Genome Sequencer FLX 系统进行环境基因组学分析

Tom Jarvie[1] and Tim Harkins[2]
[1]454 Life Sciences, Branford, USA; [2]Roche Applied Science, Indianapolis, USA
联系作者：tjarvie@454.com; tim.harkins@roche.com

罗氏应用科学部的 Genome Sequencer FLX 测序系统(GS FLX)做为一个通用的测序技术平台，适用于多领域范围的应用：包括基因组DNA从头测序和拼接、转录组测序、小分子 RNA 的分析和扩增产物测序等。建立在454公司测序技术平台之上GS FLX 系统具有较长的读长和极高的单一读长准确性；而其在环境基因组学研究上的应用更是加速了人类对这一领域的认识。

简介

环境基因组学（Metagenomics）是对混合微生物进行的基因组水平上的研究（见table 1）。此方法的两个基本的目标是在特定的环境中会存在哪些共同的微生物种群（水平筛选），然后鉴定每一种微生物在这特定的环境中发挥怎样的作用（垂直筛选）。环境基因组学的样品几乎在任何地方都可以找到，包括人体的不同微环境，土壤样品，极端环境比如深海和海洋的不同分层中。因此，微生物的多样性被认为是比它的种类数量（成亿的）还要多上成千到几百万倍。据估计，单是海洋本身就有 3.6×10^{29} 种微生物细胞。如果再考虑到微生物已经经过的35亿年的进化及普遍存在的不同种类之间的基因转移，那么就很容易理解为什么会有数量如此巨大的种类之分了。

近期的文献显示了微生物世界的多样性

在一篇文献中，作者发现来源于两个深海的样品组成了两个代谢完全不同的群落，即使这两个样品互相具有极其相近的亲缘关系【1】。有意思的是，所研究的微生物和来源与其他已经测序的微生物群落的样品是完全不同的。另一篇研究海洋微生物多样性的文献中，描述了一种利用r RNA 序列标签进行筛选的方法【2】。在此研究中，据估计微生物的多样性比先前预计的 10^6 种还要多 3 个数量级。

虽然大多数微生物的作用仍然没有被发现，但是已经有很多近来发表的文献表明，微生物互相之间是同时具有竞争和协同关系的，并且这种互相之间的作用会随着它们所在的环境条件的变化而变化。比如，一篇关于人内脏的研究发现有两种主要的细菌种群，拟杆菌和厚壁菌，它们的相对丰度的大小会随着个体身体内脂肪含量的变化而变化【3】。当体内的脂肪增加时，厚壁菌的丰度随之增加，这反过来增加了对能量的摄取能力，也就导致了更高的肥胖率。

表 1 环境基因组学词组表

- **Metagenomics：**不需要培养，克隆样品，对来自于环境样品中的 DNA 进行测序和分析。
- **Environmental Genomics：**metagenomics 的同义词 Random Community Genomics：经常作为metagenomics的同义词。此词被许多研究者定义为：经过克隆或没有经过克隆的来自于环境中的整个群落的测序研究，但是没有对功能性成分进行筛选研究。
- **Microbial Diversity：**利用一种测序方法对基因组的高度变化的区域（16s rRNA 的 V6 区域）进行的一种环境基因组学研究，以此来对环境中不同种类的数量进行评定。
- **Virome：**环境中病毒和它们的遗传分子的总称。
- **Microbiome：**环境中微生物和他们的遗传分子的总称；
- **Metatranscriptomics：**环境样品中cDNA的测序和分析，而不需要经过个体培养和克隆。

Genome Sequencer 的技术优势

GS FLX 系统（Figure 1）的几个优势使其非常适合于环境基因组学研究：首先是它不需要将 DNA 片段克隆到细菌中，因此，GS FLX 系统就避免了 Sanger 技术样品制备过程中所具有的克隆偏差的问题。

GS FLX 系统每个反应可以得到超过 40 万个读长，促进了对鉴定大量不同基因、代谢途径和可能存在的微生物种类的深入分析，同时大大降低了每个项目的成本。这让研究者也可以进行原先只有大规模的基因组研究中心才能进行的研究。总之，GS FLX 系统为支持科学研究来解答环境和生态问题提供了强大的技术保证。

测序读长的长度也是环境基因组研究中需要考虑的一个非常重要的因素，因为大多数环境基因组学研究中的样品基因组都是未知的，并且没有很多可以参考的相关参考序列。具有较长的读长可以让研究者既能以从头测序的方式进行基因组的拼接，又能将特异的测序数据排列到基因或基因组上。其他的下一代的测序技术选用叫做“microreads”的技术，读长在15-40 bp之间，这

Figure 1: The Genome Sequencer FLX Instrument.

