. Cork, Ireland

[2] School of Microbiology, University College Cork, Co. Cork, Ireland

[3] APC Microbiome Institute, University College Cork, Co. Cork, Ireland

Corresponding Author: Paul D Cotter
Email address: paul.cotter@teagasc.ie

The human microbiota plays a key role in health and disease, and bacteriocins, which are small, bacterially produced, antimicrobial peptides, are likely to have an important function in the stability and dynamics of this community. Here we examined the density and distribution of the subclass I lantibiotic modification protein, LanB, in human oral and stool microbiome datasets using a specially constructed profile Hidden Markov Model (HMM).

The model was validated by correctly identifying known lanB genes in the genomes of known bacteriocin producers more effectively than other methods, while being sensitive enough to differentiate between different subclasses of lantibiotic modification proteins. This approach was compared with two existing methods to screen both genomic and metagenomic datasets obtained from the Human Microbiome Project (HMP).

Of the methods evaluated, the new profile HMM identified the greatest number of putative LanB proteins in the stool and oral metagenome data while BlastP identified the fewest. In addition, the model identified more LanB proteins than a pre-existing Pfam lanthionine dehydratase model. Searching the gastrointestinal tract subset of the HMP reference genome database with the new HMM identified seven putative subclass I lantibiotic producers, including two members of the Coprobacillus genus.

These findings establish custom profile HMMs as a potentially powerful tool in the search for novel bioactive producers with the power to benefit human health, and reinforce the repertoire of apparent bacteriocin-encoding gene clusters that may have been overlooked by culture-dependent mining efforts to date.

1 **Author Cover Page**

2 A Profile Hidden Markov Model to investigate the distribution and frequency of LanB-encoding

3   lantibiotic modification genes in the human oral and gut microbiome

5 Calum J. Walsh[1,3], Caitriona M. Guinane[1], Paul W. O'Toole[2,3], Paul D. Cotter[1,2]*

7 Teagasc Food Research Centre, Moorepark, Fermoy, Cork, Ireland[1];

8 APC Microbiome Institute, University College Cork, Cork, Ireland[2];

9 School of Microbiology, University College Cork, Cork, Ireland[3].

11 *To whom correspondence should be addressed: paul.cotter@teagasc.ie

19 **Abstract**

20 **Background**

21 The human microbiota plays a key role in health and disease, and bacteriocins, which are small,

22 bacterially produced, antimicrobial peptides, are likely to have an important function in the

23 stability and dynamics of this community. Here we examined the density and distribution of the

24 subclass I lantibiotic modification protein, LanB, in human oral and stool microbiome datasets

25 using a specially constructed profile Hidden Markov Model (HMM).

26 **Methods**

27 The model was validated by correctly identifying known *lanB* genes in the genomes of known

28 bacteriocin producers more effectively than other methods, while being sensitive enough to

29 differentiate between different subclasses of lantibiotic modification proteins. This approach was

30 compared with two existing methods to screen both genomic and metagenomic datasets obtained

31 from the Human Microbiome Project (HMP).

32 **Results**

33 Of the methods evaluated, the new profile HMM identified the greatest number of putative LanB

34 proteins in the stool and oral metagenome data while BlastP identified the fewest. In addition,

35 the model identified more LanB proteins than a pre-existing Pfam lanthionine dehydratase

36 model. Searching the gastrointestinal tract subset of the HMP reference genome database with

37 the new HMM identified seven putative subclass I lantibiotic producers, including two members

38 of the *Coprobacillus* genus.

Conclusions

These findings establish custom profile HMMs as a potentially powerful tool in the search for novel bioactive producers with the power to benefit human health, and reinforce the repertoire of apparent bacteriocin-encoding gene clusters that may have been overlooked by culture-dependent mining efforts to date.

Background

Bacteriocins are ribosomally synthesised peptides produced by bacteria that inhibit the growth of other bacteria. Some classes of bacteriocins are post-translationally modified to provide structures beyond those possible by ribosomal translation alone. These modifications are typically key to the peptide's functionality, stability and target recognition (Arnison et al. 2013). Class I bacteriocins, also known as lantibiotics, are one such class of small (<5 kDa) modified bacteriocins, possessing the characteristic thioester amino acids lanthionine or methyllanthionine (Perez et al. 2014). Lantibiotics form a subgroup within the larger lantipeptide family, which also includes peptides that lack antimicrobial activity. Lantipeptides can be divided into four different subclasses (I – IV) based on the distinct biosynthetic enzymes responsible for their posttranslational modification (Arnison et al. 2013).

The most commonly studied lantibiotic, Nisin, is a subclass I lantibiotic, meaning that the linear prepeptide is processed by a LanBC modification system (Arnison et al. 2013). The core peptide undergoes a two-step posttranslational modification catalysed by two distinct enzymes - the dehydratase LanB and the cyclase LanC (Xie & van der Donk 2004). The leader sequence, necessary for recognition by the modification enzymes in the two previous steps, is then removed by the protease LanP to produce the active lantibiotic (Xie & van der Donk 2004). The

61   gene-encoded nature of bacteriocins and bacteriocin-like peptides makes them ideal candidates

62   for genome mining. In the case of modified bacteriocins, the structural prepeptide coding

63   sequence often appears alongside the genes encoding proteins responsible for its modification

64   and export from the cell. However, as more bacteriocins are discovered, the heterogeneous

65   nature of these prepeptides is becoming ever more apparent. This diversity, coupled with their

66   small sequence length, makes bacteriocin prepeptides much more difficult to detect using

67   sequence-homology based searches like BLAST (Altschul et al. 1990). In an effort to address

68   these obstacles, shifting the focus to the detection of bacteriocin-associated proteins opens up

69   more avenues of discovery than simply searching for prepeptide homologs. This provides

70   opportunities to better determine the frequency with which specific types of bacteriocin gene

71   clusters can be found in different environmental niches, such as the human microbiota, through

72   the investigation of metagenomic data.

73   It has been estimated that the human microbiota comprises approximately 100 trillion bacterial

74   cells, outnumbering our own cells by a factor of 10 or more (Bäckhed et al. 2005). A recent

75   publication, however, has argued that the ratio is actually more likely to be one-to-one, with the

76   numbers being similar enough that each defecation event may alter the ratio to favour human

77   cells over bacteria (Sender et al. 2016). Of greater consequence than bacterial numbers, however,

78   is the collection of genes encoded in this metagenome, thought to be approximately 150 times

79   larger than that of the human genome, with a functional potential far broader than that of its host

80   (Qin et al. 2010). Regardless of absolute numbers, this dynamic community is thought to contain

81   100-1000 phylotypes (Faith et al. 2013; Qin et al. 2010) and play an integral role in human

82   health and disease (Clemente et al. 2012; Flint et al. 2012). The human microbiota exhibits

83   robust temporal stability (Belstrøm et al. 2016; Jeffery et al. 2016) perhaps due, in part, to the

84 protection against invading bacteria conferred by bacteriocins and other antimicrobials produced

85 *in situ* (Corr et al. 2007; Moroni et al. 2006; Rea et al. 2011a). As such, investigation of the

86 density and diversity of bacteriocins produced in the microbiome of healthy individuals may

87 shed light on beneficial and harmful members of this community, and key organisms for

88 maintaining typical, i.e. health-associated, microbiota composition.

89 Mining the human microbiota, especially for antimicrobial compounds, has become a popular

90 area of research in recent years (Donia et al. 2014; Walsh et al. 2015). Due to the availability of

91 metagenomic data generated by large public funding initiatives such as the Human Microbiome

92 Project in the U.S. (The Human Microbiome Project Consortium 2012) and the European

93 MetaHIT consortium (Dusko Ehrlich 2010), *in silico* mining of data has emerged as a new tool

94 that has the potential to identify antimicrobial-producing probiotics that can modulate the gut

95 microbiota (Erejuwa et al. 2014; Walsh et al. 2014), or address the increasingly serious threat to

96 public health caused by antimicrobial resistance. There are many available tools for mining the

97 microbiome for antimicrobials, including BAGEL3 (van Heel et al. 2013), antiSMASH (Weber

98 et al. 2015), and traditional sequence-based approaches like BLAST (Altschul et al. 1990). A

99 feature commonly integrated into these tools are Hidden Markov Models (HMM) (Morton et al.

