Diversity and dynamics of bacteriocins from human microbiome

Jinshui Zheng[1], Michael G. Gänzle[2,3], Xiaoxi B Lin[2], Lifang Ruan[1] and Ming Sun[1*]

[1]State Key Laboratory of Agricultural Microbiology, College of Life Science and Technology, Huazhong Agricultural University, Wuhan 430070, China, [2]Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, Canada, and [3]Department of Bioengineering, Hubei University of Technology, Wuhan 430064, China

*Corresponding author: Sun Ming, E-mail: m98sun@mail.hzau.edu.cn; Tel: 86-27-87283455; Fax: 86-27-87280670.

Running title: Bacteriocins of human microbiome

Summary

Human commensal microbiota are an important determinant of health and disease of the host. Different human body sites harbour different bacterial microbiota, bacterial communities that maintain a stable balance. However, many of the factors influencing the stabilities of bacterial communities associated with humans remain unknown. In this study we identified putative bacteriocins produced by human commensal microbiota. Bacteriocins are peptides or proteins with antimicrobial activity that contribute to the stability and dynamics of microbial communities. We employed bioinformatic analyses to identify putative bacteriocin sequences in metagenomic sequences obtained from different human body sites. Prevailing bacterial taxa of the putative bacteriocins producers matched the most abundant organisms in each human body site. Remarkably, we found that samples from different body sites contain different density of putative bacteriocin genes, with the highest in samples from the vagina, the airway, and the oral cavity and the lowest in those from gut. Inherent differences of different body sites thus influence the density and types of bacteriocins produced by commensal bacteria. Our results suggest that bacteriocins play important roles to allow different bacteria to occupy several human body sites, and to establish a long-term commensal relationship with human hosts.

Keywords: Bacteriocins, human microbiome, bacterial community, bacterial ecology

Introduction

Bacteriocins are ribosomally synthesized peptides with antimicrobial activity. Bacteriocins are produced by all major lineages of eubacteria and by some Archaea (Riley and Wertz, 2002b; Cotter et al., 2005). Bacteriocins produced by Gram-negative bacteria include relatively large proteins; for example, colicins range in size from 449 to 629 amino acids (Riley and Wertz, 2002a). Most bacteriocins from Gram-positive are peptides with typically less than 70 amino acids (Jack et al., 1995).

Bacteriocins are classified on the basis of the peptide structure. Class I bacteriocins, also termed lantibiotics, contain the polycyclic thioether amino acids lanthionine or methyllanthionine which are formed by post-translational modification (McAuliffe et al., 2001). Lantibiotics were grouped depending on the biosynthetic enzymes that install the thioether motifs (Arnison et al., 2013). The dehydration for the four groups of Class I bacteriocins is carried out by LanB, LanM, LanKC and LanL, respectively. Class II bacteriocins are small and heat stable (Ennahar et al., 2000; Cotter et al., 2005; Oppegard et al., 2007; Lee et al., 2009; Nissen-Meyer et al., 2009). Class IIa or pediocin-like bacteriocins contain the conserved N-terminal amino acid sequence YGNGVXCXXXCXV; the two conserved cysteine form a disulfide bridge. Class IIb bacteriocins require two different peptides for antibacterial activity. Class IIc are circular bacteriocins. Class IId are bacteriocins that do not match the other three categories. Bacteriolysins are large, heat labile proteins which exert antimicrobial activity by cell wall hydrolysis;

large bacteriocins including bacteriolysins were previously designated as Class III bacteriocins (Cotter et al., 2005; Nes et al., 2007).

The inhibitory spectra of bacteriocins can be broad or narrow but typically include organisms that are closely related to the producer strain (Cotter et al., 2013). The mode of action can be generalized for certain classes of bacteriocins. Colicins exert their lethal action by first binding to specific receptors on the outer membrane, followed by translocation to the periplasm. Bactericidal activity is dependent on formation of a voltage-dependent channel in the inner membrane or the endonuclease activity of colicins on DNA, rRNA, or tRNA (Cascales et al., 2007). Class I bacteriocins bind to membrane lipids and subsequently inhibit cell wall synthesis, disrupt the membrane in a non-targeted fashion, or both (Hechard and Sahl, 2002). Class II bacteriocins are generally membrane active and dissipate the transmembrane proton motive force by formation of pores (Ennahar et al., 2000).

Bacteriocin production contributes to the dynamics or stability of bacterial communities. Bacteriocins serve as toxins in interference competition, preventing the invasion of other species or enabling the producer strain to establish itself in a new community. This ecological function is well documented for the colonization of the oral cavity by *Streptococcus mutans* (Hillman et al., 1987; Gronroos et al., 1998). Bacteriocin production is often quorum-sensing regulated and also participates in bacterial inter-species communication (Tabasco et al., 2009). Intercellular peptide signals are particularly relevant in dense bacterial consortia such as biofilms (Kreth et al., 2005; Majeed et al., 2011; Dobson et al., 2012). Bacteriocins also contribute to

host-microbe interactions. Plantaricin produced by *L. plantarum* modulates the production of cytokines by dendritic cells (Meijerink et al., 2010) and may thus contribute to probiotic activity. In contrast, the lantibiotic cytolysin is considered a virulence factor because it lyses eukaryotic cells as well as prokaryotic cells (Coburn et al., 2004).

Although multiple potential roles for bacteriocin formation for the microbial ecology of the human microbiome were suggested, only few studies provide evidence for an ecological relevance for bacteriocin production by human commensal microbiota. The availability of metagenome sequences can direct further studies to improve our understanding the ecological role of bacteriocins. The availability of more than 700 shotgun metagenomic datasets from the Human Microbiome Project (HMP) enables analysis of the distribution and diversity of bacteriocins among different body sites (Human Microbiome Project Consortium, 2012). It was the aim of this study to identify genes coding for putative bacteriocins on the metagenomic dataset from the HMP. Bacteriocin distribution was studied among different body sites to reveal the relationships between bacteriocins and the taxonomic structure of bacterial communities. The highly fragmented metagenome sequence data did not allow the reliable identification of the whole gene clusters of bacteriocins; therefore, this study thus focused on the identification of putative bacteriocins and bacteriolysins by using proteins sequences of bacteriocins from the database BAGEL (van Heel et al., 2013) as query sequences.

Results

Content of putative bacteriocin genes in the HMP: mapping the diversity.

The diversity of bacteriocin genes produced by the human microbiome was evaluated by metagenome mining. This method predicted 4875 putative bacteriocins, 802 Class I bacteriocins, 3048 Class II bacteriocins, and 1025 large bacteriocins (Tables S1, S2 and S3 of the online supplementary information). These bacteriocins were clustered together with reference sequences and the clusters were visualized in sequence similarity networks (Fig. 1A, 1B and 1C). Putative Class I bacteriocins clustered in 47 clusters (Fig. 1A), which can be combined into 31 groups based on similarity to query sequences (Table 1). About 20% of class I bacteriocins belong to the most abundant group containing an "L_biotic_typeA" domain (pfam04604) (Table 1). Sequences of the second and third abundant group showed similarity to the two-peptide lantibiotics haloduracin β and haloduracin α, respectively. These two groups together represented 34% of the class I bacteriocins. None of the other groups contained more than 10% of the class I bacteriocins (Table 1).

Putative Class II bacteriocins grouped in 40 MCL clusters (Fig. 1B) that can be combined into 14 groups based on their conserved domains (Table 1). Only two groups represented more than 100 sequences; these two groups represented 93% of the class II bacteriocins (Table 1). The largest group with 2446 sequences related to Bacteriocin IIc (PF10439). The second largest group with sequences from 2 MCL clusters included bacteriocins related to lactococcin 972 (PF09683). Other groups represented less than 100 sequences (Table 1).

Large bacteriocins from HMP were grouped into 10 MCL clusters (Fig. 1C) and

also 10 groups (Table 1). The most abundant groups belonged to closticin 574, colicin, linocin M18, putidacin L1, albusin B and carocin D. These groups together comprised 96% of all large bacteriocin sequences identified in this study.

