

EBIOTECH

生物通技术周刊

Focus: 第三代测序系统PacBio RS
[技术剖析]

访谈: 深入3000bp测序平均读长的单分子实时测序
测序读长2700bp! 第三代测序PacBioRS升级
纵观第三代测序之Pacific Biosciences

第89期

2012年1月12日

全文下载

[应用指南]

单分子测序灵敏检测5-hmC
单分子测序助力脆性X综合征的研究
利用基因组学来对付疾病爆发
致命大肠杆菌研究成果登上《新英格兰医学杂志》

[公司动态]

JGI购入PacBio取代454 明年测序47Tb
德国总理默克尔启动第三代测序仪
PacBio正式发售第三代测序仪

个人型测序仪
[产品大比拼]

半导体测序仪 每个实验室都触手可及
我的样本, 我的研究。MiSeq个人测序系统
三款热门个人型测序仪比拼

[它的应用]

面对越来越多的突发疫情, 我们该如何应对?
个人化操作基因组测序仪快速发现德国毒黄瓜致病原因
罗氏GS Junior助力致命大肠杆菌的深入分析
Life Tech推出取代芯片的RNA-Seq方案
罗氏 GS Junior测序系统与MagNA Pure Compact核酸纯化系统联手发现新生物

[数据分析]

测序数据分析之专家指南(上)
测序数据分析之专家指南(下)
测序数据分析之工具大全

访谈：深入 3000bp 测序平均读长的单分子实时测序

前言

测序技术已成为当今生命科学研究中发展最快的领域之一，其技术更新速度之快，用“日新月异”一词来形容也不为过。全球首个第三代测序平台——PacBio RS 单分子实时测序系统自今年 4 月底推出，受到了很多研究者的关注。Pacific Biosciences 预计将在明年一季度正式发布 PacBio RS 的新版本 C2 试剂，届时该系统的平均读长将从现在的 1300 bp 骤然提升至 2500-3000 bp！最长读长可超过 10000bp。令人惊叹的数字！经典的 Sanger 测序长度也就是 1200bp 左右，已难以匹敌，新一代测序（NGS，即二代测序）就更是望尘莫及啦。



短读长之困

风起云涌的新一代测序（NGS）市场三国鼎立，三大平台你追我赶，在测序速度和单次运行能获取的数据量上不断刷新记录，精彩纷呈，但其共同的短板也非常明显——序列读长太短——仅 100-200bp 的读长相比庞大的基因组，使得完成拼接工作变得无比艰巨，不少用户虽然获得了大量的测序数据，测序覆盖深度达到了几十倍甚至上百倍，但仍然没法完成基因组的拼接。对于新物种来说，往往还需要传统的 Sanger 测序先做 scaffold。

今年罗氏 454 平台通过升级达到平均读长 500bp 以上，最长可达到 1000bp，因此一些顶级研究院在做 de novo 测序时更倾向于用 454 做 scaffold，再配合其他速度更快的 NGS 平台完成后续的测序工作，以求提高数据处理速度。

关于单分子实时测序的各种疑问

第三代单分子实时测序技术自问世以来就备受关注，有关其工作原理，生物通早在去年的 ebiotech 期刊新一代测序专辑中已有文章介绍过（纵观第三代测序之 Pacific Biosciences <http://www.ebiotech.com/newsf/2010-10/20101014173036143.htm>）。C2 试剂对平均读长大幅提升则在“第三代测序 PacBioRS 升级（<http://www.ebiotech.com/newsf/2011-10/20111018165836418.htm>）”一文着重介绍。

但是许多疑问依然存在，比如

1) 最受大家关注的，是 PacBio RS 的准确性究竟如何？有用户反馈，PacBio 单分子测序准确度大概只有 85%，这么低的准确度怎么能保证结果的准确性？

2) PacBio RS 的长读长，对数据分析而言到底有什么优势？

3) PacBio RS 单分子测序无需 PCR 扩增，有何优势和劣势？

4) 单分子测序获得结果的速度会更快吗？运行通量有多大？一次运行需要多长时间？能获得多少可定位数据？

5) 除了速度, 读长, 通量, 精确性, PacBio 未来的可扩展性如何?

6) PacBio 的运行费用如何?

7) 已成三国鼎立之势的测序市场, PacBio RS 加入战局究竟会扮演什么样的角色? 会不会秒杀 NGS 出局呢?

为此, 生物通特邀了 PacBio RS 中国代理——基因有限公司专门负责 PacBio 的技术专家, 就大家普遍关心的这些问题进行了比较深入的探讨, 希望能使广大读者对这一最新的测序技术了解更多。

关于读长

Q: 长读长是 PacBio RS 最引人注目的特点之一, 长读长对测序来说到底有什么优势?

