Supporting Information For

Unveiling the mechanism by which microsporidian parasites prevent locust's swarm behavior

Wangpeng Shi[a,1,2], Yang Guo[a], Chuan Xu[b], Shuqian Tan[a], Jing Miao[a], Yanjie Feng[a], Hong Zhao[b], Raymond J St. Leger[c], Weiguo Fang[b,2]

1. Department of Entomology, China Agricultural University, Beijing, 100193, China

2. Institute of Microbiology, College of Life Science, Zhejiang University, Hangzhou, 310058, China

3. Department of Entomology, University of Maryland College Park, MD, 20742, USA

Summary

Supporting methods

RNA-seq and *de novo* assembly

Total RNA was prepared from fat bodies of migratory locusts using Invitrogen TRIzol® Reagent (Life technologies, USA). Enrichment of mRNA, fragment interruption, addition of adaptors, size selection and PCR amplification of RNA-Seq were performed by Beijing Berry Genomics Co., Ltd (Beijing, China). Sequencing was performed using Illumina HiSeq ™ 2500.

The raw reads were first filtered by removing the reads flagged as either adaptors or low quality using NGS QC Toolkit (1). Clean reads were *de novo* assembled using the Trinity software (Version 2013-02-25) at the parameters of "--seqTypefq --JM 60G --min_contig_length 200 --bflyHeapSpaceMax 50G " (2). Short reads from each library were first assembled into unigenes separately. These unigenes were then combined to cluster out no-redundant unigenes using cd-kit (parameter: "-c 0.95 -T 48 -n 8") (3). The clustered unigenes were blasted against all sequences of two insect parasites (*Nosema bombycis* and *Nosema locustae*) to delete suspected parasite transcripts. The filtered unigenes were subjected to further annotation.

Function annotation

Function of unigenes was annotated by BLASTxing with E-value threshold of 10^{-5} to protein databases including the NCBI non-redundant (NR) database and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (4). Based on NR annotation, Gene ontology (GO) classification was obtained by WEGO (5) (http://wego.genomics.org.cn/cgi-bin/wego/index.pl) via GO id annotated by

Blast2GO (Version 2.3.4) (6). The annotated KEGG ontology (KO) numbers of unigenes were mapped back to the KEGG database to get pathways through KEGG mapper (http://www.genome.jp/kegg/tool/map_pathway2.html).

Expression analysis

Bowtie 2 (Version 2.3.4) (7) was used to map reads back to the transcriptome at the parameters of "--end-to-end -I 0 -X 500". Then the number of mapped clean reads for each unigene was counted and then normalized into RPKM value (Reads Per Kilo base per Million reads), which was widely used to calculate the unigene expression (8). The differentially expressed unigenes were examined using DESeq software ($P <$ 0.01 and fold change > 2) (9). The GO classifications were compared between up-regulation and down-regulation unigenes using WEGO (5). Annotated KO ids of DEGs were mapped back to the KEGG database to get pathways through KEGG mapper.

Gene Ontology and KEGG pathway enrichment analyses

To assign putative biological functions and pathway involvement to the unigenes, enrichment analysis was carried out using R language. To compare these unigenes to the whole locust transcriptome background, the hypergeometric test was applied to find significantly enriched GO terms and pathways. The formula of enrichment test was:

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

N: the total number of unigenes with GO or KEGG pathway annotation; n: the

number of DEGs in N; M: the number of unigenes that were annotated to certain GO or pathway terms; m: the number of DEGs in M. The initially obtained *P*-values were then adjusted using a Bonferroni-Correction and a corrected *P*-value of 0.05 was adopted as a threshold.

Quantitative RT-PCR analysis

Total RNA was extracted from fat bodies using Invitrogen TRIzol® Reagent (Life technologies, USA), and the first-strand cDNA was synthesized using Revert Aid ™ First Strand cDNA Synthesis Kit (Fermentas, USA). The cDNA was serially diluted 3-fold and the dilutions were used for analyzing PCR efficiency of primers. The preparation of RT-PCR mixture was conducted as instructed by the manual of the SYBR® Premix Ex TaqTM (Tli RNaseH Plus) (Takara, China). The reactions were performed on a LightCycler 1.5 instrument (Roche, USA) using the two-step method, which was initiated by 30s at 95°C, then followed by 45 cycles of 95°C for 5s, 60°C for 20s, and completed with a melting curve analysis program. The specificity of qRT-PCR primers was confirmed by melting curve and sequencing of qRT-PCR products. The reference gene is β-actin gene (Genbank accession number: KC118986). The $2^{-\Delta\Delta Ct}$ method was used to calculate gene expression levels (10). All primers are presented in Table S1.

Supporting Results

RNA sequencing and *de novo* assembly

To systematically investigate the effect of *P. locustae* infection on locusts, two libraries [Control locusts (CL) and infected locusts (IL)] were constructed for RNA-Seq. Total RNA was prepared from fat bodies of control healthy locusts and infected ones 16 days post-inoculation. The locust's genome is very large (~6,500 Mb), and it has not been sequenced yet (11). In order to achieve a high coverage of the gene content of the locust, each of our libraries produced more than 89 Mb raw reads (>8.5 Gb nucleotides) with a single read length about 100bp (Table S2). We removed about 13% of these reads flagged as either adaptors or low quality, each library remaining more than 78 Mb clean reads (>7.4Gb nucleotides) with Q20 percentage (percentage of sequences with sequencing error rate lower than 1%) over 97% (Table S2).

We adopted an optimized procedure to do *de novo* assembly (12), and reads from two libraries were assembled into unigenes separately and then combined for clustering to obtain a new set of unigenes for following analysis. Short reads from two libraries were assembled by Trinity (2) into 4,680,373 (a mean length of 57bp) and 1,470,285 contigs (a mean length of 74bp) for CL and IL, respectively (Table S3). Further assembly generated 53,878 unigenes from CL and 78,362 from IL (Table S3). The unigenes from the two libraries were then combined and a cluster analysis using cd-kit (3) gave 78,343 non-redundant unigenes with a mean length of 749bp. At 16 days post-inoculation, the parasite could infect the fat body and its mRNA thus

contaminated the insect's RNA collected from fat body. In order to eliminate suspected parasite RNA, the clustered unigenes were blasted against all sequence of two insect parasites (*Nosema bombycis* and *Nosema locustae*), and 1.2% (949 sequences) suspected parasite transcripts were removed (cut off E-value: $1e^{-5}$). The final number of clean unigenes was 77,349, with a mean length of 724bp. A N50 (half of the assembled bases were incorporated into unigenes) length of 1,320bp was obtained, with one third (28,143 unigenes) having a length over 500bp (Fig. S8). All these clean unigenes were subjected to following function annotation.

