



•综述•

# DNA条形码参考数据集构建和序列分析相关的新兴技术

刘山林\*

(中国农业大学植物保护学院, 食品营养与人类健康高精尖创新中心, 北京 100193)

**摘要:** 近年来DNA条形码技术迅速发展, 产生的条形码的数量及其应用范围都呈指数性增长, 现已广泛用于物种鉴定、食性分析、生物多样性评估等方面。本文重点总结并讨论了构建条形码参考数据库和序列聚类相关的信息分析的技术和方法, 包括: 基于高通量测序(high throughput sequencing, HTS)平台以高效并较低的成本获取条形码序列的方法; 同时还介绍了从原始测序序列到分类操作单元(operational taxonomic units, OTUs)过程中的一些计算逻辑以及被广泛采用的软件和技术。这是一个较新并快速发展的领域, 我们希望本文能为读者提供一个梗概, 了解DNA条形码技术在生物多样性研究应用中的方法和手段。

**关键词:** DNA条形码; 可操作物种单元; 聚类; 宏基因条形码; 高通量测序

## DNA barcoding and emerging reference construction and data analysis technologies

Shanlin Liu\*

Beijing Advanced Innovation Center for Food Nutrition and Human Health, College of Plant Protection, China Agricultural University, Beijing 100193

**Abstract:** DNA barcoding has been growing exponentially in terms of the number of barcode generated as well as its applications, e.g. as conservation tools in: species identification for damaged specimens, diet analysis from gut content and feces, biodiversity assessment from environmental DNA (eDNA), bulk arthropod samples or invertebrate-derived DNA (iDNA). These applications often require coupling with high throughput sequencing (HTS) technologies, and when done so are referred to as metabarcoding. Here, we discuss the methods used to generate reference barcodes using cost-efficient HTS platforms, and introduce several rules-of-thumb and some widely-used tools to conduct data quality control, denoising, and Operational Taxonomic Units (OTUs) clustering. We hope this review will help readers better understand how these emerging technologies can be implemented alongside existing technologies to accelerate biodiversity assessments in an accurate and efficient way.

**Key words:** DNA barcoding; OTUs; clustering; metabarcoding; high throughput sequencing

### 1 引言

DNA条形码是指用于鉴定物种的一个或多个标准化短基因片段(Hebert et al, 2003)。标准化DNA条形码参考数据库的构建需要遵循几个特征, 包括: (1)序列长度控制在目前测序技术可读取的长度范围内; (2)种间序列差异一般应大于种内差异; (3)具

有高度保守的侧翼序列以便于扩增引物的设计, 并保证其能覆盖足够多的代表物种; (4)最重要的是其应存在于绝大部分人们感兴趣的物种中(Kress & Erickson, 2012)。目前, 有多个符合这些标准的基因片段被广泛接受并作为DNA条形码应用。例如, 用于动物的有线粒体细胞色素氧化酶亚基(cytochrome c oxidase subunit 1, *COI*)的一段650 bp的序列

收稿日期: 2018-07-30; 接受日期: 2018-12-25

基金项目: 深圳市基础研究(自由探索)(JCYJ20170817150755701)

\* 通讯作者 Author for correspondence. E-mail: shanlin1115@gmail.com

(Hebert et al, 2003), 用于植物的有质体核酮糖1,5-二磷酸羧化酶基因(ribulose 1,5-bisphosphate carboxylase gene, *rbcL*)和成熟酶K基因(maturase K, *matK*) (Hollingsworth et al, 2009)以及用于真菌的转录间隔区(internal transcribed spacer, *ITS*) (Nilsson et al, 2009; Schoch et al, 2012) (表1)。