A: 长读长在序列拼接、定位以及需要跨越重复区域的应用中有着极大优势。

例如在 De Novo Assembly 时, 目前遇到的主要困难在于如何跨越那些重复区域以及高/低 GC 含量的区域, 从而完成整个基因组的拼接。

如果把拼接工作看作是在做拼图游戏。NGS 获得的读长都很短, 就好像把一幅图打成非常小的碎片, 然后做拼图。由于碎片太小, 因此许多碎片看起来都差不多, 这样要拼出一副完整的图难度很大。PacBio RS 目前可以获得 1300bp 的平均读长, 明年初随着试剂升级, 平均读长可提升至 2500-3000bp, 这就相当于同样的一幅图, 用大的碎片来做拼图, 由于大碎片比小碎片的识别度要高, 因此完成拼图的难度就可以大幅降低。

同 NGS 通常 100-150bp 的读长相比, PacBio RS 的平均读长提高了近 20 倍。试想一下, 同样大小的一幅拼图, 10,000 片的还是 500 片的更好拼?

另外, 随着读长的增加, 拼接过程中所需要测序覆盖深度也会随之下降。

对变异检测来说, 我们首先需要的是准确定位, 如果无法准确定位, 那无论原始或者一致性准确度有多高都是没有意义的。而长读长可以帮助研究者进行更准确的定位。

Q: C2 试剂据说能将读长提高到 2500-3000bp, 这是 PacBio 实验室得出的数据, 还是用户数据?

A: 目前 PacBio 公布出来的 C2 试剂的参数是, 在 Long 模式下, 平均读长可达到 2500bp, 95%ile 读长可达到 6500bp, 在 X-Long 模式下, 平均读长可达到 3000bp, 95%ile 读长可达到 8500bp。C2 试剂明年 Q1 将实现大规模的商品化供应, 目前部分实验室已经率先在使用。例如:

C2 试剂第一次被使用是在德国大肠杆菌疫情研究中, 研究人员通过将不同测序模式混合使用, 最终获得了 2900bp 的平均读长以及 99.998% 的一致性准确度。这项研究的结果今年七月底已经发表在新英格兰医学杂志上, 题为“Origins of the E. coli Strain Causing an Outbreak of Hemolytic-Uremic Syndrome in Germany”。在这项研究中, 研究人员在 PacBio RS 平台上通过全球合作几天内就完成了对从疫情中获得的大肠杆菌样品以及近似菌株的测序和数据分析。PacBio RS 的长读长优势使得只使用 PacBio RS 数据完成致病大肠杆菌的 De Novo Assembly 成为可能, 而 PacBio 相对 NGS 平台可更快获得结果这点这对鉴定新的病原体来说也是一个极大优势。

另外美国著名基因组技术服务商 Expression Analysis (EA) 负责研发的 Pat Hurban 在今年 9 月底一次网络会议中展示了其最新获得的 PacBio 数据: 他们将大肠杆菌基因组 DNA 分别处理成 2kb 和 6kb 的片段, 其中 2kb 的模板只用 C1 试剂进行了测序, 6kb 的模板则分别用 C1 和 C2 试剂进行了测序。Hurban 发现使用 C2 试剂后, 获得了 2715bp 的平均读长, 最长的读长甚至达到了 13091bp (可能有用户会问, 插入片段大小只有

6kb, 怎么会产生将近 13kb 的读长? 这是由于在 PacBio 平台上, 样本制备完成后会形成环形的而非线性的模板结构, 因此如果影片持续拍摄, 当完成插入片段从一端到另一端的读取后, 会跨过接头, 继续读取其反义链的序列, 因此这里最长读长超过了插入片段的长度), EA 对使用 C2 试剂获得的结果非常满意。

关于准确性

Q: PacBio RS 的准确性究竟如何? PacBio 工作原理和 C2 试剂的技术文章在生物通发布后, 多位读者在评语、留言和邮件中都不约而同的提到了一篇留美中国学者的博文, 提及在今年 4 月他个人在美国实验室试用 PacBio 样机的感受, 原文这么说“首先, Library 和 SMRTBell 的准备快速而简单 (生物通注: SMRTBell is the name of the prepared template DNA. SMRTBell 的制备实际就是 PacBio 文库制备过程的一部分)。测序时间确实很短。其次, 合成的 DNA 链长度可以达到 3kb, 确实比目前所有的高通测序仪都高。最致命的是误差问题。结论是单次测序错误率 15%, 循环测序误差 8% 左右, 仪器目前的性能很令人失望。我自己写了个程序来做序列对比, 能把误差降到 3% 左右, 相比起其 454 来, 还是有很大的差距。PacBio 目前还不足以投放市场。他们的软件部分实在是需要改进。”您会怎么回应?