Function annotation

In order to get information of pathway and GO function from the unigenes determined by *de novo* assembly, NCBI NR database, the KEGG database were selected as the annotation reference. Approximately 33% of unigenes (25,312) were annotated by BLASTX with a threshold of 10^{-5}. This annotation percentage is low, but comparable to previous annotation based on the data generated from locust's RNA-Seq projects (13). Among these annotated unigenes, 23,809 were annotated in both databases, while a small number of unigenes could only be annotated to either the NR database (1,193 unigenes) or KEGG (310 unigenes) (Fig.S9A). Based on NR annotation and the E-value distribution, 70% of the mapped sequences showed strong homology (E-value < 10^{-20}), and 43% were very strong homology (E-value < 10^{-50}) to previously annotated sequences (Fig.S9B). The 25 top-hit species based on NR annotation are shown in Fig.S9C. Nearly 34% of unigenes were annotated from 5 top-hit species, i.e., the body louse *Pediculus humanus corporis,* the red flour beetle

Tribolium castaneum, the leaf cutter bee *Megachile rotundata,* the wasp *Nasonia vitripennis,* and the aphid *Acyrthosiphon pisum*. There 5,717 unigenes were classed into three gene ontology categories: cellular component, biological process and molecular function. Under the cellular component category, the top 3 subcategories were cell, cell part and organelles; in the biological process category were metabolic process (2,550 unigenes, 44.6%) and cellular process (2,710 unigenes, 47.4%); in the molecular function category were binding (2,786 unigenes, 48.7%) and catalytic activity (2,874 unigenes, 50.3%) (Fig. S10).

There were 6, 078 unigenes mapped into 329 KEGG pathways using the annotated KO. The maps with the largest number of unigenes (excluding the global pathways ko01100, ko01110 and ko01120) were ubiquitin mediated proteolysis (ko04120, 299unigenes, 4.92%), followed by RNA transport (ko03013, 294 unigenes, 4.84%) and pathways in cancer (ko05200, 283 unigenes, 4.65%) (Fig. S11).

Expression profiling

RPKM was used to calculate the expression of unigenes (8). All 77,387 unigenes were expressed in CL and/or IL. When RPKM is over 1, the number of unigenes in CL was bigger than IL, but for low expression category (RPKM < 1), the IL had more unigenes than CL. The trend is also true for annotated unigenes (Table S4). Among the expressed unigenes, 1,684 were differentially expressed between IL and CL (fold change > 2 and $P < 0.01$ based on DESeq) (9), among which 861 were up-regulated and 823 down-regulated. The proportions and comparisons between up-regulated and down-regulated unigenes were summarized in three main GO functional categories

(Fig. S8). The results obtained from GO enrichment were presented in Table S5.

References

1. Patel RK, Jain M (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 7(2):e30619.
2. Grabherr MG, *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29(7):644-652.
3. Li W & Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13):1658-1659.
4. Kanehisa M & Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28(1):27-30.
5. Ye J, *et al.* (2006) WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res* 34(Web Server issue):W293-297.
6. Conesa A, *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21(18):3674-3676.
7. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357-359.
8. Mortazavi A, Williams BA, McCue K, Schaeffer L, & Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7):621-628.
9. Anders S & Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11(10):R106.
10. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using realtime quantitative PCR and the 2−ΔΔCT method. *Methods* 25:402– 408.
11. TR G (2010) Animal Genome Size Database. http://www.genomesize.com.
12. Chen S, *et al.* (2010) De novo analysis of transcriptome dynamics in the migratory locust during the development of phase traits. *PLoS One* 5(12):e15633.
13. Wang Y, Yang P, Cui F, & Kang L (2013) Altered immunity in crowded locust reduced fungal (Metarhizium anisopliae) pathogenesis. *PLoS Pathog* 9(1):e1003102.

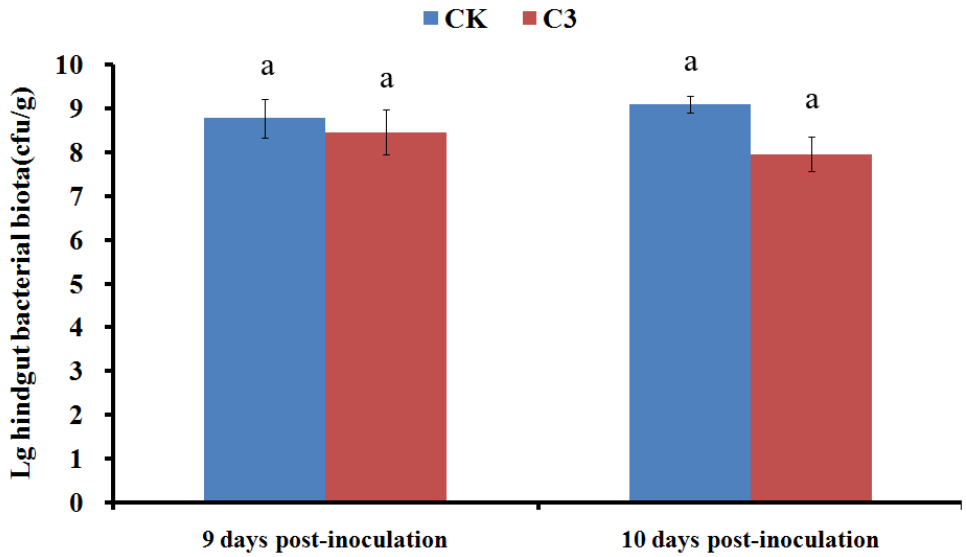

Fig. S1. Hindgut bacterial populations (mean ± SE) of *L. migratoria* 9 and 10 days post-inoculation with *P. locustae*. CK: uninfected locusts (blue); C3: locusts infected with 20,000 *P. locustae* spores (red). Values with the same letters are not significantly different ($P < 0.05$) (Kruskal-Wallis test).

