Data in brief 26 (2019) 104405

Contents lists available at ScienceDirect

Data in brief

journal homepage: www.elsevier.com/locate/dib

Data Article

# High-throughput amplicon sequencing datasets of the metacommunity DNA of the gut microbiota of naturally occurring and laboratory aquaculture green sea urchins *Lytechinus variegatus*

Joseph A. Hakim [a, *], Casey D. Morrow [b, *], Stephen A. Watts [a], Asim K. Bej [a, *]

[a] *Department of Biology, The University of Alabama at Birmingham, 1300 University Blvd., Birmingham, AL 35294, USA*
[b] *Department of Cell, Developmental and Integrative Biology, The University of Alabama at Birmingham, 1918 University Blvd., Birmingham, AL 35294, USA*

ARTICLE INFO

*Article history:*
Received 9 May 2019
Received in revised form 8 August 2019
Accepted 9 August 2019
Available online 20 August 2019

*Keywords:*
Echinoderm
Arcobacter
PhyloToAST
QIIME
MiSeq

ABSTRACT

We present high-throughput amplicon sequence (HTS) datasets of the microbial metacommunity DNA of the gut tissue and the gut digesta of naturally occurring ($n = 3$) and laboratory aquaculture ($n = 2$) green sea urchins, *Lytechinus variegatus*. The HTS datasets were generated on an Illumina MiSeq by targeting the amplicons of the V4 region of the 16S rRNA gene. After the raw sequences were quality checked and filtered, 88% of the sequence reads were subjected to bioinformatics analyses to generate operation taxonomic units (OTUs), which were then verified for saturation by using rarefaction analysis at a 3% sequence variation. Further, the OTUs were randomly subsampled to the minimum sequence count values. Then, the FASTA-formatted representative sequences of the microbiota were assigned taxonomic identities through multiple databases using the SILVA ACT: Alignment, Classification and Tree Service (www.arb-silva.de/aligner). The HTS datasets of this metagenome can be accessed from the BioSample Submission

* Corresponding authors.
*E-mail addresses:* joe21@uab.edu (J.A. Hakim), caseym@uab.edu (C.D. Morrow), abej@uab.edu (A.K. Bej).

https://doi.org/10.1016/j.dib.2019.104405

2352-3409/© 2019 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Portal (https://www.ncbi.nlm.nih.gov/bioproject/) under the BioProject IDs PRJNA291441 and PRJNA326427.

© 2019 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Specifications Table

| | |
|---|---|
| Subject area | *Biology* |
| More specific subject area | *Metagenomics* |
| Type of data | *Figures and Tables* |
| How data was acquired | *Illumina MiSeq platform with 250 paired-end kits.* |
| Data format | *Raw, analyzed* |
| Experimental factors | *Laboratory aquaculture (LAB)* Lytechinus variegatus *(n = 2) were collected from Port Saint Joseph, Florida (29.80° N 85.36° W), and held in the laboratory aquaculture condition, fed with a formulated diet for six months prior to investigation. Naturally occurring (ENV)* Lytechinus variegatus *(n = 3) were collected from the same location and sample preparation began immediately upon arrival to the University of Alabama at Birmingham (UAB)* |
| Experimental features | *Targeted high-throughput sequencing of the microbial metacommunity 16S rRNA gene (V4 hypervariable regions) using the Illumina MiSeq with 250 paired-end kits followed by bioinformatics analyses.* |
| Data source location | Lytechinus variegatus *collected from Port Saint Joseph, Florida, USA (29.80° N 85.36° W) located in the Gulf of Mexico.* Lytechinus variegatus *were maintained in laboratory aquaculture condition at the UAB Biology Department, 1300 University Blvd., Birmingham, AL 35294, USA. Microbial metacommunity DNA was prepared and sequenced at the UAB Department of Genetics, Heflin Center Genomics Core, School of Medicine, the University of Alabama at Birmingham, 705 South 20th Street, Birmingham, AL 35294, USA.* |
| Data accessibility | *Raw data corresponding to the 10 samples are available at the NCBI's BioSample database following this link:*<br>*http://www.ncbi.nlm.nih.gov/sra/?term=sea+urchin+gut+microbiome*<br>*For the LAB group, the BioProject number is PRJNA291441 and the BioSample IDs are SAMN03944319, SAMN03944320, SAMN03944321, SAMN03944322. For the ENV group, the BioProject number is PRJNA326427 and the BioSample IDs are SAMN05277844, SAMN05277845, SAMN05277846, SAMN05277847, SAMN05277848, SAMN05277849, and SAMN05277850.* |

