



•方法•

基于Nextflow构建的宏条形码 自动化分析流程EPPS

李诣远* David C. Molik Michael E. Pfrender

(Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46554, USA)

摘要: 基于宏条形码技术的物种快速检测有助于生物多样性的评估、预测和保护。本文介绍了常用宏条形码分析的步骤和参数设定方法。我们利用Nextflow搭建了一款宏条形码分析流程EPPS, 可以自动化地运行从原始数据的质量控制到环境多样性的比较。Nextflow软件还拥有流程监控的功能, 可视化输出每个进程所消耗的时间与内存。本文还使用测试数据和已发表数据证明该平台能够有效地分析宏条形码数据并可靠地分析环境生物多样性的相似性。

关键词: 环境DNA; USEARCH; Trimmomatic; 主成分分析

EPPS, a metabarcoding bioinformatics pipeline using Nextflow

Yiyuan Li*, David C. Molik, Michael E. Pfrender

Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46554, USA

Abstract: Metabarcoding helps to quickly assess biodiversity. In this study, we discuss popular metabarcoding analytical tools and parameter settings. We also develop a metabarcoding bioinformatics pipeline, EPPS, to process data from quality control of raw reads to biodiversity comparisons between samples using a pipeline building program, Nextflow. The EPPS pipeline can summarize the time and memory cost of each process in the pipeline. We also apply the pipeline on a test dataset and a public dataset from a previous study. The result suggests that this pipeline can reliably analyze metabarcoding data and facilitate pipeline sharing of metabarcoding studies.

Key words: environmental DNA; USEARCH; Trimmomatic; principal component analysis

生物多样性为人类提供了重要的生态系统服务, 包括洁净的水源和空气、食物、气候调节、碳循环、航运、休闲娱乐等等(Millennium Ecosystem Assessment, 2005)。近年来由于人类活动和环境变化, 生物多样性正受到严重的影响(Worm et al, 2006; Collen et al, 2014; Pimm et al, 2014; Newbold et al, 2015)。对生物多样性全面详细地了解有助于评价其状况, 并预测其发展趋势, 促进保护工作的开展。

随着基因组学技术的发展, 第二代测序(next-generation sequencing)技术越来越多地被应用于生物多样性的调查(Pfrender et al, 2010; Lodge et

al, 2012; Bohmann et al, 2014; Thomsen & Willerslev, 2015; Deiner et al, 2017a; Simon & Evans, 2017), 包括: 生物多样性的监测、入侵物种的监测、食性分析等等。生物多样性调查时, 往往通过检测基因组的一段或多段DNA序列(如: DNA条形码)与已知物种的DNA序列的匹配来实现物种的快速分类(Taberlet et al, 2012)。常见的技术包括: 宏条形码(metabarcoding)技术、线粒体基因组捕获技术(Dowle et al, 2016; Liu et al, 2016; Wilcox et al, 2018)和全线粒体基因组测序技术(Zhou et al, 2013; Crampton-Platt et al, 2015, 2016; Tang et al, 2015; Deiner et al, 2017b; Bista et al, 2018)。由于宏条形码

收稿日期: 2018-08-01; 接受日期: 2019-03-05

* 通讯作者 Author for correspondence. E-mail: yyli19@icloud.com

技术成本低, 实验操作简便, 所需的DNA起始量低, 因此应用最为普遍。宏条形码的一般流程是提取混合物种的DNA, 并通过PCR扩增目标DNA片段, 利用高通量测序获取PCR扩增子的DNA序列, 并通过生物信息学方法进行多样性分析(Thomsen et al, 2012; Ji et al, 2013; Liu et al, 2013; Evans et al, 2016, 2017; Olds et al, 2016; Li et al, 2018)。

基于PCR的宏条形码生物信息分析流程的基本步骤包括: 测序质量控制, PCR引物的删除, 双向序列的拼接, 分子可操作分类单元(molecular operational taxonomic unit, MOTU)的聚类分析, 生物多样性分析以及物种分类(图1)。目前宏条形码的常用分析软件有很多, 如QIIME (Caporaso et al, 2010)、DADA2 (Callahan et al, 2016)、Mothur (Schloss et al, 2009)、USEARCH (Edgar, 2010, 2013)、VSEARCH (Rognes et al, 2016)、obitools (Boyer et al, 2016)等。其中, 有很多软件开发时主要是针对微生物宏基因组16S rRNA基因的生物信息分析, 如QIIME、DADA2、Mothur和USEARCH。在这些软件中, QIIME主要基于python并且整合了各种宏基因组分析的软件。DADA2、Mothur和USEARCH则提供了一个一体化的分析流程, 自身包含了分析所需的所有命令。obitools是一款面向宏条形码生物信息

分析的软件, 与上述几个软件不同的是, obitools包含了很多为宏条形码设计的功能, 例如, 提供引物设计的帮助和针对真核生物的物种鉴定。除了通用的宏条形码分析软件, Sato等(2018)还设计了针对鱼类的宏条形码分析软件MiFish。MiFish提供了在线鱼类宏条形码序列分析的解决方案, 通过USEARCH和BLAST软件将序列与鱼类宏条形码数据库MitoFish比对获得多样性信息。虽然以上软件包含了大部分宏条形码的分析步骤, 但分析的步骤依旧需要调用众多的软件和命令, 针对不同的研究宏条形码的分析参数也有差异。因此快捷地调用和调整不同的程序和命令可以极大地提高宏条形码分析的效率。

