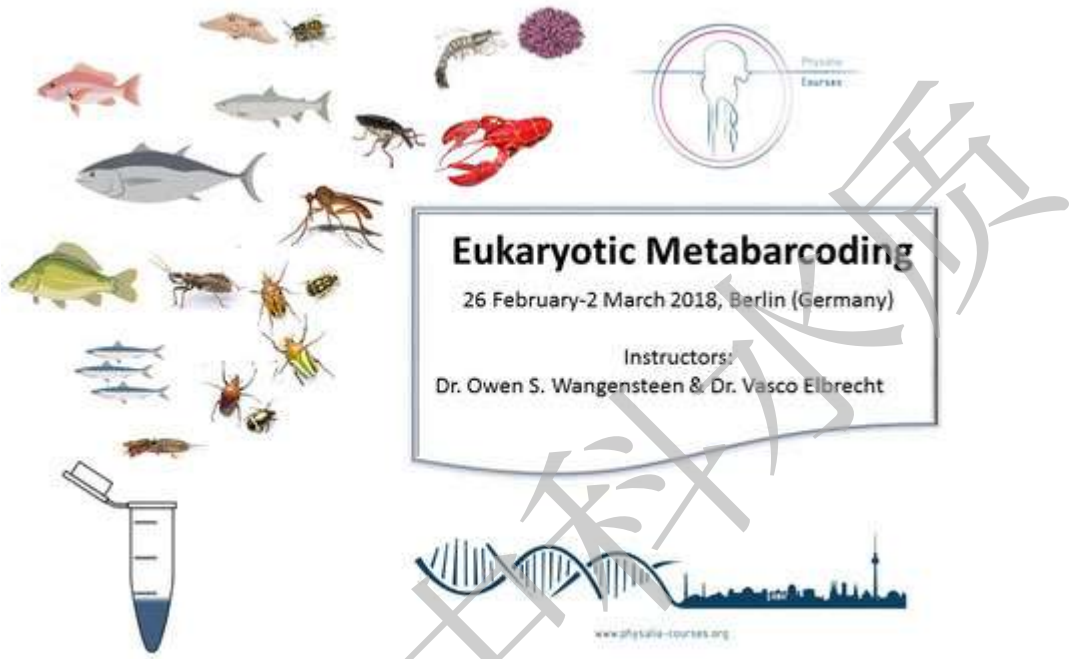
元编码(METABARCODING)：改变生物多样性调查的方式

(本文由 Elizabeth Bourne 于 2018 年 1 月 9 日在网络平台公开，由中国科学院生态环境研究中心王子健研究员编译。文章比较系统的分析了原编码(或宏编码)在生物多样性调查实践中可能存在的科学和技术问题，因此对 eDNA 技术在当前环境生物检测和评价中的应用关键技术有提示意义，所以值得推荐阅读。)



在过去的 15 年里，DNA 条形码(DNA BARCODING)已经改变了我们研究生物多样性的传统方式。现在这个领域正在从单一物种条形码转向高通量的元编码。元编码是一种生物多样性快速评估的方法，涉及高通量 DNA 测序、生物信息学、计算基础设施和实验设计。

本文是以问答方式讨论了元编码技术及其在生物多样性调查中的应用，参与本文的讨论的有 Dr. Vasco Elbrecht (加拿大 Guelph 大学) 和 Dr. Owen S. Wangensteen (英国 Salford 大学)。

问：你们是如何将元编码应用到你的研究领域的?自从你们开始这个领域的研究以来，这项技术又是如何发展的?

Vasco: 当我在 2014 年开始使用元编码时，我想使用这个工具来识别淡水大型无脊椎动物分类。当时我在做一个中宇宙的野外模拟实验，目的是测试多重应激对水生态系统的影响。当我首次想要应用这种方法做分类鉴定时，我就意识到实际上生态学家早已开始使用元编码来回答生态学问题了，只是尚没有涉及到方法学

本身的片面性。从那时起，我就一直在探索元编码中潜在的问题，如引物的选择、生物量的影响、不同标记系统的差异，以及实操流程中其他步骤的影响。我很高兴方法学验证已经成为元编码领域的讨论重点，我们现在所处的阶段已经可以将元编码技术可靠、常规地用于生物监测和生物多样性研究了。然而，元编码的许多方面仍然需要进一步改进和验证。此外，元编码方法的使用和需求的增加为有抱负的学生和研究人员提供了良好的就业机会。我的心愿仍然是保持这个不断增长的元编码科学圈更加开放、友好和相互支持。因此我乐意尽我所能分享和讨论我在这个领域的观点。