**Value of the data**

- These HTS datasets would help expand our knowledge of the source, distribution, selection, and nutritional benefit of the gut microbial communities in diverse species of marine echinoderms, and other marine invertebrates at various trophic levels.
- The metagenome data provide for the first time an insight into the modulation of gut microbiota of laboratory aquaculture sea urchins fed with a standard formulated reference diet at the highest possible taxonomic coverage.
- Access to the raw files of these HTS data permits researchers to apply their own bioinformatics analyses, based on their exploratory goals.

## 1. Data

The metagenomic datasets presented in this article describe the microbial community compositions in the gut ecosystem of a marine invertebrate echinoderm of ecological, economic, and scientific importance, *Lytechinus variegatus*, fed with formulated diet in laboratory aquaculture conditions and from their natural habitat. Fig. 1 describes the OTUs from the quality-checked and filtered HTS data of the 16S rRNA gene and visualized by rarefaction analysis, which indicated that the total quality

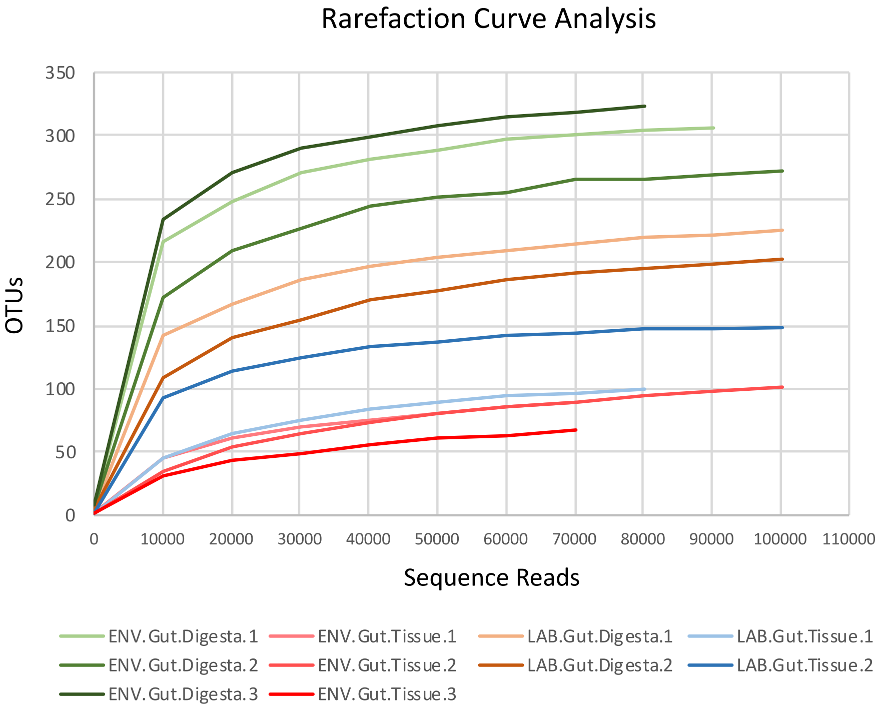

**Fig. 1.** Rarefaction curve analysis of the HTS data showing the number of OTUs (Y-axis) plotted against total number of sequences (X-axis) per sample. OTUs were determined by using the PhyloToAST (v1.4) taxonomy condensing workflow, which is integrated into QIIME (v1.9.1). Samples were rarefied to the minimum sequence count across all samples for downstream bioinformatics analysis. Data were plotted using Microsoft Excel Software (Seattle, WA, USA).

sequences from each sample are approaching saturation when constructed at a 3% sequence variation. Fig. 2 shows the microbial community profiles of both LAB and ENV gut tissue with a near-exclusive abundance of Epsilonproteobacteria, with the ENV group showing a slightly higher diversity. Table 1 presents the applicability of the HTS datasets for profiling *Lytechinus variegatus* gut microbiota at the highest possible taxonomic levels following the alignment of the representative sequences to five microbial databases.