近年来在微生物领域已经有多个研究致力于设计宏基因组(Piro et al, 2017; Urtskiy et al, 2018; Visconti et al, 2018)和16S rRNA数据(<https://github.com/h3abionet/h3abionet16S>)的分析流程。但在宏条形码研究领域, 还没有一套可以快速建立分析大量样品的分析流程。面对越来越多的第二代测序数据和宏条形码样本, 一套能够在不同平台上快速建立并根据样品情况调整参数的信息分析流程将有助于快速分析大量样品, 并可以针对不同的样品调整流程。本文利用Nextflow (Di Tommaso et al, 2017)搭建了一款宏条形码分析流程EPPS。Nextflow是一款基于Groovy语言的流程管理系统(workflow management system), 可以方便快捷地调度程序和中间文件, 降低了创建流程的复杂性。同时Nextflow支持Linux、MacOS和云计算平台, 便于将流程扩展到不同的计算平台。本文将详细介绍宏条形码分析中不同生物信息分析步骤的推荐参数并介绍如何修改参数, 方便其他宏条形码研究使用。借助于Nextflow的优势, EPPS可以让研究者快速地建立宏条形码的分析流程并调整参数, 同时支持并行计算, 能够分析和监控大量宏条形码样品的分析。EPPS的分析流程可以从GitHub下载(https://github.com/lyy005/epps_nf)。

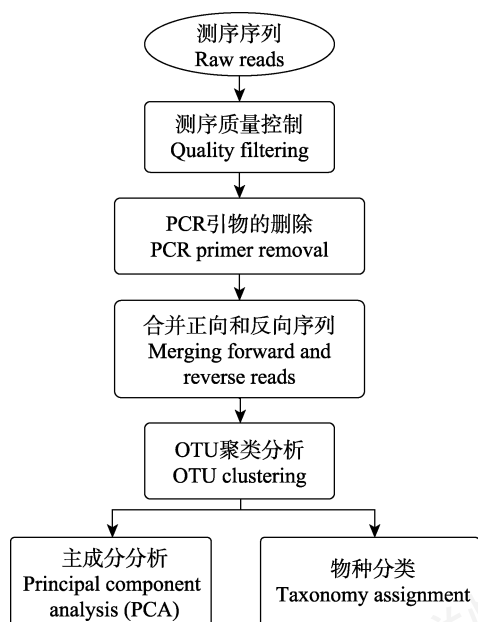


图1 EPPS的主要分析步骤。OTU聚类分析还包括去除重复序列、OTU聚类和嵌合体的检测。

Fig. 1 The workflow of EPPS. OTU clustering step includes, dereplication, OTU clustering and chimera detection.

1 EPPS流程设计

EPPS的分析流程包括: 原始测序结果的质量控制, 引物序列的删除, 正向和反向序列的合并, OTU聚类分析和群落多样性分析(主成分分析)(图1)。EPPS使用Nextflow工作流程语言, 简化了用

户在分析过程中的重复操作,使其可以一键完成大量宏条形码样品的分析,并提供了进程所占的内存和时间消耗,便于用户对进程的监控。EPPS流程提供常用的OTU表格输出格式,便于用户将结果输入其他的后期分析软件中,如QIIME和Phinch (Bik & Interactive Pitch Inc., 2014)。

(1)软件安装。EPPS流程基于Linux和MacOS操作系统。用户在运行EPPS之前需要安装Nextflow、Java 7、R (R Core Team, 2016)和USEARCH软件。R软件可以从<https://www.r-project.org>下载安装。USEARCH软件可以从<http://drive5.com/usearch/download.html>下载安装。另外, EPPS流程还包含了VSEARCH的可执行文件, 用户可以使用命令:

```
git clone git@github.com:lyy005/epps_nf.git
```

或者前往https://github.com/lyy005/epps_nf下载流程。

(2)测序文件的输入。EPPS分析流程的输入文件分为两部分: 第一部分是fastq格式的Illumina双向测序结果文件; 第二部分是一个名为primer.fas的fasta格式文件, 用来指定宏条形码使用的引物序列。EPPS下载完成后, 用户使用Linux或者MacOS终端命令进入EPPS文件夹: `cd epps_nf`。文件夹里包含3个文件夹: `./bin`、`./example`和`./input`。`./bin`文件夹包含运行脚本所需的所有程序, 下载的USEARCH可执行文件也需要添加到bin文件夹中。`example`文件夹包含有可以直接运行的测试文件, 其中包括8个fastq文件, 分别是4个测试样品的正向和反向测序文件, 并且文件需要以“样品名.1.fastq”和“样品名.2.fastq”的方式命名。`example`文件夹中还包括1个名为primer.fas的文件, 该文件包含有宏条形码测序的引物序列。宏条形码测序时, 通常是将不同样品的标签序列(index)添加到Illumina平台的测序接头(adaptor)上, 这样Illumina测序平台可以直接将不同样品分流(demultiplexing), 生成每个样品的fastq文件。具体实验方法可以参见: <http://www.earthmicrobiome.org/protocols-and-standards/16s/> (Caporaso et al, 2011)。默认的测序文件输入格式后缀为fq。如果输入文件格式为fq.gz或者fastq.gz, 用户可以将epps_v190209.nf第4行的命令改成相应的格式。例如, 如果是fastq.gz格式, 可以将命令改为: `params.reads = "$PWD/input/*{1,2}.fastq.gz"`

