

1 SUPPLEMENTARY INFORMATION

2

3 Detailed methods

4 *DNA extraction, 16S rRNA gene amplification, and sequencing*

5 Microbial DNA was extracted from collected swabs using the
6 PureLink™ Microbiome DNA Kit (Invitrogen, Carlsbad, CA,
7 USA) which utilizes chemical, heat and bead-beating cell lysis
8 prior to purification. Extracted DNA samples were stored at -
9 20°C until amplification for sequencing. DNA was also extracted
10 from two swabs that were not used to sample cows; these
11 served as negative controls. Amplification of the V3-V4
12 hypervariable region of the 16S rRNA gene for sequencing was
13 conducted using Illumina_16S_341F and Illumina_16S_805R
14 universal primers with adapter sequences. The first step PCR
15 mastermix contained 1.25µl of amplicon PCR forward and
16 reverse primers (2.5µM) and 12.5µl of NEBNext High-Fidelity
17 2X PCR Master Mix (New England Biolabs, Ipswich, MA, USA).
18 Reaction conditions were 95°C initial denaturation for 3 min,
19 followed by 12 cycles of 95°C for 30 s, 62.3°C for 30 s, and 72°C
20 for 30 s, and a final extension at 72°C for 5 min. Negative
21 (Nuclease-Free Water (not DEPC-Treated) Thermo) and
22 positive controls (ZymoBIOMICS™ Microbial Community DNA
23 Standard) were also amplified. PCR products were cleaned with
24 Agencourt AMPure XP beads (Beckman Coulter Genomics,
25 Fullerton, CA, USA) following the manufacturer's protocol. In a
26 second PCR step, dual indexes and Illumina sequencing
27 adapters were attached using 7.5µl of amplicon PCR product
28 DNA, 2.5µl of Illumina Nextera XT Index Primer 1 (N7xx), 2.5µl

29 of Nextera XT Index Primer 2 (S5xx), and 12.5µl of NEBNext
30 High-Fidelity 2X PCR Master Mix, with thermocycling at 95°C
31 for 3 min, followed by 13 cycles of 95°C for 30 s, 55°C for 30 s,
32 and 72°C for 30 s, and a final extension at 72°C for 5 min. The
33 final PCR products were cleaned with Agencourt AMPure XP
34 beads and standardised to the same concentration, purified
35 once more using Agencourt AMPure XP beads, and eluted in
36 30µl to increase the final DNA concentration. Concentration and
37 quality of the PCR amplicons was evaluated using the Qubit™
38 dsDNA HS Assay Kit (Thermo Fisher Scientific, Fair Lawn, NJ,
39 USA), and a fragment analyser (Agilent, Agilent Technologies
40 Inc., Santa Clara, CA, USA). Amplicons were sequenced using
41 the Illumina® HiSeq 2500 platform (Illumina, San Diego, CA,
42 USA) to generate 2 x 300 bp paired-end reads. 15% PhiX
43 fragment library was added to increase sample diversity.

44 *Shotgun metagenomic analysis*

45 To maximise the chances of achieving sufficient sequencing
46 depth, cows were selected at random from those whose
47 previous 16SrRNA samples had a DNA content of >5ng/µl after
48 the initial DNA extraction, as measured using the Qubit™
49 dsDNA HS Assay Kit. Microbial DNA was extracted from a
50 second set of swabs that had been collected parallel to those
51 used in the marker gene analysis. The DNA extraction method
52 was the same, using the PureLink™ Microbiome DNA Kit
53 (Invitrogen, Carlsbad, CA, USA) according to the
54 manufacturer's instructions. Agarose gel electrophoresis was
55 carried out using SYBR green as the nucleic acid stain (Thermo