The density of putative bacteriocin genes differs between different body sites

To assess the significance of bacteriocin production in different sites of the human body, the dataset was analyzed to depict the contribution of different habitats to the overall number of bacteriocin sequences. Most bacteriocins were identified in samples from the oral cavity (Table 1). About 88% (373/420) samples from the oral cavity, 73% (113/154) samples from the gut, 67% (18/27) samples from the skin and 72% (49/68) samples from the vagina contained one or more bacteriocins; for samples from the airway, only 18% (17/94) contained bacteriocin sequences. However, the metagenome sizes for samples from the gut and the oral cavity were generally larger than 10 Mbp, while metagenome sizes of other body sites were substantially smaller (Fig. S1). To correct the number of putative bacteriocin sequences for the metagenome sizes of the different body sites, we plotted the total number of bacteriocins in each body site relative to the total metagenome sequence size (Fig. 2A). Samples from vagina contained on average more than 5 putative bacteriocin genes per Mbp total sequence. In contrast, samples form the gut contained on average less than 0.1 putative bacteriocin gene per Mbp total sequence, indicating that a majority of members of colonic microbiota do not harbour bacteriocin genes. Samples from the gut contained significantly fewer putative bacteriocin genes per Mbp metagenomic DNA when compared to samples from all

other sites while sample from the vagina contained significantly more putative bacteriocin genes per Mbp metagenomic DNA when compared to samples from all other sites (Fig. 2A). Because sequences for class I and II bacteriocins are shorter than sequences for large bacteriocins, we additionally compared the ratio of the length of genes coding for Class I and II bacteriocins (bp) to the total metagenome sequence length (Mbp) of the samples (Fig. 2B). This analysis confirmed results shown in Fig. 2A. Samples from the vagina contained more than 1000 bp bacteriocin gene sequences per Mbp total sequence; samples from airway, oral cavity, and the skin contained more than 100 bp bacteriocin gene sequences per Mbp total sequence while samples from the gut has the lowest bacteriocin density and contained only about 10 bp bacteriocin gene sequences per Mbp total sequence.

The type of putative bacteriocins differs between different body sites

Analysis of the distribution of the five most abundant bacteriocin families among the body habitats revealed that the abundance of bacteriocins classes differs between the respective body sites (Table 1). A breakdown based on the individual sampling sites is provided in Fig. S2 of the online supplemental material. Class I bacteriocin sequences were mainly obtained from oral cavity samples (Table 1 and Fig. S2A). Putative haloduracin α like bacteriocins were the only group with a significant proportion of sequences from the gut; sequences for L_biotics_type A included a substantial proportion of sequences from vaginal samples.

Class II bacteriocins representing the 5 most abundant bacteriocin clusters were also identified mainly in samples from the oral cavity (Table 1 and S2B). Bacteriocin

153 sequences containing conserved domain Bacteriocin_IIa (PF01721) were an
154 exception as 80% (42/50) of these were from vaginal samples (Tabel 1). With the
155 exception of Albusin B like bacteriocins, the top 6 most abundant groups of large
156 bacteriocins were also dominated by samples from the oral cavity.

157 **Relationship of organisms producing putative bacteriocins to the community**
158 **structure in the different body sites.**

159 To compare the composition of bacterial taxa producing putative bacteriocins with
160 the overall community structure in the different body sites, bacteriocin sequences
161 were assigned to reference genomes and compared to the abundance of bacterial
162 genera in the respective body sites (Figs 3 and 4). Of the putative Class I, Class II,
163 and large bacteriocins, 60%, 92% and 80%, respectively, matched to reference
164 genomes (Table S1, S2 and S3). Among these, 39% and 91% of Class I and II
165 bacteriocins, respectively, matched to genomes of the genus *Streptococcus* (Figs 3A
166 and 3B). The other genera which referred by more than 5% of class I bacteriocins
167 were *Prevotella*, *Rothia*, *Bacteroides* and *Actinomyces*. For class II bacteriocins, only
168 *Lactobacillus* represented more than 5% of the bacteriocin sequences. Many of the
169 large bacteriocin sequences were represented by *Actinomyces* and *Prevotella*, with
170 28% and 7%, respectively (Fig. 3C); *Neisseria* and *Streptococcus* also represented
171 more than 5% of large bacteriocins.

172 To reveal differences in reference genomes of bacteriocin-producing bacteria
173 between different body sites, the reference genomes representing different body sites
174 are plotted in Figs 4A - 4C; reference genomes of all bacteriocin producing bacteria

are plotted in Fig. 4D. The reference genomes generally matched the taxonomic composition of microbiota in the respective body sites (Fig. 4E). Reference genomes for bacteriocins from the oral cavity predominantly included *Streptococcus* spp., *Prevotella* spp. and *Actinomyces* spp (Figs 4A, 4B and 4C); each of these genera were also abundant in oral cavity samples (Fig 4E). Abundant species representing reference genomes for gut bacteriocins included *Bacteroides fragilis*, *Bacteroides dorei*, *Eubacterium rectale*, *Escherichia coli* and *Blautia hansenii* (Table S4 and Fig 4); with exception of *E. coli*, these species are also abundant in colonic microbiota (Fig. 4E). Reference genomes for the bacteriocins from vaginal samples were almost exclusively from *Lactobacillus*, the most abundant genus in vaginal microbiota (Fig. 4E). Likewise, reference genomes for bacteriocins in airway samples were represented by abundant bacterial taxa in these samples (Grice and Segre, 2011; Human Microbiome Project Consortium, 2012). Most bacteriocin sequences from skin sites matched reference genomes of *Staphylococcus epidermidis*.

The large number of putative bacteriocin sequences from the oral cavity allowed the differentiation of reference genomes between specific sampling sites of the oral cavity (Fig. S3A and S3B) and differentiation at the species level (Fig S3C and S3D). *Streptococcus* spp. account for a large majority of class I and II bacteriocins reference genomes in most subsites, however, Class I bacteriocin sequences of the supragingival plaque matched predominantly to reference genomes of *Rothia* spp.. *Prevotella* spp. accounted for 45% of the reference genomes for class I bacteriocins from the tongue dorsum (Fig. S3A)

197 The reference genomes of *Streptococcus* matching bacteriocin sequences from the
198 buccal mucosa and the tongue dorsum were analyzed at the species level (Fig S3C
199 and S3D). Genomes from several species matched sequences of both classes of
200 bacteriocins from both sites. For example, genomes of *Streptococcus mitis* were the
201 most abundant reference genomes matching both classes of bacteriocins from buccal
202 mucosa, and frequently identified as reference genomes for bacteriocin sequences
203 from the tongue dorsum. Other reference genomes that were frequently matched to
204 bacteriocin sequences from the tongue dorsum or the buccal mucosa include
205 *Streptococcus pneumonia, Streptococcus salivarius Streptococcus australis*,
206 *Streptococcus pseudopneumoniae* and *Streptococcus vestibularis*.

207 The reference genomes of the query sequences the BAGEL database are plotted in
208 Fig. S4. The database comprises entries representing 145 genera producing Class I
209 bacteriocins; *Streptococcus*, *Streptomyces* and *Bacillus* together account for 47% of
210 the Class I bacteriocins (Fig. S4A). The genera *Lactobacillus*, *Enterococcus* and
211 *Streptococcus* together account for 58% of the Class II bacteriocins (Fig. S4B).
212 Almost half of the large bacteriocins in the BAGEL database are produced by
213 *Escherichia coli* (Fig. S4C); *Klebsiella* and *Pseudomonas* each account for 6% of the
214 large bacteriocins. The discrepancy in composition of bacterial taxa represented in
215 the query sequences (Fig. S4) and the HMP sequences (Fig. 3) demonstrates that this
216 study identified a large number of bacteriocins sequences in bacterial taxa that are
217 not represented in the query data set.