分析前先将输入文件(fastq文件和primer.fas文件)复制到input文件夹中, 再使用命令

```
./nextflow run epps_v190209.nf -with-timeline
```

运行分析流程并获得分析结果。EPPS流程分析测试样品的时间和内存消耗参见图2。本文将通过流程自带的4个测试样品test1到test4具体介绍工作流程和分析结果。使用“-with-timeline”命令会产生一个时间消耗文件(timeline.html), 如图3。该文件包括了EPPS流程中每个进程消耗的时间以及虚拟存储的峰值。

(3)测序质量控制。首先, 输入的测序文件(fastq格式)会使用Trimmomatic (version 0.38)软件进行严格的质量控制。Trimmomatic的输入参数如下: “ILLUMINACLIP:combined.Illumina.fasta:3:30:6:1:true SLIDINGWINDOW:10:20 MINLEN:50”。Trimmomatic首先使用16 bp的种子序列(seed)与Illumina的测序接头进行匹配。如果有≤ 3 bp的错配, 种子序列将继续延长匹配。接着, Trimmomatic会计算匹配得分, 如果双向测序序列与测序引物接头的匹配达到30分(约50 bp的匹配)或者单向测序序列与测序引物接头的匹配达到10分(大约17 bp的匹配), 序列的匹配部分将会被删除。Trimmomatic还会使用一个10 bp长的滑动窗口(sliding window)为测序序列计算平均序列质量值。如果平均序列质量值小于20, 该序列将不会被使用。最后, 任何< 50 bp的经过测序接头筛查和质量控制的序列均不会被后续分析使用。只有正向和反向序列同时通过了Trimmomatic的质量控制, 该序列才会被用于后续分析。用户可以在epps_v190209.nf文件的25行对Trimmomatic的参数或者测序接头文件进行修改。

Trimmomatic通常包含4个fastq格式的输出文件, 其中有2个配对的fastq文件, 包含正向和反向都已通过质量控制的序列。还有2个不配对的fastq文件, 包含了只有正向或者只有反向通过质量控制的序列。EPPS中只保留了正反向序列都通过Trimmomatic质量控制的fastq文件。该结果包含在output文件夹中, 文件名分别是“样品名称.pe.1.fq”和“样品名称.pe.2.fq”。

(4) PCR引物的删除。PCR引物通常加在PCR扩增子的两端。去除PCR引物有多种方法。由于引物序列通常是由合成的引物连接在扩增子的末端并进行PCR扩增, EPPS流程使用PCR引物完全匹配的方法进行引物序列的查找与删除。同时, 由于引物通常位于测序序列的开端, 且有很高的测序质量。

Processes execution timeline

Launch time: 10 Feb 2019 00:05

Elapsed time: 22.7 s

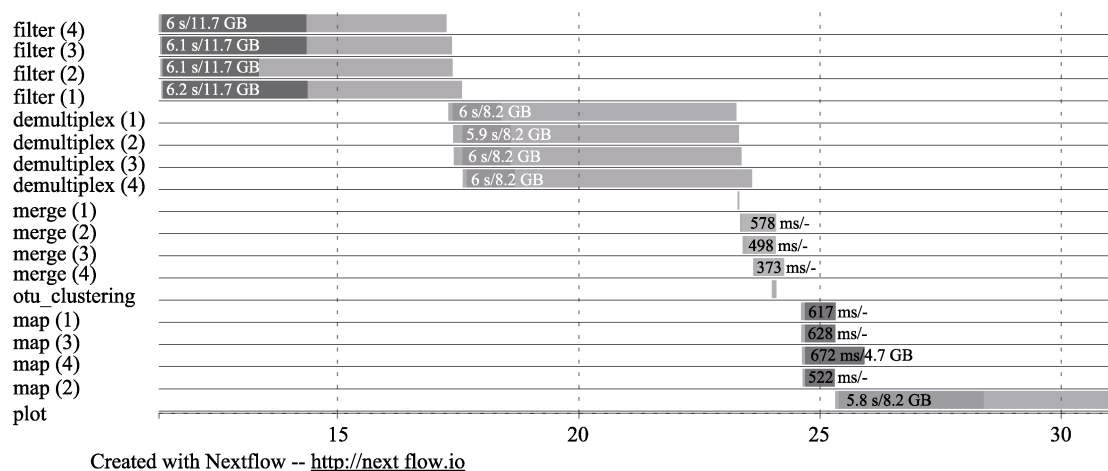


图2 EPPS流程每一个进程的时间消耗。横坐标代表时间, 单位是秒。最左列的名称分别对应了宏条形码分析的流程。**filter**: 测序质量控制; **demultiplex**: PCR引物的删除, 如果有多个引物则将各个引物分开; **merge**: 合并正向和反向序列; **otu_clustering**; **map**: OTU聚类分析; **plot**: 主成分分析。由于测试数据有4个样品, 因此每个进程的右侧括号里有1-4的序号。浅色进度条代表进程所消耗的系统时间。深色进度条代表的是每个进程的CPU时间。每个进度条包含有两个数字, 第1个数字代表每个进度的系统时间, 第2个数字代表虚拟内存的峰值。