A: 首先, 在 PacBio RS 之前, 没有任何一台测序仪能够提供单分子的准确度数据。在 NGS 平台上, 由于硬件的限制, 其所检测到的信号是基于成百上千甚至更多分子, 无法检测到单分子的信号。NGS 平台给出的准确度是将这成百上千甚至更多分子获得的信号的平均值同 reference sequence 比对后获得的结果, 也即一致性准确度。PacBio RS 在测序历史上第一次给出了“单分子的测序准确度”数据, 这是把单个模板分子的原始测序结果 (标准测序模式) 同 reference sequence 比对获得的数据, 目前单分子的原始测序准确度在 85~92%, 平均值为 87%。由于 PacBio 的单分子

测序反应是在 ZMW 中进行, 而每个 SMRT cell 含有 15W 个 ZMW, 因此当我们把 N 个 ZMW 中的单分子测序数据也做个平均, 之后再同 reference sequence 比较, 则 PacBio 的准确率也会大幅上升。另外, 在 NGS 平台上, 文库制备时必须要先进行 PCR 扩增, PCR 过程中的 bias 或者 mismatch 等将无法在测序时修正, 也就意味着这些错误会变成系统误差, 且无法通过增加测序覆盖深度来消除。PacBio 平台上, 文库制备时无需 PCR 扩增, 因此避免了 PCR 产生的 bias 等。由于 PacBio 上产生的错误是随机错误, 且错误率并不随着读长增加而升高。因此其一致性准确度可随着测序覆盖深度的增加而提高。当测序覆盖深度达到 30× 时, PacBio 的一致性准确度可以达到 99.999%。

其次, PacBio 的环形比对测序模式 (CCS) 可以帮助用户获得高准确度的单分子测序数据, 由于我们可以对每一个单分子模板都进行评估, 且无需通过 PCR 扩增, 这对于突变检测 (例如稀有 SNP 的检测) 来说非常重要。已有实验数据表明 PacBio 可检测到低至 1/100 的突变。在突变检测中, 现阶段我们建议的插入片段大小在 250-500bp, 以 500bp 的插入片段为例, 当使用环形比对测序模式, 单分子的测序准确度可以达到 99% @ 5×CCS。C2 试剂已经可以使平均读长提高到 2500bp, 最长的读长甚至可超过 10,000bp。考虑到后续 PacBio 还将通过对试剂的持续优化, 不断提高其读长。因此对突变检测来说, 未来可研究的插入片段长度将越来越长。

Q: 单分子单次测序产生错误的原理是什么?

是长读长的错误累计结果还是机器原因? 87% 这个单分子原始测序平均准确度数字, 在将来序列读长再次翻番时是否也会随之降低? 那怎么办? 再次提高测序覆盖深度吗?

A: PacBio 平台上目前的错误主要是插入和缺失, 只有大概 1% 是 substitution。缺失错误源自于有时候碱基掺入速度过快, 超过了 PacBio 相机的

拍摄帧数。插入错误源自于有的时候酶随机的选择一些碱基，但并未将这些碱基真的掺入合成链中。由于这些错误是随机的，因而可以随着测序覆盖深度的增加而消除。因此，尽管 PacBio 的单分子单次读取的原始准确度并不非常高，但随着测序覆盖深度的增加，它可以获得比 NGS 平台更高的一致性准确度。

PacBio 的错误是随机错误，并不会随着读长的增加而提高，因此，当读长翻番时，错误率并不会随之提高。未来随着试剂的不断优化，单分子测序的原始准确度也会逐步提高，且每个 SMRT cell 可获得的数据量也会进一步提高。

关于速度和运行通量

Q: PacBio RS 的测序速度有多快？样品制备需要多久？一次测序运行需要多长时间？一次运行最多能得到多少 Gb 可定位数据？运行通量能有多大？

A: 目前 PacBio 上所使用的 DNA 聚合酶的合成速度大概是 1-3 个碱基/秒，由于在该平台上，聚合酶合成的过程就是序列解读的过程，这意味着测序速度每分钟可超过 100 个碱基。

从样品制备到获得碱基序列的全部流程可在 1 天内完成。

如果使用 C2 试剂，每个 SMRT cell 可以获得 90M 的可定位数据（mappable data）。现阶段每天最多可运行 12 个 SMRT cell，因此每天可获得的数据是 $12 \times 90 = 1080\text{Mb}$ mappable data。

Q: PacBio 平台每天可获得的数据量目前来看与大型 NGS 平台（例如 HiSeq2000 每天可获得 55Gb 数据）相比还小得多，PacBio“每天最多运行 12 个 SMRT cell”这个界限几时能翻番？每个 SMRT cell 最大读取数还能继续提高吗？平台未来的扩展性如何？有的 NGS 平台强调只需升级试剂部分即可实现读长翻番或者测序通量翻番，PacBio 未来将从哪些方面扩展其性能呢？

A: 现阶段 PacBio 平台和 NGS 平台更多的是一种互为补充的关系，NGS 可以获得更多的数据量，而 PacBio 可以获得更多的信息量。接下来的发展计划中，PacBio 将通过对试剂以及软件的持续开发和优化，进一步提升读长，增加每个 SMRT cell 的数据产出量，并且会对 DNA 碱基修饰（例如 DNA 甲基化等）分析以及 RNA 直接测序等提供更多的支持（例如提供配套试剂盒和相应分析软件等）。