在Hebert等(2003)首次提出DNA条形码的概念后, DNA条形码技术在条形码数量及其应用方面都呈指数增长, 例如作为生物多样性保护的工用于: 受损标本的物种鉴定(Armstrong & Ball, 2005)、通过肠道内容物和粪便分析食性(Kunz & Whitaker Jr, 1983; Bohmann et al, 2014)、基于环境 DNA (environmental DNA, eDNA) (Baird & Hajibabaei, 2012; Taberlet et al, 2012)和生物混合样品或无脊椎动物源DNA (invertebrate-derived DNA, iDNA)样品 (Yu et al, 2012; Bohmann et al, 2013; Liu et al, 2013) 进行生物多样性评估。这些应用通常需要依赖于高通量测序(HTS)技术, 并被称为宏基因条形码技术 (metabarcoding) (Taberlet et al, 2012)。宏基因条形码技术或宏基因组学的方法最初主要应用于微生物学领域, 通过从各种环境样品中提取的DNA来分析表征微生物群落(Caporaso et al, 2011)。过去十多年的研究表明, 此种方法同样可以应用于动物和植物群落 (后文中将其称为大生物群落, macrobial community) (Hajibabaei et al, 2016; Deiner et al, 2017)。考虑到微生物宏基因组学研究已经非常成熟, 本文将重点总结和讨论与大生物群落相关的宏基因条形码的技术进展, 主要包括构建DNA条形码参考数据库和DNA序列聚类的一些技术和方法。

表1 广泛用于DNA条形码技术的标记基因  
Table 1 Marker genes widely used for barcoding

标记基因 Marker gene	目标物种 Targeted group	数据库 Database
<i>16S</i>	细菌和古细菌 Bacteria and archaea (Sogin et al, 2006)	核糖体数据库项目 Ribosomal Database Project (RDP, Cole et al, 2008); Greengenes (DeSantis et al, 2006); SILVA (Pruesse et al, 2007)
<i>ITS</i>	真菌(Schoch et al, 2012)、植物(Group et al, 2011)、原生生物 (Pawlowski et al, 2012) Fungi (Schoch et al, 2012); plant (Group et al, 2011); protist (Pawlowski et al, 2012)	UNITE (Kõljalg et al, 2005); GenBank (Benson et al, 2012)
<i>18S</i>	原生生物 Protist (Pawlowski et al, 2012)	SILVA (Pruesse et al, 2007)
<i>matK + rbcL</i>	植物 Plant (Hollingsworth et al, 2009)	生命条形码数据库 Barcode of Life Data Systems (BOLD, Ratnasingham & Hebert, 2007); GenBank (Benson et al, 2012)
<i>COI</i>	动物群(Hebert et al, 2003)、原生生物(Pawlowski et al, 2012) Fauna (Hebert et al, 2003) and protist (Pawlowski et al, 2012)	核糖体数据库项目 Ribosomal Database Project (RDP, Cole et al, 2008)

2 构建DNA条形码参考数据库

通过过去十多年中全球范围内科学家的通力合作, DNA条形码参考序列数据库, 例如BOLD (Ratnasingham & Hebert, 2007), 已经初具规模。然而, 目前参考数据库存在的一个典型问题是其数据在地理空间和物种覆盖度方面均存在很大程度上的不平衡, 这主要是由于全球各地在条形码研究方面投入差异所致, 尤其是在物种多样性热点地区, 科研投入尤其是分子生物学相关方面的投入不足, 制约了这些地区物种信息的数字化。尽管HTS平台的单碱基的测序成本显著下降, 但由于测序长度相对较短, 并不适用于对长扩增子测序(例如, *COI*基因条形码包括引物序列长~719 bp, 而最新的Miseq系统最长可以完成双端300 bp的测序, 仍然无法测通标准条形码序列), 使得Sanger测序仍然是目前获取DNA条形码序列(表1)的主流技术。

不可否认, 随着集中式和工业化的实验室流程的普及, 用于标准DNA条形码技术的分析成本自2000年以来已显著降低(Hebert et al, 2016)。Meier等(2016)通过咨询条形码研究中心, 如安大略生物多样性研究所(Biodiversity Institute of Ontario, BIO)和加拿大DNA条形码中心(Canadian Centre for DNA Barcoding, CCDB), 真实地估算了分析成本。结果表明, 对于具有高质量DNA的样品, 如果不包含额外的服务(例如重新索要收据、二次抽样、返回剩余的组织或DNA等), CCDB的每个样品的商业成本约为20美元。如果序列和标本图片共享给国际生命条形码项目(International Barcode of Life, iBOL),