Fig. 2 The timeline chart of EPPS. The x-axis is the amount of time for each process in seconds. Each row indicates the name of different stages of the analysis. filter, Data filtering; demultiplex, Primer removal and demultiplex if there are multiple primers; merge, Merging of forward and reverse reads; otu_clustering and map, OTU clustering and mapping of reads; plot, Plotting PCA plot. As there are four samples in the testing data set, there are four processes (1 to 4) for filter, demultiplex, merge, and map steps. Each bar indicates the time for each process. The dark area in each bar represents the real execution time. Each bar displays two numbers: the task duration time and the virtual memory size peak.

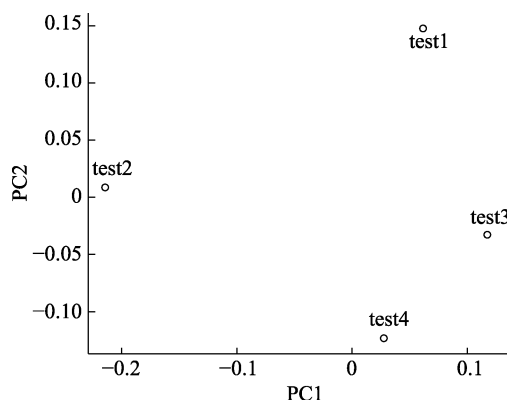


图3 基于测试数据的主成分分析结果。图中每一个点代表一个测试数据的样品。点与点之间距离越近代表样品之间的物种组成相似度越高。例如, test3和test4的相似度大于test3和test1的相似度。

Fig. 3 PCA plot based on testing data. Each dot in the figure represents a test sample. The distance between dots indicates the dissimilarity between samples. For example, the similarity between test3 and test4 is higher than test1 and test2.

如果在序列开端出现引物的错配, 说明该序列是由于PCR错误或者测序错误而导致序列的不一致。因此, EPPS通过完全匹配引物可以进一步筛选序列的

测序质量。其他的分析软件中, QIIME通过局部比对 (truncate_reverse_primer.py命令) 匹配并删除引物序列, 默认值是容许引物序列与测序结果有2个错配 (mismatch)。DADA2通过在序列的开端截取指定长度的碱基删除引物序列。除此之外, 用户也可以使用cutadapt (Martin, 2011) 或Trimmomatic (Bolger et al, 2014) 进行引物序列的匹配。

EPPS流程中使用了Perl语言脚本对PCR引物序列进行匹配和删除。Perl语言脚本搜索正向和反向测序序列中完全匹配的引物序列。通过匹配的PCR引物序列, 所有的扩增子均被调整为同一个方向, 以方便下一步OTU聚类分析。测序的结果同样可以在output文件夹中找到。命名的格式为“样品名称.demul.F.fq”和“样品名称.demul.R.fq”。“样品名称.demul.F.fq”包含所有匹配正向PCR引物的序列, “样品名称.demul.R.fq”包含了所有匹配反向PCR引物的序列。

(5)合并正向和反向序列。合并正反向序列是宏条形码分析的重要一步。当PCR扩增子小于两端测

序读长时,宏条形码分析流程通常会利用正向和反向测序序列的重叠部分将它们合并为1个序列。这样做的优点有两个:(1)合并测序序列可以保证PCR扩增子的完整,不会出现由于扩增子部分的缺失影响OTU聚类;(2)合并测序序列可以利用正反向序列重叠的部分纠正测序序列末端的错误,从而提高测序序列的质量值(Zhou et al, 2011; Masella et al, 2012; Edgar, 2013)。EPPS流程使用USEARCH (version 10.0.240) `fastq_mergepairs`命令合并正向和反向的测序结果,并利用USEARCH针对合并的正反向序列检测测序质量。如果测序结果的期望测序错误率高于0.5,该序列将被删除。相比于USEARCH默认的期望错误率(1.0),EPPS流程选用了相对严格的错误率筛选以提高最终物种多样性分析的可靠度。合并之后的结果存放在output文件夹中,命名的方式为“样品名称.merged.rename.fasta”。在不同分析流程中都有合并这一步,例如:QIIME的`vsearch join-pairs`命令(Rognes et al, 2016),DADA2的`mergePairs`命令,Mothur的`make.contigs`命令,USEARCH的`fastq_mergepairs`命令,obitools的`obiojoinpairedend`命令。这些程序通常会重新计算合并后序列的错误率(Edgar, 2013; Rognes et al, 2016)。

