Dix-seq: An integrated pipeline for fast amplicon data analysis

Yongjun wei¹, Tianqi Ren², Lei Zhang²

¹School of Pharmaceutical Sciences, Key Laboratory of State Ministry of Education, Key Laboratory of Henan Province for Drug Quality Control and Evaluation, Collaborative Innovation Center of New Drug Research and Safety Evaluation, Zhengzhou University, 100 Kexue Avenue, Zhengzhou, Henan 450001, PR China

²DeepBiome. Co. Ltd., NO.38 Debao Road, China (Shanghai) Pilot Free Trade Zone, Shanghai, 200031, PR China

Corresponding authors

Yongjun Wei

Email: yongjunwei@zzu.edu.cn

Lei Zhang

Email: zhanglei@logicinformatics.com

Abstract

The amplicon derived from 16S rRNA genes, 18S rRNA genes, internal transcribed spacer sequences or other functional genes can be used to infer and evaluate microbial diversity or functional gene diversity. With the development of sequencing technologies, large amounts of amplicon data were generated. Several different software or pipelines had been developed for amplicon data analyses. However, most current software/pipelines require multistep and advanced programming skills. Moreover, they are often complex and time-consuming. Here, we introduced an integrated pipeline named Dix-seq for high-throughput amplicon sequence data processing (<https://github.com/jameslz/dix-seq>). Dix-seq integrates several different amplicon analysis algorithms and software for diversity analyses of multiple samples. Dix-seq analyzes amplicon sequences efficiently, and exports abundant visual results automatically with only one command in Linux environment. In summary, Dix-seq enables the common/advanced users to generate amplicon analysis results easily and offers a versatile and convenient tool for researchers.

Keywords

Amplicon; 16S rRNA; Microbiome; Pipeline; Next-generation sequencing

1. Introduction

The microbial diversity and functional information can be inferred with the amplicon sequencing generated from marker genes, such as 16S rRNA gene, 18S rRNA gene, internal transcribed spacer sequences, and other functional genes. The low cost of next-generation sequencing technology and popularization of amplicon study, as well as the easy application of amplicon technology, lead to the generation of large amount of amplicon data. To show and compare microbial communities, several different amplicon analysis tools and pipelines have been developed. These tools or pipelines often require the users to master in-depth bioinformatics knowledge and understanding of parameters in each step during amplicon analyses. Moreover, the obtained output needs to be visualized with additional software/tools. Besides the requirement of computer professional knowledge, the whole amplicon analysis process is complex and time-consuming. Here, we develop an integrated pipeline named Dix-seq for convenient amplicon data analyses (<https://github.com/jameslz/dix-seq>). Dix-seq generates abundant visual results automatically with only one command line in Linux environment, from the raw data demultiplex, chimeric sequence removal, sequence classification, alpha and beta diversity, functional prediction, variance analyses, to the final visual results (Figure 1). The pipeline is easy and efficient for further improvement, for diverse third-party procedures can be integrated and pipeline optimization are allowed (Table 1).

2. Raw sequence data analyses

Dix-seq can extract raw fastq files directly from the sequencing machines in short time. Next, raw fastq files are demultiplexed with the corresponding barcode sequences by our developed atlas-utils tool (<https://github.com/jameslz/biostack-suits> version: 0.0.1). After discarding barcode sequences in each read, low-quality sequences were filtered by Trimmomatic (version 0.38) [1]. During filter process, the start and end of a read below a threshold quality of score < 3 were removed, and the reads were truncated at any site receiving an average quality score < 15 over a 4 bp sliding window. Moreover, truncated reads which were shorter than 36bp were discarded. Paired-end reads were merged using Usearch fastq_mergepairs command with the default parameters (usearch11.0.667, <http://drive5.com>) [2]. Reads were discarded if they could not be merged. The primer sequences would affect further operational taxonomic unit (OTU) table construction. Therefore, merged reads without full primer sequences or with more than two nucleotide mismatches in the primer region were discarded, and the clean merged reads (clean tags) without primer sequences were obtained.

3. Bioinformatic analyses

3.1 ZOTU Construction

In the pipeline, Usearch fastq_filter was used to filter all the low-quality sequences in the clean tags with default parameters (fastq_maxee 1), and the high-quality sequences of the clean tags were obtained. The obtained high-quality reads were sorted in non-redundant abundance order using USEARCH fastx_uniques. The USEARCH unoise3 algorithm filters the non-redundant sequences with minsize = 8, and generate the zero-radius operational taxonomic unit (ZOTU) table with 97% identity. The representative sequences of each ZOTU was identified. During the process, the chimeric sequences were removed in the denoising procedure. All the high-quality sequences were binned to the ZOTU representative sequences, and the relative abundance of each ZOTU were obtained.