218 **Putative cell wall hydrolases (bacteriolysins) in the human microbiome.**

Sequences that were deposited were edited to exclude bacteriolysins, which are not consistently classified as bacteriocins. The analysis of sequences of putative bacteriolysins retrieved numerous sequences related to the streptococcal peptidoglycan hydrolase zoocin A (2756 sequences). This large number of sequences with homologies to zoocin A likely overestimates the number of antibacterial proteins because the antibacterial enzyme zoocin A is homologous to housekeeping enzymes involved in cell wall turnover (Meisner and Moran, 2011; Simmonds et al., 1997). The cell wall degrading, antibacterial enzymes enterolysin A (124 sequences) and helveticin J (62 sequences) were also identified (Fig. S5 of the online supplemental material). Putative bacteriolysin sequences were primarily identified in samples from oral cavity and vaginal samples (Fig. S5). Remarkably, a majority of sequences matching the bacteriolysins enterolysin A and helvetivin J were obtained from vaginal samples (Fig. S5).

Discussion

Past studies to identify bacteriocins with bioinformatic tools used genome sequence data to identify conserved sequences of putative bacteriocins or bacteriocin biosynthetic enzymes (Dirix et al., 2004; Nes and Johnsborg, 2004; Murphy et al., 2011). Our approach to mine a metagenomic library has the advantage of making a much larger dataset accessible for (meta-)genome mining. Moreover, the dataset includes sequence data from uncultured organisms for which genome sequence data is not available. Because many of the bacteriocins predicted in this study are unrelated to known bacteriocins, and match to bacterial taxa for which bacteriocin

241 production has not been described, this approach likely predicted novel bacteriocins
242 with biologically relevant antimicrobial activity. Bacteriocins are not toxic to
243 humans and were suggested to be a viable alternative to antibiotics (Cotter et al.,
244 2013). The identification of novel putative bacteriocins from the human microbiome
245 greatly expands the number of compounds for potential therapeutic use. This study
246 also analyzed the density and distribution of putative bacteriocins in different body
247 sites. The estimation of the abundance of putative bacteriocin sequences in different
248 body sites provides insight into the role of bacteriocins in different microbial
249 ecology.

250 Bacteriocin sequences were retrieved on the basis of homologies to sequences
251 deposited in the BAGEL database (van Heel et al., 2013). The comparison of
252 putative bacteriocin sequences to reference genomes demonstrated that bacteriocin
253 producing strains matched the taxonomic composition of commensal microbiota in
254 the respective body sites (Aas et al., 2005; Bik et al., 2010; Arumugam et al., 2011;
255 Human Microbiome Project Consortium, 2012; Ma et al., 2012; Santagati et al.,
256 2012) but not the taxa represented in the BAGEL database (van Heel et al., 2013).
257 For example, *Bacteroidetes* are not represented in the BAGEL database but a
258 substantial proportion of sequences of putative bacteriocins from the oral cavity and
259 the gut were assigned to *Prevotella* and *Bacteroides* (Fig. 4). Conversely,
260 bacteriocins from *Escherichia coli* (microcins and colicins) are well represented in
261 the BAGEL database, however, very few predicted bacteriocins were assigned to this
262 species. Of note, bacteriocins produced by *Bifidobacterium* spp. were identified in

samples of the oral cavity, where bifidobacteria are only a minor component, but not in samples from the gut, where bifidobacteria account for up to 10% of the microbiome.

The role of bacteriocins in the ecology of human commensal microbiota is well described for oral microbiota. Oral bacteria form biofilms to persist despite the constant flow of host secretions (Kolenbrander et al., 2010). Bacteriocin production by oral streptococci was linked to their ability to colonize the oral cavity (Hillman et al., 1987; Gronroos et al., 1998; Hillman, 2002), and to a reduced frequency of colonization by respiratory pathogens (Santagati et al., 2012). This study predicted a high density of putative bacteriocin genes in samples from the oral cavity, confirming the role of bacteriocins in the microbial ecology of the oral cavity. The density of bacteriocin sequences was comparable for samples from the airway and the skin; the highest density of bacteriocin sequences was observed in samples from the vagina. Bacteriocin sequences from vaginal samples predominantly matched the most abundant *Lactobacillus* species in human vaginal microbiota (Ma et al., 2012). Our predictions from HMP sequence data thus confirm and extend previous reports that bacteriocin production is a frequent physiological trait of vaginal lactobacilli (Stoyancheva et al., 2014) and thus may play a comparable ecological role as in the oral cavity.

Past studies provided proof of concept that bacteriocin production by bacteria that are allochthonous to the gut modulate populations of other allochthones, i.e. listeria (Corr et al., 2007), or food-derived vancomycin resistant enterococci (Millette et al.,

2008). In contrast to the oral cavity of adults, which can be permanently colonized by bacteriocin producing streptococci (Hillman et al., 1987; Gronroos et al., 1998), colonic microbiota are “colonization resistant” and a role of bacteriocin production by autochthonous members of colonic microbiota remains to be established. This study observed that the density of bacteriocin sequences was 10- 100 fold lower in samples from the gut when compared to samples from the oral cavity and the vagina. The low density of genes coding for production of putative bacteriocins in colonic microbiota may reflect inherent differences in the ecology of different body sites.

Host immunity and host genetics influence the composition of colonic microbiota and the secretion of antimicrobial peptides is an important component of control of colonic microbiota by the host (Benson et al., 2010; Hansen et al., 2010; Willing et al., 2011). Competition and cooperation among microbes also is a major factor shaping the microbial ecology of the gut (Lahti et al., 2014). Competition can be divided two broad categories: exploitative competition, which indirectly affects competitors by limiting the abundance of resources, and interference competition which directly harms other strains and species through production of antimicrobial compounds (Mitri and Foster, 2013). In the gut, the stability of ecosystem is tightly maintained by various syntrophic links within a community (Fischbach and Sonnenburg, 2011). Microbial competitiveness of colonic microbiota is linked to the efficient exploitation of the limited nutrient (carbohydrate) supply (Louis et al., 2007; Derrien et al., 2008; van den Broek et al., 2008; Walter and Ley, 2011). The high density of genes for putative glycosyl hydrolases in genomes of colonic bacteria (Flint et al., 2008)

corresponds to exploitative competition and thus relates to the low density of putative bacteriocin genes as indicators of interference competition. At other body sites where the selection by nutrient availability and host secretions is less strict, interference competition might play a bigger role.

The spatial structure of the microbial habitat may also affect the benefit of antimicrobial production in the gut. Habitats where antimicrobial producers are abundant generally have highly structured microbial communities. For example, specific temporal and spatial distribution is crucial for the development of dental biofilms (Rickard et al., 2003). The highly structured spatial distribution allows for localized interaction between bacterocin producer and its target organisms even when the producer species is a minority in the total microbiota. For example, *S. mutans*, a minor species on healthy teeth, has been known to produce a diversity of bacteriocins. In contrast, the gut environment is less localized. The semi-liquid state of the digesta and frequent mixing by intestinal peristalsis facilitate diffusion of bacteriocin. Here, production of bacteriocins and related antimicrobials is favored only when the producer organisms account for a high proportion of the overall population, allowing the product to accumulate to active concentrations, and thus compensating for the metabolic cost of bacteriocin production by inhibiting competitors (Abrudan et al., 2012). If the producer organisms constitutes only a minority in the total population, bacteriocins will be diluted below their active concentrations and the producer organisms is not compensated for the cost of production (Abrudan et al., 2012).

In conclusion, we have identified thousands of putative novel bacteriocin sequences produced by human commensal microbiota. Our approach identified novel clusters of Class I and Class II bacteriocins from Gram-positive and Gram-negative members of colonic microbiota, including bacterial taxa for which bacteriocin production has not been described. Moreover, the different density of bacteriocin sequences in samples from different body habitats indicates that bacteriocin production is less relevant for the ecology of colonic microbiota but provides an competitive advantage to members of the microbiome in other body habitats, particularly the oral cavity and the vagina.