100 2015; van Heel et al. 2013; Weber et al. 2015), a statistical method often used to model

101 biological data such as speech recognition, disease interaction and changes in gene expression in

102 cancer (Gales & Young 2007; Seifert et al. 2014; Sherlock et al. 2013). Profile HMMs, a specific

103 subset of HMMs, represent the patterns, motifs and other properties of a multiple sequence

104 alignment by applying a statistical model to estimate the true frequency of a nucleotide or amino

105 acid at a given position in the alignment from its observed frequency (Yoon 2009). Profile

106 HMMs differ from general HMMs as they move strictly from left to right along the alignment

107   and do not contain any cycles, a feature that makes them suitable for modelling nucleotide and

108   protein sequence data, and have been notably utilized to detect viral protein sequences in

109   metagenomic sequence data (Skewes-Cox et al. 2014). For each column in the multiple sequence

110   alignment, the profile uses one of three types of hidden state - a match state, an insert state, or a

111   delete state, to describe residue frequencies, insertions, and deletions, respectively (Yoon 2009).

112   Profile HMMs are potentially more sensitive than sequence homology approaches for identifying

113   more distantly related proteins as they focus on function-dependent conserved motifs that are

114   theoretically slower-evolving, as opposed to focusing on overall sequence similarity. Indeed,

115   profile HMMs are known to typically outperform pairwise sequence comparison methods (such

116   as BLAST) in the detection of distant homologs (Park et al. 1998), at the cost of greater

117   computational requirements – particularly in alignment scoring and E-value calculation (Madera

118   & Gough 2002). Correspondingly, speed is the main advantage of BLAST over profile HMMs,

119   however, as it is a heuristic algorithm it does not guarantee identification of the optimal

120   alignment between query and database sequences.

121   In this study we designed, validated and implemented a Profile HMM to search for putative

122   subclass I lantibiotic gene clusters in the HMP metagenomes and compared its performance to

123   some of the tools mentioned above.

124   **Methods**

125   **Data Collection**

126   HMASM (HMP Illumina WGS Assemblies) and HMRGD (HMP Reference Genomes Data)

127   were downloaded from the Data Analysis and Coordination Centre for the HMP. 835 bacterial

128   RefSeq protein sequences annotated as "lantibiotic dehydratase" were downloaded from NCBI

129   Protein website (13 Apr 2015) in FASTA format (listed in Supplemental Table 1).

130   **Building and Validating the new Profile Hidden Markov Model**

131   A global multiple sequence alignment was generated in the aligned-FASTA format using

132   MUSCLE (v3.8.31) (Edgar 2004), and a profile HMM was built from the MSA aligned-FASTA

133   file using the HMMER tool hmmbuild (v3.1b1 May 2013). For comparison of the new model's

134   performance, HMMER3's hmmsearch tool was used, with default parameters, to search the Pfam

135   lantibiotic dehydratase model PF04738 against the same stool and oral HMASM assemblies (the

136   sequences used to build this model are listed in Supplemental Table 2). Positive and negative

137   controls (listed in Table 1) were used to evaluate the model's ability to 1) accurately identify

138   LanB protein sequences, and 2) distinguish LanB protein sequences from other, related,

139   lantibiotic modification proteins (i.e. LanM, LanKC, and LanL). The controls were also screened

140   using the PF04738 model, the web-based bacteriocin genome mining tool BAGEL3 (van Heel et

141   al. 2013), and a traditional BlastP using the nisin-associated lanthionine dehydratase, NisB, as

142   the driver sequence (GenBank accession number CAA79468.1) to compare the sensitivity and

143   specificity of each approach. A flowchart of the steps involved in building, validating and

144   applying a profile HMM is depicted in Supplemental Figure 1.

145   **Target Sequence Translation**

146   The HMMER3 hmmsearch tool only accepts protein sequences as targets for comparison to

147   protein profile HMMs so a python script was created to translate the nucleotide sequences into

148   protein sequences. The DNA nucleotide sequences were translated in six frames using the

149   standard genetic code.

Metagenomic Screen

The HMMER3 tool hmmsearch was used, with default parameters, to search both the new LanB profile HMM and the Pfam PF04738 profile HMM (Punta et al. 2012) against the stool and oral subsets of the Human Microbiome Project's whole metagenomic shotgun sequencing assemblies (HMASM). 139 stool communities and 382 communities from eight different body sites within the oral cavity were screened from the HMP database. These are listed in Table 2. As an additional comparison of performance, a traditional BlastP screen was performed on the same metagenomic samples using the previously mentioned nisin-associated lanthionine dehydratase, NisB, driver sequence. A significance cutoff of $E \leq 1\text{x}10^{-5}$ was chosen for both profile HMM and BlastP methods.

Manual Examination of Randomly Selected Gene Neighbourhoods

A subset of sixty hits were randomly selected and the surrounding region examined to identify other proteins involved in lantibiotic biosynthesis. Open Reading Frames were identified using Glimmer v3.02 (Delcher et al. 1999), which were then visualised using Artemis (Carver et al. 2012) and blasted against the nr database using BlastP.

Genomic Screen

HMMER3's hmmsearch tool was used, with default parameters, to search the new profile HMM against the draft genomes comprising the gastrointestinal tract subset of the Human Microbiome Project's reference genome database.

169   **Taxonomic Classification of LanB-encoding contigs**

170   Taxonomy was assigned to LanB-encoding contigs, as assigned by our pHMM, using Kaiju

171   (Menzel et al. 2016). Analysis was performed in MEM run mode using default parameters and

172   the NCBI non-redundant protein database.

173   **Statistical Analysis**

174   Statistical analysis was performed in R (v. 3.1.3) (R Core Team 2015).

175   **Results**

176   **Validation of the Profile Hidden Markov Model**

177   The ability of the newly developed profile HMM and the Pfam lantibiotic dehydratase model

178   PF04738 to detect LanB-encoding genes were compared using the positive and negative controls

179   listed in Table 1. The positive controls were selected based on a relevant book chapter (Rea et al.

180   2011b) and all are previously characterised bacteriocin producers for which the sequence of the

181   relevant biosynthetic gene cluster was available. None of the positive control sequences were

182   used in the building of the model and a graphical representation of these clusters is presented in

183   Figure 1. *Lactococcus lactis* subsp. *lactis* KF147 was chosen as a negative control because it is

184   of the same subspecies as three of the positive controls (*Lactococcus lactis* subsp. lactis S0,

185   *Lactococcus lactis* subsp. *lactis* CV56 and *Lactococcus lactis* subsp. *lactis* IO-1) but does not

186   produce a bacteriocin. *Streptococcus mutans* GS-5, *Streptomyces cinnamoneus cinnamoneus*

187   DSM 4005, the *Lactococcus lactis* subsp. *lactis* IL1835 plasmid pES2, the *Streptococcus*

188   *salivarious* plasmid pSsal-K12, and the newly characterised formicin producer *Bacillus*

189   *paralicheniformis* APC 1576 were chosen as negative controls to evaluate the ability of the

190   model to differentiate between LanB (subclass I) proteins and the LanM proteins-from these

191   strains, which perform a similar, but distinct, function in the posttranslational modification of

192   subclass II lantibiotics. *Streptomyces venezuelae* ATCC 10712 was chosen as the final negative

193   control as it has been reported to produce a LanL-type lantipeptide (Goto et al. 2010).

194   Examination of the ATCC 10712 genome using BAGEL3 identified several other orphan

195   lantibiotic modification genes, including those encoding putative LanL, LanM, LanD and LanB

196   proteins. The genome also appeared to encode a subclass III lantipeptide cluster comprised of

197   genes potentially encoding a structural protein, two ABC-type transporters and a LanKC

198   modification protein. Notably, there have been no reports of subclass I lantibiotic production by

199   ATCC 10712 despite an in-depth investigation into the strain's lantipeptide producing capability

200   (Goto et al. 2010), and BAGEL3 identified no other lantibiotic-related genes in the area of

201   interest leading us to determine that this was a false positive.