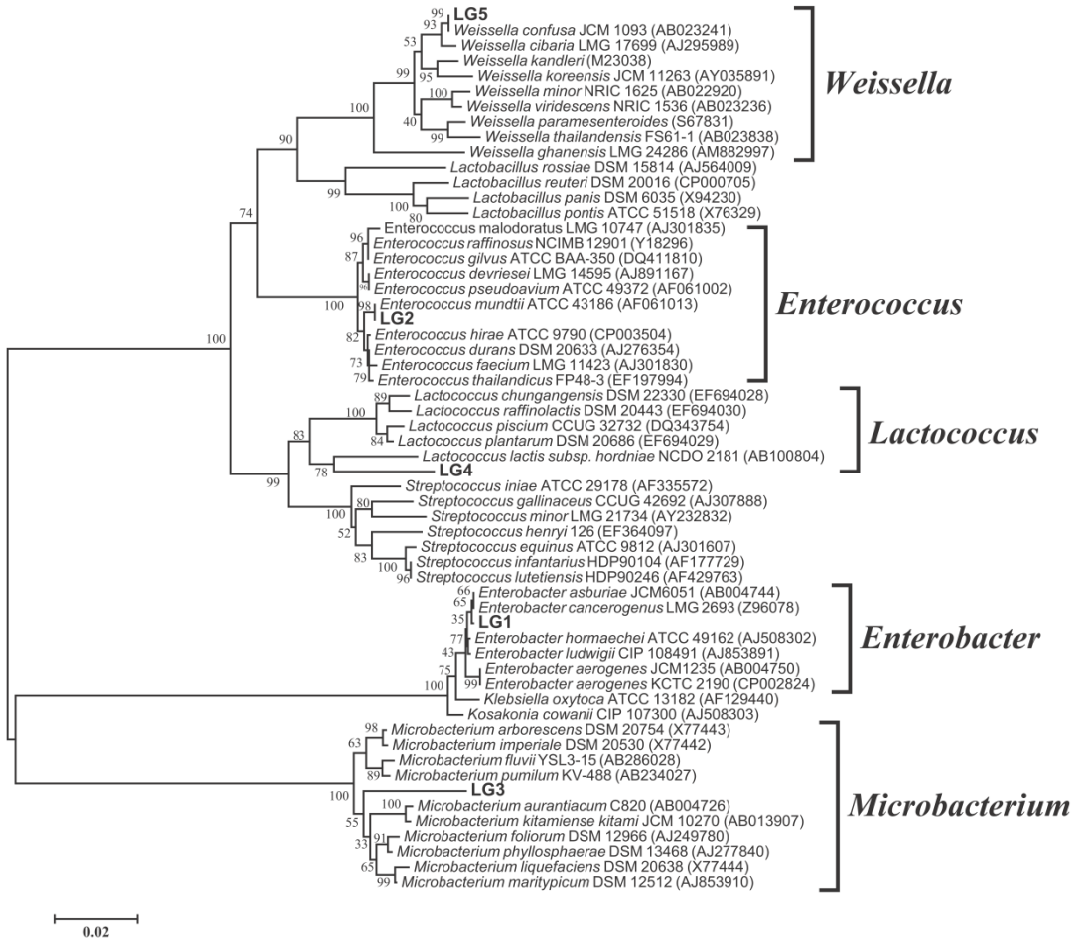

Fig. S2. Phylogenetic analysis using neighbor-joining algorithm provided by MEGA 5.1 of 16s rRNA sequences from the 5 primiary hindgut bacteria [LG1(locust gut bacterium), LG2, LG3, LG4 and LG5] and bacterial type strains. The 16s rRNA sequences of the type strains were retrieved from NCBI database. Bootstrap values are adjacent to each internal node, representing the percentage of 1,000 bootstrap replicates. Bar: 0.02 substitution per nucleotide positions. The 16s rRNA sequences of the hindgut bacteria are shown in Table S7.

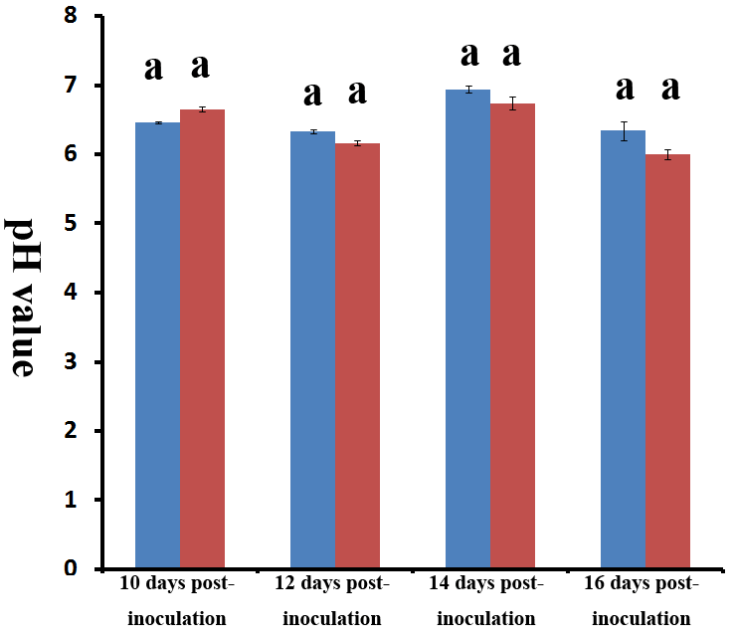

Fig. S3. pH values (mean ± SE) of the supernatant of midgut contents of healthy locusts (blue) and *P. locustae*-infected ones (red). The inoculation dose was 20,000 spores/3rd instar nymph, and pH measurement was conducted 10, 12, 14 and 16 days post-inoculation. Values with different letters are significantly different ($P < 0.05$) (Kruskal-Wallis test). No significant difference in pH value of the midguts was found between control healthy locusts and the ones infected by *P. locustae*.

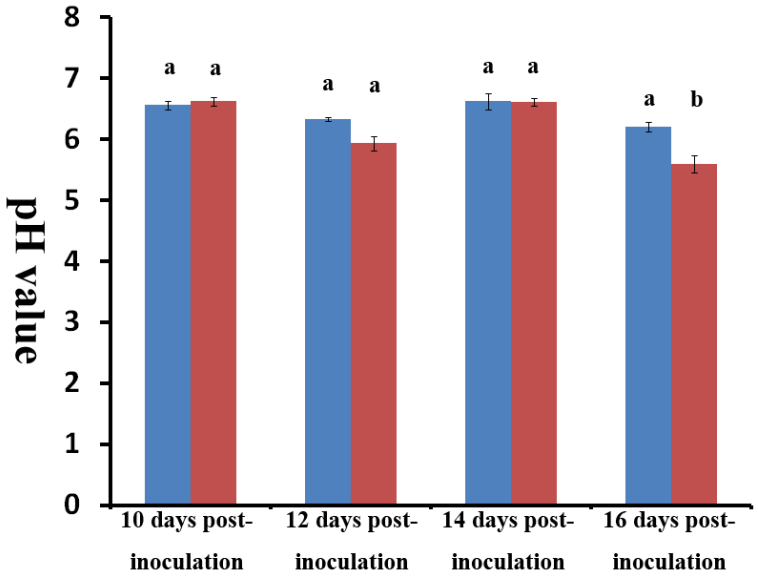

Fig. S4. pH values (mean ± SE) of the supernatant of hindgut contents of healthy locusts (blue) and *P. locustae*-infected ones (red). The inoculation dose was 20,000 spores/3rd instar nymph, and pH measurement was conducted 10, 12, 14 and 16 days post-inoculation. Values with different letters are significantly different ($P < 0.05$) (Kruskal-Wallis test).