(6)去除重复序列和OTU聚类分析。在合并正反向序列之后,EPPS会把相同的序列合并以提升聚类分析的效率。合并的序列的丰度(read number)会被保留,序列将会按照丰度从高到低排序。EPPS使用USEARCH (version 10.0.240) `usearch_global`命令或者VSEARCH (version 2.10.4)的`cluster_size`命令对前一步获得的序列进行聚类分析。聚类分析使用97%的全局序列相似度。如果需要使用不同的聚类相似度,可以修改`epps_v190209.nf` 88行`-id 0.97`。1个OTU里丰度最高的序列会被作为OTU的代表序列,用于后续的比较分析和物种分类。OTU聚类的结果存放在`step1_otu_clustering/otus.fasta`文件中。每个样品的测序序列再通过USEARCH或者VSEARCH的`usearch_global`命令利用全局比对方法和OTU的代表序列进行比对,以获取OTU在每个样品中的丰度。每个样品的比对结果存放在`step2_mapping`文件夹中,命名的方式为“样品名称.uc”。最后,EPPS使用perl脚本将不同的比对结果合并,结果存放在“combined.uc.table”文件中。该文件的每一列代表1个样品,每一行是1个OTU的丰度。

通常在合并正反向序列之后,可以选择直接进行序列比对或者OTU的聚类分析。在有相对完整的参考序列数据库(reference database)时,通过将序列与数据库进行比对可以直接获得样品中的物种多样性信息。常见的比对软件包括BLAST, UCLUST和QIIME。当数据库不完备时,为了获得相对完整的物种信息,通常信息分析流程使用OTU聚类进行物种多样性分析。常用的聚类软件包括QIIME、USEARCH、CROP、SWARM。根据样品中物种亲缘关系的差异,聚类的相似度一般设为97%–99%。

(7)检测嵌合体。宏条形码的PCR过程中会导致PCR嵌合体(chimera)的产生。检测嵌合体通常有两种方法:基于参考序列;从头检测。EPPS流程里USEARCH `cluster_otus`命令自带从头检测嵌合体的方法。而VSEARCH不能自动检测嵌合体。如果用户使用VSEARCH,需要额外运行从头检测嵌合体`uchime_denovo`命令(`epps_v190209.nf` 92行)。

(8)主成分分析。用户获得OTU的序列和OTU在样品中的分布之后,就可以进行样品内部的多样性(α 多样性)、样品之间的多样性(β 多样性)或者整体样品的多样性(γ 多样性)分析。其中, β 多样性分析可以通过主成分分析完成。通过比较样品之间物种多样性组成的差异将样品之间的相似性可视化。计算样品与样品间的距离(β 多样性)有多种方法(Cardoso et al, 2009),包括Jaccard相似性指数、Shannon相似性指数、Sørensen相似性指数等等。由于PCR的影响,基于PCR的宏条形码的序列数量往往不被作为物种丰度和生物量的参考(Tang et al, 2015; Deiner et al, 2017a; Bista et al, 2018),因此EPPS流程使用Jaccard相似性指数计算样品间的相似性。

EPPS流程使用R语言和R分析包ggplot2 version 3.0.0.9 (Wickham, 2016), vegan version 2.4.6 (Oksanen et al, 2013)和ggrepel version 0.8.0 (Slowikowski, 2018)计算样品间的差异并作图。主成分分析的结果为PDF格式的文件,存放在`step3_profiling_table/plot.pdf`中。该文件中不同的点代表不同的宏条形码样品,样品间的距离越近代表相似度越高。

(9)物种分类。由于不同的物种类群常常有不同的参考序列集和分类方法,EPPS流程本身不包含物种分类的软件。用户可以基于EPPS流程产生的OTU代表序列,并根据研究的需要采用不同的方法进行分类,常用的有三种:基于相似度分类、基于基因

序列特征分类、基于系统发育树分类(Bazinet & Cummings, 2012)。(1)基于相似度分类。通过将测序序列与已知物种的序列比对来获得物种分类结果。常见的分类软件包括BLAST (Camacho et al, 2009)、MEGAN (Huson et al, 2007)、MetaPhyler (Liu et al, 2010)和CARMA (Gerlach & Stoye, 2011)。该方法的优点在于在参考序列集完整的时候准确性非常高,所以参考序列集的完整性对物种分类的影响非常显著。(2)基于基因序列特征的分类。通常是利用DNA序列的特征或者kmer频率进行物种分类。常见的分类软件包括: SINTAX (Edgar, 2016)、RDP (Wang et al, 2007)、NBC (Rosen et al, 2010)、PhyloPythiaS (Patil et al, 2012)、Phymm和PhymmBL (Brady & Salzberg, 2009)。这一类方法的优势在于一旦模型建立,对测序序列分类的速度非常快,缺点在于如果基因序列长度较短会限制有效的序列特征的数量,从而影响分类的准确性。(3)基于系统发育树的分类。该方法是三种方法中最准确的,并且由于分类的方法建立在系统发育树的基础上,可以精确到不同分类阶元。该方法的限制也很明显,为了构建系统发育树,测序序列与参考序列集需要进行多序列联配和系统发育树的构建,这两个步骤往往需要消耗较多的计算资源和时间。基于系统发育树的软件包括pplacer (Matsen et al, 2010)、EPA (Berger et al, 2011)和FastTree (Price et al, 2009)。同时,也有软件同时利用多种方法进行分类,例如Statistical Assignment Package (SAP) (Munch et al, 2008)。不同研究可以根据参考序列集的完整情况和分析数据量的大小进行选择。