The phylogenetic assignment of the representative sequences of each ZOTU was determined by **Usearch syntax** algorithm [4] with the RDP training set (version **v16**) 16S rRNA database (Other functional gene database can be used, such as nifH sequence database) as reference database (the confidence threshold is 0.8). The ZOTUs annotated as mitochondrial or chloroplast rRNA sequences were discarded. As a result, ZOTU counts table and ZOTU relative abundance table were generated. The **stacked bar charts** and **KRONA** software (<http://sourceforge.net/projects/krona/>) were used for visualization of the taxonomic composition and phylogenetic annotation. The heatmap plots at different taxonomic levels (ZOTU, genus, family, order, class, phylum) were generated using **heatmap** in R environment (version 3.6.0) [5]. In the meanwhile, the clustering analyses between different samples or ZOTUs were applied.

The ZOTU table was normalized using USEARCH otutab_rare with specific sample size, and the normalized sample size was the sample harboring the minimum sequences among all the samples. The diversity and other analyses were based on the normalized ZOTU table.

3.2 Diversity and functional analyses

Based on ZOTU abundance table, the alpha diversity indices of richness, chao1, simpson and shannon diversity were calculated with Usearch alpha_div. The ZOTU representative sequences were aligned with the reference sequences using HMMER and EPA-NG method in PICRUST2 (version 2.3.0-b), and the multiple alignment results was used to generate the phylogenetic tree.

The beta diversity distance was calculated with **Usearch beta_div**, and the frequently used distance metrics were weighted_unifrac, unweighted_unifrac and bray_curtis. The principal coordinates analysis (PCoA) and Non-metric Multidimensional Scaling (NMDS) analyses based on weighted_unifrac, unweighted_unifrac and bray_curtis distances were used to evaluate the variation/difference between samples based on the homogenized OTU table. The gmodels in R software and nmets.py were used to generate the final figures PCoA and NMDS figures. For unweighted and weighted UniFrac distance metrics, the required phylogenetic tree was generated using **PICRUST2 in previous step**.

The **Vegan** 2.4.2 package [6] in R was used to generate Anosim based on the distances calculated with homogenized OTU tables. The clean tags were used for differential abundance analysis at different taxonomic levels (phylum, order, class, family, genus, ZOTU) using **DEseq2 with the parameters of test=Wald and fitType=parametric**.

The R package **Tax4Fun** (version 0.3.1) [6] was used to predict the functional profiles with the information of ZOTU representative sequences and their relative abundance. During the process, **SILVA_123_SSURef_Nr99** database in Tax4Fun was used to identify the best hit for representative ZOTU sequences, and the predicted KO (KEGG Orthology) abundance was used to infer pathway, enzyme, module and catalog information in KEGG database. The predicted functional information of samples was compared and shown.

Demo amplicon analyses

Dix-seq was used to analyze one amplicon data with six samples (Accession number of PRJNA436765). The amplicon sequences were V3-V4 fragments of 16S rRNA genes, and the six samples were divided into three groups. Dix-seq generates visual results in several minutes. Part of the visual outputs were shown in Figure 2. The phylum-level microbial abundance and the phylogenetic classification of these six samples can be generated (Figure 2a), and other phylogenetic level abundance results can also be generated. The alpha diversity indices, including richness, were calculated (Figure 2b). Based on the principal components, the relationship between these six samples were analyzed (Figure 2c). Moreover, other microbial community differences, such as the unweighted pair group method with arithmetic mean (UPGMA), were analyzed (Figure 2d). Other beta diversity distances were visualized with Dix-seq. Principal component analysis diagram shows microbial composition differences of the six samples based on prediction of Tax4Fun at pathway-level was expressed with the principal component analysis (Figure 2e), and other functional levels such as KO, modules and catalog, were predicted and expressed. In addition, the pair-wise comparison between two groups at OTU-level was calculated with Deseq (Figure 2f), other phylogenetic level and variance between each two group were analyzed.

4. Conclusion

We developed an integrated pipeline for efficient amplicon data analyses. There are several advantages of this pipeline. Firstly, the developed Dix-seq pipeline is based on command line model, and all the visual output can be easily realized with one command line. Secondly, Dix-seq pipeline is composed with different functional workflow modules, and these modules can work together or separately. Moreover, the functional modules are independent PERL scripts and can be added or deleted easily. Thirdly, the pipeline integrates three command-line tools: atlas-utils, tsv-utils, fastx-utils (packed in biostack-suits), which can generate results efficiently. Finally, Dix-seq pipeline

integrates several third-party procedures, including functional prediction, variance analysis. In summary, our integrated Dix-seq pipeline offers a versatile and efficient amplicon analysis tool for different-level researchers.