## 2. Experimental design, materials and methods

### 2.1. Sample description

The sea urchins were collected from Saint Joseph Bay Aquatic Preserve of the U.S. Gulf of Mexico (29.80° N 85.36° W). For the laboratory aquaculture (LAB) group [1], adult sea urchins ($n = 2$) were kept in a recirculating saltwater tank system for six months, and fed a formulated feed *ad libitum* once every 24–48 h that consisted of 6% lipid, 28% protein, and 36% carbohydrate relative percentages [2]. The aquaria were maintained at 22 ± 2 °C with a pH of 8.2 ± 0.2 and salinity of 32 ± 1 ppt. For the naturally occurring (ENV) group [3], adult sea urchins ($n = 3$) were collected from within the same 1 $m^2$ area and transported to the laboratory at the University of Alabama at Birmingham (UAB) for sample collection. Water conditions were recorded as 20 ± 2 °C with a pH of 7.8 ± 0.2 and salinity of 28 ± 1 ppt. For both groups, the Illumina MiSeq high throughput-sequencing (HTS) platform was used with the 250 bp paired-end kits targeting the V4 hypervariable region [4,5]. The paired-end raw sequence data were

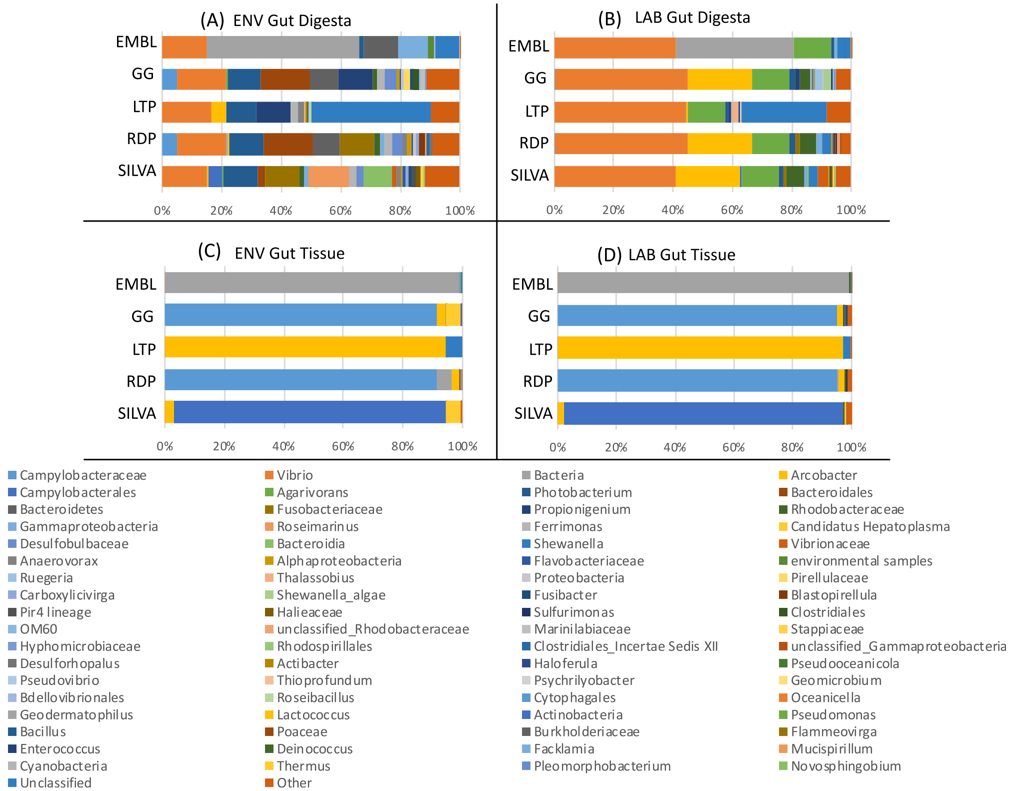

**Fig. 2.** Relative abundance distribution of taxa at the highest resolution determined for the merged biological replicates using multiple taxonomic databases. The FASTA-formatted representative sequences determined by the PhyloToAST (v1.4) workflow integrated into QIIME (v1.9.1) were aligned to multiple databases using the SILVA ACT: Alignment, Classification and Tree Service (www.arb-silva.de/aligner). Taxonomic assignments were performed using the SSU (Small Subunit) category and the Least Common Ancestor (LCA) method with the following databases: SILVA, Ribosomal Database Project (RDP), The All-Species Living Tree (LTP) project, Greengenes (GG), and the European Molecular Biology Laboratory (EMBL). Sequences aligned with a similarity threshold below 70% were discarded. The top 25 taxa from each database were merged based on their common taxonomic assignments at the specific level of classification.