费用将降至10美元左右。由于Sanger测序技术即将接近其通量和相关化学成本的极限,因此基于此的条形码测序成本不太可能进一步大幅降低。据估计,在全球范围内对物种进行条形码登记(Hebert et al, 2016)需要测序1亿个样本。这意味着仅仅构建全球条形码参考序列的预算就需要约10亿美元。参考数据库的不足导致基于HTS的宏基因条形码研究的结论常常只能限于使用分类操作单元(operational taxonomic units, OTUs),无法确定其具体的物种信息(Linnaean species name),因而不能将现有的生物学和生态学的知识与用此方法得出的多样性分布规律结合起来。

科学家一直在尝试使用HTS平台以更低的成本和人工投入获取DNA条形码参考序列。为了解决读长的限制,不同的学者采用了不同的方法,包括:获取标准条形码序列的局部序列;用多轮PCR的方法扩增全长条形码的不同区域;又或者利用三代单分子高通量测序技术,如Pacific Biosciences的长读长单分子实时(Single Molecular Real Time, SMRT)测序系统;通过生物信息学算法来拼装填补由于读长限制而产生的序列中部的缺口(gap)。例如,早期的研究分别对不同物种单独进行PCR后再混合,通过Roche 454平台获取物种条形码序列(Shokralla et al, 2014),但是由于测序通量的限制,化学试剂成本高,454平台被迫退出市场,而且此方法同Meier等(2016)的方法(基于Illumina平台)一样无法获取标准全长的条形码序列。研究人员还试图应用两轮

PCR扩增,每一轮分别扩增全长条形码的部分序列,随后通过简单拼接获取全长序列(Shokralla et al, 2015; Cruaud et al, 2017)。另外,单分子测序平台SMRT技术可以获取环形一致序列(circular consensus sequences, CCSs),对一个分子多次测序,可以有效校正该平台固有的高测序错误率(10%–16%) (Eid et al, 2008)。因此,有研究测试了将SMRT技术应用于条形码数据库构建的可行性(Liu et al, 2017; Hebert et al, 2018),研究测试的混合样品中,最复杂的样品有来自多达10,000个不同样本的DNA中的COI扩增子。结果表明随着其测序成本的进一步下降,SMRT技术将会是构建DNA条形码数据库方向的强有力的方法之一。Liu等(2017)还提出了一套新的解决方案(HIFI-Barcode),这是一套准确高效的生物信息学替代方案。该方法可以使研究者基于目前最经济高效的Hiseq平台,以现有成本的1/10获取全长条形码序列,而且不需要额外的PCR步骤(表2)。与此同时,该研究组还利用BGISEQ-500平台最新的单端400 bp (SE 400)测序技术开发了一套简单有效的DNA条形码数据库的实验和分析流程(HIFI-SE) (Yang et al, 2018)。此方法利用测序读长的优势,可以通过简单的两端序列比对连接从而获得全长的DNA条形码序列。

基于高通量技术的条形码参考序列构建方法还体现了更高的灵敏度,进一步降低了操作成本。虽然目前有对Sanger测序峰图自动读取和识别的程序,但是Sanger测序往往要求对结果峰图文件进行