56 Fisher Scientific Fair Lawn, NJ, USA) to ensure presence of
57 clear DNA bands. Library preparation was carried out on gDNA
58 samples using the Nextera XT kit (Illumina). gDNA input was
59 quantified using Qubit™ to ensure 1ng of each sample was
60 submitted for tagmentation. Whole samples were used in a
61 limited amplification step (12 cycles) which incorporates the
62 indexes at the priming step. Libraries were cleaned using 0.6x
63 AMPure beads. After elution, the quantity of the library was
64 checked by Qubit™ assay and the quality checked on a DNA
65 high sensitivity Bioanalyser chip. The final libraries were pooled
66 in equimolar ratio and the quantity and quality of the final pool
67 was assessed by Qubit™ and Bioanalyser (Agilent 2100
68 Bioanalyser, Agilent Technologies Inc, Santa Clara, CA, USA),
69 and subsequently by qPCR using the Illumina Library
70 Quantification Kit from KAPA (KK4854) on a Roche Light Cycler
71 LC480II according to manufacturer's instructions. The final
72 loading concentration of 300 pM was reached by adding 35µl
73 exclusion amplification enzyme mix. The libraries were
74 sequenced on an Illumina HiSeq 4000 platform using
75 sequencing by synthesis (SBS) technology to generate 2 x 150
76 bp paired-end reads.

77 *Quality control and filtering of shotgun metagenomic sequences*

78 Data files were demultiplexed and converted to FASTQ format
79 using Casava v.1.8.2 (Illumina). FASTQ files were trimmed
80 using option _O3 Cutadapt version 1.2.1 (1) to exclude those
81 matching Illumina adaptor sequences by ≥ 3 bp at the 3' end.
82 The reads were further trimmed to remove low quality bases,

83 using Sickle version 1.200 with a minimum window quality score
84 of 20. After trimming, reads shorter than 20 bp were removed,
85 and single reads were excluded as length distributions showed
86 they were of poor quality. Host reads were removed following
87 alignment against the host *Bos Taurus* genome using Bowtie2
88 v2.2.6 (2): read pairs where one or both reads aligned were
89 removed. The remaining reads in pairs were merged using
90 PEAR v0.9.11 (3) to form a single long read based on
91 overlapping homology. Those that could not be merged in this
92 way were concatenated with an intervening N-base. The
93 resulting sequences underwent taxonomic assignment using
94 Kraken v0.10.6 (4) and results were filtered using a confidence
95 threshold of 0.1. Results were analysed using Linear
96 discriminant analysis effect size (LefSe) (5) to determine the
97 taxa most likely to explain differences between the two classes
98 HtHt and HtIn. The HUMAnN2 search strategy (6) was used to
99 functionally annotate read data and abstracts to show biological
100 pathway abundance and completeness. Finally, reads that did
101 not align to their pangenomes using this strategy were
102 submitted to a protein database (UniRef) for translated
103 searching (6). The gene families identified were further analysed
104 using the MetaCyc database to reconstruct and quantify
105 complete metabolic pathways (7).

106 Samples were renormalized for sequence depth and terms
107 converted using GO slim (EMBL-EBI, Cambridgeshire, UK)
108 where possible to generate heatmaps representing gene

109 functions categorised as biological processes, molecular
110 functions, and cellular components.

111 *Genome Wide Association and regional heritability mapping*
112 *study of foot skin microbiota related traits*

113 Animal sampling and genotyping are described by Sánchez-
114 Molano et al. (8). The phenotypic traits (10 traits) analysed here
115 were three different alpha diversity indices; Chao1, Shannon,
116 Simpson indices, and relative abundances of seven genera;
117 *Porphyromonas* spp., Clostridiales Family XI, *Fastidiosipila*
118 spp., *Peptoclostridium* spp., *Macrococcus* spp., *Treponema*
119 spp., and genera of the family Bacteroidetes.

120 A genotype call rate threshold of 95%, minor allele frequency
121 (MAF<0.01), Hardy-Weinberg equilibrium (threshold of 1.45E-6
122 Bonferroni corrected), sample call rate threshold of 95% were
123 applied. 236 out of 242 cows with foot microbiota records were
124 eventually analysed because 6 cows were lost in QC steps.
125 Wald tests using ASReml software package (9) were used to
126 determine statistically significant (P -value= 0.05) fixed effects to
127 be included in the analysis as described in Sánchez-Molano et
128 al. (2019). After performing this analysis for all foot microbiota
129 phenotypic traits, concordant models were chosen including
130 farm, parity and season as fixed effects.