demultiplexed and formatted into FASTQ files [6]. The raw data were deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under BioProject #PRJNA291441 and #PRJNA326427 for the LAB and ENV group, respectively. The paired-end sequence data for the gut microbial communities can be accessed under the following NCBI BioSample Ids: SAMN03944319 -

**Table 1**
Statistical analysis of the representative sequences aligned to multiple databases using the SILVA ACT: Alignment, Classification and Tree Service (www.arb-silva.de/aligner). Taxonomic assignments were performed using the SSU (Small Subunit) category and the Least Common Ancestor (LCA) method with the following databases: SILVA, Ribosomal Database Project (RDP), The All-Species Living Tree (LTP) project, Greengenes (GG), and the European Molecular Biology Laboratory (EMBL). Sequences aligned at a similarity threshold below 70% were discarded. For each database, the total number of uniquely assigned sequences were determined, and the fraction of those assignments to the family and the genus level were listed.

| Level | SILVA | RDP | LTP | GG | EMBL |
|---|---|---|---|---|---|
| Family | 234 | 193 | 128 | 219 | 18 |
| | 82.98% | 80.08% | 100.00% | 76.04% | 54.55% |
| Genus | 167 | 147 | 121 | 132 | 12 |
| | 59.22% | 61.00% | 94.53% | 45.83% | 36.36% |
| Total Unique | 282 | 241 | 128 | 288 | 33 |

SAMN03944322 (LAB group) and SAMN05277845 - SAMN05277850 (ENV group). Subgroups for the laboratory-fed group are as follows: LAB.Gut.Tissue ($n = 2$), LAB.Gut.Digesta ($n = 2$), ENV.Gut.Tissue ($n = 3$), and ENV.Gut.Digesta ($n = 3$).

### 2.2. *Quality assessment and filtering*

The raw and demultiplexed paired-end sequence datasets were initially assessed by FastQC [7], and only reads showing 80% of bases at a Q score of >33 were retained by using the "fastx_trimmer" command from the FASTX Toolkit [5,8] and merged using USEARCH [9]. Paired-end reads with <50 base overlap and/or >20 mismatching nucleotides were filtered from the analysis, and chimeric sequences were removed using USEARCH [9].