表2 利用高通量测序平台批量获取DNA条形码的方法

Table 2 High throughput methods to achieve barcode sequences

目标序列长度 Targeted region length (bp)	优势 Advantages	劣势 Disadvantages	参考文献 Reference
~300	—	无法处理较长的目标序列; Roche 454平台 Can not work on long fragments; Roche 454 platform	Shokralla et al, 2014
~180	简单, 易操作, 成本低 Straightforward, easy to operate, cost-efficient	目标序列偏短, 只能用于物种初筛 Short targeted region; can only be used for species pre-clustering	Meier et al, 2016
~650	标准DNA条形码全长 Standard full-length COI	普适性差; 需要多轮PCR过程 Poor universality; multiple rounds of PCR	Shokralla et al, 2015; Cruaud et al, 2017
~650	易操作, 标准DNA条形码全长 Easy to operate, standard full-length COI	相对较高的计算资源 Relatively high requirement for computational resources	Liu et al, 2017
~650	易操作, 标准DNA条形码全长 Easy to operate, standard full-length COI	SMRT平台成本高 High cost of SMRT platform	Hebert et al, 2018
~650	易操作, 标准DNA条形码全长 Easy to operate, standard full-length COI	测序平台暂时不够普及 Not a mass production	Yang et al, 2018



肉眼观察甄别以优化数据质量,因此很难开展高效的自动化分析流程。而基于HTS的方法就不存在这样的问题,使得高效自动化成为可能。此外,这些基于HTS的方法可以检测微量的PCR扩增子,因而可以获取那些“失败”的PCR扩增子(电泳凝胶上没有明显条带),进一步提高了总体条形码成功率(Liu et al, 2017)。虽然所有这些方法都需要在引物上添加样品特异的标签序列,这会导致在最开始订购引物时产生一次性的费用,然而一次性的引物合成可以进行数以千次的反应,每次反应中的引物成本可以少到忽略不计。

虽然上述所有的方法目前都只是在动物类群的COI相关基因中通过验证,但是根据其方法原理,SMRT技术和HIFI-barcode技术可以很容易转移应用于其他类型的标记基因,如用于植物的*rbcL*和*matK*基因,但是需要进一步的实验证明。总之,我们相信,如果这些新方法在分类学中得到迅速而广泛的应用,将为全球生物DNA条形码的生成记录开辟新纪元,最终完成全球生物物种条形码信息的数字化。

### 3 DNA序列聚类

对DNA分子测序时,HTS平台不同于Sanger测序,它对结合在测序芯片上的几乎所有DNA分子测序,分别产出单独的序列。因此,源自于PCR扩增的错误如单碱基替换和嵌合体,因为在扩增产物中比例较低,在Sanger测序中不会被测到,但是这些问题序列在HTS测序过程中将会被测序并输出为有效序列。这个问题在基于扩增子的生物多样性研究中尤为显著,导致OTU预测数目大大增加,产生的序列多样性远远高于抽样的生物群落的“真实”丰富度,进而高估多样性,产生不可信的生态发现。如何区分有意义的生物学变异(种内和种间变异)与PCR和测序过程中的错误,是目前相关分析工具包开发的主要问题,也是最核心的挑战之一。考虑到现有条形码参考数据库的完整性不足(在物种多样性及单一物种的空间覆盖度方面都有很大的偏向性),大部分的研究分析都是根据序列相似度采用从头聚类的方法,以便于真实评估所研究样点的物种多样性信息(alpha diversity),而非通过比较一个预先准备好的数据库进行多样性分析(Quast et al, 2012; Zhang et al, 2013)。本文将简要介绍几种广泛使用的

工具以及其背后的运算逻辑,以帮助读者更好地理解如何将大规模HTS序列聚类成有生态学意义的OUT(图1)。由于早期很多为Roche 454焦磷酸测序开发的软件已经不再被广泛使用,将不在本文中进行讨论。

首先,在进行聚类分析之前,通过Illumina HiSeq或Miseq测序产生的原始序列需要进行质量过滤,主要包括去除建库或测序过程中产生的错误序列,如:(1)含有任何测序接头的序列。在文库制备过程中,接头序列会添加到插入片段两端。如果模板长度小于测序长度,测序序列就会含有接头序列。接头也可能出现在序列中间,这是由于建库过程中引物结合到模板错误的位置,或者PCR反应中退火,延伸不足所致。对于前者,目前常见的可用工具有AdapterRemoval (Schubert et al, 2016)、Skewer (Jiang et al, 2014)和SOAPnuke (Chen et al, 2017)等,其中部分软件还可以去除PCR引物,如果提供标签序列,还可以进一步拆分样品。如果是后者,需要将这样的序列整条去除。大多数测序服务提供商都提供接头报告文件帮助用户去除接头序列。(2)含低质量(Q)碱基的序列。Q值平均值虽然是一个广泛采用的参数,但是在OTU聚类分析中被证明并不是一个可取的办法(Edgar, 2013)。假设有2个相同长度为150 bp的序列,其质量值分别为 $140 \times Q35 + 10 \times Q2$ 和 $150 \times Q25$ ,那么它们的平均Q值是一样的,但是前一条序列的预期错误碱基数为6.4,而后者为0.5。因此,对于有低Q值的序列,正确的