么短的读长在环境基因组学研究中是非常受限的，因为在同一基因组中和不同样品的基因组之间都存在着相同或者重复区域。当一个样品中有大量的基因组存在时，将一个非常短的读长特异性的序列排列到一个基因组上是一项非常具有挑战性的工作。除此之外，短的读长阻止了许多读长互相之间的拼接。从以前已经发表过的文献中数据的有效性来看，100 bp已经非常接近最短的有效测序长度了【1,4】，GS FLX的读长长度在200-300 bp之间，而读长长度是依赖于特殊的序列特性的。

环境基因组学生物信息学分析

GS FLX 系统每个反应产生超过 100 M的数据。对于环境基因组学研究而言，这是一个巨大的数据库，对于数据的分析有许多挑战不得不指出。为帮助理解可用的生物信息学资源，我们列出了几个公共的可用的网址供大家参考(Table 2)。需要特别指出的是，分析的第一个目的是识别出哪些序列读长和已知的基因组相关，哪些是未知的。利用MEGAN方法，研究者可以将他们的数据进行归类，并且依据数据的分类水平可以对他们的结果进行总结和整理。使用这种方法，研究者可以对他们样品的复杂性和存在的微生物的多样性进行评估。除此之外，仅需少量的测序读长就可以帮助说明研究样品中存在哪个种或者亚种。

GS FLX 支持不同的应用格式，使得用户可以在每个反应选择所测样品的数目和每个样品的读长数目。一个单一的反应可以分成2、4、16个样品，相对应的，每个样品的读长数目依次是210,000、70,000和12,000。

大多数环境基因组学的研究的下一个目标是识别样品中微生物的代谢功能。GS FLX平均读长长度为250 bp，因此用它来搜索同源序列数据库是可行的，虽然搜索成功率可能是非常低，在5-10%范围内（主要由于微生物序列数据库的不完整性）。但是，仍可提供20,000-40,000个序列对已知的功能性进行鉴定。通过使用CPU加强型应用的BLASTX，可以进行功能性分析，来帮助识别环境基因组样品中存在的微生物的代谢功能。

环境基因组学的最终目标是序列拼接，至少能够得到全长的基因，最好是完整的基因组。GS FLX产生的250 bp的读长使环境基因组的从头拼接成为可能。测序读长能够拼接对于全长基因的识别、发现新的基因和帮助最终从序列一致性上建立新的微生物群落是非常重要的。

表2 环境基因组学分析可用的 URL

- **http://www-ab.informatik.uni-tubingen.de/software/megan**
  MEGAN: 环境基因组分析工具，可以允许一个科学家对一个大数据库进行分析，对数据进行分类分析
- **http://seed.sdsu.edu/FIG/index.cgi**
  SEED 数据库是一个全基因组和基因组草图的公共资源库，用来帮助对从序列到代谢功能进行相关性分析
- **http://img-jgi.doe.gov/cgi-bin/m/main.cgi**
  IMG/M 提供了基于环境基因组序列进行微生物群落功能性分析的工具，同时也可以使用大量的公共性的功能和路径资源进行基因组的分离
- **http://camera.calit2.net**
  CAMERA 是一个提供环境基因组学数据和分析工具的站点

总结

迄今为止，环境基因组学研究进展受到了Sanger技术成本、低通量和克隆偏差等种种限制。GS FLX 系统由于具有高通量、无克隆偏差、读长足够长的优势，为环境基因组学样品开辟了一份广阔的研究天地，也使微生物群落的多样性和功能性分析成为可能。

参考文献

1. Edwards RA et al., 2006, BMC Genomics 7:57
2. Sogin M et al.,2006, Proc Natl Acad Sci USA103:12115-12120
3. Turnbaugh PJ et al., 2006, Nature,444:1027-1031
4. Goldberg SM et al., 2006, Proc Natl Acad Sci USA 103:11240-11245
5. Sambrook J, Fritsch EF, and Maniatis T: Molecular cloning: A Laboratory Manual, 2.74 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1989)
6. Ausubel FM et al., Short Protocols in Molecular Biology 【5th Ed.】,Vol.1:2-11(John Wiley and Sons, Inc.,2002).

产品名称	包装规格	序列号
Genome Sequencer FLX Instrument	1 instrument plus accessories	04 896 548 001
Genome Sequencer PicoTiterPlate Kit(70*75)	1 kit (1plate with accessories)	04 852 427 001
Genome Sequencer LR 70 Sequencing Kit	1 kit (for 1 sequencing run)	04 932 315 001
Genome Sequencer emPCR kit l (Shootgun)	1 kit (for 16 amplification reactions)	04 852 290 001
Genome Sequencer emPCR ll (Amplicon A, Paired End)	1 kit (for 16 amplification reactions)	04 891 384 001
Genome Sequencer emPCR lll (Amplicon B)	1 kit (for 16 amplification reactions)	04 891 392 001
Genome Sequencer DNA Library Preparation Kit	1 kit (for 10 library preparations)	04 852 265 001
Genome Sequencer Paired End Adaptor Kit	1 kit (for 10 library preparations)	04 891 457 001