关于价格和市场

Q: 这是个敏感问题！精明的用户肯定会关心性价比。特别是后续运行费用，往往是决定采购的一个关键因素。PacBio RS 每次运行费用大概多少？

A: PacBio 的后续消耗品主要包含试剂和耗材两部分。试剂有 3 种：模板制备试剂盒，结合试剂盒以及测序试剂盒，耗材就是 SMRT cell。具体的费用取决于样本数、所要研究的基因片段的大小、测序方案的选择、测序模式等，很难一概而论。详细情况可以联系基因有限公司咨询。

你说的性价比，研究者更加关注的是从样品开始到可发表的最终结果——你不能只看单价，或者单次运行的成本——因为不是运行结束，软件自动出来的结果就可以打印发表。如果考虑到长读取的优势和在数据拼接上能节省的时间和费用，就这一点而言，PacBio RS 有其他方法不可比拟的优势。关键是它是否能满足你的需要，帮助你快人一步达到目标。

Q: 已成三国鼎立之势的测序市场，PacBio RS 加入战局究竟会扮演什么样的角色？能否秒杀 NGS 出局呢？

A: 这取决于用户的具体应用。以基因组拼接为例，对于基因组较大的物种，例如植物，现阶段 PacBio 平台和 NGS 平台是一种互为补充的关系，NGS 可以获得更多的数据量，而 PacBio 可以获得

更多的信息量（例如 NGS 平台很难获得的高 GC 含量区域的信息等）。通过 PacBio RS 的配套分析软件，我们可以实现同 NGS 数据（兼容三大 NGS 平台）的混合拼接，从而大幅提高 Genome Finishing 的速度。对于基因组较小的物种，例如微生物和病毒等，则可以仅通过 PacBio RS，独立完成 De Novo Assembly。

Q: PacBio RS 除了在 De Novo Assembly、突变检测等领域有优势，还在其他哪些方面有更多应用？

A: PacBio RS 可以对高 GC 含量区域测序，例如美国 UC Davis 医学院利用单分子实时测序技术，对脆性 X 染色体综合征的关键基因 FMR1 中的 CGG 三核苷酸重复区域进行了测序，并在第 15 届脆性 X 和早发性认知缺损国际研讨会上公布研究成果。在所有人的 X 染色体上都有一段 CGG 三核苷酸重复序列，正常人的 CGG 重复次数为 5-44 次。过长的重复次数会对 FRM1 基因转录或翻译出 FRM1 蛋白不利，当重复次数超过 200 次时，就会导致脆性 X 综合征。所以检测 CGG 的重复次数非常有意义。一般 CGG 重复在 200 次以上，被认为是具有临床意义的，但这个长度不管对于 Sanger 法测序或者新一代测序来说，都是很困难的，而 PacBio 则很好的解决了这个问题，利用环形比对测序模式，UC Davis 医学院获得了超过 10kb 的原始读长，覆盖了 CGG 重复超过 750 次的三核苷酸重复区域。

另外一个例子。甲基化研究如今开展得如火如荼。除了人们熟知的 5-mC，另一种修饰方式——5-mC 的羟基化形式 5-hmC 也引起人们注意。但现有的测序方法如亚硫酸氢盐测序，无法区分 5-mC 和 5-hmC。若想深入了解 5-hmC 的生物学功能，必须开发出一种灵敏的测序方法，以揭示它在基因组中的位置。美国芝加哥大学利用第三代单分子 SMRT 测序技术和 5-hmC 的选择性化学标记方

法来高通量检测 5-hmC。通过聚合酶动力学带来的宝贵信息，研究人员可直接检测 DNA 甲基化，包括 N6-甲基腺嘌呤、5-mC 和 5-hmC。（详细阅读：单分子测序灵敏检测 5-hmC，<http://www.ebiotrade.com/newsf/2011-11/20111123171704271.htm>）

PacBio RS 还可对连续的 A 或者 T 区域测序，有研究者曾成功的对含有 poly A 的序列（含 111 个连续的 A）测序。PacBio 可以获得动力学信息，因此可以用于研究 DNA 甲基化等 DNA 碱基修饰情况。

另外，由于 PacBio 从样本制备到获得序列信息所需时间非常短（<1 天），在具有时效性的病原微生物鉴定中（例如生物反恐、流行病爆发监控等）也非常有优势。（详细阅读：利用基因组学来对付疾病爆发

<http://www.ebiotrade.com/newsf/2011-11/20111123170930924.htm>）

其他的应用文献，欢迎联系 PacBio 独家代理商基因有限公司索取。[点击索取资料](#)

至此，我们对 PacBio 第三代单分子测序技术有了更深入的了解。令人期待的技术，更令人期待的是 PacBio 能更快的升级，尽快将测序成本降低——如他们自己所预测的：到 2013 年，个人基因组的测序能在 15 分钟内完成，费用低于 1000 美元，人人都可以消费得起。。。。。（访问结束，感谢基因有限公司提供的协助）

