The Flowchart of Oligotyping and an Example Analysis to Highlight Concepts and Best Practices

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

Here we elaborate the steps of the oligotyping analysis by providing a flowchart and demonstrate the workflow by analyzing a mock dataset with the oligotyping pipeline version 0.96. Please see the oligotyping paper for read preparation and noise filtering.

Flowchart

A successful oligotyping analysis 1) identifies only those nucleotide positions necessary to explain the maximum amount of biological diversity represented by a dataset of closely related sequences, and 2) generates *converged* oligotypes.

Fig. 1 illustrates a flowchart for oligotyping analysis inclusive of three critical steps (1-3) as detailed below.

Box 1: Unexplained entropy?

The first oligotyping step performs an initial entropy analysis on the dataset of closely related sequences to determine whether it potentially contains information for decomposing the data into distinct oligotypes. If the entropy analysis does not identify clear entropy peaks, it suggests that either all reads for the assemblage derive from one identical template that occurs in all genomes of that taxon, or the templates that give rise to distinct sequences correspond to rare genomes that cannot be confidently distinguished from random sequencing errors based on entropy values. Most sequencing errors will randomly distribute along the length of the alignment and appear as white noise that increases towards the end of the reads in entropy profiles. Our empirical observations indicate that random sequencing errors generate entropy values that hover at or below 0.2 for Illumina platforms. Therefore, if all entropy values are below 0.2 for a group of Illumina reads, oligotyping will most likely not help recover ecologically meaningful oligotypes.

Box 2: How to choose "n" for the initial oligotyping?

The oligotyping process should begin by setting the n parameter to select only the highest entropy positions in the first round of oligotype analysis. We recommend initially setting n between 1 and 3, because use of a single site will sometimes resolve multiple entropy peaks into discrete oligotypes.

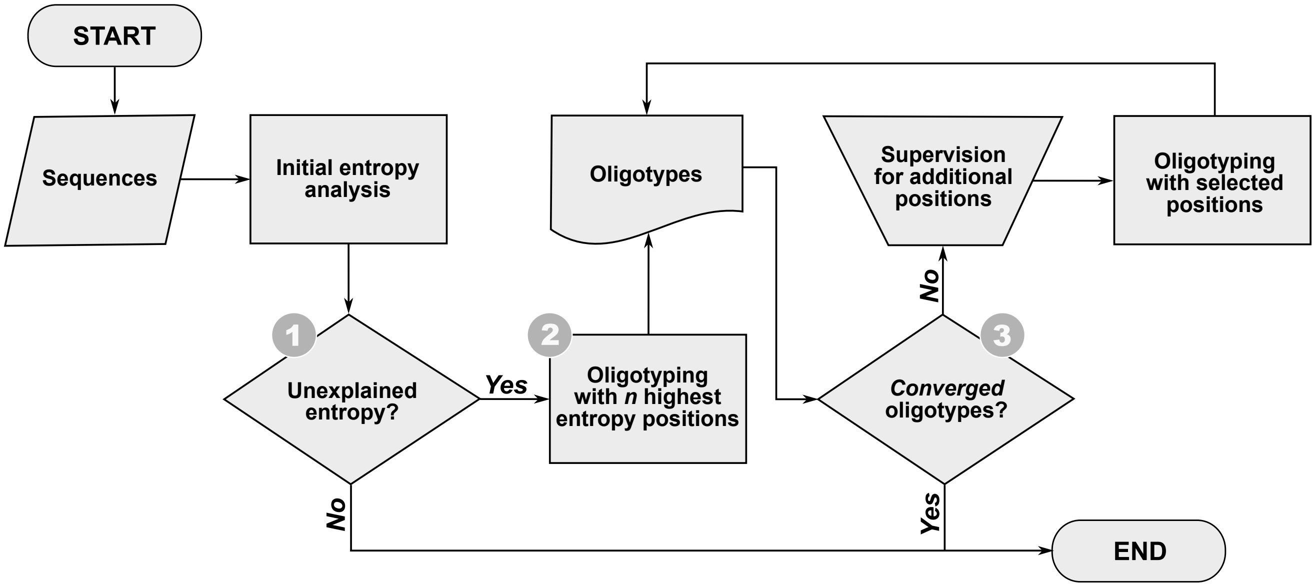
START
Sequences
Initial entropy analysis
1
Unexplained entropy?
Yes
No
2
Oligotyping with n highest entropy positions
Oligotypes
3
Converged oligotypes?
No
Yes
Supervision for additional positions
Oligotyping with selected positions
END

Fig. 1. Flowchart of an oligotyping analysis

Box 3: Converged oligotypes?