202   The newly developed LanB profile HMM correctly identified the LanB protein in all nine

203   positive controls, while the PF04738 profile HMM correctly identified the LanB protein in eight

204   of the nine positive controls, failing to detect the Bsa-associated LanB protein in *Staphylococcus*

205   *aureus* subsp. *aureus* USA300_FPR3757. Both the LanB and PF04738 profile HMMs returned

206   no false positives when searched against the seven negative controls used in this study.

207   The web version of BAGEL3 correctly identified the lantibiotic modification proteins in all

208   positive and negative controls, excepting the aforementioned ATCC 10712-encoded LanB

209   concluded to be a false positive. Interestingly, examination of these controls with the BlastP

210   method described previously, failed to correctly identify the LanB proteins encoded by

211   *Streptococcus mutans* CH43 and *Streptococcus mutans* UA787, although the former ($E = 3\text{x}10^{-4}$)

212   fell just short of the significance cutoff ($E \leq 1\text{x}10^{-5}$). BlastP also incorrectly identified a LanB

213   protein in the negative control *Lactococcus lactis* subsp. *lactis* KF147.

214   **Metagenomic Screen**

215   A search with the newly developed profile HMM against the HMASM database identified 399

216   hits from the stool metagenomes and 1169 hits from the oral metagenomes. In contrast, the

217   PF04738 model identified 288 hits from the stool metagenomes and 686 from the oral

218   metagenomes. Our model reported at least one putative lantibiotic gene cluster in 81% of oral

219   metagenomes and 86% of stool metagenomes, compared to 73% and 76%, respectively,

220   identified by the Pfam model. The distribution of hits per sample is presented in Figure 2. BlastP

221   identified 231 hits from the stool metagenomes and 374 hits from the oral metagenomes. The

222   results of these three approaches were compared to ascertain what proportion of significant hits

223   was common to more than one search method. The results of this comparison are summarised in

224   Figure 3 and show that the newly developed profile HMM identified the greatest number of

225   lantibiotic modification genes in datasets from both body sites, while the BlastP approach

226   identified the fewest.

227   The overall results of these combined screening approaches, illustrated in Figure 4 and

228   summarised in Supplemental Table 3, show a higher number and density of hits in the oral

229   metagenomes than in the stool metagenomes (Welch's two sample t-test, $p = 5.422e-08$) and they

230   also reveal a large variation in density of hits between the different sites within the oral

231   metagenomes. This pattern was also reflected in four of the Oral subsites, namely Buccal

232   Mucosa, Subgingival Plaque, Supragingival Plaque and Tongue Dorsum, all of which had a

233   significantly higher LanB density than the Stool metagenomes ($p = 0.00304$, 0.02467, 5.5e-08,

234   and 0.00024, respectively). Within the Oral samples, our model revealed a large variation in

235   density of hits between different subsites. The throat metagenomes had the lowest LanB density,

236   and exhibited significantly lower densities than Buccal Mucosa ($p = 0.00581$), Subgingival

237   Plaque ($p = 0.00587$), Supragingival Plaque ($p = 0.00943$), and Tongue Dorsum ($p = 0.02219$)

238   subsites.

239   **Manual Examination of Selected Gene Neighbourhoods**

240   Sixty hits, listed in Supplemental Table 4, were randomly selected from those identified by the

241   new profile HMM, 45% (27/60) of which were identified by at least one of the other two

242   methods, and manually examined to determine if a bacteriocin gene cluster could be identified.

243   42% (25/60) of these were not further analysed because the often relatively short regions

244   assembled from the shotgun data prevented the identification of a full lantibiotic gene cluster.

245   However, of the 35 remaining clusters, 28 (80%) appeared to encode multiple genes involved in

246   the biosynthesis of bacteriocins and thiopeptides. These genes encode proteins involved in

247   posttranslational modification, bacteriocin transport, leader cleavage and regulation

248   (Supplemental Figure 2).

249   81 hits identified by BlastP were missed by both profile HMM approaches. 50 of these originated

250   in the stool metagenomes and were selected for manual annotation to determine if an overall

251   structure or similarity could be observed. 29 of these 50 were part of clusters whose components

252   showed relatively low sequence identity (39-50%) with proteins responsible for the biosynthesis

253   of thiopeptides and lantibiotics, including a putative lanthionine dehydratase, a radical

254   SAM/SPASM domain-containing protein, a thiopeptide-type bacteriocin biosynthesis domain-

255   containing protein, an S41 family peptidase, and a protein of unknown function (DUF4932)

256   predicted to be a putative metalloprotease. All 50 manually annotated gene clusters are available

257   in GENBANK format and an example of this cluster architecture is summarised in Supplemental

258   Table 5.

259   **Genomic Screen**

260   The draft genomes of the gastrointestinal tract subset of the HMRGD were also used as a

261   database and searched using the new profile HMM. This resulted in the identification of seven

262   hits, including two strains of *Coprobacillus*, a potentially probiotic genus (Stein et al. 2013; Yan

263   et al. 2012) (Table 3). From these seven genomes, only three lantibiotic gene clusters were

264   identified by BAGEL3, these are illustrated in Figure 5. Although this low frequency of

265   lanthionine dehydratase proteins in the genomic dataset (0.006 hits/Mb) contrasts with the

266   findings of the metagenome screen reported above, it is in agreement with previous reports of

267   relatively low subclass I lantibiotic density within the human microbiota (Walsh et al. 2015;

268   Zheng et al. 2014). A possible explanation for this significantly lower gene density (Welch's two

269   sample t-test, $p = 1.232e\text{-}10$) is that the subclass I lantibiotic clusters identified in the

270   metagenomic data by the new profile HMM are present in the genomes of rarer members of the

271   gut microbiota, which are not represented in the HMP reference genome database.

272   **Taxonomic Classification of LanB-encoding contigs**

273   The MEM run mode of Kaiju works by searching for exact matches of given length between the

274   query and database sequences, in the case of multiple hits of the same length in different taxa, a

275   lowest common ancestor is inferred. Kaiju classified 378 of 399 LanB-encoding contigs. Of

276   these, 232 were classified to the species level - however 68 were removed as their exact species

277   was ambiguous. Of the remaining 164 classified contigs, 66 (40.2%) were represented at the

278   species-level in the previously screened HMRGD database. The most abundant genus was

279   *Alistipes*, accounting for 14.03% of LanB-encoding cotigs identified by our model, followed by

280   *Blautia* (7.77%), *Clostridium* (4.51%), and *Bacteroides* (3.76%) (Supplemental Table 6).

Discussion

281 **Discussion**

282 Bacteriocin production enhances the competitiveness of bacteria living in complex communities

283 and has the potential to be harnessed for the benefit of human health. The goal of this study was

284 to develop a profile HMM and to assess its ability, in comparison with several other approaches,

285 to detect putative subclass I lantibiotic gene clusters in human metagenomic datasets. Through

286 this process it was also possible to evaluate the potential frequency and distribution of these

287 bacteriocin gene clusters in the human microbiota.

288 To validate the model, nine positive controls and five negative controls were selected to evaluate

289 its sensitivity and specificity. These controls were selected based on reported bacteriocin

290 production; all positive controls were known producers of subclass I lantibiotics while the

291 negative controls produced either different subclasses of lantibiotics or none at all. Following

292 validation, genomic and metagenomic data corresponding to two niches within the human

293 microbiome were chosen as the focus of this study. The first of these niches was human stool and

294 was selected as the corresponding samples were most likely to yield bacteriocin producers with

295 the potential to modulate undesirable microbiota profiles associated with obesity, colorectal

296 cancer, type 2 diabetes or inflammatory bowel diseases due to their ability to survive and

297 colonise this environment. Secondly, human oral communities were examined as a previous

298 study showed that they contained, by far, the greatest proportion of bacteriocin structural genes

299 across a number of human metagenome samples (Zheng et al. 2014). Zheng *et al.* reported that

300 80% of class I bacteriocins (lantibiotics) and 89% of all bacteriocins identified using their

301 method originated in the oral metagenomes, while the stool metagenomes contained just 15%

302 and 7%, respectively. The same study reported that 88% of samples from the oral cavity and

303 73% of samples from the gut contained at least one bacteriocin (regardless of class), while the

304 new profile HMM reported these statistics as 81% and 83%, respectively for subclass I

305 lantibiotics alone. The *in silico* screen carried out with the profile HMM is consistent with the

306 observation by Zheng *et al.* (Zheng et al. 2014) by yielding a higher number and density of hits

307 from the oral, compared to the stool, metagenomic data. Furthermore, the large variation in

308 density of hits between sites within the oral environment suggests that lantibiotic production

309 confers a greater advantage in buccal musoca, subgingival plaque, supragingival plaque, and

310 tongue dorsum communities compared to communities from the throat. This may be due to the

311 direct benefits of antimicrobial activity but could also involve the intra- and interspecies

312 signalling roles attributed to lantibiotic peptides (Upton et al. 2001), particularly in the intensely

313 competitive microbial biofilm environment of dental plaque.