Experimental procedures

Data collection

Data of HMGI (HMP Gene Index), HMASM (HMP Illumina WGS Assemblies) and HMRGD (HMP Reference Genomes Data) were downloaded from the Data Analysis and Coordination Center (DACC) for the Human Microbiome Project (HMP) (http://www.hmpdacc.org/). Amino acid sequences of bacteriocins were obtained from the Bacteriocin databases of BAGEL (http://bagel2.molgenrug.nl/index.php/bacteriocin-database) (van Heel et al., 2013). On 21 April 2013, the BAGEL database listed 158 class I, 235 class II and 91 class III bacteriocins (large bacteriocins and bacteriolysins). Bacteriolysins (proteins matching to zoocin A, enterolysin A, or helveticin J) were eliminated from the class III bacteriocins in the query dataset; remaining bacteriocins were termed “large bacteriocins”.

Bacteriocins prediction

The process for identification class I and II bacteriocins with anmino acid lengths less than 100 is depicted in Fig. S6. Amino acid sequences from HMGI were used to build BLAST database for each human body sites. Using amino sequences of bacteriocins obtained from BAGEL as query dataset, different sets of parameters for PSI-BLAST against these databases were evaluated to improve the recognition of class I and II bacteriocins. The optimum set of parameters was as follows: matrix: PAM70; word size: 2; expect threshold <0.001; no low complexity filtering; no composition based statistics (Schaffer et al., 2001). PSI-BLAST with iterations = 4 was performed for each of the query bacteriocin sequence against amino sequences from each human body sites. Target sequences (raw bacteriocin amino acid sequences) were extracted and used as query sequences for BLASTP against the NR database of NCBI. With an E-value of $< 10^{-6}$, sequences were excluded if the length of the top target sequences in NR was longer than 100 amino acids or annotated as with functions other than those referring to bacteriocins. After filtering, 802 class I and 3048 class II bacteriocins were predicted. For predicting large bacteriocins with more 100 amino acids, we combined the results of the following two methods. Firstly, we collected the conserved domains for all the sequences from large bacteriocins form BAGEL whose conserved damain were available. And then all the sequences containing these domains from HMP were obtained through the HMMER software. Secondly, using those BAGEL large bacteriocins sequences without any known conserved domain as query, amino acid sequences from HMG as database,

372 homologous sequences with more than 30% identity and 70% coverage were
373 obtained.

374 **Sequences clustering and grouping**

375 The amino acid sequences from the BAGEL dataset and the sequences retrieved
376 from the HGMI were combined to form three fasta-format files, one for class I, one
377 for class II and the other for large bacteriocins. An all against all BLASTP was
378 performed with E-value < 0.001 for each class of bacteriocins, respectively. Before
379 sequences clustering, we obtained the best E-value pairwise comparison for each
380 pair of sequences with similarity to get the best BLAST result files. For class I and II,
381 an E-value of 10^{-6} corresponding to a median of 30% amino acid identity was used
382 as the upper limit for both of the best BLAST result files. MCL (Markov Cluster
383 Algorithm) algorithm (Enright et al., 2002) with an inflation value of 1.6 was used
384 for clustering amino acid sequences for both class I and II bacteriocin sequences. For
385 large bacteriocins, the clustering was similar as class I and II with the E-value cutoff
386 being 10^{-10}, and the inflation for MCL being 2.

387 Protein sequences in each MCL cluster were analyzed with respect to conserved
388 domains and the most similar sequences in the query bacteriocins. Different clusters
389 with the same conserved domain or the same most similar query bacteriocin were
390 treated as the same group.

391 For all the three classes of bacteriocins, protein sequences similarity network for
392 each MCL cluster were built as reported (Atkinson et al., 2009). Clusters for both

393 classes and sub-clusters for huge clusters of class II bacteriocins were visualized
394 using Cytoscape (Smoot et al., 2011) with the organic layout and colored the nodes
395 by which sites those sequences from.

396 **Reference taxonomic assignment**

397 Nucleotide sequences for each of the scaffolds or contigs which contained the
398 predicted bacteriocins were extracted from the HMASM dataset. Each of these
399 nucleic acid sequences was BLASTN against the HMRGD, and the reference
400 genome taxonomy was assigned to the scaffold or contig and the bacteriocin(s) it
401 contained if the identity between the query and target sequences was higher than
402 90%.

403 **Acknowledgements**

404 Ming Sun acknowledges support from the National High Technology Research and
405 Development Program (863) of China (2011AA10A203), and China 948 Program of
406 Ministry of Agriculture (2011-G25). Michael Gänzle acknowledges support from the
407 Canada Research Chairs Program.

408 **Reference**

409 Aas, J.A., Paster, B.J., Stokes, L.N., Olsen, I., and Dewhirst, F.E. (2005) Defining the normal bacteria1
410 flora of the oral cavity. *J Clin Microbiol* **43**: 5721-5732.

411 Abrudan, M.I., Brown, S., and Rozen, D.E. (2012) Killing as means of promoting biodiversity.
412 *Biochem Soc Trans* **40**: 1512-1516.

413 Arnison, P.G., Bibb, M.J., Bierbaum, G., Bowers, A.A., Bugni, T.S., Bulaj, G. et al. (2013)

414 Ribosomally synthesized and post-translationally modified peptide natural products: overview and
415 recommendations for a universal nomenclature. *Nat Prod Rep* **30**: 108-160.
416 Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R. et al. (2011)
417 Enterotypes of the human gut microbiome. *Nature* **473**: 174-180.
418 Atkinson, H.J., Morris, J.H., Ferrin, T.E., and Babbitt, P.C. (2009) Using sequence similarity networks
419 for visualization of relationships across diverse protein superfamilies. *PLoS One* **4**: e4345.
420 Benson, A.K., Kelly, S.A., Legge, R., Ma, F., Low, S.J., Kim, J. et al. (2010) Individuality in gut
421 microbiota composition is a complex polygenic trait shaped by multiple environmental and host
422 genetic factors. *Proc Natl Acad Sci U S A* **107**: 18933-18938.
423 Bik, E.M., Long, C.D., Armitage, G.C., Loomer, P., Emerson, J., Mongodin, E.F. et al. (2010)
424 Bacterial diversity in the oral cavity of 10 healthy individuals. *ISME J* **4**: 962-974.
425 Cascales, E., Buchanan, S.K., Duche, D., Kleanthous, C., Lloubes, R., Postle, K. et al. (2007) Colicin
426 biology. *Microbiol Mol Biol Rev* **71**: 158-229.
427 Coburn, P.S., Pillar, C.M., Jett, B.D., Haas, W., and Gilmore, M.S. (2004) *Enterococcus faecalis*
428 senses target cells and in response expresses cytolysin. *Science* **306**: 2270-2272.
429 Corr, S.C., Li, Y., Riedel, C.U., O'Toole, P.W., Hill, C., and Gahan, C.G. (2007) Bacteriocin
430 production as a mechanism for the antiinfective activity of *Lactobacillus salivarius* UCC118. *Proc*
431 *Natl Acad Sci U S A* **104**: 7617-7621.
432 Cotter, P.D., Hill, C., and Ross, R.P. (2005) Bacteriocins: developing innate immunity for food. *Nat*
433 *Rev Microbiol* **3**: 777-788.
434 Cotter, P.D., Ross, R.P., and Hill, C. (2013) Bacteriocins - a viable alternative to antibiotics? *Nat Rev*
435 *Microbiol* **11**: 95-105.