The decision whether to stop or continue oligotyping requires the examination of the entropy profiles of each individual oligotype recovered to make sure that the second criterion for a successful oligotyping analysis is met: all oligotypes have *converged*. An oligotype has converged if additional decomposition does not generate new oligotypes that exhibit differential abundances in different samples (or environments). In general, a converged oligotype will not display entropy peaks, however, there may be some high entropy positions within a converged oligotype that do not reflect ecological variation. For instance, if a microbial genome has 7 copies of the 16S rRNA gene and one varies from all others by a single nucleotide, entropy analysis will identify one potentially information-rich site that can resolve into two oligotypes with abundance ratios of 1:6 in every sample. Yet, these oligotypes will not contribute to beta diversity estimates when comparing multiple samples because the 1:6 ration is fixed by genomic content rather than differences in microbial community structures. Another case may be the existence of an entropy peak due to a homopolymer region-associated error. If remaining entropy in an oligotype appears to reflect systematic sequencing errors, the user can abort attempts to further resolve it. The oligotyping pipeline provides tools to examine the convergence of oligotypes including graphical distributions of divergent sequences in an oligotype among samples.

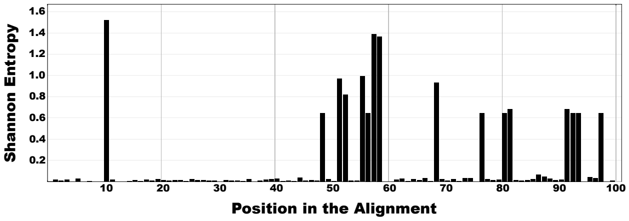

Fig. 2. Entropy analysis results for *mock.fa*

An Example Oligotyping Analysis

The mock dataset we will use here to illustrate the concepts laid out in the previous section derives from a subsampling of a large publicly available human microbiome dataset. Details of this dataset are provided in [1]. The subsampled dataset contains 859,302 reads that were taxonomically classified to the genus *Bacteroides*, and can be retrieved in FASTA format from the URL `http://goo.gl/dpzJ9`. To further simplify the mock dataset, we merged all reads collected from subjects from the same geographical location into one sample (e.g., all data collected from Malawi were merged into one sample named "Malawi"), and the final FASTA file (hereinafter called *mock.fa*) contained three samples: "Malawi" (53,476 reads), "Venezuela" (101,683 reads) and "USA" (704,143 reads).

The entropy profile computed for *mock.fa* using the command "`entropy-analysis mock.fa`" (Fig. 2) shows that most positions are highly conserved for all reads, hence exhibiting very low entropy values. In contrast, the large entropy peaks reveal the existence of several highly variable nucleotide positions that have the potential to identify oligotypes within the dataset.

The initial round of oligotyping in this analysis set n=1 (see Fig. 1, "Box 2: Oligotyping with n highest entropy positions"). The oligotyping pipeline will use the location of the highest entropy peak (position 10) to generate the first round of oligotypes. The exact command to perform this operation is "`oligotype mock.fa mock.fa-ENTROPY -c 1 -M 50 --gen-html`", where mock.fa-ENTROPY is the file that was generated after the initial entropy analysis. The "`--gen-html`" flag generates output that is required for further supervision; see the Methods section in the oligotyping manuscript for the explanation of "`-M`" parameter.

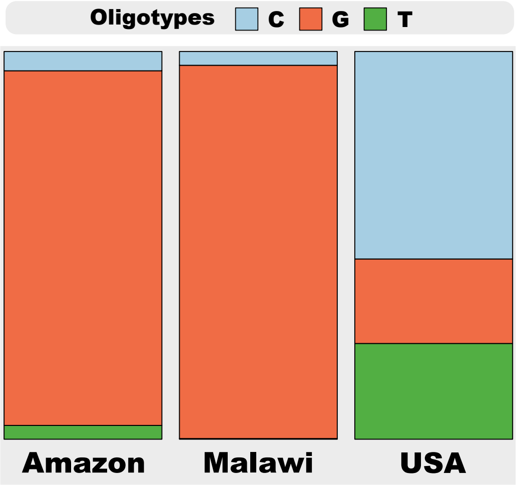

Fig. 3. Oligotype distributions among samples.