131 The Genomic relationship matrix (GRM) was computed using
132 GEMMA (10) and principal components analysis (PCA) was
133 used to find out any genetic structure of the cow population. This
134 population structure was accounted for in GWA models by

135 automatically fitting the GRM as part of the polygenic effect,
 136 whereas in RHM analysis the first 7 PCs were fitted to account
 137 for this structure (RHM analyses failed to converge when the
 138 GRM was fitted); further correction for the inflation factor (λ) was
 139 applied as described by Amin et al. (11). REACTA (12) was first
 140 used to assess the full genomic variance for each trait with a
 141 general explanatory analysis. GWA was performed using
 142 GEMMA (10) and the following linear mixed model:

$$143 \quad \mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{x}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\varepsilon}$$

144 where \mathbf{y} represents the vector of foot microbiota phenotypes, \mathbf{W}
 145 is the incidence matrix for $\boldsymbol{\alpha}$, which is the vector of associated
 146 fixed effects, \mathbf{x} is the vector of genotypes which were coded as
 147 0/1/2, $\boldsymbol{\beta}$ is phenotype's regression on genotypes, and \mathbf{Z} is the
 148 design matrix for the vector \mathbf{u} of random polygenic effects
 149 (distributed as $MVN(0, V_g\mathbf{G})$, with \mathbf{G} being the GRM matrix and
 150 V_g being the genetic variance), and $\boldsymbol{\varepsilon}$ represents the vector of
 151 residual errors (distributed as $MVN(0, V_e\mathbf{I})$, with \mathbf{I} being the
 152 identity matrix and V_e being the residual variance). Further λ
 153 correction was applied to lower any potential inflation.

154 The RHM was performed using the following model:

$$155 \quad \mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{X}\mathbf{u}_{(i)} + \boldsymbol{\varepsilon}$$

156 where \mathbf{y} corresponds to the vector of phenotypes, \mathbf{W}
 157 corresponds to the design matrix for the fixed effects $\boldsymbol{\alpha}$
 158 (including the principal components), \mathbf{X} is the corresponding
 159 design matrices for the effects $\mathbf{u}_{(i)}$ of the corresponding
 160 region i (distributed as $MVN(0, V_{g(i)}\mathbf{G}_{(i)})$, with $V_{g(i)}$ and $\mathbf{G}_{(i)}$ being

161 the genomic variance and the GRM corresponding to the SNPs
162 in the i th region, respectively) and ϵ being the error.

163 The significance of the region effect was assessed using the
164 likelihood ratio test statistic. A total of 1733 regions were
165 analysed, leading to a genome-wide significant threshold
166 ($P = 0.05$) defined at $P = 2.89E-5$ with Bonferroni correction for
167 multiple regions ($-\log_{10}(P) = 4.54$) and a suggestive threshold
168 (one false positive per genome scan) defined at $P = 5.77E-4$ ($-\log_{10}(P) = 3.24$).

170 GWA and RHM results were compared to determine common
171 significant/ suggestive regions, and the proportion of variance
172 explained by each region was worked out as a percentage of
173 the total genomic variance.

