Supplemental information.

Gut symbionts from distinct hosts exhibit genotoxic activity via divergent colibactin biosynthetic pathways

Philipp Engel, Maria I. Vizcaino, and Jason M. Crawford

Content

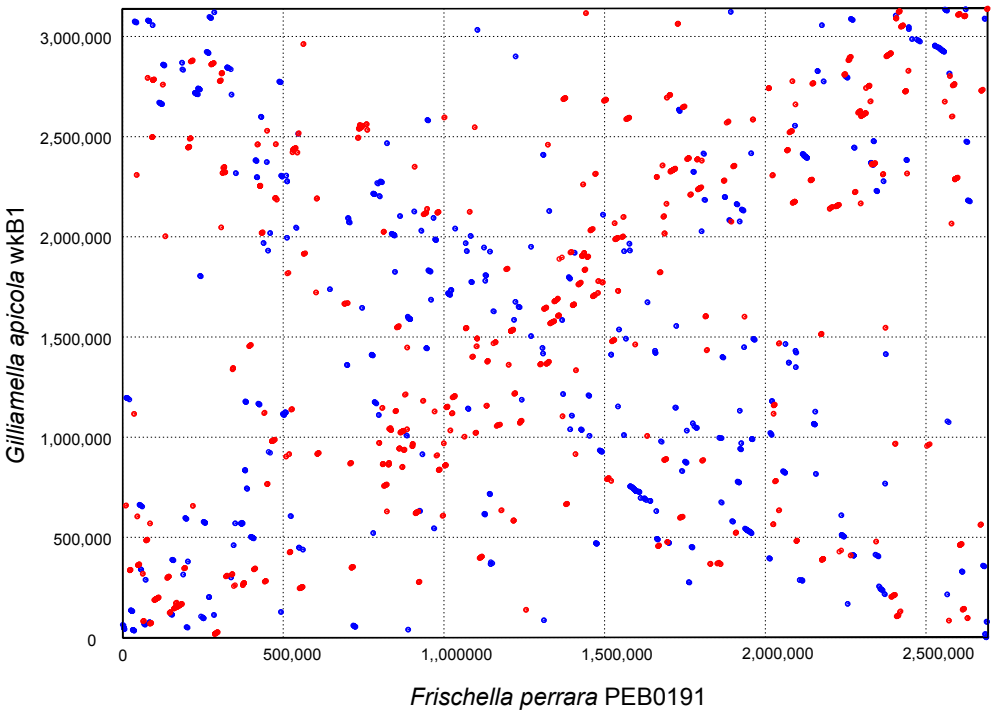


Figure S1. Dotplot analysis of the genomes of *G. apicola* wkB1 and *F. perrara* PEB0191. Plots were generated with Promer, a command of the software package MUMmer v3.23 (1). Promer generates local alignments based on amino acid sequences and plots aligned regions onto the x-axis and y-axis representing the genome positions of *F. perrara* PEB0191 and *G. apicola* wkB1, respectively. Default parameters were used. Red and blue dots represent alignments on the same strand and on the opposite strand, respectively.

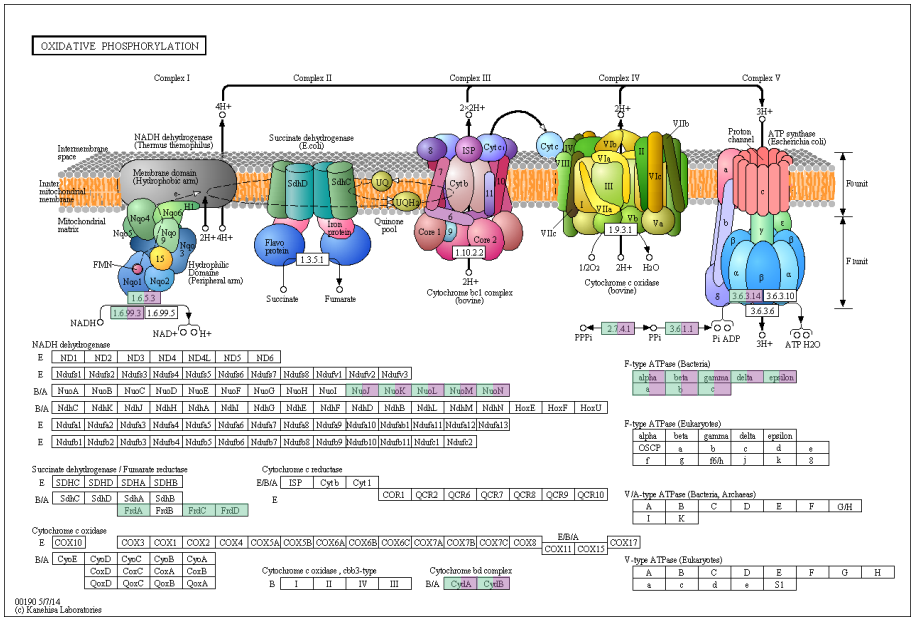


Figure S2. Oxidative phosphorylation chains in the genomes of the two honey bee gut symbionts *F. perrara* and *G. apicola* wkB1. Gene functions identified in the genomes of *F. perrara* and *G. apicola* wkB1 are highlighted in magenta and green, respectively. Other gene functions are either absent or could not be identified.

