Oligotyping analysis of *E. coli* K12 strain

A. Murat Eren [*], **Lois Maignien, Woo Jun Sul, Leslie G. Murphy, Sharon L. Grim, Hilary G. Morrison and Mitchell L. Sogin**

[*] Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543 USA

Appendix S2 for the manuscript titled "Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data" published in Methods in Ecology and Evolution; doi:10.1111/2041-210X.12114

We performed oligotyping and OTU clustering analyses on 43,182 16S rRNA gene tag reads obtained by sequencing the V6 through V4 hypervariable region of Escherichia coli K12, substrain DH10B. For oligotyping analysis we used the five highest entropy components with no noise reduction. For OTU clustering we used 95% through 99% similarity thresholds to recover two groups of reads representing the two variants of operons in the E. coli K12 genome.

OTU clustering analysis of 43,182 quality-controlled V6-V4 reads resulted in 28, 53, 100, 223, and 839 clusters at 95%, 96%, 97%, 98%, and 99% similarity thresholds respectively (Fig. 1). The percentage of reads retained in the two most abundant clusters at each similarity threshold were 99.7%, 98.9%, 97.1%, 96.2%, and 81.9% respectively. At 99% similarity the third largest OTU contained 6.8% of reads.

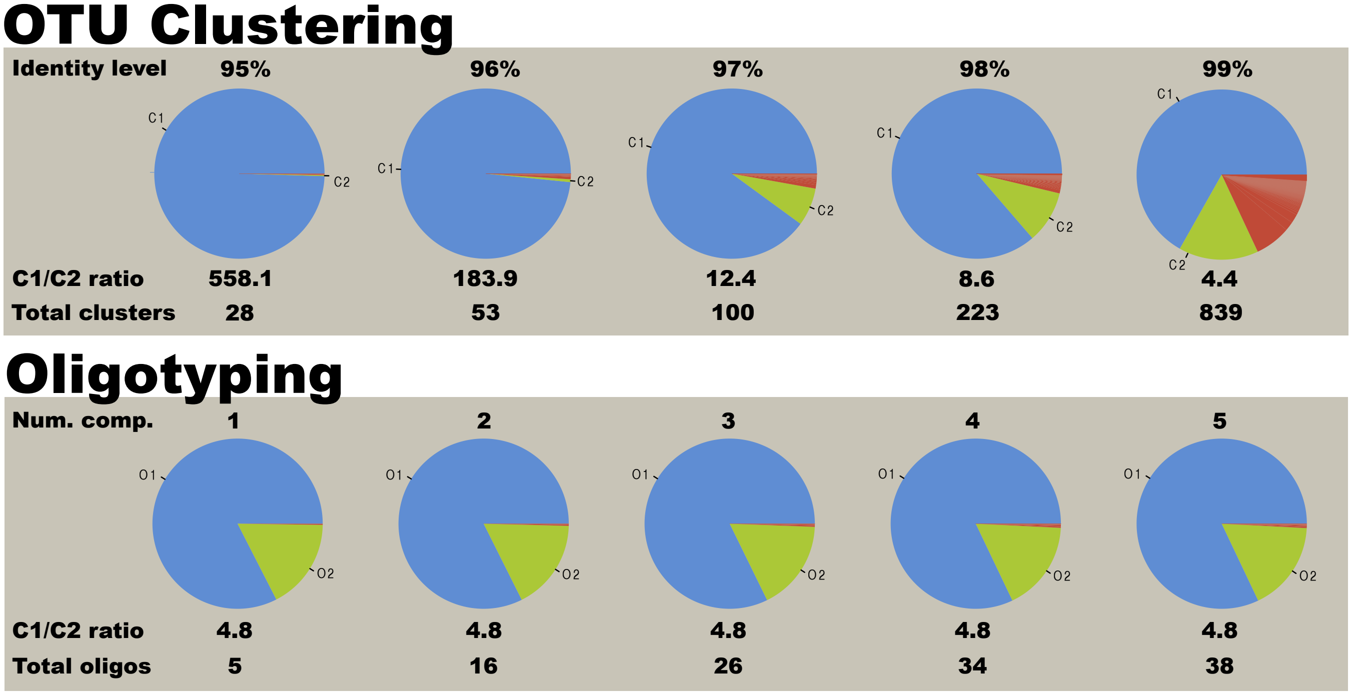

Fig. 1. OTU clustering and oligotyping results for 43,182 reads obtained from *E. coli* strain K12 substrain DH10B. Reads were clustered at 95%, 96%, 97%, 98%, and 99% clustering identity levels. Oligotyping was conducted iteratively using the five highest entropy components. In each pie chart blue and green color codes mark the most abundant and second most abundant clusters. All other clusters are colored as red. C1/C2 ratio is the ratio of the proportions of the most abundant cluster to the second most abundant cluster. Total operational units obtained through each iteration are also shown below each pie chart.

Oligotyping analysis of the same dataset resulted in 5, 16, 26, 34, and 48 oligotypes for 1, 2, 3, 4, and 5 entropy components used to identify different groups of reads respectively. The percentage of reads retained in the two most abundant oligotypes were 99.8%, 99.6%, 99.4%, 99.2% and 99.2% (Fig. 1).

The published genome of *E. coli* strain K12 substrain DH10B has 7 copies of 16S rRNA gene in its genome, 6 of which are identical and possess more than 98% sequence similarity with the rrnH copy at the V4-V6 region (NCBI accession: CP000948.1, [5]). Within this fragment of the 16S gene, heterogeneity between the two variants is located in the V6 hypervariable region. Quantitative PCR determined our strain has 7 copies of the 16S rRNA gene. Tag sequencing of our *E. coli* strain recovered the expected two variants of 16S.