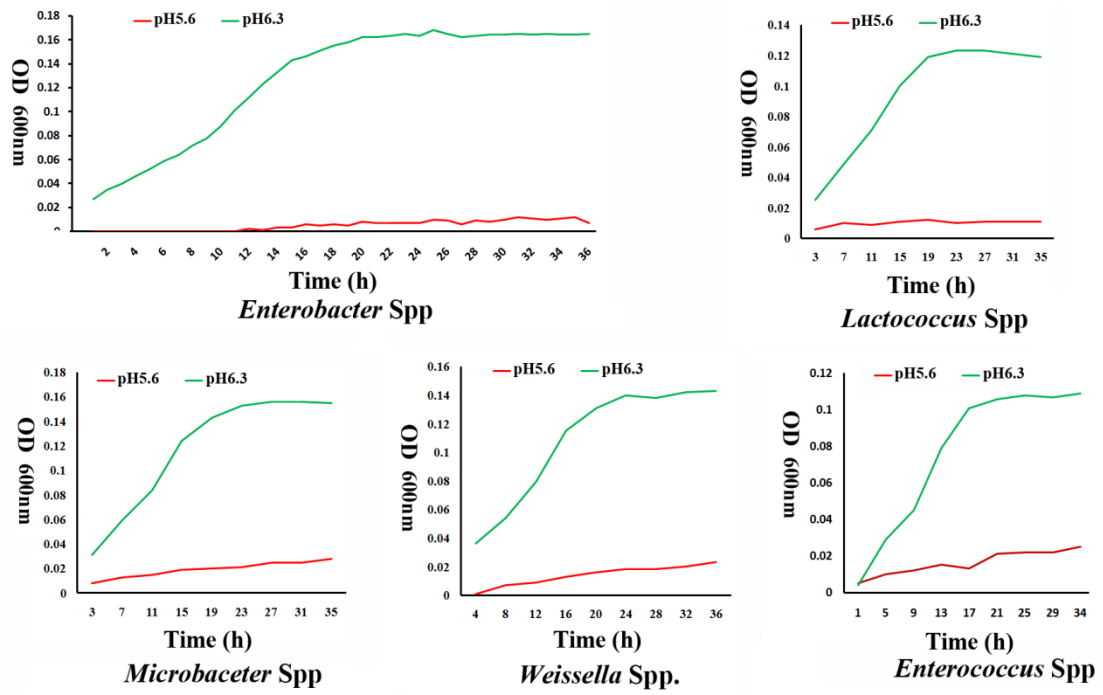

Fig. S5. The effect of pH on the growth of 5 primary hindguts bacteria from the migratory locust. Bacteria grew in Luria-Bertani broth with pH of 5.6 (red) and 6.3 (green).

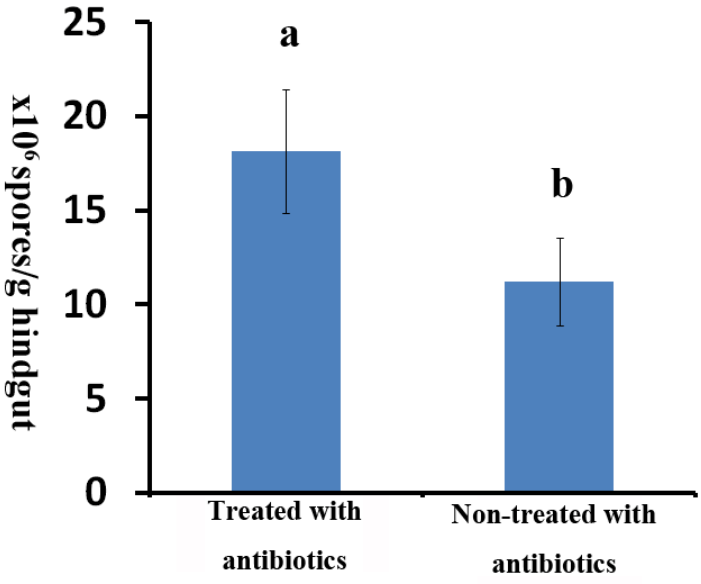

Fig. S6. Colonization resistance of gut bacteria against *P. locustae* infection. The insects were fed with an antibiotic cocktail for 2 days before inoculation with *P. locustae* (20,000spores/3rd instar nymph). At 16 days post-inoculation, these locust hindguts were dissected, weighed and homogenized in sterile water for spore counting. Values with different letters are significantly different ($P < 0.05$) (Kruskal-Wallis test).