174

175 References

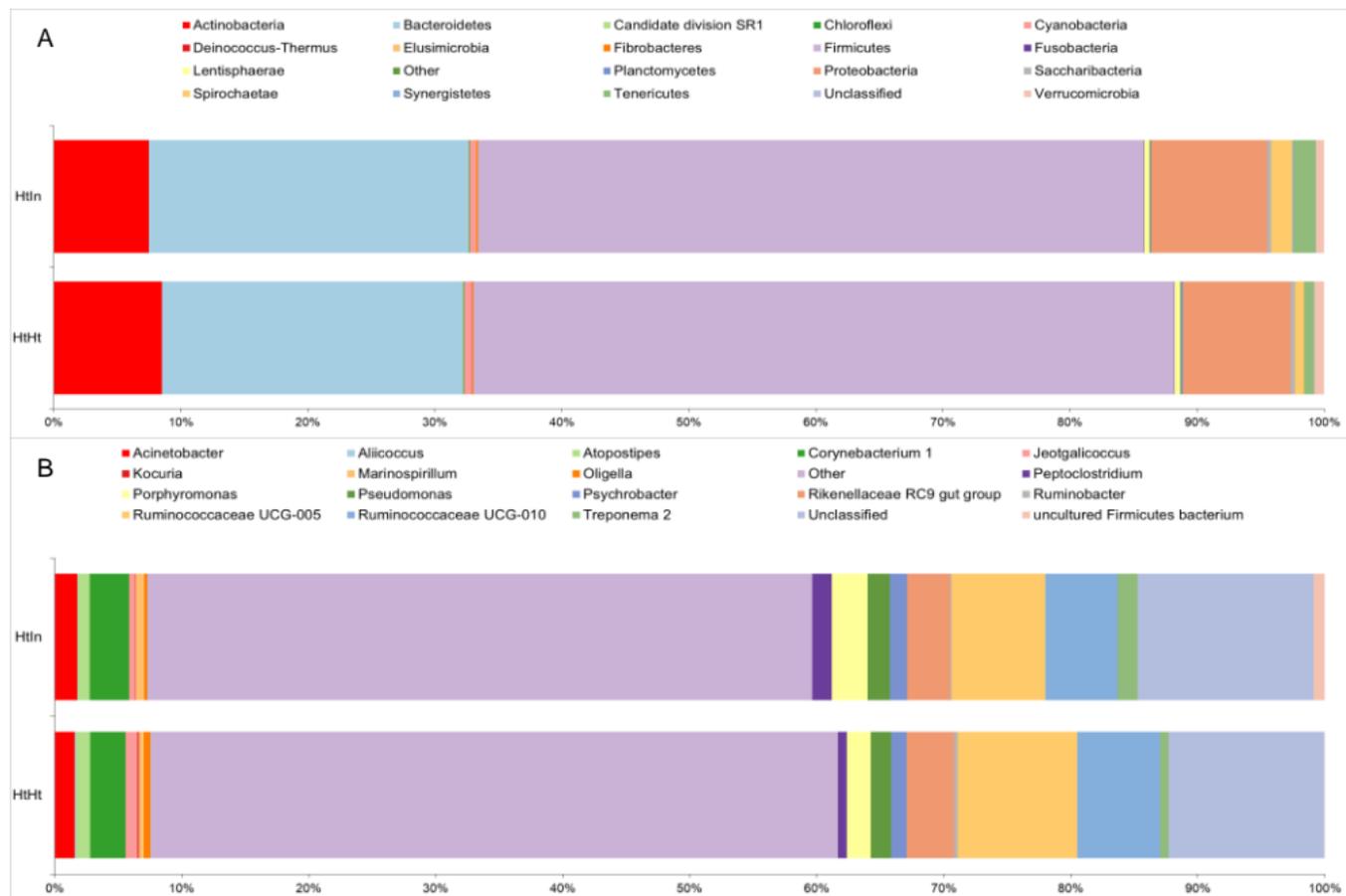
- 176 1. Marcel Martin. Cutadapt removes adapter sequences
177 from high-throughput sequencing reads. EMBnet. 2011;
178 17:5–7.
- 179 2. Langmead B, Salzberg SL. Fast gapped-read alignment
180 with Bowtie 2. Nat Methods. 2012;9(4):357–9.
- 181 3. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: A fast
182 and accurate Illumina Paired-End reAd mergeR.
183 Bioinformatics. 2014;30(5):614–20.
- 184 4. Wood DE, Salzberg SL. Kraken: Ultrafast metagenomic
185 sequence classification using exact alignments.
186 Genome Biol. 2014;15(3):R46.
- 187 5. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L,
188 Garrett WS, et al. Metagenomic biomarker discovery
189 and explanation. Genome Biol. 2011:R60.
- 190 6. Franzosa EA, McIver LJ, Rahnava G, Thompson LR,
191 Schirmer M, Weingart G, et al. Species-level functional
192 profiling of metagenomes and metatranscriptomes. Nat
193 Methods. 2018;15:962–8.
- 194 7. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA,
195 Keseler IM, et al. The MetaCyc database of metabolic