后继：另外一个感受，今后的科学研究，除了依靠思维创新，某种程度上拼的就是装备了。。。装备好，用得好（这个还得靠研究人的想法啦），出成果快，文章多，影响力大——面包（经费）就会有的，一切都会有的。。。 （循环 N 次放大，一如 PCR）。

测序读长 2700bp! 第三代测序 PacBioRS 升级

Pacific Biosciences 公司自今年 4 月底正式推出全球首个第三代测序平台 PacBioRS 单分子实时测序系统后，一直致力于进一步优化其试剂的性能并努力开拓更多的应用，据最新消息，其新版本的 C2 试剂将在今年年底或明年年初正式发布，C2 试剂可使该系统的平均读长从现在的 1300bp 提升至 2700bp! 基因有限公司作为 Pacific Biosciences 产品在中国大陆和香港地区的唯一授权经销商，将一如既往的秉承“Let Professionals Serve Professionals”的宗旨，为用户提供更好、更便利、更专业的服务。



由Pacific Biosciences公司开发研制的革命性的单分子实时测序（SMRT™, Single Molecule, Real Time）技术在测序历史上首次实现了人类观测单个DNA聚合酶聚合过程的梦想。该技术通过光学方法直接记录单个聚合酶在不受干扰的情况下的连续合成，其读长之长，测序速度之快，灵活性之强大，已经使许多极富挑战性的基因组学研究成为可能。该系统可应用于基础研究、临床诊断、传染病研究、卫生防疫等多个领域。

目前，Pacific Biosciences 公司已携手基因有限公司，致力于开拓中国市场。基因有限公司凭借其遍及各大中城市的服务网络以及包括销售、市场和技术支持、售后服务等多个部门的完整的服务体系，将一如既往的为广大用户提供包括技术咨询、产品选配、售后培训及维护的专业服务。

[索取PacBioRS单分子实时测序系统的详细资料](#)

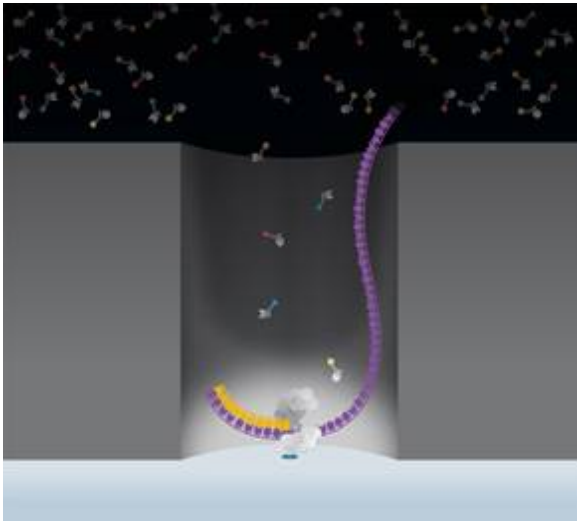
PacBioRS 系统介绍:

PACBIO RS 系统采用了 SMRT 测序技术，系统包含主机和用于数据存储和运算的 Blade center。主机内置的机械臂能够自动将 SMRT Cell 在存储区、准备区以及测序区之间进行转移，实现自动化的单个或批量测序实验。系统的 RS Touch 触摸屏实时显示系统反馈信息，如系统的运行状态，每个 SMRT Cell 的制备过程，测序过程，碱基判定结果以及对应 Q 值等，并实时播放测序过程中记录的影片信息。系统的 RS Remote 远程监控软件，允许客户远程设计和监控测序过程并进行初级数据分析。PACBIO RS 系统的测序分析软件包括 SMRT Pipe、SMRT Portal 和 SMRT View 三部分计算模块，可与用户的生物信息学平台实现无缝整合，轻松实现测序数据浏览、数据过滤和比对、诸如单核苷酸多态性（SNP）等稀有突变的可视化筛查等数据分析操作。系统还自带条形码阅读器，能够提供样品和试剂的信息，便于用户进行实验设计和数据管理。

SMRT™技术原理

3 个核心创新技术的完美结合。

1. The SMRT Cell



测序在专利的 SMRT cell 中进行,每个 SMRT cell 含有 150,000 个 ZMW (Zero-mode waveguide, 零模波导孔)。

ZMW 是直径为几十纳米的小孔,被固定在玻璃基片表面的金属薄膜上。由于 ZMW 的尺寸原因,可见光进入 ZMW 后会迅速衰减,这样当用激光透过玻璃照射 ZMW 时,只有底部 30nm 左右的区域能被照亮。

DNA 聚合酶同模板结合后,可通过专利技术锚定在 ZMW 的底部玻璃表面,不同种类的核苷酸用不同颜色的荧光团标记,随后核苷酸涌入 ZMW 中,并在阵列表面扩散。当聚合酶检测到正确的核苷酸时,便将其掺入新生链中,这个过程需要几毫秒,而单纯的扩散只需要几微秒。这种时间差使掺入的核苷酸产生了很高的信号强度,因此,通过 ZMW,可以在荧光标记核苷酸的背景下检测单个掺入事件。