436 Derrien, M., Collado, M.C., Ben-Amor, K., Salminen, S., and de Vos, W.M. (2008) The Mucin
437 degrader Akkermansia muciniphila is an abundant resident of the human intestinal tract. *Appl Environ*
438 *Microbiol* **74**: 1646-1648.
439 Dirix, G., Monsieurs, P., Dombrecht, B., Daniels, R., Marchal, K., Vanderleyden, J., and Michiels, J.
440 (2004) Peptide signal molecules and bacteriocins in Gram-negative bacteria: a genome-wide in silico
441 screening for peptides containing a double-glycine leader sequence and their cognate transporters.
442 *Peptides* **25**: 1425-1440.
443 Dobson, A., Cotter, P.D., Ross, R.P., and Hill, C. (2012) Bacteriocin production: a probiotic trait?
444 *Appl Environ Microbiol* **78**: 1-6.
445 Ennahar, S., Sashihara, T., Sonomoto, K., and Ishizaki, A. (2000) Class IIa bacteriocins: biosynthesis,
446 structure and activity. *FEMS Microbiol Rev* **24**: 85-106.
447 Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002) An efficient algorithm for large-scale
448 detection of protein families. *Nucleic Acids Res* **30**: 1575-1584.
449 Fischbach, M.A., and Sonnenburg, J.L. (2011) Eating for two: how metabolism establishes
450 interspecies interactions in the gut. *Cell Host Microbe* **10**: 336-347.
451 Flint, H.J., Bayer, E.A., Rincon, M.T., Lamed, R., and White, B.A. (2008) Polysaccharide utilization
452 by gut bacteria: potential for new insights from genomic analysis. *Nat Rev Microbiol* **6**: 121-131.
453 Grice, E.A., and Segre, J.A. (2011) The skin microbiome. *Nat Rev Microbiol* **9**: 244-253.
454 Gronroos, L., Saarela, M., Matto, J., Tanner-Salo, U., Vuorela, A., and Alaluusua, S. (1998) Mutacin
455 production by *Streptococcus mutans* may promote transmission of bacteria from mother to child.
456 *Infect Immun* **66**: 2595-2600.
457 Hansen, J., Gulati, A., and Sartor, R.B. (2010) The role of mucosal immunity and host genetics in

458 defining intestinal commensal bacteria. *Curr Opin Gastroenterol* **26**: 564-571.

459 Hechard, Y., and Sahl, H.G. (2002) Mode of action of modified and unmodified bacteriocins from
460 Gram-positive bacteria. *Biochimie* **84**: 545-557.

461 Hillman, J.D. (2002) Genetically modified *Streptococcus mutans* for the prevention of dental caries.
462 *Antonie Van Leeuwenhoek* **82**: 361-366.

463 Hillman, J.D., Dzuback, A.L., and Andrews, S.W. (1987) Colonization of the human oral cavity by a
464 Streptococcus mutans mutant producing increased bacteriocin. *J Dent Res* **66**: 1092-1094.

465 Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy
466 human microbiome. *Nature* **486**: 207-214.

467 Jack, R.W., Tagg, J.R., and Ray, B. (1995) Bacteriocins of gram-positive bacteria. *Microbiol Rev* **59**:
468 171-200.

469 Kolenbrander, P.E., Palmer, R.J., Jr., Periasamy, S., and Jakubovics, N.S. (2010) Oral multispecies
470 biofilm development and the key role of cell-cell distance. *Nat Rev Microbiol* **8**: 471-480.

471 Kreth, J., Merritt, J., Shi, W., and Qi, F. (2005) Competition and coexistence between *Streptococcus*
472 *mutans* and *Streptococcus sanguinis* in the dental biofilm. *J Bacteriol* **187**: 7193-7203.

473 Lahti, L., Salojarvi, J., Salonen, A., Scheffer, M., and de Vos, W.M. (2014) Tipping elements in the
474 human intestinal ecosystem. *Nat Commun* **5**: 4344.

475 Lee, K.D., Gray, E.J., Mabood, F., Jung, W.J., Charles, T., Clark, S.R. et al. (2009) The class IId
476 bacteriocin thuricin-17 increases plant growth. *Planta* **229**: 747-755.

477 Louis, P., Scott, K.P., Duncan, S.H., and Flint, H.J. (2007) Understanding the effects of diet on
478 bacterial metabolism in the large intestine. *J Appl Microbiol* **102**: 1197-1208.

479 Ma, B., Forney, L.J., and Ravel, J. (2012) Vaginal microbiome: rethinking health and disease. *Annu*

480 *Rev Microbiol* **66**: 371-389.

481 Majeed, H., Gillor, O., Kerr, B., and Riley, M.A. (2011) Competitive interactions in *Escherichia coli*

482 populations: the role of bacteriocins. *ISME J* **5**: 71-81.

483 McAuliffe, O., Ross, R.P., and Hill, C. (2001) Lantibiotics: structure, biosynthesis and mode of action.

484 *FEMS Microbiol Rev* **25**: 285-308.

485 Meijerink, M., van Hemert, S., Taverne, N., Wels, M., de Vos, P., Bron, P.A. et al. (2010)

486 Identification of genetic loci in *Lactobacillus plantarum* that modulate the immune response of

487 dendritic cells using comparative genome hybridization. *PLoS One* **5**: e10632.

488 Millette, M., Cornut, G., Dupont, C., Shareck, F., Archambault, D., and Lacroix, M. (2008) Capacity

489 of human nisin- and pediocin-producing lactic Acid bacteria to reduce intestinal colonization by

490 vancomycin-resistant enterococci. *Appl Environ Microbiol* **74**: 1997-2003.

491 Mitri, S., and Foster, K.R. (2013) The genotypic view of social interactions in microbial communities.

492 *Annu Rev Genet* **47**: 247-273.

493 Murphy, K., O'Sullivan, O., Rea, M.C., Cotter, P.D., Ross, R.P., and Hill, C. (2011) Genome mining

494 for radical SAM protein determinants reveals multiple sactibiotic-like gene clusters. *PLoS One* **6**:

495 e20852.

496 Nes, I.F., and Johnsborg, O. (2004) Exploration of antimicrobial potential in LAB by genomics. *Curr*

497 *Opin Biotechnol* **15**: 100-104.

498 Nes, I.F., Yoon, S.-S., and Diep, D.B. (2007) Ribosomally Synthesiszed Antimicrobial Peptides

499 (Bacteriocins) in Lactic Acid Bacteria: A Review. *Food Sci Biotechnol* **15**: 15.

500 Nissen-Meyer, J., Rogne, P., Oppegard, C., Haugen, H.S., and Kristiansen, P.E. (2009)

501 Structure-function relationships of the non-lanthionine-containing peptide (class II) bacteriocins

produced by gram-positive bacteria. *Curr Pharm Biotechnol* **10**: 19-37.

Oppegard, C., Rogne, P., Emanuelsen, L., Kristiansen, P.E., Fimland, G., and Nissen-Meyer, J. (2007) The two-peptide class II bacteriocins: structure, production, and mode of action. *J Mol Microbiol Biotechnol* **13**: 210-219.

Rickard, A.H., Gilbert, P., High, N.J., Kolenbrander, P.E., and Handley, P.S. (2003) Bacterial coaggregation: an integral process in the development of multi-species biofilms. *Trends Microbiol* **11**: 94-100.

Riley, M.A., and Wertz, J.E. (2002a) Bacteriocin diversity: ecological and evolutionary perspectives. *Biochimie* **84**: 357-364.

Riley, M.A., and Wertz, J.E. (2002b) Bacteriocins: evolution, ecology, and application. *Annu Rev Microbiol* **56**: 117-137.

Santagati, M., Scillato, M., Patane, F., Aiello, C., and Stefani, S. (2012) Bacteriocin-producing oral streptococci and inhibition of respiratory pathogens. *FEMS Immunol Med Microbiol* **65**: 23-31.

Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I. et al. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res* **29**: 2994-3005.

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., and Ideker, T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**: 431-432.

Stoyancheva, G., Marzotto, M., Dellaglio, F., and Torriani, S. (2014) Bacteriocin production and gene sequencing analysis from vaginal *Lactobacillus* strains. *Arch Microbiol* **196**: 645-653.