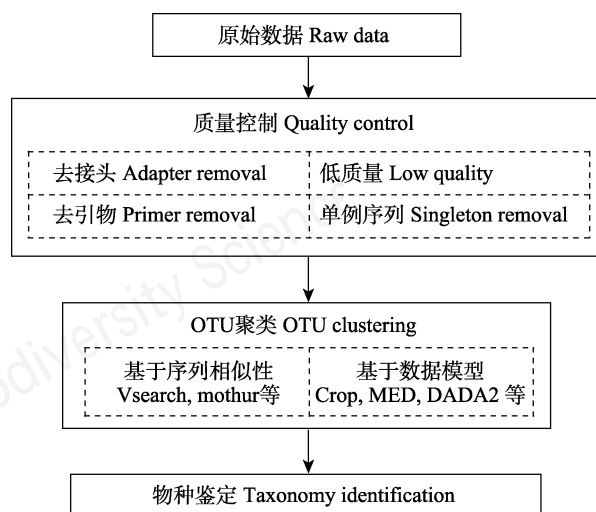


图1 条形码分析的数据处理流程图

Fig. 1 Diagram of DNA barcode data analysis

做法是删除整条的序列,或删减其末端低质量的序列。

经过序列的预处理之后,大多数OTU聚类流程,如U/VSEARCH (Edgar, 2010; Rognes et al, 2016)、DADA2 (Callahan et al, 2016)、UPARSE (Edgar, 2013)等,都会进行去除单例序列(singletons)的过程:首先将相同序列合并为一条代表序列并记录其丰度,随后去除单例序列(丰度为1的序列)。单例序列被普遍认为是错误序列。此外,这一步降噪处理还会大大降低序列数量,从而减轻后续分析的计算负荷。一些软件还包含额外的处理步骤,如将序列删除为长度一致的序列(Edgar, 2013)或进行氨基酸翻译检查(Liu et al, 2013)。然而这些额外的降噪处理只适合某些特定的基因,在使用时需要谨慎,确认其使用范围。另外,尽管不同的研究在嵌合体分析中,有的分析流程倾向于在去除单例序列之后马上进行,有的倾向于在OTU聚类之后进行,但是其嵌合体鉴定的原理基本相似,它们都试图找出双源嵌合体(bimera, two-parent chimera),即嵌合体序列头尾两端分别来自于同一混合样品中高丰度的其他序列(Edgar et al, 2011; Callahan et al, 2016)。

现已发表的OTU聚类算法大致可以分为两类:(1)首先计算两两序列相似度,然后以高丰度序列为根序列,用一个预先设定的相似度(一般为97%)将这些序列进行分组。(2)用数学模型中心化和表征每个聚类单元。传统聚类方法需要计算比较所有的序列来计算距离矩阵,已经很难处理现阶段研究的数据量(Matias Rodrigues & von Mering, 2013)。而目前广泛流行的USEARCH (Edgar, 2010),亦或是开源的VSEARCH (Rognes et al, 2016),或其特定版本的UPARSE (Edgar, 2013)都是基于快速启发式聚类算法。这个算法比传统方法能更快找到与目标序列相似的一个或几个代表序列,极大降低了计算复杂度。因此目前主流的很多软件采取了类似的聚类方法,包括:QIIME (UCLUST) (Caporaso et al, 2010)和mothur (Schloss et al, 2009)。CROP是一个基于高斯混合模型(Gaussian Mixture Model)的聚类方法(Hao et al, 2011)。它将高斯分布的平均值替换为一条中心序列,以此来代表一个特定的组并利用高斯分布来处理测序错误和种间变异。DADA2 (Callahan et al, 2016)开发了一套基于测序质量值的模型,用于估测Illumina扩增子测序中的错误。首先,