314 One of the most interesting observations from the study was the large variation in the numbers of

315 *lanB* genes reported by the three different approaches. The BlastP approach identified, by far, the

316 lowest number of significant hits overall and the lowest in every body site examined, except for

317 the saliva microbiome. Our model identified more than double the number of hits provided by

318 the BlastP-based approach, in line with the aforementioned knowledge that profile HMMs can

319 detect as much as three times as many distant homologs than pairwise methods (Park et al.

320 1998). Our model also identified a greater number of LanB proteins than the Pfam PF04738

321 model when used to search the same data using the same parameters. While the PF04738 model

322 relates to the C-terminus of the lanthionine dehydratase protein, responsible for the glutamate

323 elimination step of lantibiotic modification (Ortega et al. 2015), the newly developed profile

324 HMM takes the full length of the LanB protein into consideration, thereby providing greater

325 predictive power. Our model, in addition to identifying more potential LanB proteins, also

326 exhibited greater sensitivity and specificity during validation than all other methods used to

327   analyse the controls. As stated above, profile HMMs are already known to be particularly

328   sensitive, the validation step, however, also suggests that they are more specific than the other

329   methods evaluated as they were the only approach which did not return any false positives.

330   When selecting the controls used to examine the performance of the different approaches, greater

331   consideration was given to the quality of these controls than their quantity. Only controls with

332   experimentally characterised lantibiotic production were included in the validation dataset. This

333   relatively small control group means that, although the results of the validation step may explain

334   the contrasting numbers of LanB proteins reported by our model and the PF04738 model, it

335   cannot be said for certain that our model performed better.

336   Zheng *et al.*, using the same metagenomic data that was the focus of this study, identified 17

337   potential subclass I lantibiotics from stool samples and 76 from oral samples, a much lower

338   frequency of detection than in this study, probably due to the different methodologies used. That

339   study focused on searching for proteins similar to those in BAGEL3's manually curated

340   database, an approach which likely lost sensitivity because bacteriocin precursor peptides can

341   differ considerably at primary sequence level. Furthermore, the screen employed a BLAST-

342   based approach which, as demonstrated here, exhibited the lowest number of significant hits

343   reported.

344   To investigate the areas surrounding the LanB-encoding genes identified by our model we

345   randomly selected thirty positive hits from the oral and stool metagenome screens for manual

346   examination. This approach revealed that several of the hits were on scaffolds that were either

347   too small to contain a full gene or did not contain the gene's start codon. This was most likely as

348   a consequence of the fragmented nature of the metagenomic data, as opposed the identification

349   of true false positives by the model and would probably occur regardless of the method

350   employed. 42% (25/60) of hits selected for manual examination were discarded based on these

351   criteria. It also revealed that a considerable number of hits exhibited low (~30%) similarity to

352   putative thioesterases in the nr protein sequence database, highlighting that lanthionine

353   dehydratases are relatively-closely related to proteins involved in the posttranslational

354   modification of thiopeptides, most likely those responsible for dehydration of serine and

355   threonine residues (Garg et al. 2013). The similarity between these dehydratase proteins suggests

356   a possible common ancestor protein (Kelly et al. 2009). Another possible explanation relates to

357   the fact that all of the proteins annotated as thiopeptide modification proteins are putative

358   annotations and none, to our knowledge, have been confirmed as such *in vitro*. It is possible,

359   therefore, that these may simply be lanthionine dehydratases which have been incorrectly

360   annotated due to automatic software and incomplete/under-curated databases. The majority of

361   clusters identified contained genes encoding both LanB and LanC modification proteins, while

362   many also contained a leader cleavage and activation peptidase and/or ABC transporter proteins

363   for export of the mature peptide, suggesting that these have the potential to encode a functional

364   lantibiotic.

365   To evaluate the model's performance in a genomic context we applied it to the gastrointestinal

366   tract subset of the HMP's reference genome database and compared the results to our previously

367   published study which used the online bacteriocin genome mining tool BAGEL3 to screen this

368   same database (Walsh et al. 2015). The results of the two screens were startlingly different and

369   served to highlight the variation in results that can arise from applying different methods to the

370   same data. Interestingly, the gastrointestinal tract reference genomes encoded a significantly

371   lower frequency of LanB hits than the stool metagenomic samples. Taxonomic classification of

372   the 399 LanB-encoding contigs identified by our new model from the stool metagenomes

373 revealed that only 40.2% of these potential lantibiotic producing strains were represented in the

374 reference genome database, suggesting that the majority of these lantibiotics were encoded by

375 species rarer members of the gut microbiota or those that have not previously been identified as

376 important. Taxonomic classification of these LanB-encoding contigs also served to highlight

377 patterns in the results of the three approaches used (Supplemental Figure 3), for example out

378 model identified *Allokutzneria*, *Coprococcus*, *Enterovibrio*, *Paenibacillus,* and *Tenicibaculum-*

379 encoded LanB proteins that were completely missed by the Pfram and BlastP approaches.

Conclusions

381 Across the oral and stool communities examined, this study identified 2007 unique putative

382 subclass I lantibiotic biosynthetic gene clusters by three different methods, further emphasising

383 the tremendous potential that the human microbiota has as a source of therapeutic compounds.

384 As this study was performed entirely *in* silico, the next challenge lies in experimentally

385 identifying and characterising these putative bacteriocins to identify those with the ability to

386 desirably modulate the microbiota for the treatment of disease.

Acknowledgements

388 The authors would like to thank Manimozhiyan Arumugam for helpful discussion.

List of Abbreviations

Abbreviation	Description
HMASM	Human Microbiome Project's Illumina Whole Genome Shotgun Assemblies
HMM	Hidden Markov Model

HMP	Human Microbiome Project
HMRGD	Human Microbiome Project's Reference Genome Data

References

390 *Hmmer. Available at* http://hmmer.org.

392 Human Microbiome Project; Data Analysis and Coordination Center. *Available at*
393   http://hmpdacc.org/.

394 Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search
395   tool. *J Mol Biol* 215:403-410. 10.1016/s0022-2836(05)80360-2

396 Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, Bulaj G, Camarero JA, Campopiano DJ,
397   Challis GL, Clardy J, Cotter PD, Craik DJ, Dawson M, Dittmann E, Donadio S, Dorrestein
398   PC, Entian K-D, Fischbach MA, Garavelli JS, Goransson U, Gruber CW, Haft DH,
399   Hemscheidt TK, Hertweck C, Hill C, Horswill AR, Jaspars M, Kelly WL, Klinman JP, Kuipers
400   OP, Link AJ, Liu W, Marahiel MA, Mitchell DA, Moll GN, Moore BS, Muller R, Nair SK, Nes
401   IF, Norris GE, Olivera BM, Onaka H, Patchett ML, Piel J, Reaney MJT, Rebuffat S, Ross RP,
402   Sahl H-G, Schmidt EW, Selsted ME, Severinov K, Shen B, Sivonen K, Smith L, Stein T,
403   Sussmuth RD, Tagg JR, Tang G-L, Truman AW, Vederas JC, Walsh CT, Walton JD, Wenzel
404   SC, Willey JM, and van der Donk WA. 2013. Ribosomally synthesized and post-
405   translationally modified peptide natural products: overview and recommendations for a
406   universal nomenclature. *Natural Product Reports* 30:108-160. 10.1039/C2NP20085F

407 Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA, and Gordon JI. 2005. Host-Bacterial Mutualism
408   in the Human Intestine. *Science* 307:1915-1920. 10.1126/science.1104816