Tabasco, R., Garcia-Cayuela, T., Pelaez, C., and Requena, T. (2009) *Lactobacillus acidophilus* La-5 increases lactacin B production when it senses live target bacteria. *Int J Food Microbiol* **132**: 109-116.

van den Broek, F.J., Fockens, P., van Eeden, S., Reitsma, J.B., Hardwick, J.C., Stokkers, P.C., and Dekker, E. (2008) Endoscopic tri-modal imaging for surveillance in ulcerative colitis: randomised comparison of high-resolution endoscopy and autofluorescence imaging for neoplasia detection; and evaluation of narrow-band imaging for classification of lesions. *Gut* **57**: 1083-1089.

van Heel, A.J., de Jong, A., Montalban-Lopez, M., Kok, J., and Kuipers, O.P. (2013) BAGEL3: Automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. *Nucleic Acids Res* **41**: W448-453.

Walter, J., and Ley, R. (2011) The human gut microbiome: ecology and recent evolutionary changes. *Annu Rev Microbiol* **65**: 411-429.

Willing, B.P., Russell, S.L., and Finlay, B.B. (2011) Shifting the balance: antibiotic effects on host-microbiota mutualism. *Nat Rev Microbiol* **9**: 233-243.

Figure legends

Fig. 1 Metagenomic samples of human body sites contain abundant putative bacteriocins. Sequence similarity networks of bacteriocin protein sequences from BAGEL database and HMP for class I, II and large bacteriocins are presented in (A) - (C). The networks are thresholded at a BLAST E-value of 10^{-6} (for class I and II) or 10^{-10} (for large bacteriocins) and the worst edges displayed correspond to 30% identity for each pair-wise comparison. Nodes were colored to the same color if their presented sequences from the same body site.

Fig. 2 The density of putative bacteriocin genes differs between different body sites. (A) Number of bacteriocins normalized to the amount of sequence data for each body site. The number of bacteriocins per Mbp total metagenome sequence of gut samples were significantly smaller than those of the airway ($p = 1.1*10^{-10}$), the oral cavity ($p < 2.2*10^{-16}$), the skin ($p = 1.2*10^{-09}$) and the vagina ($p < 2.2*10^{-16}$), respectively (Mann-Whitney test). The number of bacteriocins per Mbp total metagenome size of vaginal samples were significant higher than those of airway ($9.3*10^{-9}$), gut ($p < 2.2*10^{-16}$), oral cavity ($p < 2.2*10^{-16}$), and skin ($9.1*10^{-14}$). (B) The proportion of the sequence length of putative class I and class II bacteriocin genes (total of class I and II) normalized to the metagenome sequence length.

Fig. 3 Reference microbial genomes of the putative bacteriocins from HMP. Reference microbial genomes of class I (A), II (B) and large (C) bacteriocins. The same genus from different panels was in the same color.

Fig. 4 Reference genera for class I (A), II (B) and large (C) bacteriocins among

559 different human body sites. Panel D summarizes the reference genera for all
560 bacteriocins combined and panel (H) indicates the phylogenetic composition at the
561 genus level of each site based on 16S rRNA abundance. The same genus from
562 different panels was in the same color.

563

564

565

Supporting information

Fig. S1 Sample metagenome size of different body sites.

Fig. S2 Distributions of dominant bacteriocins among different sampling sites. (A) Distributions of top 5 abundant class I bacteriocins among different sampling sites. (B) Distributions of top 5 abundant class II bacteriocins among different sampling sites. (C) Distributions of top 6 abundant large bacteriocins among different sampling sites.

Fig. S3 Reference microbial genomes of bacteriocins. (A) Reference genera for class I bacteriocins among different sampling sites of oral cavity. (B) Reference genera for class II bacteriocins among different sampling sites of oral cavity. (C) Reference species for class I bacteriocins of samples from buccal mucosa and tongue dorsum. (D) Reference species for class II bacteriocins of samples from buccal mucosa and tongue dorsum.

Fig. S4 Producer of bacteriocins from BAGEL. (A) Producers of class I bacteriocins. (B) Producers of class II bacteriocins. (C) Producers of large bacteriocins.

Fig. S5 Distribution of putative bacteriolysins (A - C) and putative producer of these bacteriocins (D - F).

Fig. S6 A diagram of the analysis in the study. aa, Amino Acid; gff, Generic Feature Format; HMGI, HMP Gene Index; HMASM, HMP Illumina WGS Assemblies; HMRGD, HMP Reference Genomes Data; MCL, Markov Cluster Algorithm; nt, Nucleotide; nr, Non-redundant protein sequences.

587 **Table S1.** Class I bacteriocins predicted in this study.

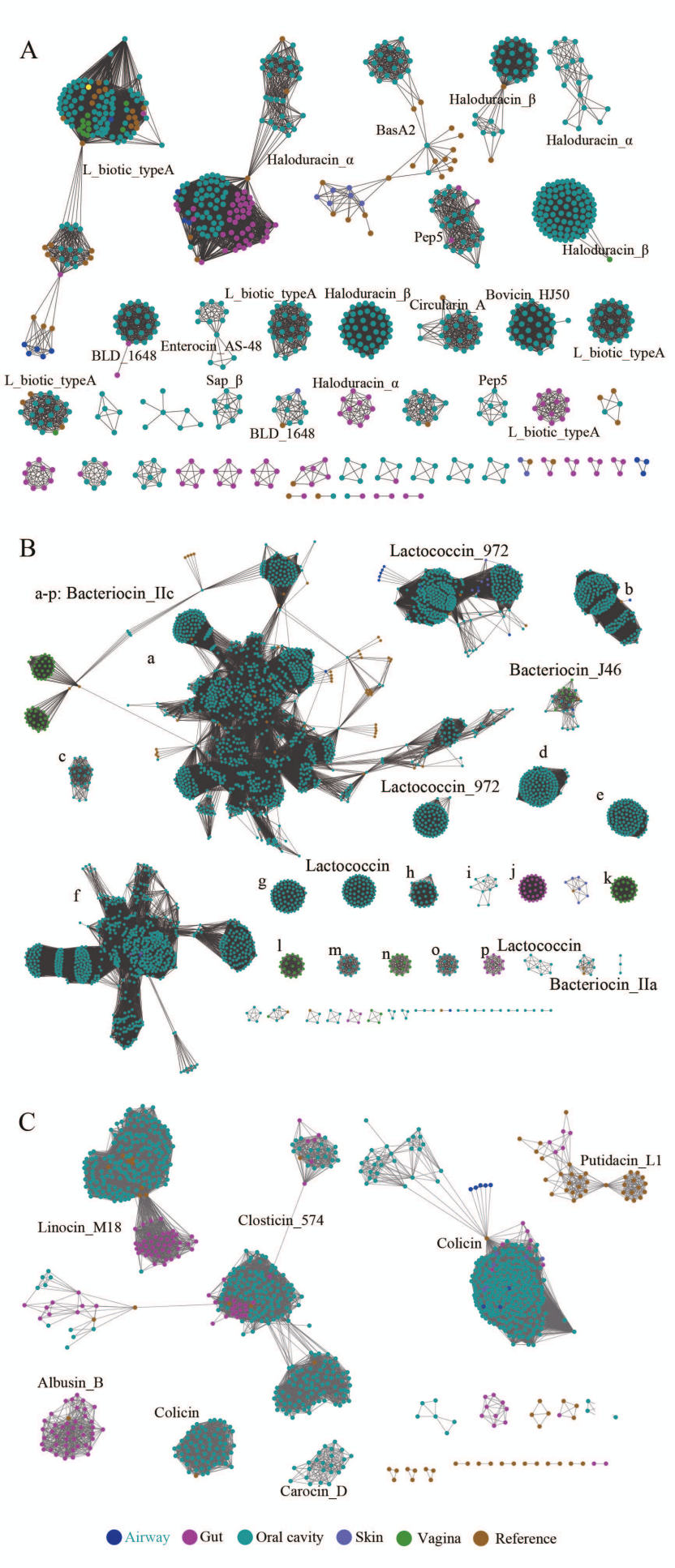
588 **Table S2.** Class II bacteriocins predicted in this study.