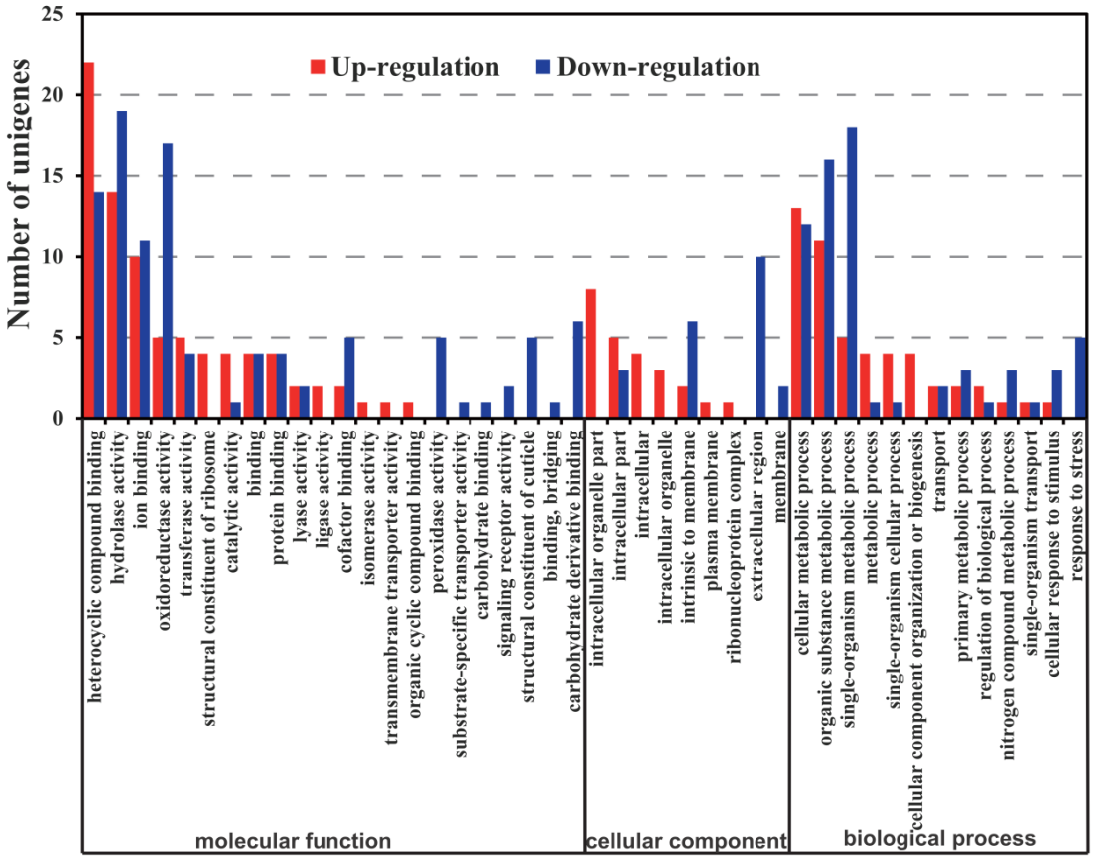

Fig. S7. Go classification for up-regulated and down-regulated unigenes

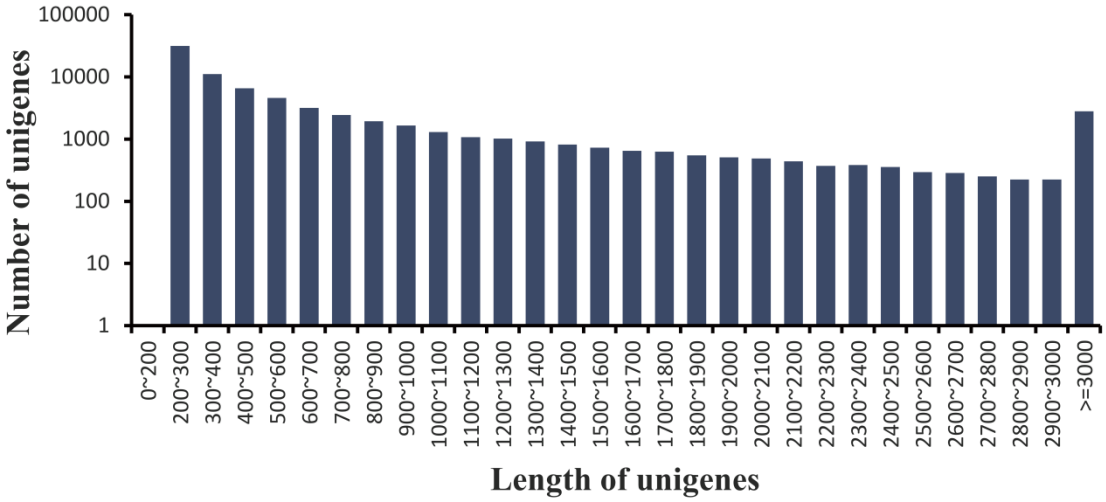

Fig. S8. Length distribution of clean unigenes

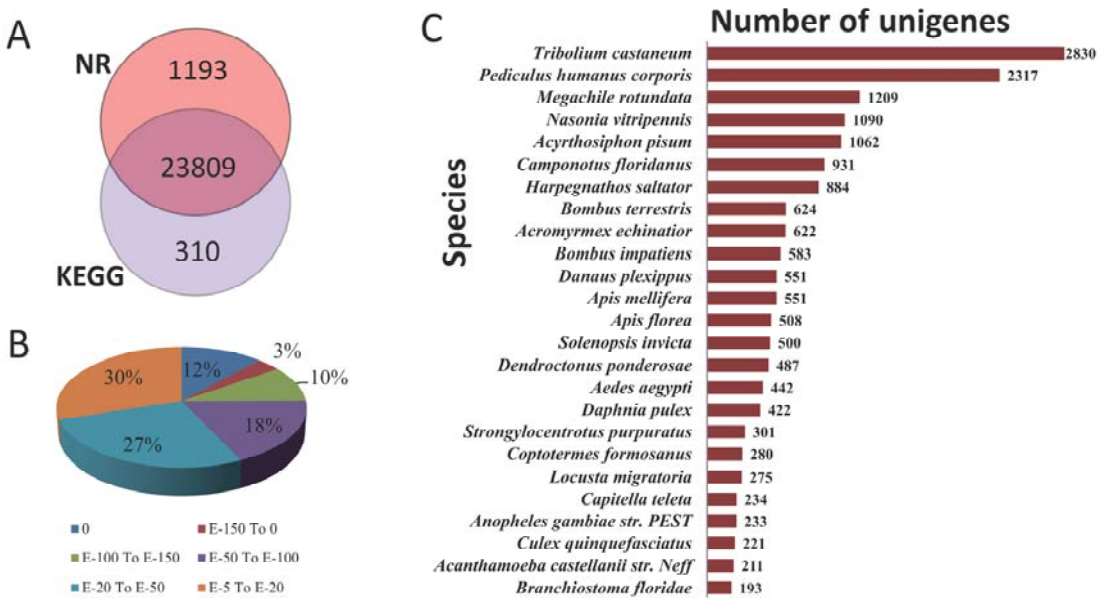

Fig. S9. Functional annotation of unigenes. (A) Venn diagram of number of unigenes annotated by BLASTX with an E-value threshold of 10^{-5} against protein databases (NR and KEGG); (B) E-value distribution of the BLASTX hits of unigenes against the NR database; (C) The Number of unigenes that have significant similarity to each of the top 25 species using BLASTX in the NR database.

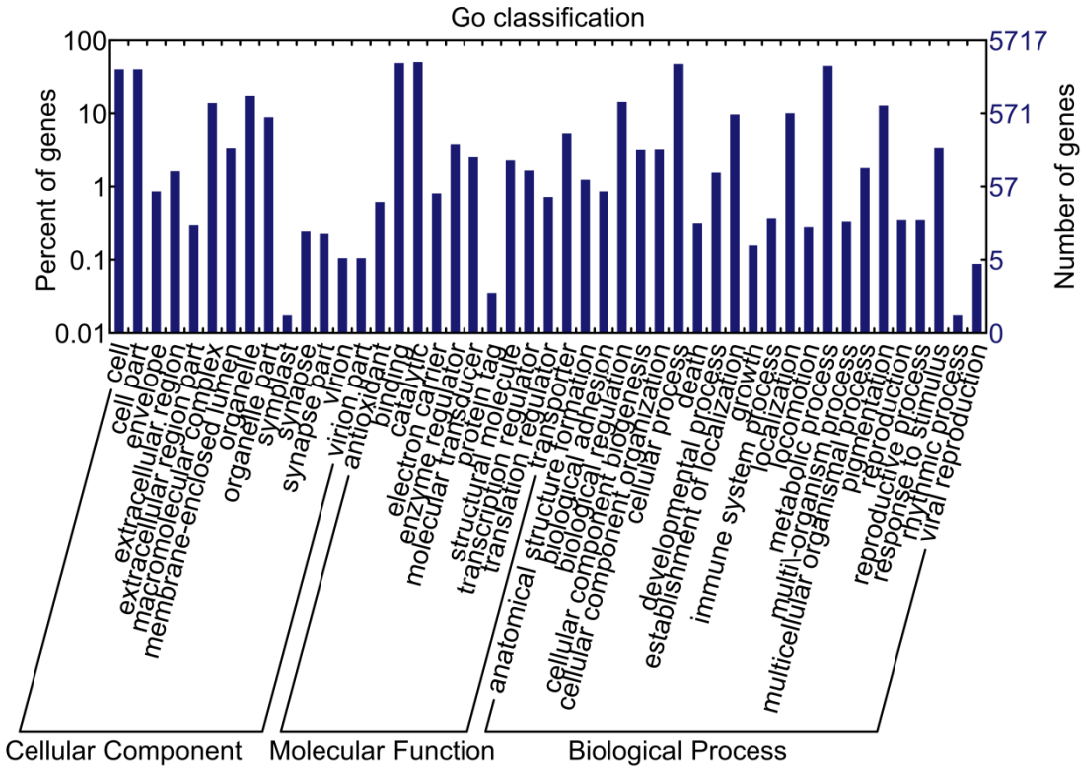

Fig. S10. Histogram presentation of GO classification. The results are summarized in three main categories: biological process, cellular component and molecular function. The right y-axis indicates the number of genes in a category. The left y-axis indicates the percentage of a specific category of genes in that main category.