Owen: 我是从研究非常复杂的群落生态系统开始的，目的是分析整个海洋真核底栖生物的多样性，那时候没人想到用细胞色素氧化酶(COI)来做标记。当时大部分人都在用 18S 研究真核微生物样本,普遍认为并没有真正用来开发 COI 的通用引物。与 COI 标记相比，常用 18S 标记的变异性较小，因此拥有比 COI 更加高的分类分辨率。从那时起逐步开发出近乎通用的 COI 引物和分类特异性的 COI 引物，开始能够检测大多数真核生物多样性(包括几乎所有的后生动物)，并且降低了由引物所导致的偏差。我们现在意识到对低丰度物种的体外环境 DNA 进行元编码与对富含组织样品的整个群落 DNA 进行元编码之间存在巨大的实践差异。这些差异必须从项目一开始就要认真考虑，因为这两种技术路线在取样设计、可重复性、所需的重复次数、样品前处理或样品沾污的可能性方面非常不同。

问：当设计基于元编码的野外调查或实验时，需要牢记在心的最重要关键点是什么？

Vasco: 当开始一个项目时，采用原编码工具是否能够解决所研究的问题及其确定性如何是关键。此外，项目设计过程中应该经过几轮来自你的同事(还有一些不熟悉这个项目的人)的意见反馈，并且应该验证在你的实验室建立的元编码方法对于你的研究问题是否适用和对你所研究的目标分类群是否足够有效。如果不是，可以考虑先进行一个小规模的验证实验，这样写高影响的论文也会容易一些。一旦你确信计划使用的元编码方法是可靠和适当的，请记住在你的研究中一定要同步包括阴性/阳性对照，可能还需要重复实验。这将使你在后续的生物信息学分析步骤容易一些，结论更可靠一些，同时给读者(或审稿人)更强的信心。同时，确保每个样本的测序要有足够的深度；从工作量角度，你可以减少需要深度测序的样本数量，而不是采用大样本量和弱化测序深度的方式来应对科学问题。

Owen: 其实有许多需要关注的重要问题。对我而言，最重要的一点是要清楚地认识到是依据非常低的离体环境 DNA 浓度来调查低丰度物种；亦或是根据对保存很好且富含组织整体样本的群落 DNA 进行评估。对研究目的的选择会对标记物的选择、采样设计(以及所需要的重复样本量)、分析工作的前处理和所采用的生物信息学分析流程产生深刻的影响。例如，如果我希望分析水样中的鲨鱼种群，我可能需要非常特异的一组引物，否则就会得到非特异扩增的细菌或微型真核生物 DNA。此时我不会对定量感兴趣，而只对样品中是否存在鲨鱼物种感兴趣。同时我可能会需要既有生态学意义上的，亦有技术层面的多个重复 PCR 实验，并通过对数据的分析得到期望的高水平计量化学证据。反之，如果我希望从昆虫

混合物匀浆来扩增和检测昆虫的整体 DNA，我则需要更加通用的引物，以便能够以较低引物偏差方式扩增得到大部分昆虫物种信息。此时我对定量数据的更加感兴趣(此处应该指多样性等参数的定量，而不是物种丰度的定量)，会需要较少的重复实验。这是因为采用通用引物对一个富含组织的混合样本进行扩增得到大部分昆虫分类群，随机性比从广袤的大海中检测一段鲨鱼 DNA 要小得多。

问：有很多的聚类方法或算法可用于生成操作性分类单元(OTUs)，你们能提供一些指导或概述一下其中某些方法或算法，有什么值得推荐的技术路线吗?

Vasco: 这个问题很难回答(扩增子的 OTU 聚类和降噪处理在生物信息学分析中必不可少，其中 UPARSE, DADA 和 UNOISE 比较常用，各有优缺点)。我个人倾向采用 Robert Edgar 开发的 UPARSE-OTU 算法(网络版本地址是 http://www.drive5.com/usearch/manual/uparseotu_algo.html)，因为这个算法包括有先进的去除嵌合体方法(指实验过程中产生的假序列)，而且计算速度相当快。但需要注意的是，固定阈值算法(UPARSE-OUT 采用的聚类相似性阈值是 97%)可能会将多个低遗传多样性的物种合并为一个 OUT，或者由于高遗传多样性或测序错误而将单一物种的的测序数据聚类成多个 OUT。解决这个问题，诸如 SWARM 这样的聚类方法可以提供帮助，因为这种聚类工具采用了更灵活的聚类阈值(但也可能会过度拆分了 OTUs)。最后需要注意，认识对所选聚类算法的潜在偏见和局限性是关键。没有一种算法是完美的，其结果高度依赖于样品中实际存在的靶标生物体、实验室流程和生物信息数据过滤步骤。除了聚类，OTU 数量还受到许多其他参数的影响，因此应该谨慎对待。即使在将 OTU 分配给引物数据库时也应该记住，并非所有序列都可能被分配到正确的(形态学)分类群。找到和选择一个合适的聚类算法至关重要，但元编码步骤相关的聚类步骤同样重要。此时，采用比较灵活的聚类方法如 SWARM 可以提供帮助(但是可能导致 OUT 的过度分裂)。最后，选择聚类算法时还是要记住方法本身潜在的偏差和不足。当然，任何一种算法都有缺陷，结果高度依赖靶标生物，实操规则和生物信息学中的数据过滤步骤。由于 OUT 的数目受到除聚类外太多参数的影响，因此需要小心对待。即使将 OUT 设定到参考数据库(这个做法很普遍，但是依赖于数据库的完整性)，也需要牢记并不是所有的序列都对应正确的设定类群。找到和选择一种恰当的聚类算法是关键，但是先于聚类的原编码步骤也同样重要。