2. Phospholinked nucleotides

将荧光染料标记在核苷酸的磷酸链而不是碱基上,这样,一旦核苷酸掺入到新生的 DNA 链中,标记基团就会自动脱落。

3. PacBioRS

能够实时监测并分析单分子测序反应。配备大孔径物镜和单光子照相机来收集荧光所发射的光脉冲,观察整个过程;通过经过优化的算法,将光学系统所捕获的荧光信号翻译成碱基序列信息。一旦测序开始,实时的数据就传送到系统的初级分析中心,在识别碱基“身份”的同时给出对应的 Q 值。

PacBioRS 的应用进展:

1) PacBioRS 用于快速鉴定引起海地霍乱爆发的菌株来源

2010 年 10 月,霍乱席卷了海地的 10 个省,还传播到邻国多米尼加。据海地卫生部报告,93000 人未能幸免,超过 2100 人在这场瘟疫中死亡。美国波士顿的研究人员采用 PacBioRS 对从海地爆发的霍乱中分离出的 2 株霍乱弧菌、1 株来自 1991 年拉丁美洲发生霍乱的菌株和 2002-2008 年从南亚分离的 2 株菌种的基因组进行了测序研究。利用早期的序列数据库,研究人员对比了这 5 株菌和其他 23 株霍乱弧菌的基因组,从而对发生于海地的霍乱疫情可能的渊源做出了评定。

结果发现:海地霍乱菌株与 2002 年和 2008 年在孟加拉国分离得到的变异霍乱弧菌 El Tor O1 菌株之间关系密切。而与南美分离株的关系较远。

在该应用中,研究人员在进行霍乱弧菌的基因组拼接时,采用了杂交拼接 (Hybrid Assembly) 策略,分别将从疾控中心获得的 contigs 和基于 Illumina 测序仪的短序列同 PacBioRS 上获得的长序列进行拼接,大幅减少了 contigs 数目,增加了 N50 长度。从而显著提高了基因组拼接的效率。

2) PacBio RS 在德国大肠杆菌疫情中一展身手

七月底在新英格兰医学杂志上发表了一篇名为“Origins of the E. coli Strain Causing an Outbreak of Hemolytic-Uremic Syndrome in

Germany”的文章，该文对引起德国大肠杆菌疫情的高度致命性的细菌的病原性以及进化起源提供了有价值的观点。这篇文章描述了研究者们是如何在 PacBioRS 平台上通过全球合作几天内就完成了对从疫情中获得的大肠杆菌样品以及近似菌株的测序和数据分析。PacBioRS 的长读长优势使得只使用 PacBio RS 数据完成 de novo 拼接成为可能，这对鉴定新的病原体来说也是一个极大优势。该实验的结果对于致病菌株提供了包括医学信息在内的迄今为止最详细的遗传“肖像”（Genetic profile）。

3) Expression Analysis 发布其使用 PacBio 新的 C2 试剂的初步结果

Expression Analysis 是美国著名基因组技术服务商，该公司可为用户在 Pacific Biosciences 和 Illumina 平台上提供测序服务。9 月底，Pat Hurban（Expression Analysis 负责研发的 Vice President）在一次网络会议中展示了其最新的 PacBio 数据。他们将大肠杆菌基因组 DNA 分别处理成 2kb 和 6kb 的片段，其中 2kb 的模板只用 C1 试剂进行了测序，6kb 的模板则分别用 C1 和 C2 试剂进行了测序。Hurban 对于使用 C2 试剂获得的结果非常高兴。他们发现使用 C2 试剂后，测序平均读长提高了 80%，可达到 2700bp 左右。

C2 试剂第一次被使用是在德国大肠杆菌疫情研究中，研究人员通过将不同测序模式混合使用，最终获得了 2900bp 的平均读长以及 99.998% 的准确度一致性。

Expression Analysis 的研究团队在 C2 试剂的帮助下，获得了 2715bp 的平均读长，最长的读长甚至达到了 13091bp。和同样模板使用 C1 试剂完成的数据比较后发现，C2 试剂将测序平均读长从 1519bp 提高了 80%，最长读长则从 6512bp 提高了一倍。

4) 加拿大安大略癌症研究所（OICR）正在使用 PacBioRS 对转移或者复发的癌症病人进行靶向测序研究

OICR 正在使用 PacBio RS 系统对 19 个已知癌症基因进行测序，位于多伦多的 Princess Margaret 医院的 CLIA 实验室则用 Sequenom 的 OncoCarta panel 对 OICR 的测序结果进行验证。

OICR 癌症基因项目的负责人 John McPherson 表示，他们的目标是验证 PacBioRS 系统并且试图建立起一套对癌症病人的测序体系，这样可以帮助医生获取更多信息从而做出更恰当的治疗决定或者临床试验安排。OICR 希望最终能将这套体系应用于对癌症病人的标准护理流程中。