In an ideal case, a similarity-based analysis of *E. coli* V4-V6 region at 99% sequence similarity level should result in two major OTUs with an abundance ratio close to 1 to 6, while any integer below this similarity threshold should recover only one OTU. Expectedly, the first recovery of two clusters with a ratio reasonably close to the expected range occurs at clustering *E. coli* reads 99% (Fig. 1). Yet, there is a substantial identification of artificial clusters persist even at 95% similarity threshold. The striking variation in the number of clusters and abundances of recovered groups at different identity levels highlight the disadvantage of performing such pairwise comparisons of closely related taxa.

Oligotyping analysis recovered both versions of the 16S rRNA gene. Using only a fraction of reads and five nucleotide locations with the highest variation to identify meaningful groups, oligotyping delivers more reasonable results with remarkably smaller inflation of artificial sequence types compared to OTU clustering. The impact of increasing the number of components is small, and oligotyping retains a stable ratio of the most abundant two oligotypes closer to the expected ratio. Though both versions of the 16S operon were recovered through the two most abundant oligotypes, we calculated the relative abundances of the variants to be close to 5:1 instead of 6:1. The skewed abundance of these copies naturally present in the *E. coli* genome as well as the high degree of similarity may have led to chimeric formations in the PCR reaction to generate the V4-V6 amplicons. Chimeric formation rates were observed to be higher for less abundant copies of 16S rRNA genes than for more abundant copies[6]. Additionally, the extent of intercopy chimeric formation can be substantially and directly related to the level of sequence similarity[6][7]. The 16S rRNA gene primer set used in qPCR amplifies an identical target in both variants, so chimeric formation would not be a concern in this assay. Since our input into amplicon PCR was a pure culture with slight mismatch between copy variants across the target region, it is likely that chimeras were a contributing factor to the altered 16S rRNA gene variant proportion observed in both clustering methods. Compared to OTU clustering, oligotyping yields a more reasonable ratio of operon variants by retaining substantially more reads in higher-resolved and fewer clusters, suggesting that the oligotyping is more suitable to explain the diversity of very closely related taxa.