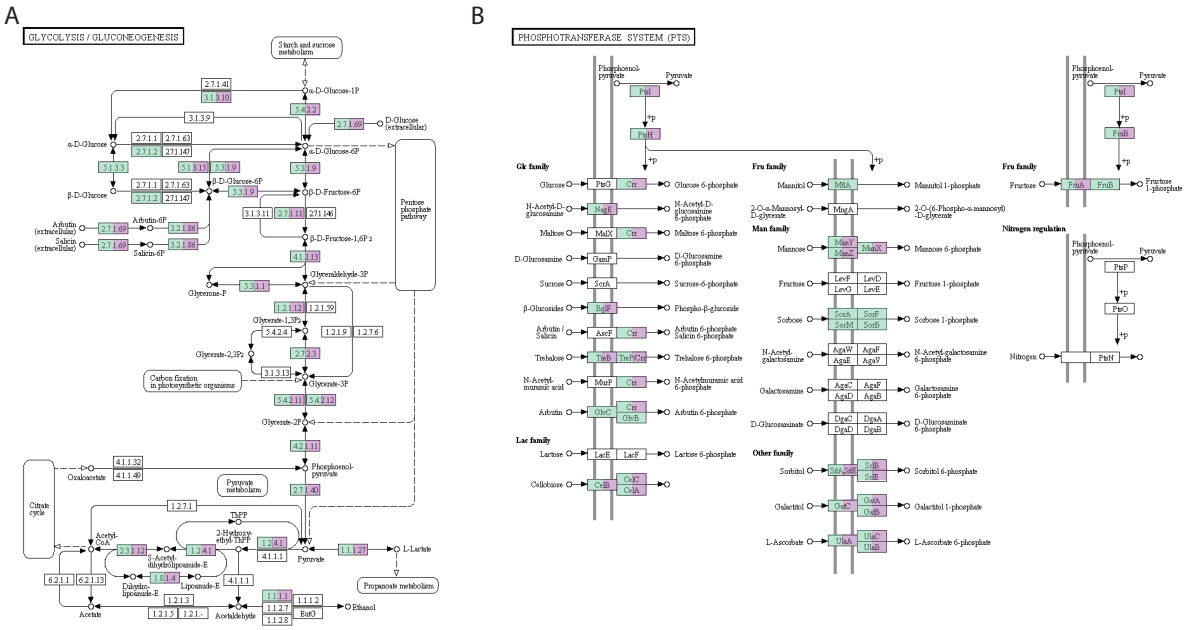


Figure S3. Glycolysis (A) and phosphotransferase system (B) genes in the genomes of the two honey bee gut symbionts *F. perrara* and *G. apicola* wkB1. Gene functions identified in the genomes of *F. perrara* and *G. apicola* wkB1 are highlighted in magenta and green, respectively. Other gene functions are either absent or could not be identified.