References

1. Bolger, A. and F. Giorgi, *Trimmomatic: A Flexible Read Trimming Tool for Illumina NGS Data*. URL <http://www.usadellab.org/cms/index.php>.
2. Edgar, R.C. and H. Flyvbjerg, *Error filtering, pair assembly and error correction for next-generation sequencing reads*. Bioinformatics, 2015. **31**(21): p. 3476-3482.
3. Edgar, R.C., *UPARSE: highly accurate OTU sequences from microbial amplicon reads*. Nature methods, 2013. **10**(10): p. 996-998.
4. Edgar, R., *SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences*. bioRxiv, 2016: p. 074161.
5. Team, R.C., *R: A language and environment for statistical computing*. 2013.
6. Oksanen, J., et al., *The vegan package*. Community ecology package, 2007. **10**: p. 631-637.
7. Dixon, P., *VEGAN, a package of R functions for community ecology*. Journal of Vegetation Science, 2003. **14**(6): p. 927-930.
8. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2*. Genome biology, 2014. **15**(12): p. 550.
9. Segata, N., et al., *Metagenomic biomarker discovery and explanation*. Genome biology, 2011. **12**(6): p. R60.
10. Aßhauer, K.P., et al., *Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data*. Bioinformatics, 2015. **31**(17): p. 2882-2884.

Table 1 Details of softwares and tools used in Dix-seq pipeline.

Module	tools	task	version	url
Data preparation	atlas-utils	demultiplex	0.0.1	https://github.com/jameslz/biostack-suits
	Trimmomatic	-	0.38	http://www.usadellab.org/cms/?page=trimmomatic
	USEARCH	fastq_mergepairs	1.0.667	https://drive5.com/usearch
ZOTU	USEARCH	search_pcr2	0.0.1	https://github.com/jameslz/biostack-suits
	USEARCH	unoise3	1.0.667	https://drive5.com/usearch
	USEARCH	sintax	1.0.667	https://drive5.com/usearch
Taxonomy	atlas-utils	filter	0.0.1	https://github.com/jameslz/biostack-suits
	atlas-utils	annotation	0.0.1	https://github.com/jameslz/biostack-suits
	atlas-utils	krona	0.0.1	https://github.com/jameslz/biostack-suits
	atlas-utils	levels	0.0.1	https://github.com/jameslz/biostack-suits
	atlas-utils	counts2freqs	0.0.1	https://github.com/jameslz/biostack-suits
Diversity	R	ggplot2	3.3.0	https://www.r-project.org/
	picrust2	place_seqs	2.3.0-b	https://github.com/picrust/picrust2
	USEARCH	otutab_rare	1.0.667	https://drive5.com/usearch
	USEARCH	alpha_div	1.0.667	https://drive5.com/usearch
	USEARCH	beta_div	1.0.667	https://drive5.com/usearch
	Rarefaction tool kit	rtk	0.93.2	https://github.com/hildebra/Rarefaction
	R	vegan	2.5-6	https://www.r-project.org/
Statistics	R	DESeq2	1.26.0	http://bioconductor.org/
	R	vegan(anosim)	2.5-6	https://www.r-project.org/
Function prediction	blast	blastn	2.10.0	https://ftp.ncbi.nlm.nih.gov/blast/executables/
	R	Tax4fun	0.3.1	http://tax4fun.gobics.de/
	atlas-utils	annotation	0.0.1	https://github.com/jameslz/biostack-suits

picrust2	metagenome_pipeline	2.3.0-b	https://github.com/picrust/picrust2
atlas-utils	kann	0.0.1	https://github.com/jameslz/biostack-suits

Figures

Figure 1 Dix-seq: An integrated pipeline for fast amplicon data analysis. The steps and softwares used in the Dix-seq pipeline were shown.

Figure 2 Some outputs generated with the Dix-seq pipeline were shown. (a) Microbial abundance of six samples at phylum-level. (b) The richness estimation of the six samples. (c) Principal component analysis diagram shows microbial composition differences of the six samples. (d) UPGMA diagram shows microbial composition differences of the six samples. The microbial compositions at phylum-level of the six samples are shown. (e) Principal component analysis diagram shows microbial composition differences of the six samples based on prediction of Tax4Fun at pathway-level. (f) Pair-wise comparison in Deseq at ZOTU-level between Group_A and Group_B.