通过对错配碱基和其测序质量值之间进行加权LOESS局部建模(weighted loess fit),然后通过对数据模型和观测数据的最佳拟合结果将序列聚类。MED (Minimum Entropy Decomposition) (Eren et al, 2015)采用了最小熵分解的算法,利用序列间的信息不确定性来迭代分解数据集,直到每个最终待解释的单元满足最大熵标准。

这些基于模型的方法不需要像前述基于序列相似性的聚类方法那样需要预先设定一个临界值(如97%),而是通过数据本身特性对序列进行聚类。此外,Frøslev等(2017)提出了一个聚类后处理的方法(LULU),结合序列相似度和共现(co-occurrence)模式从群落数据中去除错误的OTUs。此方法采用了一个类似于MED的算法,但是其特点在于OTU聚类之后的数据处理,将在涵盖多时空样本的研究中发挥重要作用。相比于传统基于序列相似度矩阵的方法,这些基于数学模型的聚类方法可以降低OTU和 $\alpha$ 多样性高估的状况,有望找到更多真实的变异,同时减少错误序列,从而获得更具有生态学意义的分类单元。得到OTU的参考序列之后,大部分研究会进一步探讨其物种分类。本文不过多讨论与OTU物种鉴定相关的主题。简而言之,物种鉴定可以通过使用BLAST或其他类似比对工具,将序列比对到已建立的参考数据库,如BOLD (Ratnasingham & Hebert, 2007)、Genbank (Benson et al, 2012)或其他用户定制的数据库。物种分类可以通过最佳匹配以及其序列相似度确定(Shi et al, 2018),也可以采用基于系统发育树的方法,根据序列在系统发育树中的位置来确定其物种分类(Zhang et al, 2013)。这两个方法都需要可靠的数据库。如果不能保证数据库的完整度和正确性,目标序列和数据库数据比对的不确定性将导致模糊的,甚至是错误的物种鉴定。

#### 4 总结和展望

我们正处于一个物种灭绝速度超过发现速度的时代,很多物种在得到描述之前就灭绝了。完善正确地描述生物界所有物种的时空分布才能使我们全面了解生物多样性的现状及其驱动力(人为的、自然的,或两者兼而有之),进而才能给我们提供有价值的保护和管理策略,缓解甚至遏止生物多样性持续降低的趋势。条形码技术旨在协助分类学家进

行物种鉴定并加速这一过程,而不是取代传统分类方法(Hebert et al, 2003)。条形码技术具有广泛的适用性(可以应用于生命之树上的几乎所有物种,无论个体的大小),而且所需的专业训练不及传统分类学繁重复杂,使得其不但成为分类学家的工具,也扩展繁衍出一系列为生态学家和公众服务的工具,应用于如受损样本和走私货物物种鉴定,结合 HTS 技术的生物多样性评估等。另外,条形码技术不能也不应该局限于基于扩增子的研究分析,随着测序成本的降低,条形码技术也在不断发展。略过 PCR 步骤(Zhou et al, 2013; Tang et al, 2015)和基于捕获芯片的技术(Liu et al, 2016)有望帮助科学家们更准确地理解生物多样性的组成及其驱动的生态过程。基于 eDNA 和 iDNA 的相关研究的顺利开展,将有助于揭示许多物种多样性的分布规律,尤其是那些小型的、隐秘的物种(Schnell et al, 2012; Mahon et al, 2013; Turner et al, 2014)。然而需要注意的是,条形码技术及其相关应用更多的是简化了对生态系统中物种鉴定和多样性评估的方法,但并不能取代合理的生态学设计,也不能减少样品采集的工作量。研究人员需要根据自身研究的特点设计样品采集方案,使其最终的研究具有统计学及生物学的意义,包括采集足够的样点分布和重复个数,以及涵盖相应的环境参数,如气候特征和 pH 梯度等。