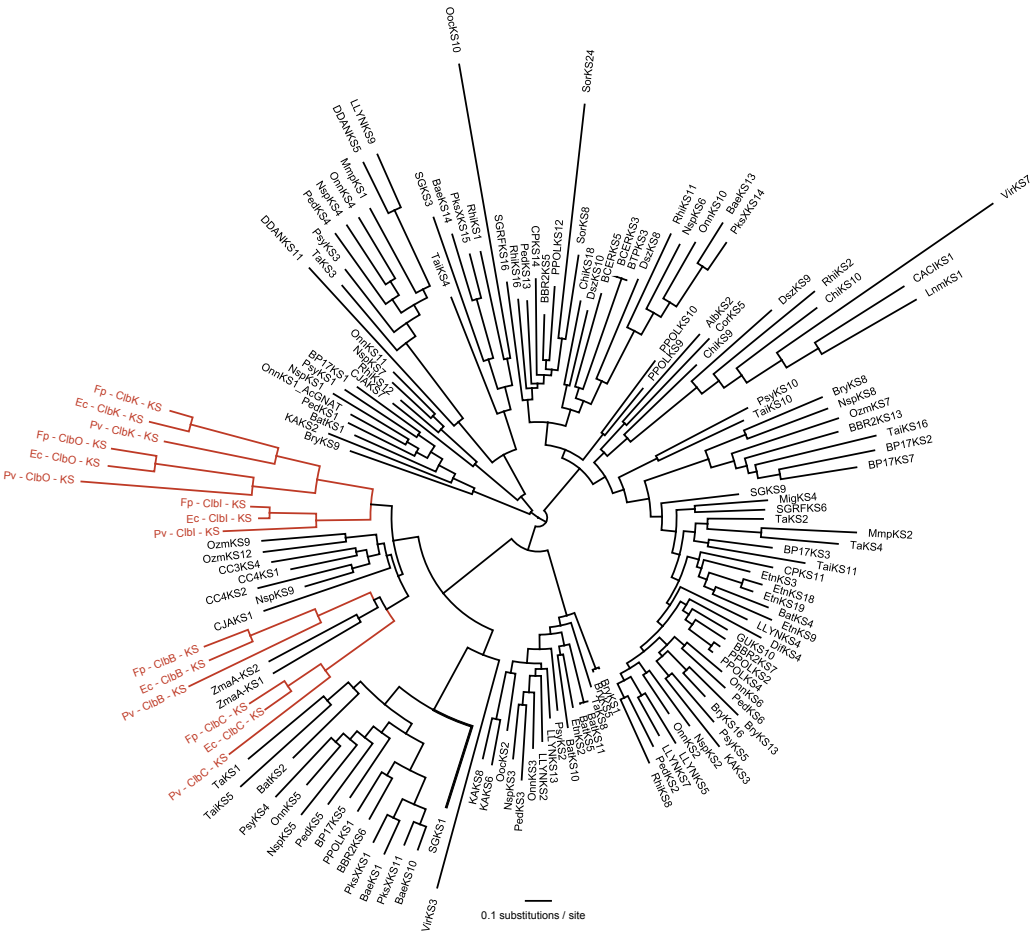


Figure S4. Maximum likelihood tree of ketosynthase (KS) domains of 154 polyketide synthases (PKS) proteins. Amino acid sequences of KS domains were obtained from a previous study (2) and aligned together with the KS domains of ClbC, ClbK, and ClbO (*trans*-AT PKS), and ClbB and ClbI (*cis*-AT PKS) of *E. coli* IHE3034 (Ec), *F. perrara* (Fp), and *Pseudovibrio* FO-BEG1 (Pv). The phylogenetic tree was obtained with PhyML (3) using default parameters. KS domains of homologous Clb PKS proteins, highlighted in red, cluster together. The next most closely related KS domains are found in the following *cis*-AT PKS: ZmaA of *Bacillus cereus* UW85 involved in the synthesis of zwittermicin (4), OzmK of *Streptomyces albus* JA3453 involved in the synthesis of oxazolomycin (5), NspD of *Nostoc* sp. 'Peltigera membranacea cyanobiont' involved in the synthesis of

48 nosperin (2), VirA of *Streptomyces virginae* involved in the synthesis of
49 virginiamycin A (6), several PKSs of *Clostridium cellulolyticum* H10 (CC), and one
50 PKS of *Cellvibrio japonicus* Ueda107 (CJA).
51