409 Belstrøm D, Holmstrup P, Bardow A, Kokaras A, Fiehn N-E, and Paster BJ. 2016. Temporal
410   Stability of the Salivary Microbiota in Oral Health. *PLoS ONE* 11:e0147472.
411   10.1371/journal.pone.0147472

412 Carver T, Harris SR, Berriman M, Parkhill J, and McQuillan JA. 2012. Artemis: an integrated
413   platform for visualization and analysis of high-throughput sequence-based experimental
414   data. *Bioinformatics* 28:464-469. 10.1093/bioinformatics/btr703

415 Clemente Jose C, Ursell Luke K, Parfrey Laura W, and Knight R. 2012. The Impact of the Gut
416   Microbiota on Human Health: An Integrative View. *Cell* 148:1258-1270.
417   http://dx.doi.org/10.1016/j.cell.2012.01.035

418 Corr SC, Li Y, Riedel CU, O'Toole PW, Hill C, and Gahan CGM. 2007. Bacteriocin production as a
419   mechanism for the antiinfective activity of Lactobacillus salivarius UCC118. *Proceedings*
420   *of the National Academy of Sciences* 104:7617-7621. 10.1073/pnas.0700440104

421 Delcher AL, Harmon D, Kasif S, White O, and Salzberg SL. 1999. Improved microbial gene
422   identification with GLIMMER. *Nucleic Acids Res* 27:4636-4641.

423 Donia Mohamed S, Cimermancic P, Schulze Christopher J, Wieland Brown Laura C, Martin J,
424   Mitreva M, Clardy J, Linington Roger G, and Fischbach Michael A. 2014. A Systematic
425   Analysis of Biosynthetic Gene Clusters in the Human Microbiome Reveals a Common
426   Family of Antibiotics. *Cell* 158:1402-1414. 10.1016/j.cell.2014.08.032

427   Dusko Ehrlich S. 2010. [Metagenomics of the intestinal microbiota: potential applications].
428   *Gastroenterol Clin Biol* 34 Suppl 1:S23-28. 10.1016/s0399-8320(10)70017-8
429   Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput.
430   *Nucleic Acids Res* 32:1792-1797. 10.1093/nar/gkh340
431   Erejuwa OO, Sulaiman SA, and Ab Wahab MS. 2014. Modulation of gut microbiota in the
432   management of metabolic disorders: the prospects and challenges. *Int J Mol Sci*
433   15:4158-4188. 10.3390/ijms15034158
434   Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, Clemente JC,
435   Knight R, Heath AC, Leibel RL, Rosenbaum M, and Gordon JI. 2013. The Long-Term
436   Stability of the Human Gut Microbiota. *Science* 341. 10.1126/science.1237439
437   Flint HJ, Scott KP, Louis P, and Duncan SH. 2012. The role of the gut microbiota in nutrition and
438   health. *Nat Rev Gastroenterol Hepatol* 9:577-589.
439   Gales M, and Young S. 2007. The application of hidden Markov models in speech recognition.
440   *Found Trends Signal Process* 1:195-304. 10.1561/2000000004
441   Garg N, Salazar-Ocampo LMA, and van der Donk WA. 2013. In vitro activity of the nisin
442   dehydratase NisB. *Proceedings of the National Academy of Sciences of the United States*
443   *of America* 110:7258-7263. 10.1073/pnas.1222488110
444   Goto Y, Li B, Claesen J, Shi Y, Bibb MJ, and van der Donk WA. 2010. Discovery of Unique
445   Lanthionine Synthetases Reveals New Mechanistic and Evolutionary Insights. *PLoS Biol*
446   8:e1000339. 10.1371/journal.pbio.1000339
447   Jeffery IB, Lynch DB, and O'Toole PW. 2016. Composition and temporal stability of the gut
448   microbiota in older persons. *ISME J* 10:170-182. 10.1038/ismej.2015.88
449   Kelly WL, Pan L, and Li C. 2009. Thiostrepton biosynthesis: prototype for a new family of
450   bacteriocins. *J Am Chem Soc* 131:4327-4334. 10.1021/ja807890a
451   Madera M, and Gough J. 2002. A comparison of profile hidden Markov model procedures for
452   remote homology detection. *Nucleic Acids Res* 30:4321-4328.
453   Menzel P, Ng KL, and Krogh A. 2016. Fast and sensitive taxonomic classification for
454   metagenomics with Kaiju. *Nature Communications* 7:11257. 10.1038/ncomms11257

455   http://www.nature.com/articles/ncomms11257#supplementary-information
456   Moroni O, Kheadr E, Boutin Y, Lacroix C, and Fliss I. 2006. Inactivation of Adhesion and Invasion
457   of Food-Borne Listeria monocytogenes by Bacteriocin-Producing Bifidobacterium Strains
458   of Human Origin. *Applied and Environmental Microbiology* 72:6894-6901.
459   10.1128/aem.00928-06
460   Morton JT, Freed SD, Lee SW, and Friedberg I. 2015. A large scale prediction of bacteriocin gene
461   blocks suggests a wide functional spectrum for bacteriocins. *BMC Bioinformatics* 16:1-9.
462   10.1186/s12859-015-0792-9
463   Ortega MA, Hao Y, Zhang Q, Walker MC, van der Donk WA, and Nair SK. 2015. Structure and
464   mechanism of the tRNA-dependent lantibiotic dehydratase NisB. *Nature* 517:509-512.
465   10.1038/nature13888
466   Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, and Chothia C. 1998. Sequence
467   comparisons using multiple sequences detect three times as many remote homologues
468   as pairwise methods. *J Mol Biol* 284:1201-1210. 10.1006/jmbi.1998.2221

469 Perez RH, Zendo T, and Sonomoto K. 2014. Novel bacteriocins from lactic acid bacteria (LAB):
470   various structures and applications. *Microbial Cell Factories* 13:S3-S3. 10.1186/1475-
471   2859-13-S1-S3
472 Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G,
473   Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, and Finn RD. 2012.
474   The Pfam protein families database. *Nucleic Acids Res* 40:D290-D301.
475   10.1093/nar/gkr1065
476 Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F,
477   Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J,
478   Lepage P, Bertalan M, Batto J-M, Hansen T, Le Paslier D, Linneberg A, Nielsen HB,
479   Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y,
480   Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Dore J, Guarner F, Kristiansen K,
481   Pedersen O, Parkhill J, Weissenbach J, Bork P, Ehrlich SD, and Wang J. 2010. A human
482   gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464:59-
483   65.
484   http://www.nature.com/nature/journal/v464/n7285/suppinfo/nature08821_S1.html
485 R Core Team. 2015. R: A Language and Environment for Statistical Computing.
486 Rea MC, Dobson A, O'Sullivan O, Crispie F, Fouhy F, Cotter PD, Shanahan F, Kiely B, Hill C, and
487   Ross RP. 2011a. Effect of broad- and narrow-spectrum antimicrobials on Clostridium
488   difficile and microbial diversity in a model of the distal colon. *Proceedings of the*
489   *National Academy of Sciences* 108:4639-4644. 10.1073/pnas.1001224107
490 Rea MC, Ross RP, Cotter PD, and Hill C. 2011b. Classification of Bacteriocins from Gram-Positive
491   Bacteria. In: Drider D, and Rebuffat S, eds. *Prokaryotic Antimicrobial Peptides: From*
492   *Genes to Applications*. New York, NY: Springer New York, 29-53.
493 Seifert M, Abou-El-Ardat K, Friedrich B, Klink B, and Deutsch A. 2014. Autoregressive Higher-
494   Order Hidden Markov Models: Exploiting Local Chromosomal Dependencies in the
495   Analysis of Tumor Expression Profiles. *PLoS ONE* 9:e100295.
496   10.1371/journal.pone.0100295
497 Sender R, Fuchs S, and Milo R. 2016. Revised estimates for the number of human and bacteria
498   cells in the body. *bioRxiv*. 10.1101/036103
499 Sherlock C, Xifara T, Telfer S, and Begon M. 2013. A coupled hidden Markov model for disease
500   interactions. *Journal of the Royal Statistical Society Series C, Applied Statistics* 62:609-
501   627. 10.1111/rssc.12015
502 Skewes-Cox P, Sharpton TJ, Pollard KS, and DeRisi JL. 2014. Profile Hidden Markov Models for
503   the Detection of Viruses within Metagenomic Sequence Data. *PLoS ONE* 9:e105067.
504   10.1371/journal.pone.0105067
505 Stein RR, Bucci V, Toussaint NC, Buffie CG, Rätsch G, Pamer EG, Sander C, and Xavier JB. 2013.
506   Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of
507   Intestinal Microbiota. *PLoS Comput Biol* 9:e1003388. 10.1371/journal.pcbi.1003388
508 The Human Microbiome Project Consortium. 2012. Structure, function and diversity of the
509   healthy human microbiome. *Nature* 486:207-214. 10.1038/nature11234
510 Upton M, Tagg JR, Wescombe P, and Jenkinson HF. 2001. Intra- and Interspecies Signaling
511   between Streptococcus salivarius and Streptococcus pyogenes Mediated by SalA and