## 参考文献

- Armstrong K, Ball S (2005) DNA barcodes for biosecurity: Invasive species identification. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360, 1813–1823.
- Baird DJ, Hajibabaei M (2012) Biomonitoring 2.0: A new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, 21, 2039–2044.
- Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2012) GenBank. *Nucleic Acids Research*, 41, D36–D42.
- Bohmann K, Evans A, Gilbert MTP, 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.
- Bohmann K, Schnell IB, Gilbert MTP (2013) When bugs reveal biodiversity. *Molecular Ecology*, 22, 909–911.
- Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13, 581–583.
- Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7, 335–336.
- 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.
- Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, Li S, Li Y, Ye J, Yu C, Li Z (2017) SOAPnuke: A MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Giga-Science*, 7, gix120.
- Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen A, McGarrell DM, Marsh T, Garrity GM (2008) The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, 37, D141–D145.
- Cruaud P, Rasplus J-Y, Rodriguez LJ, Cruaud A (2017) High-throughput sequencing of multiple amplicons for barcoding and integrative taxonomy. *Scientific Reports*, 7, 41948.
- Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, Vere N (2017) Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, 26, 5872–5895.
- DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* 72, 5069–5072.
- 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, 996–998.
- Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27, 2194–2200.
- Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B (2008) Real-time DNA sequencing from single polymerase molecules. *Science*, 323, 133–138.
- Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML (2015) Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *The ISME Journal*, 9, 968–979.
- Frøslev TG, Kjølner R, Bruun HH, Ejrnæs R, Brunbjerg AK, Pietroni C, Hansen AJ (2017) Algorithm for post-clustering



- curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*, 8, 1188.
- Group CPB, Li DZ, Gao LM, Li HT, Wang H, Ge XJ, Liu JQ, Chen ZD, Zhou SL, Chen SL (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (*ITS*) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences, USA*, 108, 19641–19646.
- Hajibabaei M, Baird DJ, Fahner NA, Beiko R, Golding GB (2016) A new way to contemplate Darwin's tangled bank: How DNA barcodes are reconnecting biodiversity science and biomonitoring. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 371, 20150330.
- Hao X, Jiang R, Chen T (2011) Clustering 16S rRNA for OTU prediction: A method of unsupervised Bayesian clustering. *Bioinformatics*, 27, 611–618.
- Hebert PD, Braukmann TW, Prosser SW, Ratnasingham S, Ivanova NV, Janzen DH, Hallwachs W, Naik S, Sones JE, Zakharov EV (2018) A Sequel to Sanger: Amplicon sequencing that scales. *BMC Genomics*, 19, 219.
- Hebert PD, Cywinska A, Ball SL (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*, 270, 313–321.
- Hebert PD, Hollingsworth PM, Hajibabaei M (2016) From writing to reading the encyclopedia of life. *Proceedings of the Royal Society of London B: Biological Sciences*, 371, 20150321.
- Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences, USA*, 106, 12794–12797.
- Jiang H, Lei R, Ding SW, Zhu S (2014) Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, 15, 182.
- Köljal U, Larsson KH, Abarenkov K, Nilsson RH, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjoller R, Larsson E (2005) UNITE: A database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist*, 166, 1063–1068.
- Kress WJ, Erickson DL (2012) DNA barcodes: Methods and protocols. In: *DNA Barcodes* (eds Kress WJ, Erickson DL), pp. 3–8. Humana Press, Totowa.
- Kunz TH, Whitaker JO Jr (1983) An evaluation of fecal analysis for determining food habits of insectivorous bats. *Canadian Journal of Zoology*, 61, 1317–1321.
- Liu S, Li Y, Lu J, Su X, Tang M, Zhang R, Zhou L, Zhou C, Yang Q, Ji Y (2013) SOAPBarcode: Revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. *Methods in Ecology and Evolution*, 4, 1142–1150.
- Liu S, Wang X, Xie L, Tan M, Li Z, Su X, Zhang H, Misof B, Kjer KM, Tang M (2016) Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. *Molecular Ecology Resources*, 16, 470–479.
- Liu S, Yang C, Zhou C, Zhou X (2017) Filling reference gaps via assembling DNA barcodes using high-throughput sequencing—Moving toward barcoding the world. *GigaScience*, 6, 1–8.
- Mahon AR, Jerde CL, Galaska M, Bergner JL, Chadderton WL, Lodge DM, Hunter ME, Nico LG (2013) Validation of eDNA surveillance sensitivity for detection of Asian carps in controlled and field experiments. *PLoS ONE*, 8, e58316.
- Matias Rodrigues JF, von Mering C (2013) HPC-CLUST: Distributed hierarchical clustering for large sets of nucleotide sequences. *Bioinformatics*, 30, 287–288.
- Meier R, Wong W, Srivathsan A, Foo M (2016) \$1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. *Cladistics*, 32, 100–110.
- Nilsson RH, Ryberg M, Abarenkov K, Sjökvist E, Kristiansson E (2009) The ITS region as a target for characterization of fungal communities using emerging sequencing technologies. *FEMS Microbiology Letters*, 296, 97–101.
- Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, Berney C, Bowser SS, Cepicka I, Decelle J, Dunthorn M (2012) CBOL protist working group: Barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biology*, 10, e1001419.
- Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO (2007) SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 35, 7188–7196.
- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2012) The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, 41, D590–D596.
- Ratnasingham S, Hebert PD (2007) BOLD: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Molecular Ecology Notes*, 7, 355–364.
- Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, 4, e2584.
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ (2009) Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75, 7537–7541.
- Schnell IB, Thomsen PF, Wilkinson N, Rasmussen M, Jensen LRD, Willerslev E, Bertelsen MF, Gilbert MTP (2012) Screening mammal biodiversity using DNA from leeches. *Current Biology*, 22, R262–R263.
- Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL,