如果用户没有事先准备高质量参考序列集进行分类,可以使用在线SAP软件(<https://services.birc.au.dk/sap/server>)。SAP首先通过NetBlast将每个OTU的代表序列比对到NCBI NR数据库。从NCBI比对获得的同源序列会进行多序列联配,并通过对代表序列和同源序列构建系统发育树计算代表序列属于特定物种的贝叶斯后验概率。

2 测试数据以及EPPS的输出结果

为了方便测试, EPPS自带了模拟数据。该数据模拟了12,500条Illumina测序结果, 每条序列的读长为300 bp。另外, 我们还使用了已发表的公共数据(Li et al, 2018)来评估该流程分析较大样本量的表现。

(1)模拟数据

模拟数据存放于./example文件夹中。开始分析之前, 用户首先需要将example文件夹里所有fastq文件和primer.fas文件复制到input文件夹中:

```
cp ./example/* ./input/
```

运行命令:

```
nextflow run epps_v180726.nf -with-timeline
```

重要的结果文件存放在文件夹: ./output/step3_profiling_table/。combined.uc.table文件是OTU表格文件。plot.pdf文件是主成分分析文件, 该文件主要是基于OTU表格进行的主成分分析。从图3可以看出, test3和test4的物种组成相似度较高, test1和test2的物种组成相似度较低。在资源消耗方面, EPPS流程需要大约15s (图2), 消耗54 M的硬盘空间, 大约是原始测序数据所占空间(63 M)的0.86倍。

(2)公共数据

为了进一步证明流程的可靠性, 我们分析了已经发表的宏条形码数据(Li et al, 2018)。Li等(2018)沿Kalamazoo河支流Eagle溪的上游至下游选择了8个样品采集点(分别以Location 1–8命名)。每个采集点采集3个水体样品(分别以a, b, c命名), 用来研究水体中鱼类组成的多样性。8个采集点中, 采集点1是河流的最上游, 水流量非常小。采集点2–6的水流量逐渐增加。采集点7–8位于Kalamazoo河的干流中, 水流量最大。该研究发现从上游到下游, 物种多样性的相似度逐渐下降。采集点内的样品的相似性显著大于采集点间的相似性。8个样品的数据可以分别从NCBI SRA SRS2894037–SRS2894043, SRS2894045中下载。

将下载的fastq文件存放在./input文件夹后, 运行EPPS并获得结果(图4)。从图4可以看出, 最上游的采样点有独特的鱼类多样性。位于河流中游的采样点2–6有相似的鱼类多样性组成。采样点7–8也有较为类似的鱼类多样性, 与文中的结论一致(Li et al, 2018)。在资源消耗方面, 由于EPPS全部为自动化, 省去了人工的时间, 流程仅需要12 min即可完成24个样品的分析, 消耗18 G的硬盘空间, 大约是原始测序数据(6.8 G)的2.6倍。

EPPS有助于研究结果的重复分析及不同研究分析方法的共享。通过分析测试数据和已发表的公共数据, 我们获得了可靠的多样性分析结果, 证明了EPPS是一款快速可靠的宏条形码分析流程。

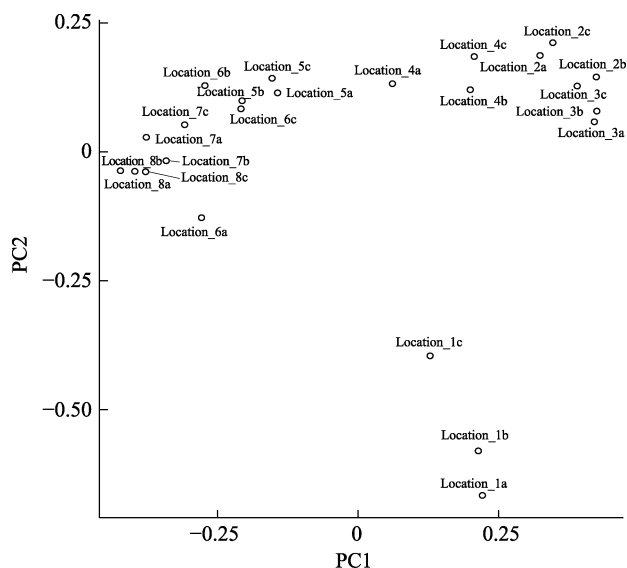


图4 公共数据的分析结果。样品的名称编号1–8分别代表从最上游到下游的8个采样点。编号的后缀a, b, c分别代表同一个采样地点的3次独立的重复取样。基于图中的结果, 最上游的样品**Location 1**有独特的鱼类多样性组成。**Location 3–6**有类似的鱼类多样性组成。最下游的样品**Location 7–8**有类似的鱼类多样性组成。

Fig. 4 EPPS result based on public data set. Samples are named from 1 to 8 from upstream to downstream. The suffix “a”, “b” and “c” indicate three replicates of the same sampling location. Based on the PCA, the most upstream sample (Location 1) has unique fish composition. Location 3–6 have similar fish composition. The downstream samples (Location 7–8) share similar fish composition.