- 196 pathways and enzymes and the BioCyc collection of
197 pathway/genome databases. *Nucleic Acids Res.*
198 2015;44(2):471–80.
- 199 8. Sánchez-Molano E, Bay V, Smith RF, Oikonomou G,
200 Banos G. Quantitative Trait Loci Mapping for Lameness
201 Associated Phenotypes in Holstein-Friesian Dairy
202 Cattle. *Front Genet.* 2019;10:926.
- 203 9. Gilmour a R, Gogel BJ, Cullis BR, Thompson R.
204 ASReml user guide release 3.0. VSN International Ltd.
205 2009. p. 275.
- 206 10. Zhou X, Stephens M. Genome-wide efficient mixed-
207 model analysis for association studies. *Nat Genet.*
208 2012;44(7):821–4.
- 209 11. Amin N, van Duijn CM, Aulchenko YS. A genomic
210 background based method for association analysis in
211 related individuals. *PLoS One.* 2007;2(12):e1274.
- 212 12. Cebamanos L, Gray A, Stewart I, Tenesa A. Regional
213 heritability advanced complex trait analysis for GPU and
214 traditional parallel architectures. *Bioinformatics.*
215 2014;30(8):1177–9.
- 216

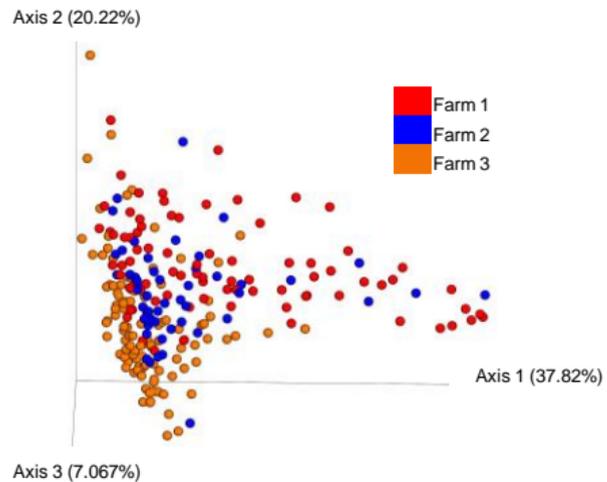
Supplementary Table 1. Classification of cows into groups according to their foot health group and farm of origin. (HtHt: The cows which remained healthy during the study, HtIn: The cows which were healthy at sampling, then developed DD, InIn: The cows which had DD in all checkpoints, InHt: The cows which had DD at initial check point then recovered, Unknown: The cows which could not be followed adequately, died or were sold during the study.)

<i>Farm</i>	<i>HtHt</i>	<i>HtIn</i>	<i>InIn</i>	<i>InHt</i>	<i>Unknown</i>
1	40	16	14	12	1
2	13	18	14	1	5
3	59	14	30	4	1
<i>Total</i>	112	48	58	17	7

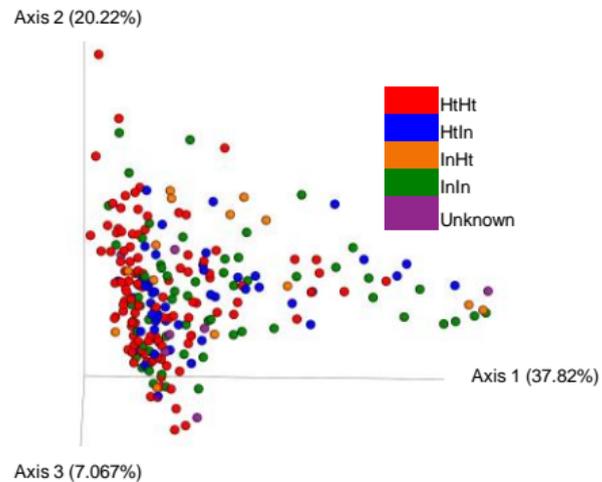
Supplementary Figure 1 Taxonomy displayed by A. Twenty most abundant phyla identified in HtIn compared to HtHt samples; and B. Twenty most abundant genera identified in HtHt compared to HtIn samples (HtHt: The cows which remained healthy during the study, HtIn: The cows which were healthy at sampling, then developed DD)



