Genomic adaptations to chemosymbiosis in the deep-sea seep-dwelling tubeworm *Lamellibrachia luymesi*

Additional file 3: commands and parameters used in this study

1. Quality trimming of raw sequencing reads

1.1 paired-end libraries

Reads were trimmed using `trimmomatic-0.36` to remove low quality regions and adaptor for each paired-end library.

```bash
module load trimmomatic/0.36
java -jar /tools/trimmomatic-0.36/trimmomatic-0.36.jar PE -threads 20 -phred33 SL84795_1_1.fastq.gz SL84795_1_2.fastq.gz SL84795_1_1.trim.R1.fq OUT.trim.unp1.fq SL84795_1_2.trim.R1.fq OUT.trim.unp2.fq ILLUMINACLIP:/tools/trimmomatic-0.36/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
java -jar /tools/trimmomatic-0.36/trimmomatic-0.36.jar PE -threads 20 -phred33 SL84795_2_1.fastq.gz SL84795_2_2.fastq.gz SL84795_2_1.trim.R1.fq OUT.trim.unp1.fq SL84795_2_2.trim.R1.fq OUT.trim.unp2.fq ILLUMINACLIP:/tools/trimmomatic-0.36/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
java -jar /tools/trimmomatic-0.36/trimmomatic-0.36.jar PE -threads 20 -phred33 SL84796_1_1.fastq.gz SL84796_1_2.fastq.gz SL84796_1_1.trim.R1.fq OUT.trim.unp1.fq SL84796_1_2.trim.R1.fq OUT.trim.unp2.fq ILLUMINACLIP:/tools/trimmomatic-0.36/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
java -jar /tools/trimmomatic-0.36/trimmomatic-0.36.jar PE -threads 20 -phred33 SL84796_2_1.fastq.gz SL84796_2_2.fastq.gz SL84796_2_1.trim.R1.fq OUT.trim.unp1.fq SL84796_2_2.trim.R1.fq OUT.trim.unp2.fq ILLUMINACLIP:/tools/trimmomatic-0.36/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

1.2 Mate-pair libraries

Mate-pair reads were trimmed and sorted using `NxTrim v0.3.1`, and only “mp” (true mate-pair reads) and “unknown” (mostly large insert size reads) reads were used for downstream scaffolding analysis.

```bash
./nxtrim -1 ~/Lamellibrachia/mate/C62UMANXX_s5_1_7bp_Index_2_SL85812.fastq.gz -2 ~/Lamellibrachia/mate/C62UMANXX_s5_2_7bp_Index_2_SL85812.fastq.gz -rf --separate -O SL85812
./nxtrim -1 ~/Lamellibrachia/mate/C62UMANXX_s5_1_7bp_Index_2_SL85813.fastq.gz -2 ~/Lamellibrachia/mate/C62UMANXX_s5_2_7bp_Index_2_SL85813.fastq.gz -rf --separate -O SL85813
cat /scratch/yzl0084/clean/SL85812_mp_R2.fq ~/Lamellibrachia/SL85812_unknown_R2.fq >SL85812_nxtrim_R2.fq
cat /scratch/yzl0084/clean/SL85813_mp_R2.fq ~/Lamellibrachia/SL85813_unknown_R2.fq >SL85813_nxtrim_R2.fq
```

2. Genome assembly

Given high heterozygosity in non-model species, all reads were assembled using `Platanus v1.2.4`
```bash
~/bin/platanus assemble -k 31 -o Lamellibrachia_31 -f ./test/SL*/SL8479{5-6}* -u 0.2 -d 0.3 -t 60 -m 1000
```

Gaps in scaffolds were filled with **GapCloser v1.12**.

```
GapCloser -b config -a Lamellibrachia.fa -o Lamellibrachia_gapclosed.fa -t 20
```

Redundant allele scaffolds were removed using **Redundans v0.13c**.

```
python2.7 ~/programs/redundans/redundans.py -v --nogapclosing --noscaffolding -f final_gapclosed.fa -o Lamellibrachia_redundant_final
```

Genome assembly quality was assessed with **QUAST v4.5**.

```
python ~/programs/quast-4.5/quast.py -o ark_quast -t 1 assembly.fa
```

Genome completeness with **BUSCO v3** using the Metazoa_odb9 database (978 Busco genes).

```
python2.7 ~/programs/busco/scripts/run_BUSCO.py -c 20 -o $FILENAME.out -i $FILENAME -l ~/programs/busco/metazoa_odb9/ -m geno
```

### 3. Genome annotation

Gene models were constructed following the Funannotate pipeline 1.3.0 ([https://github.com/nextgenusfs/funannotate](https://github.com/nextgenusfs/funannotate))
4. Orthology groups

Following all-to-all Diamond v1.0, BLASTP searches against 22 selected lophotrochozoan proteomes, orthology groups (OGs) were identified using Orthofinder with a default inflation parameter (I=1.5).

```
~/programs/OrthoFinder-2.2.1/orthofinder -f . -S diamond -t 20 -M msa -os
```

Gene families were annotated using pantherscore perl script.

```
~/programs/pantherScore2.1/pantherScore2.1.pl -l /tools/funannotate-1.5.0/db/PANTHER13.1/ -D B -V -i final.fa -o OG0000040.fa.tsv -n -s -c 20
```

Gene family expansion was followed the tutorial of CAFE (Here you can find the CAFE TUTORIAL: http://evomicsorg.wpengine.netdna-cdn.com/wp-content/uploads/2016/06/cafe_tutorial-1.pdf).

5. Manual annotation of gene families with potential interest

For gene family with particular interests, targeted genes were additionally processed through the Extract_Homologs2 script used in (Tassia et al. 2017) (available at https://github.com/mtassia/Homolog_identification (https://github.com/mtassia/Homolog_identification)).

We used MAFFT 7.2.15 to align Hb amino acid sequences.
for FILENAME in *.fasta; do mafft --auto --localpair --maxiterate 1000 $FILENAME > $FILENAME.aln; done

Maximum likelihood analyses were performed in IQTree v1.5.

```bash
iqtree -nt 8 -m [best-fit_model] -s [alignment_file] -pre [output_name]
```

### 6. Potential genes under positive selection.

We used TranslaterX to align nucleotide sequences of genes based on codon positions of each orthogroup.

```bash
for FILENAME in *.fasta; do translaterx.pl -p F -i $FILENAME -o $FILENAME -c 4; rm -rf *.aa_based_codon_coloured.html; rm -rf *nt[1-3]*; rm -rf *aaseqs*; rm -rf *.log; rm -rf *.html; done
```

Maximum likelihood analyses were performed in IQTree v1.5

```bash
iqtree -nt 8 -m [best-fit_model] -s [alignment_file] -pre [output_name]
```

Positive selection was calculated using the adaptive branch-site random effects likelihood (aBS-REL) model implemented in HyPhy.

### 7. Molecular clock of Siboglinidae.

Supermatrix dataset was constructed using Agalma.
Putative orthologous groups (OGs) were retrieved from each transcriptome following bioinformatics pipelines of Kocot et al. (2011) and Whelan et al. See Li et al. (2017) for more details (https://doi.org/10.1111/zsc.12201).