参考文献

- Bazinet AL, Cummings MP (2012) A comparative evaluation of sequence classification programs. *BMC Bioinformatics*, 13, 92.
- Berger SA, Krompass D, Stamatakis A (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic Biology*, 60, 291–302.
- Bik HM, Interactive Pitch Inc. (2014) Phinch: An interactive, exploratory data visualization framework for–Omic datasets. *bioRxiv*, 009944.
- Bista I, Carvalho GR, Tang M, Walsh K, Zhou X, Hajibabaei M, Shokralla S, Seymour M, Bradley D, Liu S, Christmas M (2018) Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. *Molecular Ecology Resources*, 18, 1020–1034.
- Bohmann K, Evans A, Gilbert MT, Carvalho GR, Creer S, Knapp M, Douglas WY, De Bruyn M (2014) Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, 29, 358–367.
- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–2120.
- Boyer F, Mercier C, Bonin A, Le Bras Y, Taberlet P, Coissac E (2016) obitools: A unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, 16, 176–182.
- Brady A, Salzberg SL (2009) Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, 6, 673.
- Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13, 581.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: Architecture and applications. *BMC Bioinformatics*, 10, 421.
- Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences, USA*, 108, 4516–4522.
- Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7, 335.
- Cardoso P, Borges PA, Veech JA (2009) Testing the performance of beta diversity measures based on incidence data: The robustness to undersampling. *Diversity and Distributions*, 15, 1081–1090.
- Collen B, Whitton F, Dyer EE, Baillie JE, Cumberlidge N, Darwall WR, Pollock C, Richman NI, Soulsby AM, Böhm M (2014) Global patterns of freshwater species diversity, threat and endemism. *Global Ecology and Biogeography*, 23, 40–51.
- Crampton-Platt A, Timmermans MJ, Gimmel ML, Kutty SN, Cockerill TD, Vun Khen C, Vogler AP (2015) Soup to tree: The phylogeny of beetles inferred by mitochondrial metagenomics of a Bornean rainforest sample. *Molecular Biology and Evolution*, 32, 2302–2316.
- Crampton-Platt A, Douglas WY, Zhou X, Vogler AP (2016) Mitochondrial metagenomics: Letting the genes out of the bottle. *GigaScience*, 5, 15.
- Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME (2017a) Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, 26, 5872–5895.
- Deiner K, Renshaw MA, Li Y, Olds BP, Lodge DM, Pfrender ME (2017b) Long-range PCR allows sequencing of mitochondrial genomes from environmental DNA. *Methods in Ecology and Evolution*, 8, 1888–1898.
- Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C (2017) Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35, 316.
- Dowle EJ, Pochon X, Banks JC, Shearer K, Wood SA (2016) Targeted gene enrichment and high-throughput sequencing

for environmental biomonitoring: A case study using freshwater macroinvertebrates. *Molecular Ecology Resources*, 16, 1240–1254.

Edgar RC (2016) SINTAX: A simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv*, 074161.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26, 2460–2461.

Edgar RC (2013) UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, 10, 996.

Evans NT, Li Y, Renshaw MA, Olds BP, Deiner K, Turner CR, Jerde CL, Lodge DM, Lamberti GA, Pfreder ME (2017) Fish community assessment with eDNA metabarcoding: Effects of sampling design and bioinformatic filtering. *Canadian Journal of Fisheries and Aquatic Sciences*, 74, 1362–1374.

Evans NT, Olds BP, Renshaw MA, Turner CR, Li Y, Jerde CL, Mahon AR, Pfreder ME, Lamberti GA, Lodge DM (2016) Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources*, 16, 29–41.

Gerlach W, Stoye J (2011) Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Research*, 39, e91.

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Research*, 17, 377–386.

Ji Y, Ashton L, Pedley SM, Edwards DP, Tang Y, Nakamura A, Kitching R, Dolman PM, Woodcock P, Edwards FA, Larsen TH (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, 16, 1245–1257.

Li Y, Evans NT, Renshaw MA, Jerde CL, Olds BP, Shogren AJ, Deiner K, Lodge DM, Lamberti GA, Pfreder ME (2018) Estimating fish alpha- and beta-diversity along a small stream with environmental DNA metabarcoding. *Metabarcoding and Metagenomics*, 2, e24262.

Liu B, Gibbons T, Ghodsi M, Pop M (2010) MetaPhyler: Taxonomic profiling for metagenomic sequences. In: *Bioinformatics and Biomedicine (BIBM)*, 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 95–100.

Liu S, Wang X, Xie L, Tan M, Li Z, Su X, Zhang H, Misof B, Kjer KM, Tang M, Niehuis O (2016) Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. *Molecular Ecology Resources*, 16, 470–479.

Liu S, Li Y, Lu J, Su X, Tang M, Zhang R, Zhou L, Zhou C, Yang Q, Ji Y, Yu DW (2013) SOAPBarcode: Revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. *Methods in Ecology and Evolution*, 4, 1142–1150.

Lodge DM, Turner CR, Jerde CL, Barnes MA, Chadderton L, Egan SP, Feder JL, Mahon AR, Pfreder ME (2012)

Conservation in a cup of water: Estimating biodiversity and population abundance from environmental DNA. *Molecular Ecology*, 21, 2555–2558.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal*, 17, 10–12.

Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD (2012) PANDaseq: Paired-end assembler for Illumina sequences. *BMC Bioinformatics*, 13, 31.

Matsen FA, Kodner RB, Armbrust EV (2010) pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11, 538.

Millennium Ecosystem Assessment (2005) *Ecosystem and Human Well-being: Biodiversity Synthesis*. World Resources Institute, Washington, DC.