A

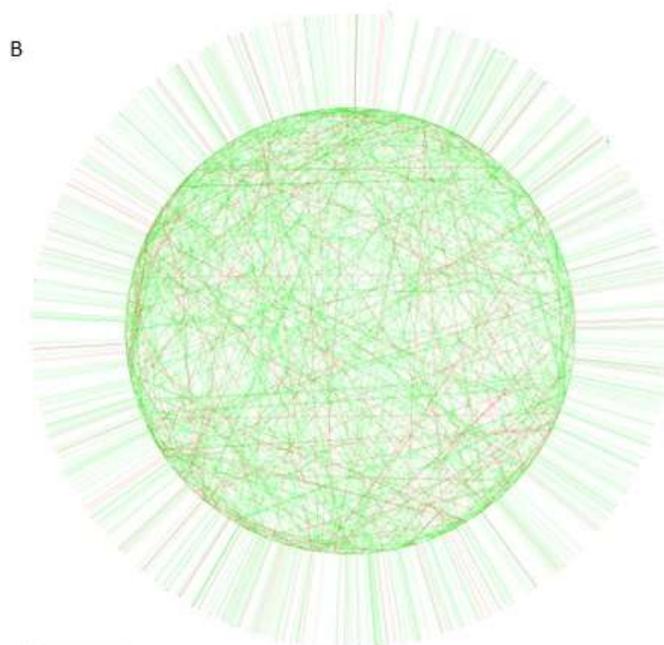
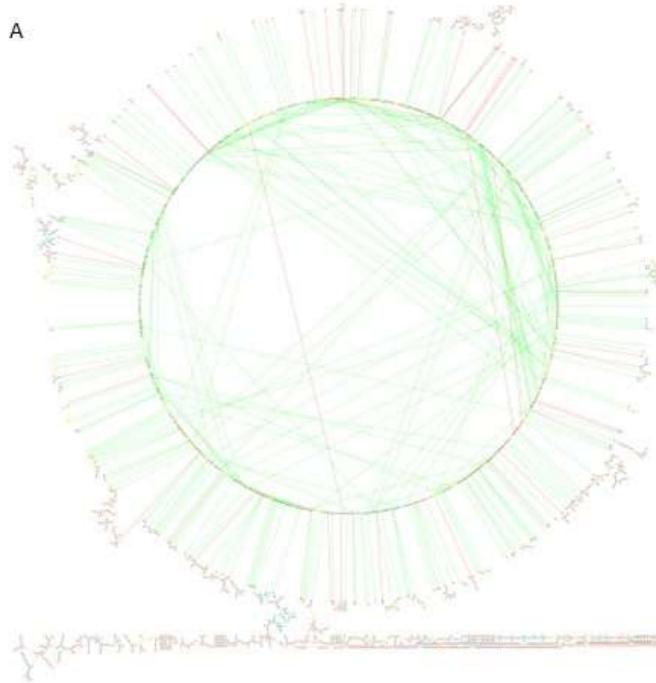
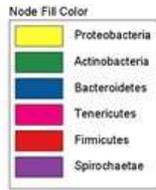