512   SalA1 Lantibiotic Peptides. *J Bacteriol* 183:3931-3938. 10.1128/JB.183.13.3931-
513   3938.2001

514   van Heel AJ, de Jong A, Montalban-Lopez M, Kok J, and Kuipers OP. 2013. BAGEL3: Automated
515   identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally
516   modified peptides. *Nucleic Acids Res* 41:W448-453. 10.1093/nar/gkt391

517   Walsh CJ, Guinane CM, Hill C, Ross RP, O'Toole PW, and Cotter PD. 2015. In silico identification
518   of bacteriocin gene clusters in the gastrointestinal tract, based on the Human
519   Microbiome Project's reference genome database. *BMC Microbiol* 15:183.
520   10.1186/s12866-015-0515-4

521   Walsh CJ, Guinane CM, O'Toole PW, and Cotter PD. 2014. Beneficial modulation of the gut
522   microbiota. *FEBS Lett* 588:4120-4130. 10.1016/j.febslet.2014.03.035

523   Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Müller R,
524   Wohlleben W, Breitling R, Takano E, and Medema MH. 2015. antiSMASH 3.0—a
525   comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic*
526   *Acids Res* 43:W237-W243. 10.1093/nar/gkv437

527   Xie L, and van der Donk WA. 2004. Post-translational modifications during lantibiotic
528   biosynthesis. *Curr Opin Chem Biol* 8:498-507. 10.1016/j.cbpa.2004.08.005

529   Yan X, Gurtler JB, Fratamico PM, Hu J, and Juneja VK. 2012. Phylogenetic identification of
530   bacterial MazF toxin protein motifs among probiotic strains and foodborne pathogens
531   and potential implications of engineered probiotic intervention in food. *Cell &*
532   *Bioscience* 2:1-13. 10.1186/2045-3701-2-39

533   Yoon B-J. 2009. Hidden Markov Models and their Applications in Biological Sequence Analysis.
534   *Current Genomics* 10:402-415. 10.2174/138920209789177575

535   Zheng J, Gänzle MG, Lin XB, Ruan L, and Sun M. 2014. Diversity and dynamics of bacteriocins
536   from human microbiome. *Environ Microbiol*:n/a-n/a. 10.1111/1462-2920.12662

537

Controls used in validation of the profile HMM, listing the lantibiotic produced and the subclass of modification protein responsible for lanthionine dehydration for each strain.

[a] Lanthionine dehydratase protein identified by our model

[b] Lanthionine dehydratase protein identified by PF04738 model

[c] Lanthionine dehydratase protein identified by BlastP

[d] Lanthionine dehydratase protein identified by BAGEL3

Strain	Bacteriocin	Subclass
Lactococcus lactis ssp. *lactis* S0 [a,b,c,d]	Nisin Z	LanB
Lactococcus lactis ssp. *lactis* CV56 [a,b,c,d]	Nisin A	LanB
Lactococcus lactis ssp. *lactis* IO-1 [a,b,c,d]	Nisin Z	LanB
Bacillus subtilis subsp. *spizizienii* ATCC 6633 [a,b,c,d]	Subtilin	LanB
Staphylococcus aureus subsp. *aureus* USA300_FPR3757 [a,c,d]	Bsa	LanB
Streptococcus mutans CH43 [a,b,d]	Mutacin I	LanB
Streptococcus mutans UA787 [a,b,d]	Mutacin III	LanB
Streptococcus pyogenes [a,b,c,d]	Streptin	LanB
Staphylococcus epidermidis [a,b,c,d]	Pep5	LanB
Lactococcus lactis subsp. *lactis* KF147 [c]	None	-
Streptococcus mutans GS-5	Mutacin GS-5	LanM
Lactococcus lactis subsp. *lactis* plasmid pES2	Lacticin 481	LanM
Streptomyces cinnamoneus cinnamoneus DSM 4005	Cinnamycin	LanM
Bacillus paralichenformis APC 1576	Formicin	LanM
Streptococcus salivarius plasmid pSsal-K12	Salivaricin B	LanM
Streptomyces venezuelae ATCC 10712 [d]	Venezuelin	LanL

Number of metagenomic samples per body site screened.

Site	Number of Samples
Attached Keratinized Gingiva	6
Buccal Mucosa	107
Palatine Tonsils	6
Saliva	3
Stool	139
Subgingival Plaque	7
Supragingival Plaque	118
Throat	7
Tongue Dorsum	128

Detailed information of all lanthionine dehydratase proteins identified in the gastrointestinal tract subset of the Human Microbiome Project's reference genome database using our profile HMM.

Accession	Strain	E Value
JH414709	*Bacillus* sp. 7_6_55CFAA_CT2	9.0E-16
GL636578	*Coprobacillus* sp. 29_1	3.7E-67
AKCB01000002	*Coprobacillus* sp. D6	4.5E-68
JH126516	*Dorea formicigenerans* 4_6_53AFAA	2.3E-81
ACEP01000029	*Eubacterium hallii* DSM3353	9.4E-27
KI391961	*Fusobacterium nucleatum* subsp. *animalis* 3_1_33	2.2E-09
GG657999	*Fusobacterium* sp. 4_1_13	7.1E-09

Figure 1

BAGEL3 output of putative bacteriocin gene clusters identified in the positive controls used for validation of our new profile HMM.

Each predicted open reading frame is colour-coded based on the role it plays in lantibiotic biosynthesis.
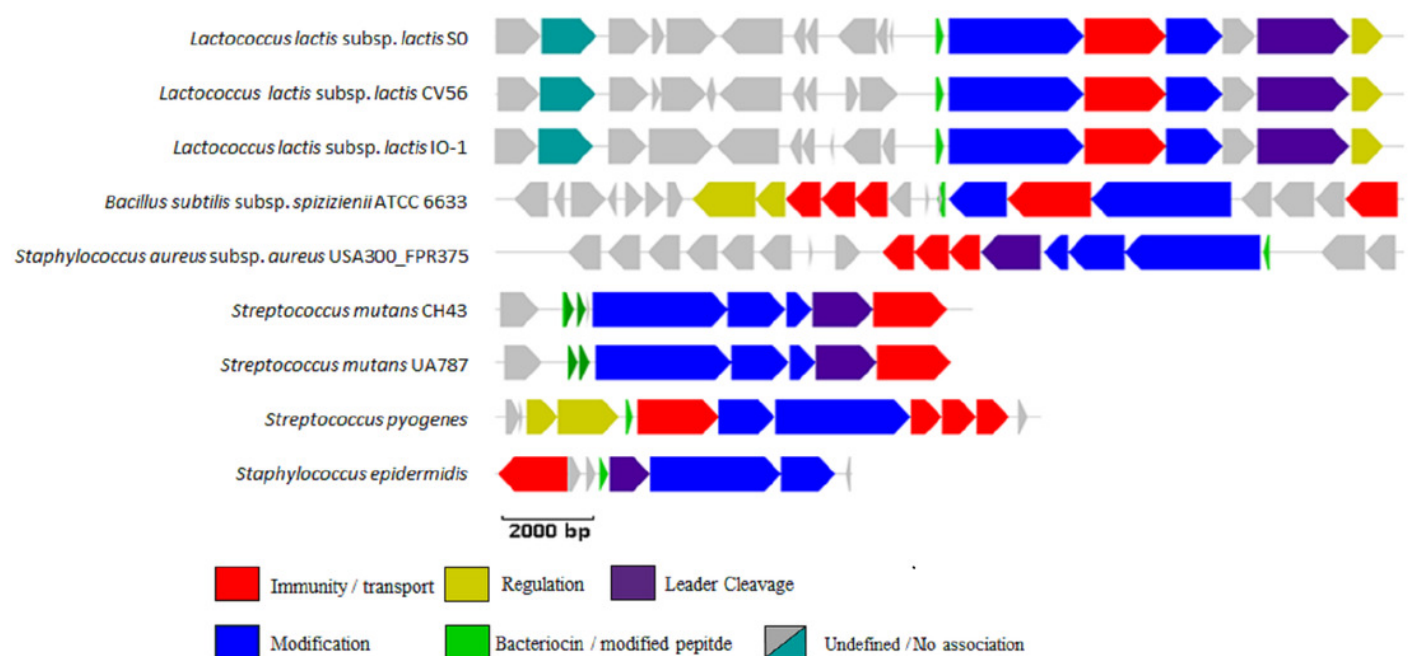
Figure 2