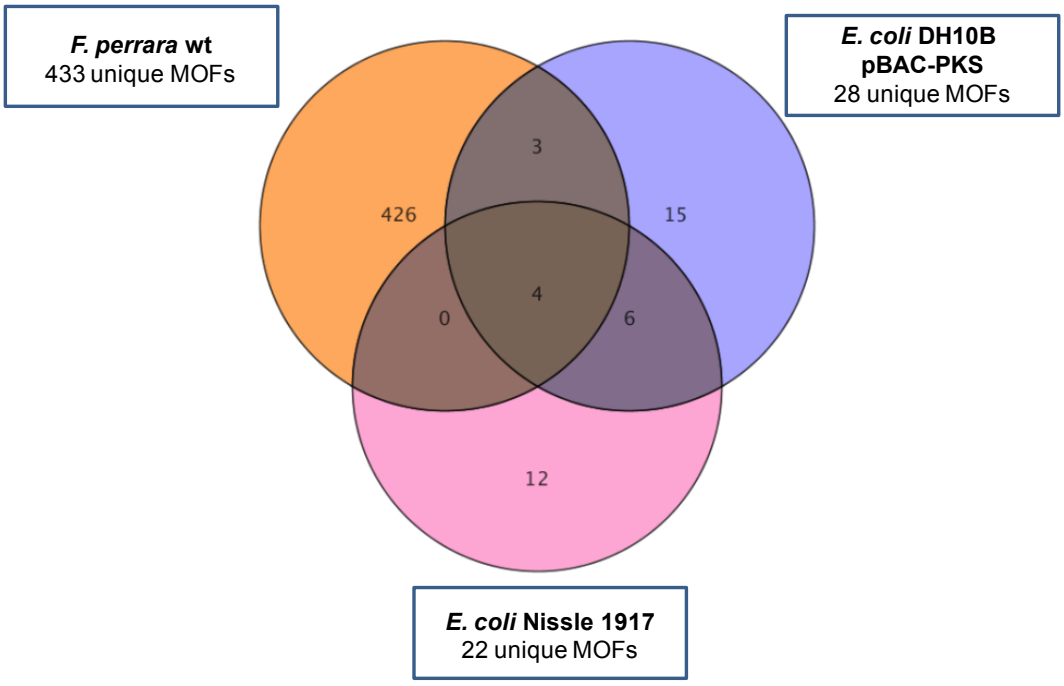
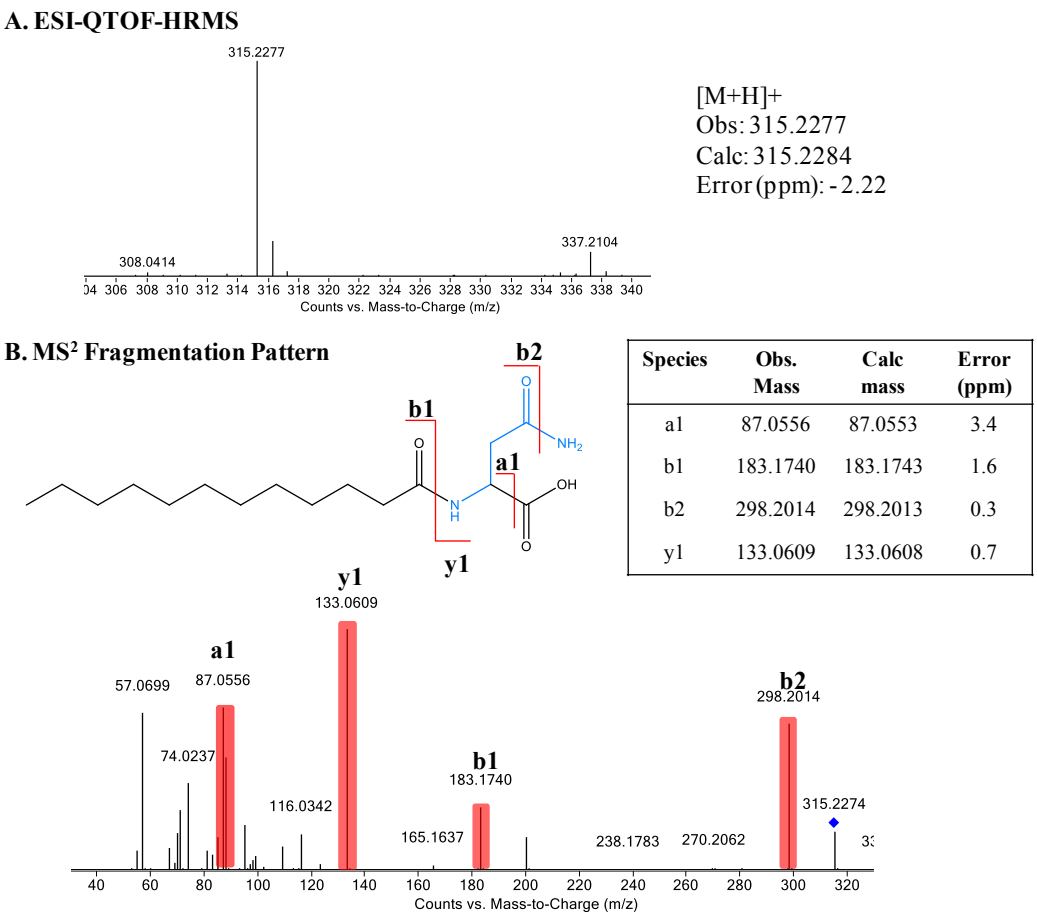


Figure S5. Venn Diagram comparison of the unique molecular features (MOFs) found in wild-type colibactin-bearing bacteria. The orange circle shows all MOFs of *F. perrara* PEB0191 (wt), from which MOFs found in the media controls were removed. The purple circle shows all MOFs of *E. coli* DH10B pBAC-PKS, from which MOFs found in *E. coli* DH10B pBAC-control were removed. The magenta circle shows all MOFs of *E. coli* Nissle 1917, from which MOFs found in *E. coli* Nissle 1917 *Δclb* (complete deletion of the genomic island) were removed. There are seven metabolites of *F. perrara*, shared among the different strains, four shared among all three (see Table S4).



Species	Obs. Mass	Calc mass	Error (ppm)
a1	87.0556	87.0553	3.4
b1	183.1740	183.1743	1.6
b2	298.2014	298.2013	0.3
y1	133.0609	133.0608	0.7

Figure S6. (A) ESI-QTOF-HRMS analysis and (B) MS² fragmentation pattern of *F. perrara* metabolite **1**. Fragmented species shown in the table are highlighted in the proposed structure of **1** and the corresponding MS² spectra. Observed mass (Obs mass) and calculated mass (Calc. mass) are shown for both MS and MS² ions, and all ppm errors were calculated to be less than 5.0.