B



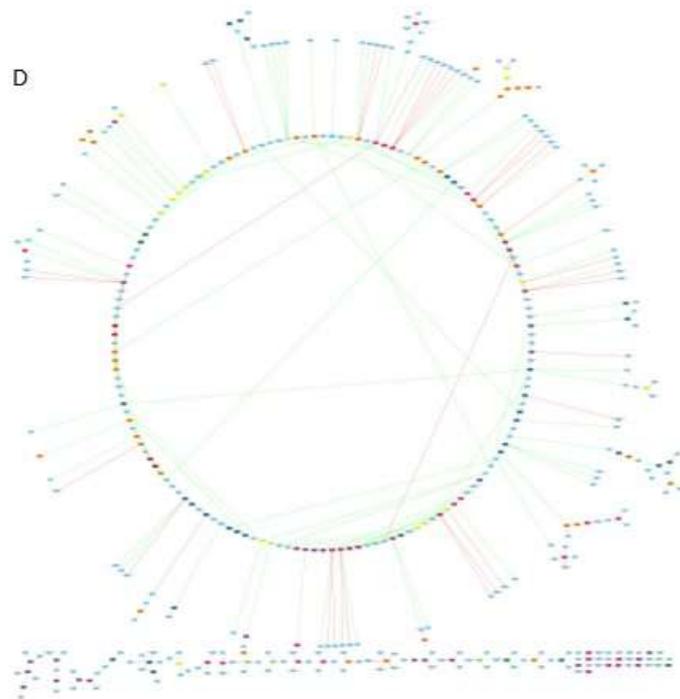
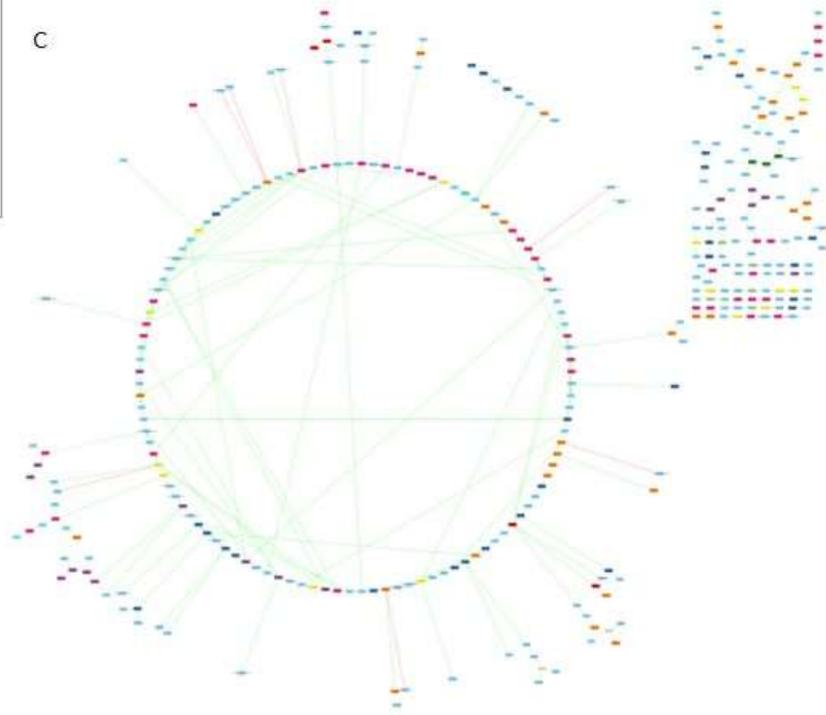
Supplementary Figure 2. Weighted unifrac distances showing beta diversity A. by farm, and B. by foot health group.

Supplementary Figure 3. Bacterial networks showing (A) HtHt samples including the six phyla contributing more than 1% of OTUs (B) HtIn samples including the six phyla contributing more than 1% of OTUs (C) Ten genera in HtHt samples which differ in relative abundance between HtHt and HtIn groups with adjacent nodes shown in light blue and (D) Ten genera in HtIn samples which differ in relative abundance between HtHt and HtIn groups with adjacent nodes shown in light blue.



Node Fill Color

Brachybacterium
Porphyromonas
Peptoclostridium
Acholeplasma
Prevotella
Fastidiosipila
Macrococcus
Succinidlasticum
Anaerococcus
Treponema



Supplementary Table 2. Simple Parameters from Network Analyses for HtHt and HtIn groups carried out using the six phyla that accounted for more than 1% of total OTUs.

Sample Group	HtHt n=112	HtIn n=48	Definition
Parameter			
<i>Number of nodes</i>	2339	2802	Nodes represent OTUs
<i>Number of edges</i>	2311	3564	Edges represent correlations (positive or negative) between Nodes (OTUs)
<i>Isolated nodes</i>	18	18	Nodes (OTUs) that are not correlated with any others in the network
<i>Connected components</i>	244	66	Lower = stronger connectivity
<i>Network diameter</i>	51	28	Largest distance between two nodes
<i>Network centralisation</i>	0.003	0.002	How central the most central node is compared to how central all the other nodes are.
<i>Network heterogeneity</i>	0.561	0.503	Reflects the tendency of the network to contain hub nodes: 1 would mean uniformity
<i>Characteristic path length</i>	17.228	9.821	Shorter = stronger connectivity
<i>Average number of neighbours</i>	1.976	2.559	More = stronger connectivity
<i>Clustering coefficient</i>	0.024	0.008	Nodes with <2 neighbours are assumed to have a clustering coefficient of 0
<i>Network density</i>	0.001	0.001	Solely isolated nodes would score 0, cliques would score 1

Supplementary Figure 4. Quantity of functional pathways identified in samples using the Gene Ontology database. A. biological processes, B. cellular components and C. molecular function