### 2.3. *Taxonomic distribution and alpha diversity*

The merged sequence data was analyzed using Quantitative Insights into Microbial Ecology (QIIME; v1.9.1) along with Phylogenetic Tools for Analysis of Species-level Taxa (PhylotoAST; v1.4.0) [10,11]. The initial OTUs were clustered at a 97% similarity through UCLUST in QIIME (v1.9.1) [9], and representative sequences were established by the "most_abundant" option. Then, OTUs with <0.0005% average abundance across all samples were filtered. The redundant OTUs were merged by using the "condense_workflow.py" command through PhyloToAST (v1.4.0) [11]. The OTUs per sample were plotted against the filtered sequence read counts as rarefaction curves, and the data was subsampled to the minimum value using "single_rarefaction.py" in QIIME (v1.9.1). The representative sequences were then assigned taxonomy using the SILVA ACT: Alignment, Classification and Tree Service (www.arb-silva.de/aligner), which utilizes the SILVA Incremental Aligner (SINA; v1.2.11) to align rRNA gene sequences and classify based on Least Common Ancestor (LCA) methods [12]. For this, the SSU (Small Sub-Unit) option selected at a minimum similarity of 0.7 with 20 neighbors per query sequence, and the databases selected were as follows: SILVA database [13], Ribosomal Database Project (RDP) [14], All-Species Living Tree (LTP) project [15], Greengenes (GG) [16,17], and European Molecular Biology Laboratory (EMBL) [18]. Biological replicates were validated and merged according to their subgroup assignment based on significant Analysis of Similarity (ANOSIM) [19] and Adonis [20] measurements ($p = 0.001$) using the weighted Unifrac distances [21] calculated for each sample. The top 25 taxa at the highest resolution from each database were combined and plotted as relative abundance graphs using Microsoft Excel Software (Seattle, WA, USA). The taxonomic data derived from each of the five databases is summarized in Table 1 showing the total number of OTUs that were assigned a taxonomy, including the proportion that was resolved to the family and the genus level.

## Acknowledgments

We would like to thank Dr. Peter Eipers of the Department of Cell, Developmental and Integrative Biology, and Dr. Michael Crowley of the Heflin Center for Genomic Sciences at the University of Alabama at Birmingham (UAB), for their assistance in high-throughput sequencing. Graduate Research Assistant funding to J.A.H. was provided from grant support to C.D.M by the UAB School of Medicine. The following are acknowledged for their support of the Microbiome Resource at the University of Alabama at Birmingham: School of Medicine, Comprehensive Cancer Center (P30AR050948; C.D.M.), Center for Clinical Translational Science (UL1TR000165; C.D.M.), UAB Microbiome Center (C.D.M.), and Heflin Center. Animal husbandry supported in part by NIH P30DK056336 (S.A.W.).

## Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] J.A. Hakim, H. Koo, L.N. Dennis, R. Kumar, T. Ptacek, C.D. Morrow, E.J. Lefkowitz, M.L. Powell, A.K. Bej, S.A. Watts, An abundance of Epsilonproteobacteria revealed in the gut microbiome of the laboratory cultured sea urchin, *Lytechinus variegatus*, Front. Microbiol. 6 (2015) 1047.

[2] H. Hammer, B. Hammer, S. Watts, A. Lawrence, J. Lawrence, The effect of dietary protein and carbohydrate concentration on the biochemical composition and gametogenic condition of the sea urchin *Lytechinus variegatus*, J. Exp. Mar. Biol. Ecol. 334 (1) (2006) 109–121.

[3] J.A. Hakim, H. Koo, R. Kumar, E.J. Lefkowitz, C.D. Morrow, M.L. Powell, S.A. Watts, A.K. Bej, The gut microbiome of the sea urchin, *Lytechinus variegatus*, from its natural habitat demonstrates selective attributes of microbial taxa and predictive metabolic profiles, FEMS Microbiol. Ecol. 92 (9) (2016) fiw146.

[4] J.J. Kozich, S.L. Westcott, N.T. Baxter, S.K. Highlander, P.D. Schloss, Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform, Appl. Environ. Microbiol. 79 (17) (2013) 5112–5120.

[5] R. Kumar, P. Eipers, R.B. Little, M. Crowley, D.K. Crossman, E.J. Lefkowitz, C.D. Morrow, Getting started with microbiome analysis: sample acquisition to bioinformatics, Curr. Protoc. Hum. Genet. 82 (1) (2014) 18.8.1–18.8.29.