- Levesque CA, Chen W, Bolchacova E, Voigt K, Crous PW (2012) Nuclear ribosomal internal transcribed spacer (*ITS*) region as a universal DNA barcode marker for fungi. *Proceedings of the National Academy of Sciences, USA*, 109, 6241–6246.
- Schubert M, Lindgreen S, Orlando L (2016) AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Research Notes*, 9, 88.
- Shi ZY, Yang CQ, Hao MD, Wang XY, Ward RD, Zhang AB (2018) FuzzyID2: A software package for large data set species identification via barcoding and metabarcoding using hidden Markov models and fuzzy set methods. *Molecular Ecology Resources*, 18, 666–675.
- Shokralla S, Gibson JF, Nikbakht H, Janzen DH, Hallwachs W, Hajibabaei M (2014) Next-generation DNA barcoding: Using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular Ecology Resources*, 14, 892–901.
- Shokralla S, Porter TM, Gibson JF, Dobosz R, Janzen DH, Hallwachs W, Golding GB, Hajibabaei M (2015) Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports*, 5, 9687.
- Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proceedings of the National Academy of Sciences, USA*, 103, 12115–12120.
- 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 (2015) High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution*, 6, 1034–1043.
- Turner CR, Miller DJ, Coyne KJ, Corush J (2014) Improved methods for capture, extraction, and quantitative assay of environmental DNA from Asian bigheaded carp (*Hypophthalmichthys* spp.). *PLoS ONE*, 9, e114329.
- Yang C, Tan S, Meng G, Bourne DG, O'Brien PA, Xu J, Liao S, Chen A, Chen X, Liu S (2018) Access COI barcode efficiently using high throughput Single End 400 bp sequencing. *bioRxiv*, doi: 10.1101/498618.
- Yu DW, Ji YQ, Emerson BC, Wang XY, Ye CX, Yang CY, Ding ZL (2012) Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, 3, 613–623.
- Zhang J, Kapli P, Pavlidis P, Stamatakis A (2013) A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, 29, 2869–2876.
- 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: 时意专)