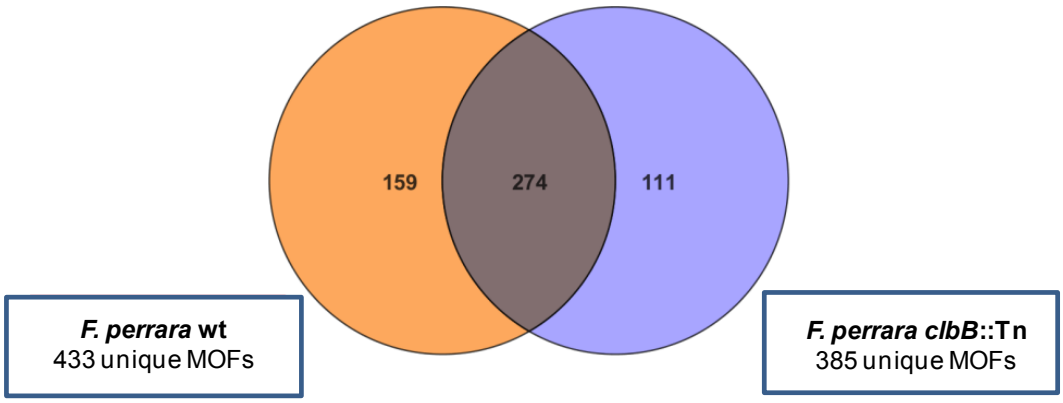


Figure S7. Venn diagram comparison of the unique molecular features (MOFs) found in *F. perrara* wt (PEB0191) and *F. perrara clbB*::Tn. MOFs found in the media control were removed from both samples.

Table S1. *F. perrara*-specific genes are found in the xls document accompanying this publication.

Table S2. Adenylation domain specificity predicted by three bioinformatic tools.

Adenylation domain[a]	AntiSmash (Stachelhaus code)	PKS/NRPS Analysis Web-site	NRPSpredictor2[b]
ClbB_Fp	valine	Pps3-M1-Glu	valine (70%)
ClbB_Pv	valine	no hit	valine (90%)
ClbB_Ec	valine	Pps3-M2-Val	valine (80%)
ClbH_Fp	serine	PhsB-M1-Ser	serine (100%)
	beta-hydroxy-tyrosine	no hit	valine (60%)
ClbG/H_Pv	serine	PhsB-M1-Ser	serine (100%)
	beta-hydroxy-tyrosine	no hit	valine (60%)
ClbH_Ec	serine	PhsB-M1-Ser	serine (100%)
	beta-hydroxy-tyrosine	no hit	valine (60%)
ClbJ_Fp	glycine	no hit	glycine (100%)
	cysteine	PchE-M1-Cys	cysteine (100%)
ClbJ_Pv	glycine	PchE-M1-Cys	glycine (100%)
	cysteine	NosC-M2-Gly	cysteine (100%)
ClbJ_Ec	glycine	no hit	glycine (100%)
	cysteine	MtaD-M1-Cys	cysteine (100%)
ClbK_Fp	cysteine	MtaD-M1-Cys	cysteine (100%)
ClbK_Pv	cysteine	PchE-M1-Cys	cysteine (100%)
ClbK_Ec	cysteine	MtaD-M1-Cys	cysteine (100%)
ClbN_Fp	asparagine	Cda2-M3-Asn	asparagine (100%)
ClbN_Pv	asparagine	Cda2-M3-Asn	asparagine (100%)
ClbN_Ec	asparagine	Cda2-M3-Asn	asparagine (100%)

[a]Fp, *F. perrara*, Pv, *Pseudovibrio* sp., Ec, *E. coli* IHE3934

[b]Percentages indicate confidence scores

Table S3. Secondary structure analysis with Phyre2 (8) identifies inactivated acyltransferase (AT) domains encoded in PKS genes *clbC, clbK,* and *clbO. ClbG* is a conserved trans-AT PKS, which served as a positive control. For each query, only the three best hits are shown.