目前，该小组已经完成了 15 个病人的测序，其中一半的病人 19 个基因中至少有 1 个发生了突变。PacBioRS 没有错过任何一个 OncoCarta 试剂盒检出的突变，并且，还发现了一个新的突变。“这就是我们为什么要做测序，我们希望能发现现有的检测试剂盒未涵盖的信息”McPherson 说。

目前该项目的周转时间为 3 周，整个流程包括：病人签署同意书，组织活检、DNA 抽提、测序、分析、验证以及同临床医生和研究人员开会给出相关报告。由于周转时间的限制，PacBioRS 的测序完成速度也是 McPherson 的研究小组选择它的一个主要原因。

McPherson 表示他们从 PacBio 上获得的结果非常好，平均读长大概在 1800bp，研究人员使用环形比对测序模式来确保测序的准确性并且获得了是扩增子数百倍覆盖深度的测序数据。

5) 最新发布“靶向重测序—EGFR-MET 基因”的技术文献

核心内容：

Fluidigm Access Array 系统同 PacBioRS 的兼容性已经被证实, 相应的靶向富集策略流程也已公布。

标准测序模式和环形比对测序模式都可以用于检测 SNP。标准测序模式的优点在于既可以维持完整的读长且并不会对 SNP 的检测造成影响。

参考文献:

1) The Origin of the Haitian Cholera Outbreak Strain, December 9, 2010, The New England Journal of Medicine.

Authors: Chen-Shan Chin, Ph.D., Jon Sorenson, Ph.D., Jason B. Harris, M.D., William P. Robins, Ph.D., Richelle C. Charles, M.D., Roger R. Jean-Charles, M.D., James Bullard, Ph.D., Dale R. Webster, Ph.D., Andrew Kasarskis, Ph.D., Paul Peluso, Ph.D., Ellen E. Paxinos, Ph.D., Yoshiharu Yamaichi, Ph.D., Stephen B. Calderwood, M.D., John J. Mekalanos, Ph.D., Eric E. Schadt, Ph.D., and Matthew K. Waldor, M.D., Ph.D.

2) Origins of the E. coli Strain Causing an Outbreak of Hemolytic-Uremic Syndrome in Germany, July 27, 2011, The New England Journal of Medicine.

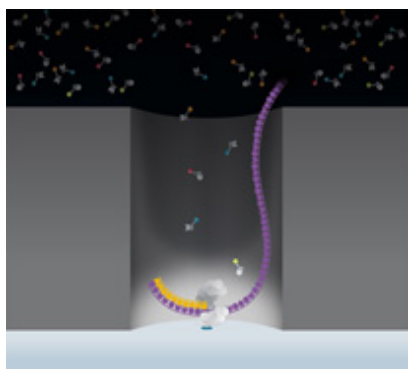
Authors: David A. Rasko, Ph.D., Dale R. Webster, Ph.D., Jason W. Sahl, Ph.D., Ali Bashir, Ph.D., Nadia Boisen, Ph.D., Flemming Scheutz, Ph.D., Ellen E. Paxinos, Ph.D., Robert Sebra, Ph.D., Chen-Shan Chin, Ph.D., Dimitris Iliopoulos, Ph.D., Aaron Klammer, Ph.D., Paul Peluso, Ph.D., Lawrence Lee, Ph.D., Andrey O. Kislyuk, Ph.D., James Bullard, Ph.D., Andrew Kasarskis, Ph.D., Susanna Wang, B.S., John Eid, Ph.D., David Rank, Ph.D., Julia C. Redman, Ph.D., Susan R. Steyert, Ph.D., Jakob Frimodt-Møller, B.S., Carsten Struve, Ph.D., Andreas M. Petersen, Ph.D., Karen A. Krogfelt, Ph.D., James P. Nataro, M.D., Ph.D., M.B.A., Eric E. Schadt, Ph.D., and Matthew K. Waldor, M.D., Ph.D.

纵观第三代测序之 Pacific Biosciences

众所周知，天然的 DNA 复制本身就是一个非常高效且准确的过程。真核生物 DNA 复制的速度达几十个核苷酸/秒。若能以这样的神速测序，那可比 Sanger 测序快了几万倍。然而，知易行难。如何在不影响聚合酶活性的前提下实时观察 DNA 合成，这可是个棘手的问题。要知道，DNA 聚合酶的直径只有 15 nm。而且，这是一个涉及酶学、表面化学和检测光学的多学科难题。

早在 10 多年前，美国康奈尔大学的研究生 Steven Turner 和 Jonas Korlach 就开始为解决这一难题而做出不断努力。如今，Korlach 和 Turner 创办了一家名为 Pacific Biosciences（以下简称 PacBio）的公司，并且终于将他们的想法转变成为现实产品。让我们看看他们主要克服了哪些挑战。