Barchart depicting the distribution of lanthionine dehydratase protein numbers identified by our new profile HMM in metagenomic samples from the stool and oral microbiota.
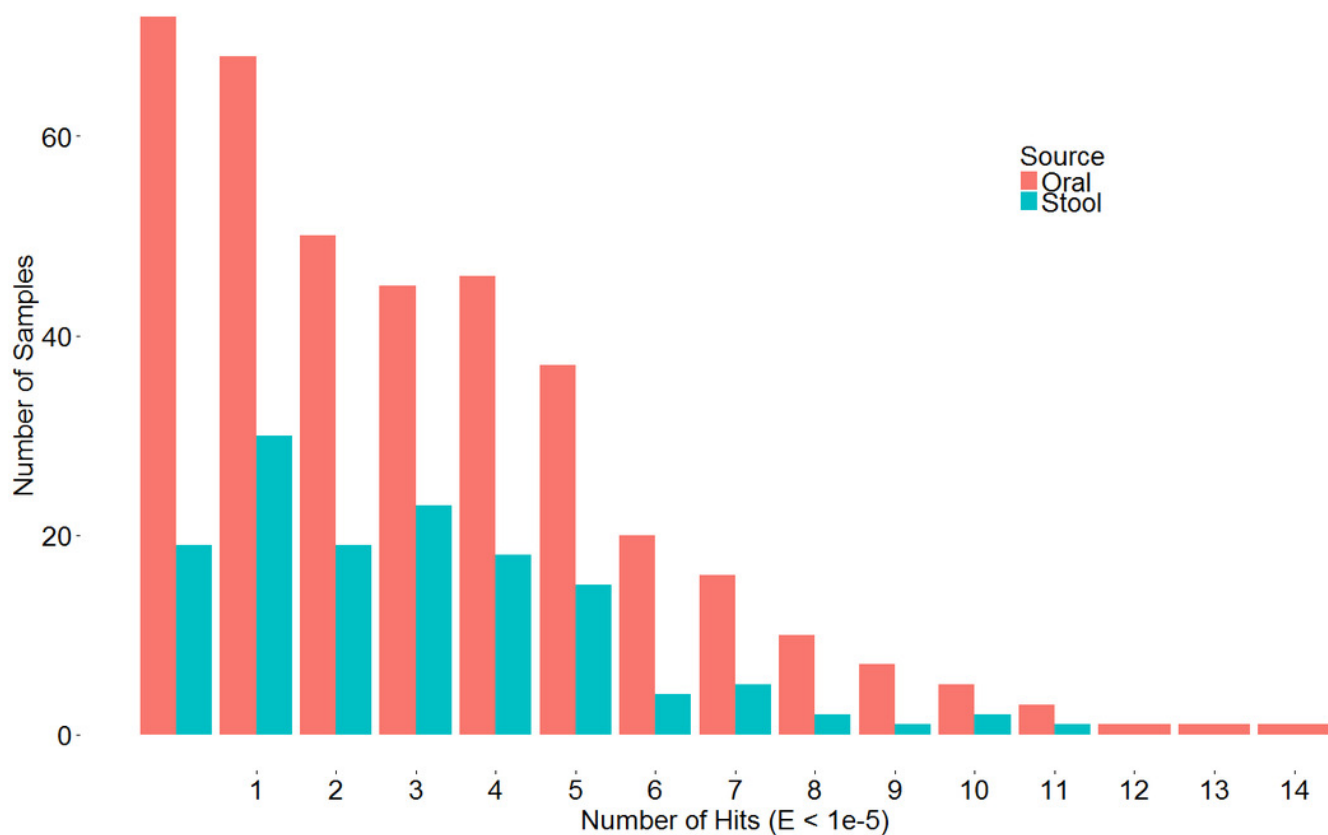
Figure 3