***F. perrara* PEB0191**	**Rank**	**Hit**	**Confidence %**	**[a]Coverage aa**	**[b]Gaps aa**	**[c]Motifs (query / template)**
ClbC	1	Transferase of PKS: c3tzzA	100%	298	149	GEGLG / GQ**S**LG
435 aa -786 aa	2	Transferase of PKS: c2qo3A	100%	308	140	GEGLG / GH**S**QG
	3	Transferase of PKS: c2hg4A	100%	304	140	GEGLG / GH**S**QG
ClbK	1	Hydrolase of PKS: c4oqjA	100%	276	6	GQGDG / ADRTE
421 aa - 724 aa	2	Transferase of PKS: c4mz0B	100%	274	125	GQGDG / GH**S**VG
	3	Transferase of PKS: c2hg4A	100%	275	184	GQGDG / GH**S**QG
ClbO	1	Transferase of PKS: c3tzzA	100%	289	155	GNQAG / GQ**S**LG
427 aa - 748 aa	2	Transferase of PKS: c2qo3A	100%	296	149	GNQAG / GH**S**QG
	3	Transferase of PKS: c2hg4A	100%	297	144	GNQAG / GH**S**QG
ClbG	1	Transferase of PKS: c2qo3A	100%	354	10	GH**S**LG / GH**S**QG
1 aa - 413 aa	2	Transferase of malonyl-CoA-ACP transacylase: c3eenA	100%	315	0	GH**S**LG / GH**S**LG
	3	Transferase of malonyl-CoA-ACP transacylase: c2qj3B	100%	314	0	GH**S**LG / GH**S**VG
***E. coli* IHE3034**	**Rank**	**Hit**	**Confidence %**	**Coverage aa**	**Gaps aa**	**Motifs (query / template)**
ClbC	1	Transferase of PKS: c4mz0B	98.70%	298	78	GAGTG / GH**S**VG
435 aa - 787 aa	2	Hydrolase of PKS: c4oqjA	98.20%	306	12	GAGTG / ADRTE
	3	Transferase of PKS: c4na3A	97.40%	301	5	GAGTG / -
ClbK	1	Hydrolase of PKS: c4oqjA	100%	266	14	GDGDG / ADRTE
420 aa - 720 aa	2	Transferase of PKS: c2qo3A	100%	265	187	GDGDG / GH**S**QG
	3	Transferase of fatty acid synthetase: c2vz8A	100%	286	170	GDGDG / GH**S**LG
ClbO	1	Transferase of fatty acid synthetase: c2vz8A	100%	306	147	GYLTG / GH**S**LG
427 aa - 746 aa	2	Transferase of PKS: c3tzzA	100%	316	58	GYLTG / GQ**S**LG
	3	Transferase of PKS: c2qo3A	100%	296	150	GYLTG / GH**S**QG
ClbG	1	Transferase of PKS: c2qo3A	100%	349	12	GH**S**LG / GH**S**QG
1 aa - 422 aa	2	Transferase of PKS: c2hg4A	100%	351	13	GH**S**LG / GH**S**QG
	3	Transferase of PKS: c3tzzA	100%	367	15	GH**S**LG / GQ**S**LG
***Pseudovibrio* FO-BEG1**	**Rank**	**Hit**	**Confidence**	**Coverage aa**	**Gaps aa**	**Motifs(query / template)**
ClbC	1	Transferase of PKS: c3tzzA	100%	271	183	GQGDM / GQ**S**LG
434 aa -752 aa	2	Transferase of fatty acid synthetase: c2vz8A	100%	293	167	GQGDM / GH**S**LG
	3	Transferase of PKS: c2qo3A	100%	280	187	GQGDM / GH**S**QG
ClbK	1	Hydrolase of PKS: c4oqjA	100%	279	11	FAGNK / ADRTE
426 aa -744 aa	2	Transferase of PKS: c3tzzA	100%	278	172	FAGNK / GQ**S**LG
	3	Transferase of PKS: c2qo3A	100%	278	166	FAGNK / GH**S**QG
ClbO	1	Transferase of fatty acid synthetase: c2jfkD	97.9%	176	14	GDGTG / GH**S**LG
444 aa - 799 aa	2	Transferase of PKS: c4mz0B	97.3%	287	59	GDGTG / GH**S**VG
	3	Transferase of PKS: c3tzzA	97.2%	188	28	GDGTG / GQ**S**LG
ClbG 1 aa - 201 aa	1	Transferase of malonyl-CoA-ACP transacylase: c3tqeA	100%	200	0	GH**S**LG / GH**S**LG
	2	Transferase of malonyl-CoA-ACP transacylase: c3eenA	100%	200	0	GH**S**LG / GH**S**LG
	3	Transferase of malonyl-CoA-ACP transacylase: c3ptwA	100%	200	0	GH**S**LG / GL**S**LG

[a]amino acids (aa) of the query sequence in the alignment

[b]number of alignment gaps in the query sequence given in amino acids (aa)

[c]aligned sequence motifs are shown with the conserved active site Serine residue highlighted in red

Table S4. Seven colibactin-pathway dependent metabolites shared among *F. perrara* and two colibactin-bearing *E. coli* strains.[a]

[M]	[M+H]+	RT	Fp *wt*	Fp *clbB*::Tn	pBAC-PKS	EcN *wt*	Proposed structure
207.0568	208.0640	1.18	8.E+05	(4/5)	4.E+05	(1/5)	
258.1221	259.1293	15.35	8.E+05	-	3.E+05	(3/5)	
314.2209	315.2281	14.44	2.E+08	1.E+06	2.E+06	8.E+05	**1**
340.2366	341.2440	15.34	7.E+06	-	3.E+06	1.E+06	**3**
342.2520	343.2593	16.67	2.E+07	(2/5)	1.E+08	4.E+07	**2**
368.2678	369.2749	17.37	5.E+05	-	5.E+05	4.E+05	**4**
370.2839	371.2903	18.95	9.E+05	-	3.E+05	(3/5)	**5**

[a]Masses not detected in all five biological replicates were represented by the number of times they were observed in five replicates). The symbol, -, indicates that the mass was not observed in any biological replicates. Abbreviations are as follows: Fp wt, *F. perrara* PEB0191, Fp *clbB*::Tn, *F. perrara clbB*::Tn, pBAC-PKS, *E. coli* DH10B pBAC-PKS, EcN wt, *E. coli* Nissle 1917.