Owen: 固定阈值的聚类算法在原核生物元编码中已经得到广泛使用，在微生物学中实际上相当于宏基因组操作分类单元(MOTU)的一个操作性定义。但是我已经确信，基于固定阈值的聚类程序不能可靠地表征真核生物样本真正的生物多样性。真核生物谱系的序列在不同的分类群、不同谱系的分化时间和进化突变率等多个层面上表现出极大的变异性。形态学的变异与序列变异是不耦合的，因此对不同物种的鉴别阈值会有很大的差异。在使用广泛的分类数据集之后，我得出了一个结论，即像 SWARM v2 这样的分步聚合算法对于反映真核生物样本的预期多样性是最有用的。用 SWARM v2 生成的 MOTU 网络可以具有低可变性(例如 99%的同一性)或广泛的可变性(例如 90%的同一性)，反映了不同谱系的自然多样性。最好的消息是，这些计算可以以一种可重复的、确定性的方式执行，使用非常短的计算时间!在我看来，SWARM v2 是目前最好的聚类解决方案。如果你选

择正确的距离参数值，这个算法可以应用于不同的标记(最近的研究表明，使用多个标记可能是有效检测广泛分类学多样性的关键, Cannon et al. 2016)。

问：在实操规程的多个步骤中，原编码技术会给数据带来许多偏差。你们是如何对待这些问题的？例如，样品收集，重复实验，引物选择，可用数据库，生物信息学分析和统计学分析?

Vasco: 原编码技术多少会受到偏差的影响，有些能够通过技术手段减少，但是不能完全避免(例如引物选择所带来的偏差)。如何处理这些偏差很大程度上取决于研究问题的属性和可用的经费和时间资源。例如，不同样品中某个具体物种的生物量可能变化很大。当提取整体样本 DNA 时，一些小的和罕见的物种只贡献了很少的 DNA，可能无法被元编码技术检测到。尤其是当提取的生物标本 DNA 并没有被所使用的一组引物很好地扩增时，就很难检测到这些物种的存在。此时，对样品先按照不同粒径做个预分类处理有助于提高总体分类检出率。虽然我们在过去的研究中清楚地证明了先做标本预分类有效，但是在实践中应用中由于大大增加了实验室的工作量而并不实际。要注意到影响元编码数据集的偏差是非常重要的，然后决定哪些应该减少，例如通过使用优化的引物集，样本重复数，更高的测序深度等方式。然而，这在很大程度上取决于研究问题所需的数据精确性。

Owen: 元编码数据的确可能有偏差，所产生的数据集与形态学鉴定结果不同。然而需要说明的是，形态学方法也不能免于其自身的偏差。我们一直认为形态学中的重复计数是生物监测和生物多样性评估的金标准，为此花了很多时间试图校准元编码数据与形态学标准数据。然而由培训不充分操作人员根据不完整的动物学特征和实操指南进行的形态学计数也同样无法检测到某些物种，这与元编码标记中引物所导入的偏差并无实质性区别。从大的形态学数据库中分析鉴定其中一小部分生物种的求解方式所导致的漏检，等同于采用测序深度较低的原编码产生的漏检，而分类学中的隐种复合体(隐种是两种或者更多从形态学无法区别的生物群组或产生了生殖隔离)则相当于原编码中采用了分辨率较低的引物。更进一步，在大多数情况下使用了不完整的动物学键值是常态，特别是对于未被充分研究的类群，如线虫、小型动物群或土壤动物群物种，或未被充分研究的地理区域(可能存在大量无形态学数据的物种)，这也相当于用了不完整的参考数据库来做元编码。在所有这些情况下，元编码方法尽管具有各种形式的偏差，仍然可能实际上比任何形态学评估方式都要优越。元编码的结果总是更可重复，更客观，更加独立于分析师的专业知识水平。