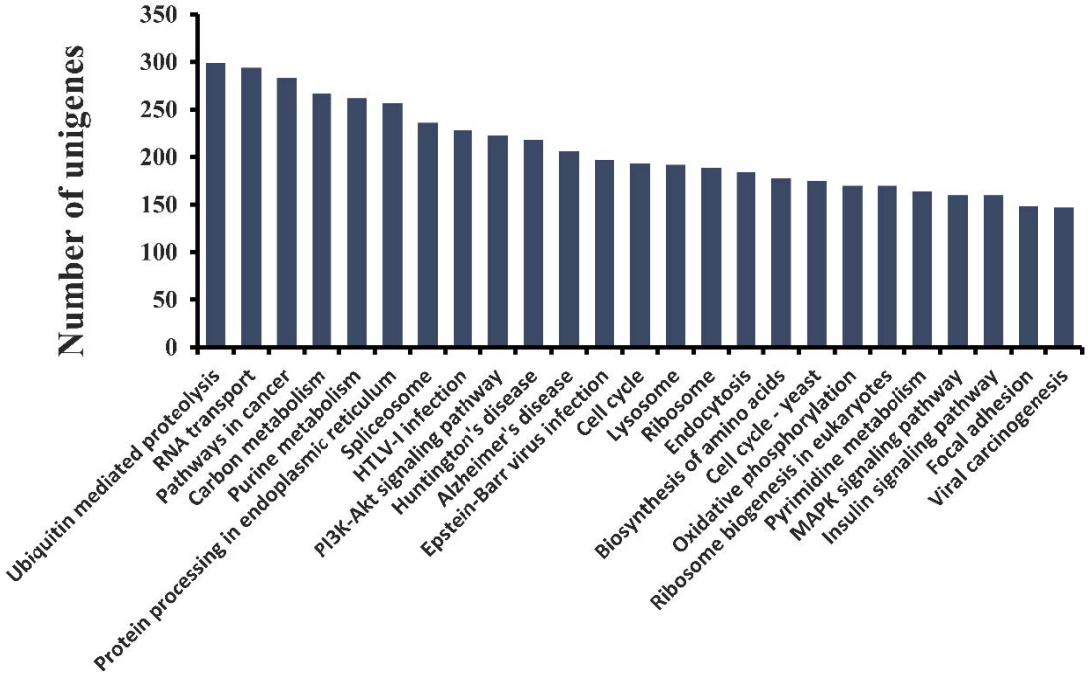

Fig. S11. The top 25 annotated pathways identified by KEGG.

Table S1. Primers for quantitative RT-PCR validation of selected differentially expressed unigenes related to tyrosine hydroxylase and primers for 16s rRNA cloning

Genes	**Fold Change (CL/IL)**		**Primers**
	RNA-Seq	qRT-PCR	
Tyrosine hydroxylase			
DEG1 (pale) (comp101707_c0_seq1)	26.8	37.4	GAGCGGGAACTATACAATGGG GCTTCTCGTCCACTCGTTC
DEG2 (comp109641_c0_seq4)	20.5	34.4	GGCATCGTTGTAGGTCTGG GAAGTCCTACAGCATCGAGAAC
β-actin			GCAAAGCTGGCTTCGCCG ATGTTCCTCGGGCGCCAC
16s rRNA			5'-CAGGCCTAACACATGCAAGTC-3' 5'-GGGCGGTGATGTACAAGGC -3'

Table S2. Overview of data and quality of RNA-seq

Sample	Raw Reads	RawData (Mb)	CleanReads	CleanData (Mb)	N[a] (%)	GC[b] (%)	Q20[c] (%)
IL	101,365,130	9,666.93	88,726,176	8,461.59	0	47.91	97.81
CL	89,417,790	8,527.54	78,232,536	7,460.84	0	47.94	97.85

Note: N[a]: the percentage of nucleotides which were not sequenced;

GC[b]: the percentage of GC content of the sequences;

Q20[c]: the percentage of sequences with sequencing error rate lower than 1%.

The three parameters (N, GC andQ20) are about clean data.

Table S3. Statistics of *de novo* assembly of locust contigs and unigenes.

Statistics	Contig		Unigene			
	CL	IL	CL	IL	Combined[a]	Filtered[b]
Number of sequence	4,680,373	1,470,285	53,878	78,362	78,343	77,394
Large sequence (>=1000bp)	9,019	10,490	17,066	12,157	14,986	14,328
Max sequence length (bp)	20,554	52,998	39,046	20,554	39,046	20,554
Mean sequence length (bp)	57	74	1,063	634	749	724
N50 length (bp)	49	62	2,039	933	1411	1320
Total bases (Mb)	258.63	104.66	54.65	47.44	55.99	53.46

Note: a: Total unigenes predicted by clustering unigenes from CL and IL;

b: Total unigenes predicted by removing suspected parasite sequences from a.

Table S4. Expression abundance measurements of CL and IL

	CL	CL (annotated)	IL	IL (annotated)
RPKM 0-1	48,718	11,339	57,127	13,403
RPKM 1-10	24,313	10,840	16,432	8,967
RPKM 10-100	3,583	2,612	3,169	2,451
RPKM >100	773	521	659	491
Total detected	77,387	25,312	77,387	25,312

Table S5. GO enrichment result of DEGs

GO Term	Go id	P-value	FDR
nitric-oxide synthase activity	GO:0004517	3.26E-06	0.0002476
nitric oxide biosynthetic process	GO:0006809	3.26E-06	0.0002476
extracellular region	GO:0005576	4.66E-06	0.0002476
heme binding	GO:0020037	6.08E-06	0.0002476
chitin binding	GO:0008061	9.89E-06	0.0002841
retinyl-palmitate esterase activity	GO:0050253	1.05E-05	0.0002841
calmodulin binding	GO:0005516	1.47E-05	0.0003425
structural constituent of cuticle	GO:0042302	3.27E-05	0.0006318
triglyceride lipase activity	GO:0004806	4.12E-05	0.0006318
retinol metabolic process	GO:0042572	4.12E-05	0.0006318
FMN binding	GO:0010181	4.26E-05	0.0006318
chitinase activity	GO:0004568	9.71E-05	0.0011786
chitin catabolic process	GO:0006032	9.71E-05	0.0011786
lipid catabolic process	GO:0016042	0.000101227	0.0011786
cell wall macromolecule catabolic process	GO:0016998	0.000137881	0.0014983
peroxidase activity	GO:0004601	0.000179096	0.0017529
flavin adenine dinucleotide binding	GO:0050660	0.000182815	0.0017529
NADP binding	GO:0050661	0.000333391	0.003019
peroxidase reaction	GO:0006804	0.000373007	0.0032
response to oxidative stress	GO:0006979	0.00043978	0.0035842
oxidation-reduction process	GO:0055114	0.000495691	0.0038475
iron ion binding	GO:0005506	0.000520964	0.0038599
peptide cross-linking	GO:0018149	0.001424996	0.0100989
fatty-acyl-CoA reductase (alcohol-forming) activity	GO:0080019	0.002808837	0.0190767
Phosphopyruvate hydratase complex	GO:0000015	0.00461391	0.0289257
Phosphopyruvate hydratase activity	GO:0004634	0.00461391	0.0289257
oxidoreductase activity	GO:0016491	0.007639834	0.0455528
cellular protein metabolic process	GO:0044267	0.007825015	0.0455528