[6] P.J. Cock, C.J. Fields, N. Goto, M.L. Heuer, P.M. Rice, The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants, Nucleic Acids Res. 38 (6) (2009) 1767–1771.

[7] S. Andrews, FastQC: a Quality Control Tool for High Throughput Sequence Data, 2010 [cited 2019]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc.

[8] A. Gordon, G. Hannon, Fastx-toolkit. FASTQ/A Short-Reads Pre-processing Tools, 2010 [cited 2019]. Available from: http://hannonlab.cshl.edu/fastx_toolkit.

[9] R.C. Edgar, Search and clustering orders of magnitude faster than BLAST, Bioinformatics 26 (19) (2010) 2460–2461.

[10] J.G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F.D. Bushman, E.K. Costello, N. Fierer, A.G. Peña, J.K. Goodrich, J.I. Gordon, G.A. Huttley, S.T. Kelley, D. Knights, J.E. Koenig, R.E. Ley, C.A. Lozupone, D. McDonald, B.D. Muegge, M. Pirrung, J. Reeder, J.R. Sevinsky, P.J. Turnbaugh, W.A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, R. Knight, QIIME allows analysis of high-throughput community sequencing data, Nat. Methods 7 (5) (2010) 335.

[11] S.M. Dabdoub, M.L. Fellows, A.D. Paropkari, M.R. Mason, S.S. Huja, A.A. Tsigarida, P.S. Kumar, PhyloToAST: bioinformatics tools for species-level analysis and visualization of complex microbial datasets, Sci. Rep. 6 (2016) 29123.

[12] E. Pruesse, J. Peplies, F.O. Glöckner, SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes, Bioinformatics 28 (14) (2012) 1823–1829.

[13] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, F.O. Glöckner, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools, Nucleic Acids Res. 41 (D1) (2012) D590–D596.

[14] J.R. Cole, Q. Wang, J.A. Fish, B. Chai, D.M. McGarrell, Y. Sun, C.T. Brown, A. Porras-Alfaro, C.R. Kuske, J.M. Tiedje, Ribosomal Database Project: data and tools for high throughput rRNA analysis, Nucleic Acids Res. 42 (D1) (2013) D633–D642.

[15] P. Yilmaz, L.W. Parfrey, P. Yarza, J. Gerken, E. Pruesse, C. Quast, T. Schweer, J. Peplies, W. Ludwig, F.O. Glöckner, The SILVA and “all-species living tree project (LTP)” taxonomic frameworks, Nucleic Acids Res. 42 (D1) (2013) D643–D648.

[16] T.Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E.L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, G.L. Andersen, Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB, Appl. Environ. Microbiol. 72 (7) (2006) 5069–5072.

[17] D. McDonald, M.N. Price, J. Goodrich, E.P. Nawrocki, T.Z. DeSantis, A. Probst, G.L. Andersen, R. Knight, P. Hugenholtz, An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea, ISME J. 6 (3) (2012) 610.

[18] C. Kanz, P. Aldebert, N. Althorpe, W. Baker, A. Baldwin, K. Bates, P. Browne, A. van den Broek, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, J. Gamble, F. Garcia Diez, N. Harte, T. Kulikova, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, M. McHale, F. Nardone, V. Silventoinen, S. Sobhany, P. Stoehr, M.A. Tuli, K. Tzouvara, R. Vaughan, D. Wu, W. Zhu, R. Apweiler, The EMBL nucleotide sequence database, Nucleic Acids Res. 33 (suppl_1) (2005) D29–D33.

[19] K.R. Clarke, Non-parametric multivariate analyses of changes in community structure, Austral Ecol. 18 (1) (1993) 117–143.

[20] J. Oksanen, R. Kindt, P. Legendre, B. O'Hara, G.L. Simpson, P. Solymos, M.H.H. Stevens, H. Wagner, The vegan package, Community Ecology Package 10 (2007) 631–637.

[21] C. Lozupone, M.E. Lladser, D. Knights, J. Stombaugh, R. Knight, UniFrac: an effective distance metric for microbial community comparison, ISME J. 5 (2) (2011) 169.