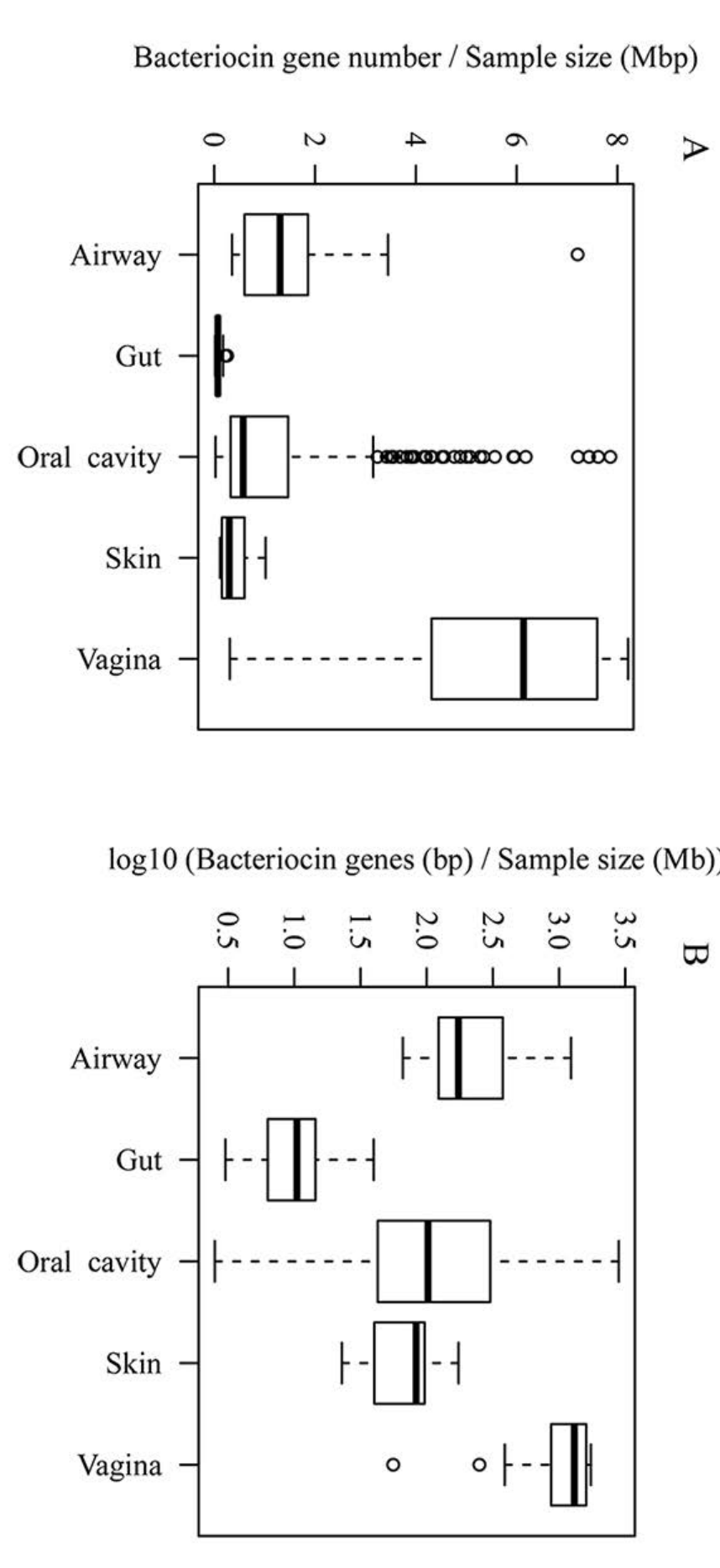
589 **Table S3.** Large bacteriocins predicted in this study.

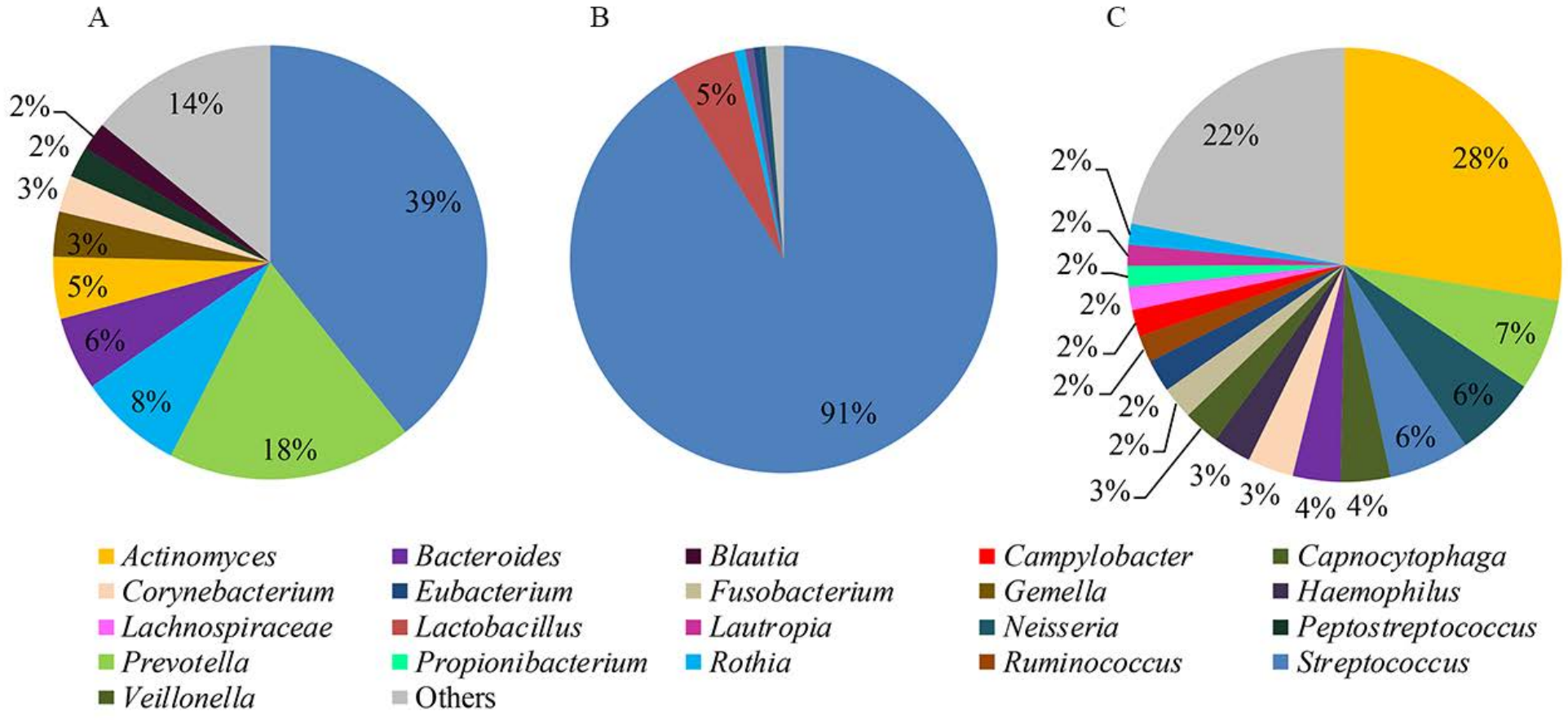
590 **Table S4.** Reference species for bacteriocins from sites other than oral cavity