Table S6. Sequences of unigenes clustered in KEGG ontology of tyrosine hydroxylase

>comp101707_c0_seq1 (**the gene *pale***)

CCGACACACTCGCCAGCGGCGCCGCTACTCGTCGGAGACGTGCTCGACTCCCGTCTGACGGAAGTCGTCACCGGTAGCAGC
CGTCTCTCCTGCACCGCTAGCGTGGGCTGGATCTGCAGTGGGCTACGAGCGGGAACTATACAATGGGTACGACGATGGGGA
CGATGCGGACGCTAAGTACTTGGGAGAGCGCGGGGGAACGAGTGGACGAGAAGCGCTCCTTAGGCGAAGTTGGCCTTGAG
CTTGGCCAGCGCGTTATTGAGGTGCAGCACCTCGAGGTTGAGCTGCGACAGGCAGCCCTCCAGCCGCTCCACCGAGTCCAG
CACCTCGACGCGCTGCGTGTGCGGGTTGTAGCGCACCTCGAACTTGCGGGACATGCCGCTGGCCACCCACCTCCTGAACTT
GTCCTTGGCGTCCTCGAAGCTCTCGGCGACGTAGTAGACGGGCTGGACGTCCTGGTCCTGGTAGGGCTGGATGGCGGTGAG
CGAGGGCTCGAAGGGCCGGTGCTCGGGCTTGTCGGACACGGCGTGCAGCAGCTCGCGCGTCGACGACAGCAGGCCGGCGC
CGTACGCCTTCAGCTGGCCCTTCTCCTTGCACAGGCCGAACTCCACCGTGAACCAGTACACCGTCGATAGCTTCTCGATCTC
CGCGTCGGACGCGCCCAGCGAAGCCAGGCCGATCTCTTGCGAGAACTGCGCGAAGCTGGGGTCAGCCAGCAGCGGCATGT
GCCCGATCAGCTCGTGGATGCAGTCGGGCTCTGGCGTGTGGGTGGGCGACGAGGTGTGGCGGACGTACTGCGTGCTCTGGA
AGATGCGGAACGCCAGGCTGGCCAGGAAGTCGCGCGCCGTCAGCAGGCCGGCGGCGGGCCGCAGCCGGAAGCCCGTGTG
CTTGCTCATGAAGTTGGACATGTCCTCGAGCTGCGGGATGCGCTCCTCGGTGAAGATGCCCGCCTCCTGCAGCAGCTTGAAG
ACGCGCCGGTACTCGGCGCACGCGTACTTGGGCAGCAGGTCGACCGTGGCGCGGAACACCTGCTTCCACGTGGCCGTCTCC
TGGGGCGTGTACTCGATGCGGG

>comp109641_c0_seq4

CGCCGCCGGCGTGGCCGTTGGTCAGGGGCTGCTCTGGTTCTTCTGGTTGCTGCGGATCGGCGACCTCGTTTGGCACCTCGCT
AGCGTCATCGGCCACTTCTGTTGCGGTAATTTCTTCTTGCGATGTGTTTTCGATTTCTTCTGTGGTAGGCAAATCTTTCTCCTG
CACCGCGTTTACCTCGGCAGAGTCAGTTGGGGCATCGTTGTAGGTCTGGCGGGCCTCCTCCAGCACGCTCTGCTTGGTCTGC
TTGACGACCAGCGTCTCGAAGCGGGCGTCGTCGACGAGCGAGCGGCGCCGCGACGGGTAGCCGTTCTCGATGCTGTAGGA
CTTCTTGATGGCGAACATCTCGCGATTCTTCTGGGCAGCGGCTACGGCCATCATGTCTGAGGGTTG

Table S7. Sequences of 16S rRNA from 5 primary hindgut bacteria

>LG1 (*Enterobacter* spp)

GACGCTGCTGCTTGCTGAGAGTGGCGGACGGGTGAGTAATGTCTGGGAAACTGCCTGATGGAGGGGGATAACTACTGGAAA
CGGTAGCTAATACCGCATAACGTCGCAAGACCAAAGAGGGGGACCTTCGGGCCTCTTGCCATCGGATGTGCCCAGATGGGAT
TAGCTAGTAGGTGGGGTAACGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAACTGA
GACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTG
TATGAAGAAGGCCTTCGGGTTGTAAAGTACTTTCAGCGGGGAGGAAGGCGATAAGGTTAATAACCTTGTCGATTGACGTTAC
CCGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGC
GTAAAGCGCACGCAGGCGGTCTGTCAAGTCGGATGTGAAATCCCCGGGCTCAACCTGGGAACTGCATTCGAAACTGGCAGG
CTAGAGTCTTGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGG
CGGCCCCCTGGACAAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCC
GTAAACGATGTCGACTTGGAGGTTGTGCCCTTGAGGCGTGGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTA
CGGCCGCAAGGTTAAAACTCAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCG
AAGAACCTTACCTACTCTTGACATCCAGAGAACTTACCAGAGATGCATTGGTGCCTTCGGGAACTCTGAGACAGGTGCTGCA
TGGCTGTCGTCAGCTCGTGTTGTGAAATG

>LG2 (*Enterococcus* spp)

TTACGGACTGCTCACCGGGAAGAGGAGTGGCGAACGGGTGAGTAACACGTGGGTAACCTGCCCATCAGAAGGGGATAACA
CTTGGAAACAGGTGCTAATACCGTATAACAATCGAAACCGCATGGTTTCGTTTTGAAAGGCGCTTTACGGTGCCGCTGATGG
ATGGACCCGCGGTGCATTAGCTAGTTGGTGAGGTAACGGCTCACCAAGGCCACGATGCATAGCCGACCTGAGAGGGTGATC
GGCCACATTGGGACTGAGACACGGCCCAAACTCCTACGGGAGGCAGCAGTAGGGAATCTTCGGCAATGGACGAAAGTCTG
ACCGAGCAACGCCGCGTGAGTGAAGAAGGTTTTCGGATCGTAAAACTCTGTTGTTAGAGAAGAACAAGGGTGAGAGTAAC
TGTTCACCCCTTGACGGTATCTAACCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTGGCAAGC
GTTGTCCGGATTTATTGGGCGTAAAGCGAGCGCAGGCGGTTTCTTAAGTCTGATGTGAAAGCCCCCGGCTCAACCGGGGAG
GGTCATTGGAAACTGGGAGACTTGAGTGCAGAAGAGGAGAGTGGAATTCCATGTGTAGCGGTGAAATGCGTAGATATATGG
AGGAACACCAGTGGCGAAGGCGGCTCTCTGGTCTGTAACTGACGCTGAGGCTCGAAAGCGTGGGGAGCAAACAGGATTAG
ATACCCTGGTAGTCCACGCCGTAAACGATGAGTGCTAAGTGTTGGAGGGTTTCCGCCCTTCAGTGCTGCAGCTAACGCATTA
AGCACTCCGCCTGGGGAGTACGACCGCAAGGTTGAACTCAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGG
TTTAATTCGAAGCAACGCGAAGAACCTTACCAGGTCTTGACATCCTTTGACACTCTAGAGATAGAGCTTCCCCTTCGGGGGC
AAAGTGACAGTGGTGCATGGTGTG