Oligotyping analysis of *mock.fa* using the highest entropy location, position 10, identifies 3 oligotypes: C (383,788 reads), G (297,937 reads) and T (177,577 reads). The HTML output for this analysis can be viewed using the web address `http://goo.gl/oY8dD` (we encourage readers to visit this web address and follow the steps described here using the analysis results provided). Fig. 3 shows the distribution profiles of these oligotypes among three samples.

Each oligotype shown in Fig. 3 is composed of reads from the mock dataset that possessed the same nucleotide at the 10th position. Naturally, the entropy is zero at the 10th location of each individual oligotype following the first round of oligotyping. As Fig. 3 indicates, some of the diversity of *Bacteroides* reads is already explained by the oligotypes.

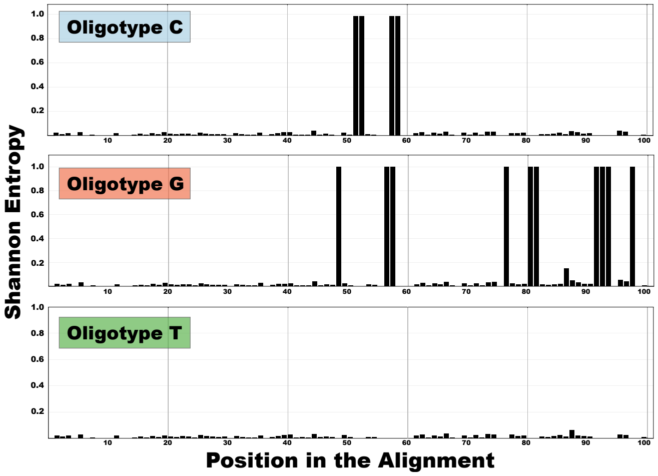

Fig. 4. Entropy profiles for oligotypes C, G and T.

The decision to stop or continue oligotyping brings us to the second criterion for a successful oligotyping analysis: determining whether the oligotypes have converged. Fig. 4 displays Entropy profiles of oligotype C, G and T (abundance curves can also be seen in the HTML output).

Fig. 4 shows that the oligotype T has minimal entropy (below 0.1), which indicates very little variance for all positions across all reads for this oligotype. In contrast, oligotypes C and G exhibit numerous entropy peaks, which indicate that the oligotyping analysis should continue with additional nucleotide positions (along with the previously selected 10th position).

The pipelines output for oligotype C (`http://goo.gl/50Ihp`) reveals additional peaks with similar entropy values at the 51st, 52nd, 57th and 58th positions. Their information content should be equally efficient for resolving the diversity confined in oligotype C. Under this condition, the pipeline facilitates user decisions about site selection for subsequent rounds of oligotyping. For example, the proximal location relative to the start of a sequencing read where quality tends to be high would favor the selection of the 51st site in combination with the 10th position for identifying a new oligotype. The next round of oligotyping could start with 51st and 10th positions. However, the entropy profiles of the remaining first-round oligotypes (in this example, oligotype G) can reveal high entropy sites that are shared among oligotypes and facilitate the convergence of oligotypes with the use of minimal number of additional nucleotide positions.

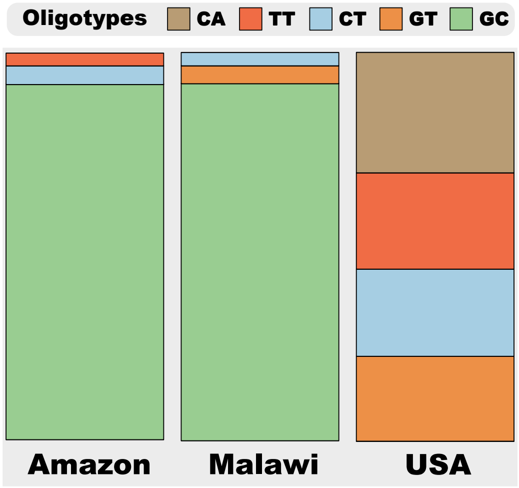

Fig. 5. Oligotype distributions among samples.