Munch K, Boomsma W, Huelsenbeck JP, Willerslev E, Nielsen R (2008) Statistical assignment of DNA sequences using Bayesian phylogenetics. *Systematic Biology*, 57, 750–757.

Newbold T, Hudson LN, Hill SL, Contu S, Lysenko I, Senior RA, Börger L, Bennett DJ, Choimes A, Collen B, Day J (2015) Global effects of land use on local terrestrial biodiversity. *Nature*, 520, 45.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'hara RB, Simpson GL, Solymos P, Stevens MH, Wagner H (2013) Package 'vegan'. *Community Ecology Package*, version. 2. (accessed on 2018-08-01)

Olds BP, Jerde CL, Renshaw MA, Li Y, Evans NT, Turner CR, Deiner K, Mahon AR, Brueseke MA, Shirey PD, Pfreder ME (2016) Estimating species richness using environmental DNA. *Ecology and Evolution*, 6, 4214–4226.

Patil KR, Roun L, McHardy AC (2012) The PhyloPythiaS web server for taxonomic assignment of metagenome sequences. *PLoS ONE*, 7, e38581.

Pfreder M, Hawkins C, Bagley M, Courtney G, Creutzburg B, Epler J, Fend S, Ferrington L Jr, Hartzell P, Jackson S, Larsen D (2010) Assessing macroinvertebrate biodiversity in freshwater ecosystems: Advances and challenges in DNA-based approaches. *The Quarterly Review of Biology*, 85, 319–340.

Pimm SL, Jenkins CN, Abell R, Brooks TM, Gittleman JL, Joppa LN, Raven PH, Roberts CM, Sexton JO (2014) The biodiversity of species and their rates of extinction, distribution, and protection. *Science*, 344, 1246752.

Piro VC, Matschkowski M, Renard BY (2017) MetaMeta: Integrating metagenome analysis tools to improve taxonomic profiling. *Microbiome*, 5, 101.

Price MN, Dehal PS, Arkin AP (2009) FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 26, 1641–1650.

R Core Team (2016) *R: A Language and Environment for Statistical Computing*. <https://www.R-project.org/>. (accessed

- on 2018-08-01)
- Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, 4, e2584.
- Rosen GL, Reichenberger ER, Rosenfeld AM (2010) NBC: The Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics*, 27, 127–129.
- Sato Y, Miya M, Fukunaga T, Sado T, Iwasaki W (2018) MitoFish and MiFish pipeline: A mitochondrial genome database of fish with an analysis pipeline for environmental DNA metabarcoding. *Molecular Biology and Evolution*, 35, 1553–1555.
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW (2009) Introducing Mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75, 7537–7541.
- Simon TP, Evans NT (2017) Environmental quality assessment using stream fishes. In: *Methods in Stream Ecology*, 3rd edn. (eds Hauer FR, Lamberti G), pp. 319–334. Elsevier, London.
- Slowikowski K (2018) ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. <https://CRAN.R-project.org/package=ggrepel>. (accessed on 2018-08-01)
- Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH (2012) Environmental DNA. *Molecular Ecology*, 21, 1789–1793.
- Tang M, Hardman CJ, Ji Y, Meng G, Liu S, Tan M, Yang S, Moss ED, Wang J, Yang C, Bruce C (2015) High-throughput monitoring of wild bee diversity and abundance via metagenomics. *Methods in Ecology and Evolution*, 6, 1034–1043.
- Thomsen PF, Kielgast JO, Iversen LL, Wiuf C, Rasmussen M, Gilbert MT, Orlando L, Willerslev E (2012) Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, 21, 2565–2573.
- Thomsen PF, Willerslev E (2015) Environmental DNA—An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, 183, 4–18.
- Uritskiy GV, DiRuggiero J, Taylor J (2018) MetaWRAP—A flexible pipeline for genome-resolved metagenomic data analysis. *bioRxiv*, 277442.
- Visconti A, Martin TC, Falchi M (2018) YAMP: A containerised workflow enabling reproducibility in metagenomics research. *GigaScience*, 7, giy072.
- Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73, 5261–5267.
- Wickham H (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York. <http://ggplot2.org>. (accessed on 2018-08-01)
- Wilcox TM, Zarn KE, Piggott MP, Young MK, McKelvey KS, Schwartz MK (2018) Capture enrichment of aquatic environmental DNA: A first proof of concept. *Molecular Ecology Resources*, 18, 1392–1401.
- Worm B, Barbier EB, Beaumont N, Duffy JE, Folke C, Halpern BS, Jackson JB, Lotze HK, Micheli F, Palumbi SR, Sala E (2006) Impacts of biodiversity loss on ocean ecosystem services. *Science*, 314, 787–790.
- Zhou HW, Li DF, Tam NF, Jiang XT, Zhang H, Sheng HF, Qin J, Liu X, Zou F (2011) BIPES, a cost-effective high-throughput method for assessing microbial diversity. *The ISME Journal*, 5, 741.
- Zhou X, Li Y, Liu S, Yang Q, Su X, Zhou L, Tang M, Fu R, Li J, Huang Q (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience*, 2, 4.

(特邀责任编辑: 周欣 责任编辑: 闫文杰)