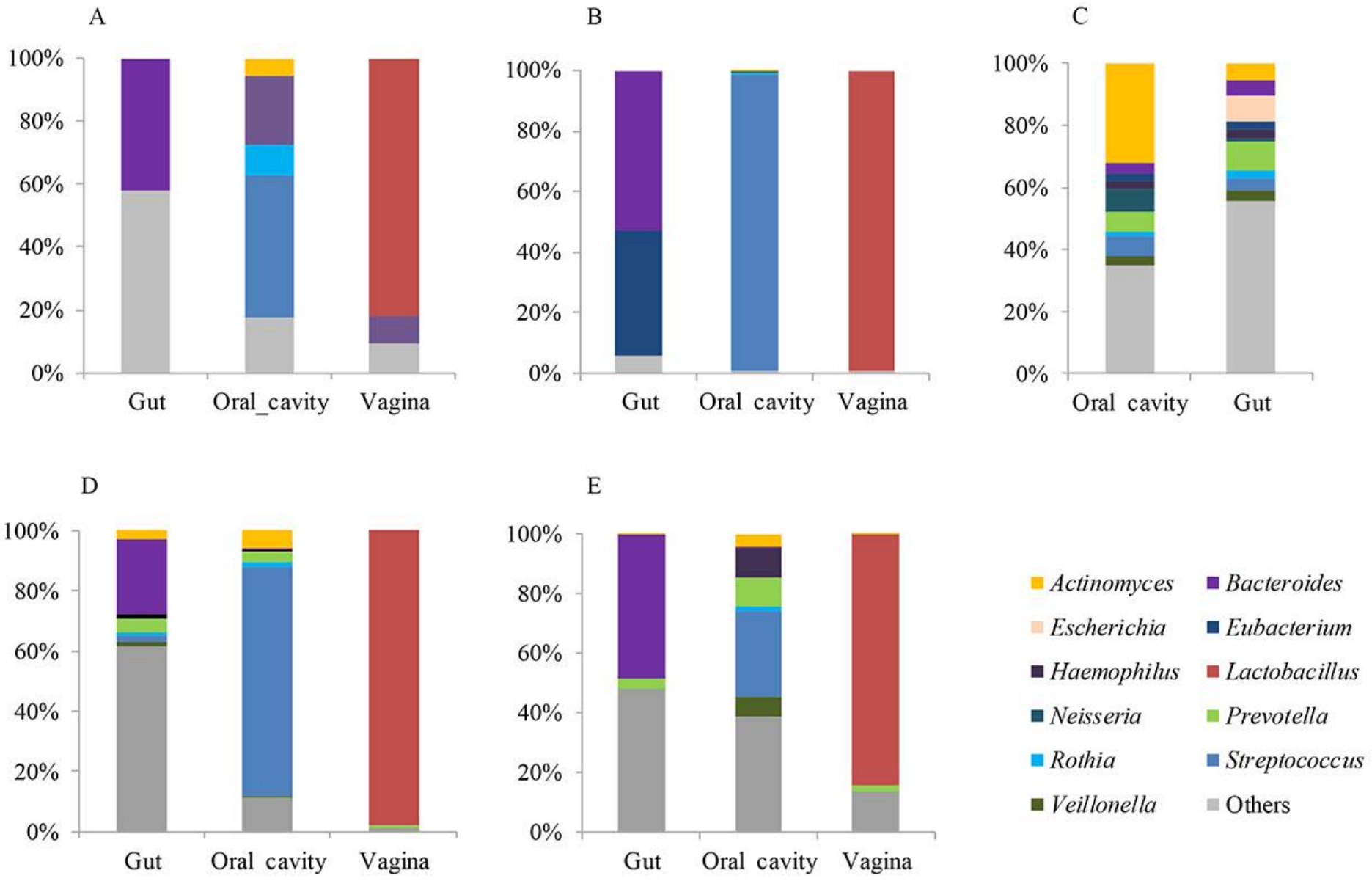
Table 1 Number and distribution of different bacteriocins among different human body sites

Protein family	Number of bacteriocin per site					
	Airway	**Gut**	**Oral cavity**	**Skin**	**Vagina**	**%Total**
			All bacteriocins			
	38	327	4321	34	155	
% Total	1%	7%	89%	1%	3%	**100%**
			Class I bacteriocins			
L_biotic_typeA	1	15	130	0	10	19%
Haloduracin_α	6	51	76	0	0	17%
Haloduracin_β	0	9	124	0	1	17%
Bovicin_HJ50	5	1	65	0	0	9%
Thuricin_CD	0	0	44	0	0	5%
BsaA2	0	0	31	6	0	5%
Geobacillin_I	0	7	26	0	0	4%
BLD_1648	0	4	24	0	0	3%
Circularin_A	0	0	23	0	0	3%
Acidocin_B	0	0	18	0	1	2%
Enterocin_AS-48	0	0	12	3	0	2%
Lacticin_3147_A2	0	0	15	0	0	2%
Pep5	0	4	10	0	0	2%
Subtilosin_A	0	1	9	1	0	1%
Others	0	31	38	0	0	9%
% Total	**2%**	**15%**	**80%**	**1%**	**2%**	**100%**
			Class II bacteriocins			
Bacteriocin_IIc	6	54	2294	4	88	80%
Lactococcin_972	9	1	389	7	0	13%
Lactococcin	0	0	85	0	0	3%
Bacteriocin_IIa	0	0	8	0	42	2%
Bacteriocin_J46	1	0	15	0	11	1%
Others	2	1	22	7	2	1%
%Total	**0**	**2%**	**92%**	**0**	**5%**	**100%**
			Large bacteriocins			
Closticin_574	8	7	286	6	0	30%
Colicin	0	37	201	0	0	23%
Linocin_M18	0	44	149	0	0	19%
Putidacin_L1	0	2	121	0	0	12%
Albusin_B	0	43	28	0	0	7%
Carocin_D	0	6	44	0	0	5%
Others	0	9	34	0	0	4%
%Total	**1%**	**14%**	**84%**	**1%**	**0**	**100%**


A
L_biotic_typeA
Haloduracin_α
BasA2
Haloduracin_β
Haloduracin_α
Pep5
Haloduracin_β
L_biotic_typeA
Haloduracin_β
Circularin_A
Bovicin_HJ50
L_biotic_typeA
BLD_1648
Enterocin_AS-48
L_biotic_typeA
Sap_β
Haloduracin_α
Pep5
BLD_1648
L_biotic_typeA
B
a-p: Bacteriocin_IIc
Lactococcin_972
b
a
Bacteriocin_J46
c
d
Lactococcin_972
e
Lactococcin
g
h
i
j
k
f
l
m
n
o
p
Lactococcin
Bacteriocin_IIa
C
Putidacin_L1
Linocin_M18
Closticin_574
Colicin
Albusin_B
Colicin
Carocin_D
Airway
Gut
Oral cavity
Skin
Vagina
Reference


Bacteriocin gene number / Sample size (Mbp)
0
2
4
6
8
A
Airway
Gut
Oral cavity
Skin
Vagina
log10 (Bacteriocin genes (bp) / Sample size (Mb))
0.5
1.0
1.5
2.0
2.5
3.0
3.5
B
Airway
Gut
Oral cavity
Skin
Vagina


A
39%
18%
8%
6%
5%
3%
3%
2%
2%
14%
B
91%
5%
C
28%
7%
6%
6%
4%
4%
3%
3%
3%
2%
2%
2%
2%
2%
2%
2%
2%
Actinomyces
Bacteroides
Blautia
Campylobacter
Capnocytophaga
Corynebacterium
Eubacterium
Fusobacterium
Gemella
Haemophilus
Lachnospiraceae
Lactobacillus
Lautropia
Neisseria
Peptostreptococcus
Prevotella
Propionibacterium
Rothia
Ruminococcus
Streptococcus
Veillonella
Others


A
100%
80%
60%
40%
20%
0%
Gut
Oral_cavity
Vagina
B
100%
80%
60%
40%
20%
0%
Gut
Oral cavity
Vagina
C
100%
80%
60%
40%
20%
0%
Oral cavity
Gut
D
100%
80%
60%
40%
20%
0%
Gut
Oral cavity
Vagina
E
100%
80%
60%
40%
20%
0%
Gut
Oral cavity
Vagina
Actinomyces
Bacteroides
Escherichia
Eubacterium
Haemophilus
Lactobacillus
Neisseria
Prevotella
Rothia
Streptococcus
Veillonella
Others