Table S5. Colibactin-pathway dependent metabolites from *F. perrara*.

[M]	*m/z*	*m/z*[a]	RT (min)	Fp *wt*	Fp *clbB*::Tn	Proposed structure[b]
227.0962[c,d]	228.1040	n.d.	14.450	3E+5	0	
258.1221	259.1293	1	15.350	8E+5	0	
286.1897	287.1970	1	12.128	1E+6	0	**6**
297.1947	298.2020	1	14.443	3E+6	0	
312.2058	313.2129	1	13.382	1E+6	0	**7**
312.2067	313.2140	1	13.655	3E+5	0	
313.0199	314.0270	1	5.648	9E+5	0	
314.2209	315.2281	1	14.444	2E+8	1E+6	**1**
328.2364	329.2443	1	15.557	4E+6	0	**8**
340.2366	341.2440	1	15.345	7E+6	0	**3**
341.1810[b,c]	683.3676	2	14.442	3E+5	0	
342.2520	343.2593	1	16.673	2E+7	(2/5)	**2**
342.9579[c]	343.9651	n.d	11.335	6E+5	0	
368.2678	369.2749	1	17.372	5E+5	0	**4**
370.2839	371.2903	1	18.953	9E+5	0	**5**
414.1452[c]	415.1529	nd	14.453	3E+5	0	
434.2220	218.1183	2	14.447	2E+6	0	
666.3902	334.2025	2	14.448	2E+6	0	
980.6098	491.3124; 981.6147	2; 1	14.443	2E+6	0	
996.5832[c]	997.5929	1	14.444	5E+5	0	

[a]n.d. = could not be determined based on low abundance

[b]See Fig. 4

[c]Raw abundance determined from diluted metabolomics samples

[d]MS^2 fragmentation was not successfully acquired

Supplemental materials and methods

Genome sequencing, assembly, and annotation. Genomic DNA was isolated from *F. perrara* PEB0191 grown for 2 days after re-streaking using the phenol-chloroform method. Libraries for SMRT sequencing (Pacific Biosciences) were constructed as recommended by the manufacturer. Sequencing of two SMRT cells was necessary to obtain 64,460 quality-filtered reads of 2.9 kb average length. Error correction, assembly, and consensus sequence polishing were carried out with the PacBio HGAP pipeline (9) resulting in two contigs, one large contig that covered the entire *F. perrara* genome and a second, smaller contig, which was removed from the assembly due to its low read coverage. The Illumina paired-end library with approximate insert sizes of 400 bp was constructed from the genomic DNA following Illumina standard protocols for genome sequencing using four PCR amplification cycles with the Bio HiFi polymerase (Kapa Biosystems, Woburn, MA, USA). Illumina sequencing was carried out on a HiSeq2000 machine in a single 2 × 100 bp lane at the Yale Center for Genome Analysis. Illumina reads were trimmed on quality with CLC Genomics Workbench (CLC Bio) as previously published (10). In total, 5,411,774 reads passed the quality filter. These Illumina reads were mapped with BWA (11) against the PacBio assembly and inspected for misassembles using the sequence assembly viewer Tablet (12). A small number of sequencing errors and one misassembly were detected and corrected. The modified assembly was again verified by read mapping and by visual inspection. Overhanging ends were trimmed and the chromosome position 1 set to the origin of replication, which was detected using the online tool Ori-Finder (13).

Transposon mutagenesis. A transposon mutant library of about ~500 clones was generated by conjugation of *F. perrara* PEB0191 with *E. coli* β2163 harboring plasmid pBT20. Cells of *F. perrara* PEB0191 grown for 30 h were harvested from two GMM (gut microbiota medium) (14) agar plates and resuspended in 100 µL of 1x PBS. An overnight culture of *E. coli* β2163 harboring plasmid pBT20 was grown for 8 h in LB supplemented with 30 µg/mL gentamicin, 100 ug/mL ampicillin, and 0.3 mM diaminopimelic acid (DAP), washed once in 1x PBS, and resuspended in 1 mL 1x PBS. 20 µL of resuspended *E. coli* cells were mixed with the resuspended *F. perrara* cells, spread onto a nitrocellulose filter (Sartorious) on a GMM agar plate supplemented with 0.1 mM DAP, and incubated for 16 h at 37° C in a 5% CO_2 incubator. Bacteria were recovered in 1 mL 1x PBS, washed once, and distributed on eight GMM agar plates supplemented with 12.5 µg/mL gentamicin. After three days of incubation under anaerobic conditions at 37° C, single colonies (transposon mutants) were picked, resuspended in 150 µL GMM supplemented with 12.5 µg/mL gentamicin and grown at 37°C under anaerobic conditions in a 96-well plate. Pools of 96 and 48 transposon mutants were generated, and genomic DNA was isolated using the DNAeasy kit (Qiagen). The remaining bacterial suspensions of the single transposon mutants were frozen at -80° C after adding 30 µL of glycerol. We screened all mutant pools by PCR with nine different primer combinations, each combination consisting of two outward-facing primers (prRND1 and prRND1rev) annealing to the two ends of the transposon and one of nine primers (prPE209-217) annealing to different regions of the *clb* GI. PCRs were analyzed by agarose gel electrophoresis to identify amplicons possibly originating from a transposon insertion into the *clb* GI. Single transposon mutants of pools suspected to contain an insertion mutant in the *clb* GI were

then individually screened by PCR. The transposon integration in the gene *clbB* was confirmed by PCR over the integration site (prPE245 and prPE246) and by Sanger sequencing. All strains, plasmids, and primers used in this study are summarized in Table 1.