Oligotype G's entropy profile (http://goo.gl/ps0LS) reveals that the 48th, 56th, 57th, 76th, 80th, 81st, 91st, 92nd, 93rd and 97th positions exhibit variation and hence information content that can further explain the diversity confined in this oligotype. Similar to the case of oligotype C, the 48th position (closest to the beginning of the alignment) might further partition oligotype G. However, the candidate sites in both oligotypes C and G include the 57th position (see Fig. S8). By using this site, the analysis will require only two positions (10th and 57th), instead of three (10th, 51st and 48th) in the second round of oligotyping. The second round of oligotyping in this example uses the command syntax "`oligotype mock.fa mock.fa-ENTROPY C 10,57 -M 50 --gen-html`" to obtain further resolution using the 10th and 57th positions to partition the sequence alignment *mock.fa*. Note the use of "`-C`" followed by the comma separated list of chosen locations, instead of "`-c`" used in the first round followed by the number of maximum entropy locations.

Oligotyping analysis of *mock.fa* using the two entropy locations identified in a supervised manner results in 5 oligotypes: CA (219,745 reads), TT (177,577 reads), CT (164,043 reads), GT (155,970 reads) and GC (141,967 reads). The HTML output of the analysis can be viewed using the web address `http://goo.gl/aMleb`. Fig. 5 shows the distribution profiles of these oligotypes among three samples.

Compared to Fig. 3, Fig. 5 shows increased separation of the USA sample from Amazon and Malawi samples. The greater dissimilarity comes from the division of oligotype G, which was shared between all three samples at the end of the first oligotyping run based on 10th position alone, into two oligotypes GC and GT with the addition of the 57th position to the second round of oligotyping. Now oligotype G partitions into an Amazon and Malawi specific oligotype GC, and another oligotype GT, that is more abundant in the USA sample.

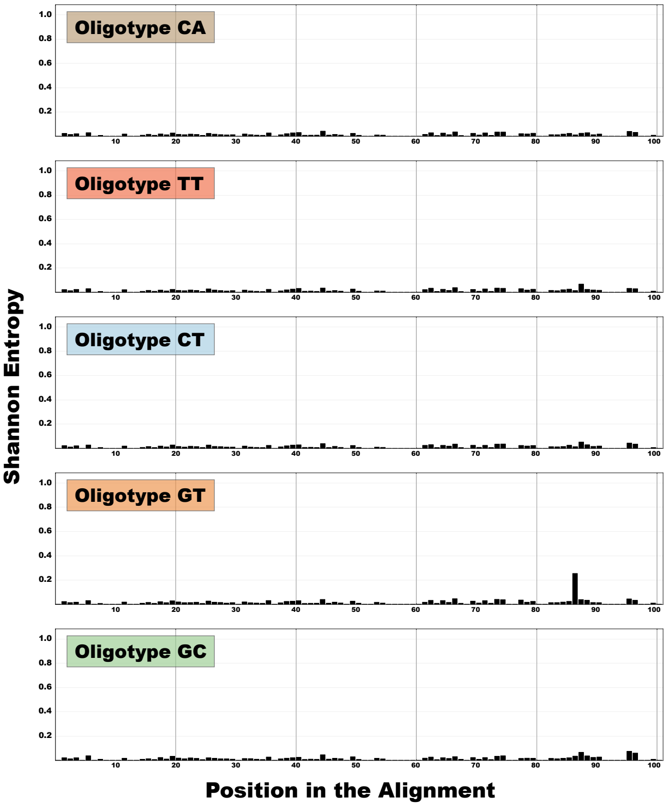

Fig. 6. Entropy profiles for oligotypes CA, TT, CT, GT and GC.

As in the previous round of oligotyping, the decision of whether to stop or continue oligotyping requires evaluation of oligotype convergence by the examination of the entropy profiles of each individual oligotype. See Fig. 6 for the entropy profiles of oligotype CA, TT, CT, GT and GC.

Fig. 6 shows that all oligotypes, with the exception of GT, are fully resolved. Oligotype GT has one entropy peak that coincides with the 86th location. Another round of oligotyping that includes the 86th location to the previous two locations could further resolve oligotype GT. Yet, when we examine the divergent sequence distribution profiles within this oligotype (profiles can be seen at the web address `http://goo.gl/Zr3OG`), we conclude that further decomposition of oligotype GT does not improve the resolution of beta diversity and thus oligotyping should end after round 2.

With this decision, oligotyping of *mock.fa* process concludes with an improved, ecologically meaningful dissection of the diversity of *Bacteroides* organisms in this dataset.

1. Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., Heath, A.C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J.G., Lozupone, C.A., Lauber, C., Clemente, J.C., Knights, D., Knight, R.Gordon, J.I. (2012) "Human gut microbiome viewed across age and geography". Nature, 486, 222-227.