Methods

A substrain of *E. coli* K12, originating from DH10B competent cells (Invitrogen, Foster City CA), was grown at 37C in Luria-Bertani broth and extracted using a Gentra Puregene Yeast/Bacteria Kit (Qiagen, Valencia, CA). Ribosomal RNA amplicons spanning the 16S V4 through V6 regions of *E. coli* K12 were amplified using fusion primers, sequenced on a Roche GS-FLX 454 instrument using Titanium protocols, and quality-filtered and trimmed as described in Marteinsson et al., 2012 [1]. For clustering analysis we used QIIME v1.5 [2] with UCLUST and the default parameters. For oligotyping analysis we aligned reads against the GreenGenes [3] alignment template (97% OTUs, October 6, 2010 release) using PyNAST [4]. Following the entropy analysis on the alignment, we conducted five separate runs of the oligotyping analysis on the same dataset. We allowed the oligotyping pipeline to choose up to five highest entropy positions to be used for oligotyping in each individual run. We did not perform any noise reduction.

We used quantitative PCR to determine 16S operon copy number in our *E. coli* genome. Specific primers were used to amplify approximately equal-length portions of the 16S gene (rrn) and a single-copy gene (dxs) [8]. For each primer set, approximately 5 nanograms of genomic DNA served as template with 0.5 uM final concentration of each primer, 1.5 mM MgSO4, 0.25 mM each dNTPs, and 1.2 U Platinum Taq DNA Polymerase Hi Fidelity (Invitrogen, Foster City, CA), in 25 uL reactions. Thermal cycling conditions were as described in Lee et al. [8]. The rrn and dxs amplicons were purified using MinElute spin columns (Qiagen), quantified with Picogreen (Invitrogen), and diluted in a ten-fold serial dilution series to construct standard curves. Duplicate 10 uL real-time quantitative PCR reactions were performed in a StepOnePlus instrument (Applied Biosystems, CA) using 0.5 ng template, 0.2 uM of each primer, and KAPA SYBR FAST Master Mix (KAPA Biosystems, Boston, MA). Cycling conditions were an initial activation step of 95C for 3 minutes, followed by 30 cycles of 60C for 20 sec and 72C for 10 sec. Fluorescence was measured at the end of each 72C extension step. A melt curve analysis of 0.1C/5 sec from 60 to 95C was performed after amplification to confirm specific amplification. Absolute quantification via the standard curves as well as relative quantification[9] verified a ratio of 6.9 copies of 16S rRNA gene to 1 copy of dxs (data not shown).

1. Marteinsson, V.T., Runarsson, A., Stefansson, A., Thorsteinsson, T., Johannesson, T., Magnusson, S.H., Reynisson, E., Einarsson, B., Wade, N., Morrison, H.G.Gaidos, E. (2013) Microbial communities in the subglacial waters of the Vatnajokull ice cap, Iceland. ISME J, 7, 427-437.
2. Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J.Knight, R. (2010b) QIIME allows analysis of high-throughput community sequencing data. Nat Methods, 7, 335-336.
3. McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R.Hugenholtz, P. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J, 6, 610-618.
4. Caporaso, J.G., Bittinger, K., Bushman, F.D., DeSantis, T.Z., Andersen, G.L.Knight, R. (2010a) PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics, 26, 266-267.
5. Durfee, T., Nelson, R., Baldwin, S., Plunkett, G., 3rd, Burland, V., Mau, B., Petrosino, J.F., Qin, X., Muzny, D.M., Ayele, M., Gibbs, R.A., Csorgo, B., Posfai, G., Weinstock, G.M.Blattner, F.R. (2008) The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse. J Bacteriol, 190, 2597-2606.
6. Wang, G.C.Wang, Y. (1997) Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. Appl Environ Microbiol, 63, 4645-4650.
7. Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D.V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S.K., Sodergren, E., Methe, B., DeSantis, T.Z., Petrosino, J.F., Knight, R.Birren, B.W. (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res, 21, 494-504.
8. Lee, C., Lee, S., Shin, S.G.Hwang, S. (2008) Real-time PCR determination of rRNA gene copy number: absolute and relative quantification assays with Escherichia coli. Appl Microbiol Biotechnol, 78, 371-376.
9. Livak, K.J.Schmittgen, T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods, 25, 402-408.