Venn diagram illustrating the numbers of lanthionine dehydratase proteins reported in stool (A) and oral (B) metagenomic data by single and multiple methods.
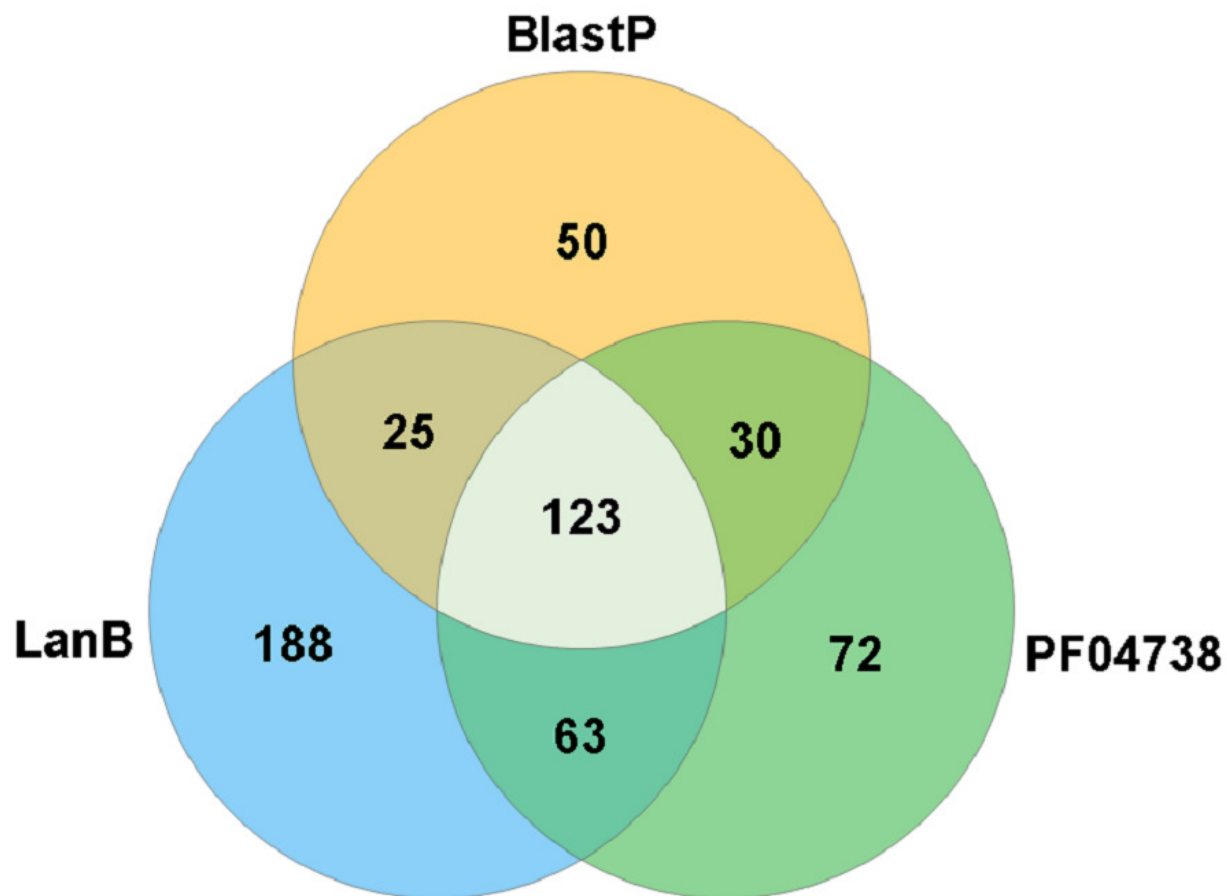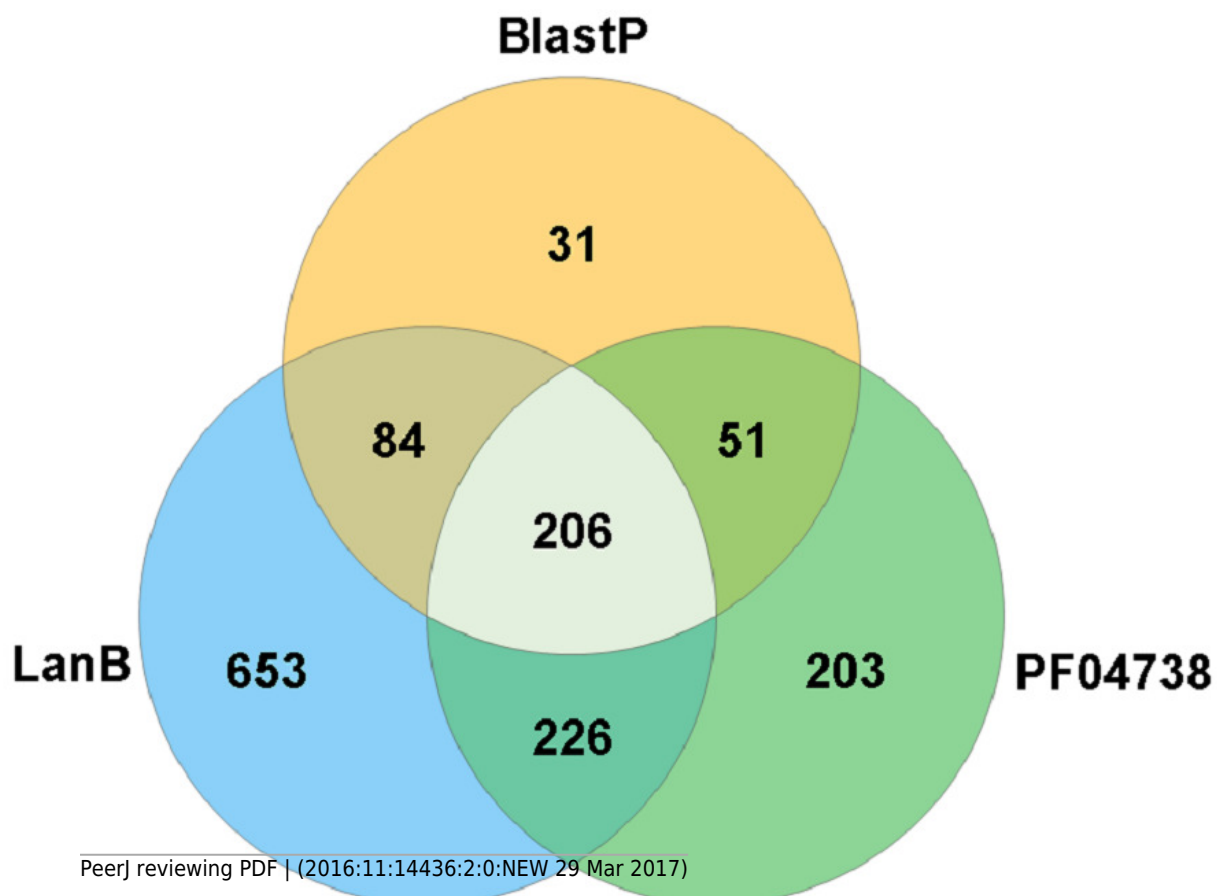
Figure 4

Comparison of lanthionine dehydratase density by body site reported by all three methods

Insert shows overall comparison between stool and oral environments.
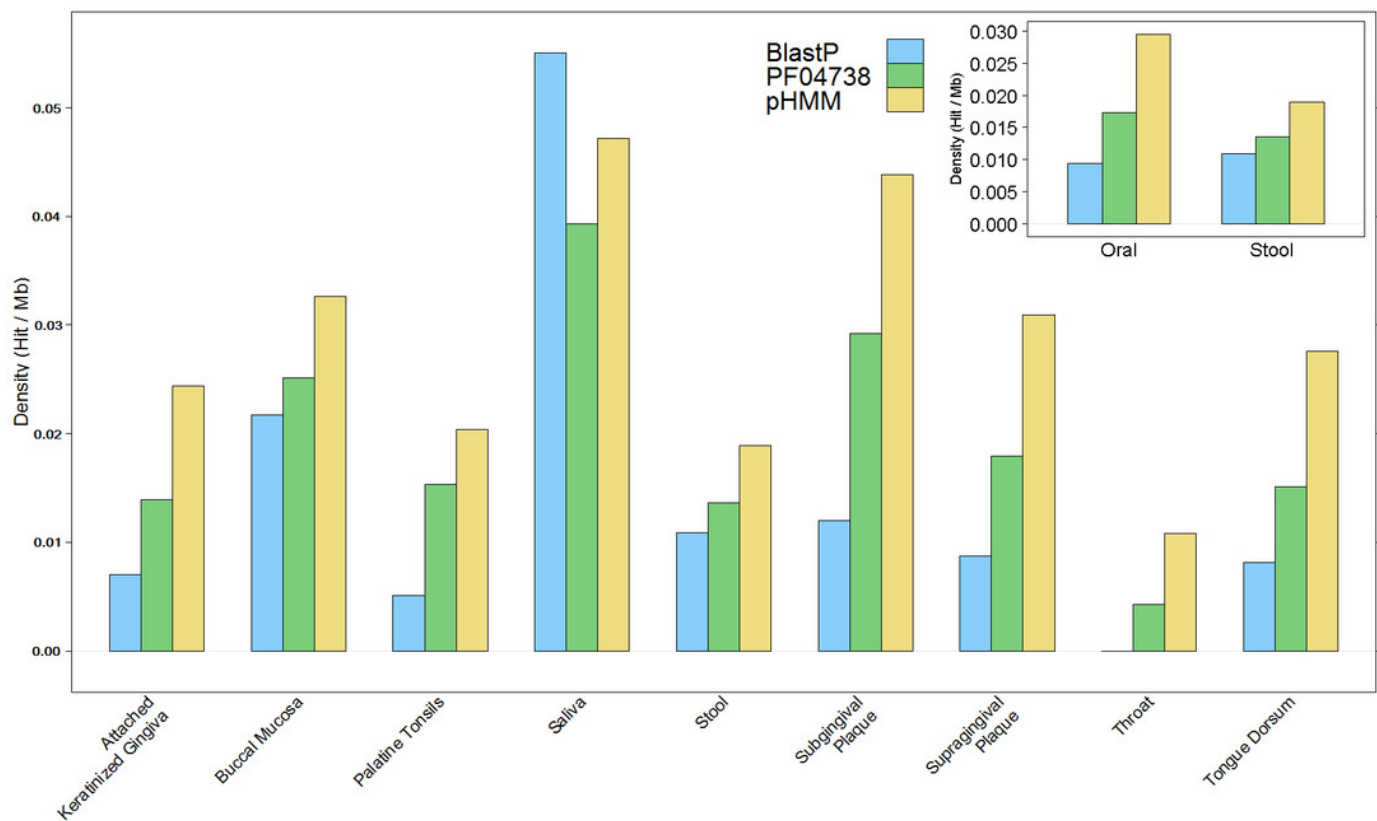
Figure 5

BAGEL3 output of three putative bacteriocin gene clusters identified from the gastrointestinal tract subset of the Human Microbiome Project's reference genome database by our new profile HMM.

A – *Coprobacillus* sp. D6

B – *Coprobacillus* sp. 29_1

C – *Dorea formicigenerans* 4_6_53AFAA

Each predicted open reading frame is colour-coded based on the role it plays in lantibiotic biosynthesis.