所以我们为什么不改变传统的生物评价模式呢(此处应该指基于形态观察的分类学方法)? 对于大多数生态和生物监测应用，元标记编码将产生比形态学更完整、客观、可重复和有用的结果。即使由于轻微的引物偏置而不能检测到某些物种，但是我们毕竟能够获得比形态学更有用的定性信息。即使我们不能为每个

定名的物种指定一个对应的序列名称，我们仍然可以直接使用这些序列作为原始或受到人类活动影响的栖息地的生物评价指标(这点非常困难，因为我们的思想已经完全被传统分类学禁锢)。这样一来，我们将得到成百上千个这样的生态信息序列！我们只需要重新思考我们对生态系统分析的最初目标是什么。如果我们的目标是以一种快速和经济有效的方式获得客观的、与生态相关的信息，而不是建立一个在该地区存在的所有形态物种的长而详尽的目录，我无疑会支持元编码方法，即使原编码和形态学方法都存在潜在偏差。

问：元编码技术距离获得定量生物评价结果还有多远？

Vasco: 视情况而定!在大多数情况下，由于样品中不同生物标本的生物量不同，用原编码是无法估算分类群绝对丰度的。此外，由于引物的偏置，并不是所有的生物标本都具有相同的扩增效率，从而极大地偏移了不同类群的原始序列丰度。这种影响可以通过使用优化的生态系统，种群特异性引物，或对元编码数据采用校正因子的方式来降低。然而尽管能够降低偏差，仍然不能解决完全未被发现的类群和标本实际生物量偏离的问题。在我看来，根据元编码数据估算生物量是棘手的事情。当然，样本之间的偏差应具有相当的可重复性，从而能够比较和使用相同实验室和生物信息学方法处理的样本之间的相对序列丰度。在我看来，这样做至少可以实现半定量的生物量估算。在某种程度上，元编码数据可以告诉你一个特定的分类单元在样本 A 中比在样本 B 中出现得更多，但不能确切地告诉你这个特定的分类单元的丰度与整个群落中的其他物种丰度相比有多大差别。然而，如何以及能够从元编码数据中派生出什么样的丰度数据仍然是一个热门话题，并引起了积极的争论，在本领域还没有形成明确的共识。

Owen: 这在很大程度上取决于你计划使用的标记类型，以及你是在评估离体环境 DNA 还是整体群落 DNA。如果你的目标是检测离体环境的 DNA(例如从海水样品检测海洋鱼类的 DNA)，那么你最好忘记所有关于定量的要求。相反，如果你针对整体群落 DNA 采用具有较低偏差的通用引物，你至少会得到一些半定量数值。当然，读长(READ 是高通量测序中一次反应获得的测序序列)永远不会与(形态学的)个体的数量成正比。但读长可以与每个物种的线粒体 DNA 生物量成正比(如果你使用的是线粒体无偏差标记)，或者与每个细胞的叶绿体基因组拷贝数成正比(如果你使用叶绿体标记)。在某些情况下，可以为每个物种计算修正系数，然后元编码结果就可以相当程度定量。例如，目前正在利用叶绿体标记对河流硅藻进行定量分析，其结果似乎非常有希望。

问：以你们的看法，当前原编码技术作为研究和生物多样性监测的能力和限制条件是什么？

Vasco: 元编码是一个神奇的工具，对在物种水平开展一般环境样品的调查研究具有快速和相对容易实施的优点。虽然元编码与形态识别方法相比也有明显的局

限性。例如，并非样品中的所有类群都能通过原编码被检测到，但这不应该阻止我们利用这个工具来回答紧迫的生态问题和环境问题调查。

Owen: 与形态学方法相比，元编码目前的优势在于其客观性(与分析人员的专业知识程度无关)、可重复性、分析所需的时间更短以及能够检测到样品中的稀有物种。如果应用于生物监测，能够表征生态信息的潜在序列数量(涵盖大范围的分类群)通常比任何仅基于一个分类单元的形态学数据高数个数量级。还有一点很重要，就是可以从元编码数据中检索得到一些群体遗传学信息(如果使用如 COI 这样的高可变标记)。局限性是大多数元编码方法无法提供样本中不同物种的某些重要的生态和生物学信息，如性成熟状态、生理条件、繁殖状态、每个物种的个体大小或总生物量。当然这些信息可以从经典的形态学分析获得精确估算。

附图(赠送)：基于形态学的生物调查与基于 DNA 的生物调查技术路线比较