Analysis of γ-H2AX phosphorylation levels in HeLa cells. γ-H2AX phosphorylation levels in HeLa cells were analyzed to detect the activation of a DNA damage response. Therefore, 3 x 10^5 cells were seeded in each well of a 6-well plate and incubated for 6 h allowing cells to adhere. Then, cells were transiently infected with bacterial cultures for 4 h at the indicated multiplicity of infection (MOI). Bacteria were removed by washing the cells 3-6 times with DMEM/5% FCS. Subsequently, HeLa cells were incubated for 12h in DMEM/5% FCS supplemented with 200 µg/mL gentamicin. Then, cells were removed from the culture dish with 100 µL 0.025% trypsin/0.01% EDTA and spun down in 1 mL cell medium in a microcentrifuge at 200 g. After washing in 1x PBS, cells were fixed in 4% paraformaldehyde for 10 min at room temperature (RT), then incubated in 20 mM NH_4Cl for 2 min, and washed again in 1x PBS. HeLa cells were resuspended in 100 µL 1x PBS, and 900 µL methanol was added during constant vortexing. After incubation for 30 min on ice, cells were washed in 0.5% BSA and incubated in 200 µl 0.5 % BSA supplemented with 0.5 µL anti-γ-H2AX primary antibody (clone 20E3, Cell Signaling) for 1h at RT. Cells were again washed in 0.5 % BSA and then incubated in 100 µL 0.5 % BSA supplemented with 2 µL of the secondary antibody conjugated to FITC (goat anti-rabbit AB97199, ABCAM) for 30 min in the dark at RT. Cells were washed once in 0.5% BSA, resuspended in 500 µL 1x PBS, and then analyzed by flow cytometry using a FACSVerse™ flow cytometer from BD Bioscience.

Supplemental references

1. **Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, *et al.*** 2004. Versatile and open software for comparing large genomes. Genome Biol **5**:R12.

2. **Kampa A, Gagunashvili AN, Gulder TAM, Morinaka BI, Daolio C, Godejohann M, *et al.*** 2013. Metagenomic natural product discovery in lichen provides evidence for a family of biosynthetic pathways in diverse symbioses. Proc Natl Acad Sci USA **110**:E3129–37.

3. **Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O.** 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol **59**:307–321.

4. **Kevany BM, Rasko DA, Thomas MG.** 2009. Characterization of the complete zwittermicin A biosynthesis gene cluster from *Bacillus cereus*. Appl Environ Microbiol **75**:1144–1155.

5. **Zhao C, Coughlin JM, Ju J, Zhu D, Wendt-Pienkowski E, Zhou X, *et al.*** 2010. Oxazolomycin biosynthesis in *Streptomyces albus* JA3453 featuring an 'acyltransferase-less' type I polyketide synthase that incorporates two distinct extender units. J Biol Chem **285**:20097–20108.

6. **Pulsawat N, Kitani S, Nihira T.** 2007. Characterization of biosynthetic gene cluster for the production of virginiamycin M, a streptogramin type A antibiotic, in *Streptomyces virginiae*. Gene **393**:31–42.

7. **Engel P, Kwong WK, Moran NA. 2013. *Frischella perrara* gen. nov., sp. nov., a gammaproteobacterium isolated from the gut of the honeybee, *Apis***

***mellifera*.** Int J Syst Evol Microbiol **63**:3646–3651.

8. **Kelley LA, Sternberg MJE.** 2009. Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc **4**:363–371.

9. **Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al.** 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods **10**: 563-569.

10. **Engel P, Martinson VG, Moran NA.** 2012. Functional diversity within the simple gut microbiota of the honey bee. Proc Natl Acad Sci USA **109**:11002–11007.

11. **Li H, Durbin R.** 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics **25**:1754-60.

12. **Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D.** 2010. Tablet--next generation sequence assembly visualization. Bioinformatics **26**:401-402.

13. **Gao F, Zhang C-T.** 2008. Ori-Finder: A web-based system for finding oriCs in unannotated bacterial genomes. BMC Bioinformatics **9**:79.

14. **Goodman AL, Kallstrom G, Faith JJ, Reyes A, Moore A, Dantas G, *et al.*** 2011. Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. Proc Natl Acad Sci USA **108**:6252–6257.