Standard Operating Procedures
Jonah Ventures

Title: NEON COIF230 AND COIBE BIOINFORMATICS

Author: Vasco Elbrecht, Joseph Craine
Version: 1.0
Authorized By: Joseph Craine
Authorized Date: 07/05/2020

COIF230 Bioinformatics

Sequencing success and read quality was verified using FastQC v0.11.8, and reads were demultiplexed by using Illumina-utils v2.6 (iu-demultiplex; https://github.com/merenlab/illumina-utils) using default settings. Sequences of each sample were then merged using the -fastq_mergepairs option in Usearch v11.0.667 [1]. The forward primer "LCO1490" (5'-GGTCAACAAATCATAAAGATATTGG -3') and reverse primer "230R" (5'-CTTATRTTRTTTATNCGNGGRAANGC -3') were removed using Cutadapt v1.18 [2]. Inosines (I) were replaced with "N" for bioinformatic processing to ensure compatibility with cutadapt. Cutadapt is also used to discard sequences with length below 219 bp and above 239 bp. Expected error filtering as implemented in Usearch is then used to discard low quality reads (max_ee=0.5)[3]. Instead of OTU clustering, reads affected by sequencing and PCR errors are then removed using the unoise3 algorithm within Usearch with an alpha value of 5 [4]. This denoising is applied to each individual sample, and Exact Sequence Variants (ESV) complied in an ESV table including ESV_ID, sequence and read counts for each sample. Taxonomy is assigned to each ESV by mapping them against GenBank reference data [5] A custom database was generated by downloading COI sequences from NCBi in October 2019. Using the primer sets and iterative mapping the targeted COI region was extracted for a total of 3,012,227 sequences which where used in the reference database. , using usearch_global with –maxaccepts 0 and –maxrejects 0 to ensure mapping accuracy. Consensus taxonomy is generated from the hit tables, by first considering 100% matches, and then going down in 1% steps until hits are present for each ESV. In the respective 1% bracket, taxonomy present in at least 90% of the hits is reported. If several taxa within a taxonomic level match the ESV, an NA is reported for that taxonomic level. All hits in the 1% bracket are also available as "detailed hits" file to manually discern ESVs matching to several taxa. If matches of 97% or higher are present but no family level taxonomy is returned, the bracket is increased to 2% to reduce the potential influence of misidentified taxa.

References:

1. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460 (2010).
2. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal 17, pp. 10 (2011).
3. Edgar, R. C., & Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics, 31(21), 3476–3482. http://doi.org/10.1093/bioinformatics/btv401
4. Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv. http://doi.org/10.1101/081257
5. D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, D. L. Wheeler, GenBank. Nucleic Acids Res 33, D34 (2005).

COIBE Bioinformatics

Sequencing success and read quality was verified using FastQC v0.11.8, and reads were demultiplexed by using Illumina-utils v2.6 (iu-demultiplex; https://github.com/merenlab/illumina-utils) using default settings. Sequences of each sample were then merged using the -fastq_mergepairs option in Usearch v11.0.667 [1]. The forward primer "B" (5'-CCNGAYATRGCNTTYCCNCG -3') and reverse primer "E" (5'- GTRATNGCNCCNGCNARNAC -3') were removed using Cutadapt v1.18 [2]. Inosines (I) were replaced with "N" for bioinformatic

processing to ensure compatibility with cutadapt. Cutadapt is also used to discard sequences with length below 303 bp and above 323 bp. Expected error filtering as implemented in Usearch is then used to discard low quality reads (max_ee=0.5)[3]. Instead of OTU clustering, reads affected by sequencing and PCR errors are then removed using the unoise3 algorithm with an alpha value of 5 [4]. This denoising is applied to each individual sample, and Exact Sequence Variants (ESV) complied in an ESV table including ESV_ID, sequence and read counts for each sample. Taxonomy is assigned to each ESV by mapping them against a GenBank reference data [5] , using usearch_global with –maxaccepts 0 and –maxrejects 0 to ensure mapping accuracy. A custom database was generated by downloading COI sequences from NCBI in October 2019. Using the primer sets and itterative mapping the targeted COI region was extracted for a total of 3,142,241 sequences which where used in the reference database. Consensus taxonomy is generated from the hit tables, by first considering 100% matches, and then going down in 1% steps until hits are present for each ESV. If several taxa within a taxonomic level match the ESV, an NA is reported for that taxonomic level. All hits in the 1% bracket are also available as “detailed hits” file to manually discern ESVs matching to several taxa. If matches of 97% or higher are present but no family level taxonomy is returned, the bracket is increased to 2% to reduce the potential influence of misidentified taxa.

References:

1. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460 (2010).
2. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal 17, pp. 10 (2011).
3. Edgar, R. C., & Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics, 31(21), 3476–3482. http://doi.org/10.1093/bioinformatics/btv401
4. Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv. http://doi.org/10.1101/081257
5. D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, D. L. Wheeler, GenBank. Nucleic Acids Res 33, D34 (2005).