>LG3 (*Microbacterium* spp)

TCCGAAGCATTCGATGGGGTGAAGATTCCTGCTCAACCTGCCCCTGCGGAAGTGAATGCCTTGGAATCACCGTCCAGACTG
GATACAAACCACAAAGGCATCTTCAGTGGTTGAAAAGAATTTCGGTTGGGGATGGGCTCGCGGCCTATCAGCTTGTTGGTGA
GGTAATGGCTCACCAAGGCGTCGACGGGTAGCCGGCCTGAGAGGGTGACCGGCCACACTGGGACTGAGACACGGCCCAGA
CTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCAACGCCGCGTGAGGGATGACGGC
CTTCGGGTTGTAAACCTCTTTTAGCAGGGAAGAAGCGAAAGTGACGGTACCTGCAGAAAAAGCGCCGGCTAACTACGTGCC
AGCAGCCGCGGTAATACGTAGGGCGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGCTCGTAGGCGGTTTGTCGCGTCTG
CTGTGAAATCCGGAGGCTCAACCTCCGGCCTGCAGTGGGTACGGGCAGACTAGAGTGCGGTAGGGGAGATTGGAATTCCTG
GTGTAGCGGTGGAATGCGCAGATATCAGGAGGAACACCGATGGCGAAGGCAGATCTCTGGGCCGTAACTGACGCTGAGGAG
CGAAAGGGTGGGGAGCAAACAGGCTTAGATACCCTGGTAGTCCACCCCGTAAACGTTGGGAACTAGTTGTGGGGTCCATTC
CACGGATTCCGTGACGCAGCTAACGCATTAAGTTCCCCGCCTGGGGAGTACGGCCGCAAGGCTAAAACTCAAAGGAATTGA
CGGGGACCCGCACAAGCGGCGGAGCATGCGGATTAATTCGATGCAACGCGAAGAACCTTACCAAGGC

>LG4 (*Lactococcus* spp)

GTGCTGCTTTTTTTGAGAGCGGCGAACGGGTGAGTAACGCGTGGGAAATCTGCCGAGTAGCGGGGGACAACGTTTGGAAA
CGAACGCTAATACCGCATAACAATGAGAATCGCATGATTCTTATTTAAAAGAAGCAATTGCTTCACTACTTGATGATCCCGCG
TTGTATTAGCTAGTTGGTAGTGTAAAGGACTACCAAGGCGATGATACATAGCCGACCTGAGAGGGTGATCGGCCACACTGGG
ACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTAGGGAATCTTCGGCAATGGGGGCAACCCTGACCGAGCAACGC
CGCGTGAGTGAAGAAGGTTTTCGGATCGTAAAACTCTGTTGTTAGAGAAGAACGTTAAGTAGAGTGGAAAATTACTTAAGT
GACGGTATCTAACCAGAAAGGGACGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTCCCAAGCGTTGTCCGGATTT
ATTGGGCGTAAAGCGAGCGCAGGTGGTTTCTTAAGTCTGATGTAAAAGGCAGTGGCTCAACCATTGTGTGCATTGGAAACTG
GGAGACTTGAGTGCAGGAGAGGAGAGTGGAATTCCATGTGTAGCGGTGAAATGCGTAGATATATGGAGGAACACCGGAGGC
GAAAGCGGCTCTCTGGCCTGTAACTGACACTGAGGCTCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCA
CGCCGTAAACGATGAGTGCTAGCTGTAGGGAGCTATAAGTTCTCTGTAGCGCAGCTAACGCATTAAGCACTCCGCCTGGGGA
GTACGACCGCAAGGTTGAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAA
CGCGAAGAACCTTACCAGGTCTTGACATACTCGTGCTATCCTTAGAGATAGGAGTTCCTTCGGGACACGGGATACATGTGGT
GCATGGTTGTCGTCAGCTCGTGTCGTGA

>LG5 (*Weissella* spp)

CGTCTCATTGAGAGCTTGCTCAGATATGACGATGGACATTGCAAAGAGTGGCGAACGGGTGAGTAACACGTGGGAAACCTA
CCTCTTAGCAGGGGATAACATTTGGAAACAGATGCTAATACCGTATAACAATGACAACCGCATGGTTGTTATTTAAAAGATGG
TTCTGCTATCACTAAGAGATGGTCCCGCGGTGCATTAGCTAGTTGGTAAGGTAATGGCTTACCAAGGCGATGATGCATAGCCG
AGTTGAGAGACTGATCGGCCACAATGGGACTGAGACACGGCCCATACTCCTACGGGAGGCAGCAGTAGGGAATCTTCCACA
ATGGGCGAAAGCCTGATGGAGCAACGCCGCGTGTGTGATGAAGGGTTTCGGCTCGTAAAACACTGTTGTAAGAGAAGAATG
ACATTGAGAGTAACTGTTCAATGTGTGACGGTATCTTACCAGAAAGGAACGGCTAAATACGTGCCAGCAGCCGCGGTAATAC
GTATGTTCCAAGCGTTATCCGGATTTATTGGGCGTAAAGCGAGCGCAGACGGTTATTTAAGTCTGAAGTGAAAGCCCTCAGC
TCAACTGAGGAATTGCTTTGGAAACTGGATGACTTGAGTGCAGTAGAGGAAAGTGGAACTCCATGTGTAGCGGTGAAATGC
GTAGATATATGGAAGAACACCAGTGGCGAAGGCGGCTTTCTGGACTGTAACTGACGTTGAGGCTCGAAAGTGTGGGTAGCA
AACAGGATTAGATACCCTGGTAGTCCACACCGTAAACGATGAGTGCTAGGTGTTTGAGGGTTTCCGCCCTTAAGTGCCGCAG
CTAACGCATTAAGCACTCCGCCTGGGGAGTACGACCGCAAGGTTGAAACTCAAAGGAATTGACGGGGACCCGCACAAGCG
GTGGAGCATGTGGTTTAATTCGAAGCACGCGAAGAACCTTACCAGGTCTGACATCCTTGACACTCCAGAGATGGAGCGTTC
CCTTCGGGGACAGGTGACA