实时观察 DNA 聚合酶的主要挑战之一是如何能在 DNA 合成期间检测到单个核苷酸的掺入。因为在显微镜实时记录 DNA 链上荧光的时候，周围众多的荧光标记的核苷酸形成了非常强大的荧光背景。这种强大的荧光背景使单分子的荧光探测成为不可能。正在他们一筹莫展之际，突然，微波炉门给了他们启发。微波炉门上的金属筛布满了小洞，它们比微波的波长要小得多。因此，这些洞能阻止微波通过并穿透玻璃。然而，波长更小的可见光就能够通过，让我们看到食物。他们的 SMRT 技术应用了相同的原理，不过规格就缩小至纳米级，称之为 ZMW（zero-mode waveguide，零模式波导）。



ZMW 是一个直径为几十纳米的小孔，它阻止可见的激光(波长大约为 600 nm)完全透过 ZMW。激光在进入 ZMW 后迅速衰减，因此，只有底部 30 nm 被照亮（下图）。在每个 ZMW 中，单个 DNA 聚合酶分子利用专利技术锚定在底部玻璃的表面。随后核苷酸涌入 ZMW 中，并在阵列表面扩散。当聚合酶检测到正确的核苷酸时，便将其掺入新生链中，这个过程需要几毫秒，而单纯的扩散只需要几微秒。这种时间差使掺入的核苷酸产生了很高的信号强度，类似于脉冲信号。因此，ZMW 有能力在荧光标记核苷酸的背景下检测单个掺入事件。

测序是在专利的 SMRT Cell 中开展的，每个 Cell 中有一个阵列，上面有 15 万个 ZMW。每个 ZMW 都能够包含一个 DNA 聚合酶及一条不同的 DNA 样品链。每次运行能够平行检测大约 75,000 个单分子测序反应。

Korlach 与 Turner 还有一项重大发明。传统的核苷酸标记方法是将荧光标记连接到核苷酸的碱基上，也就是掺入 DNA 链中。这对于实时观察 DNA 合成也是有问题的，因为大分子的染料会干扰 DNA 聚合酶的活性，或造成聚合反应提前终止。因此，Korlach 和 Turner 就开发出一种新型的核苷酸标记方法，即在核苷酸的磷酸链上进行标记，而不是碱基，这样，一旦核苷酸掺入到新生 DNA 链中，标记基团就会自动脱落。

当然，要将 ZMW 中的反应记录下来，还需要一台高精尖的仪器。PacBio RS 于是诞生了，它能够实时开展、监控并分析单分子生化反应。PacBio RS 使用一个大孔径物镜和四个单光子照相机来收集荧光所发射的光脉冲，观察整个过程；并使用一套优化的算法，将光学系统所捕获的信息翻译成 ACGT 碱基。一旦测序开始，实时的数据就传送到初步分析流水线，生成碱基身份和质量值。

以上三种新技术结合，就能让研究人员在短短的几分钟内对长片段 DNA 进行测序。它的速度、读长、灵活性和费用让其从第三代测序中脱颖而出。从样本制备到测序，所需的时间还不到一天。典型的测序运行时间低至 30 分钟。序列数据在几分钟之内就能产生，而不是几天，这对于传染病监控和分子病理学尤为重要。PacBio RS 目前的读长超过 1kb，比第二代测序要长得多。此外，试剂消耗和样本制备极少，不需要常规的 PCR 扩增，失误也大大减少，聚合酶动力学的直接观察赋予了测序之外的更多应用。凭借这些优势，PacBio 今年被美国麻省理工学院（MIT）的《技术评论》（Technology Review）杂志评为全球 50 家最具创新力的企业之一。

一系列的文章也证明了 SMRT 技术。Korlach 与 Turner 于 2009 年 2 月在《科学》杂志上发表了一篇介绍 PacBio 单分子 DNA 测序技术的文章。这篇文章代表了首个第三代测序技术的“原理验证”，因此引用率非常高。而后，他们又利用 SMRT 技术，直接测定了 DNA 的甲基化，这相对目前流行的第二代测序技术又前进了一大步。文章发表在 今年 5 月的《Nature Methods》上。

荧光脉冲的到达时间和持续时间反映了有关聚合酶动力学的信息，从而允许直接检测 DNA 模板链中的修饰核苷酸，包括 N6-甲基腺嘌呤、5-甲基胞嘧啶和 5-羟甲基胞嘧啶。各种修饰对聚合酶动

力学的影响不一，从而能够将它们区分开。研究人员使用这些动力学特征，鉴定出基因组样品中的腺嘌呤甲基化，并发现再结合 circular consensus sequencing，他们能够在单碱基分辨率上鉴定出表观遗传学修饰（mA、mC 和 hmC）。研究人员还预计，其它的表现遗传学修饰，以及各种形式的 DNA 伤害，也能用这种方法检测。

看到这里，大家可能迫不及待地想知道 PacBio 的测序仪何时上市，价格多少。其实，PacBio RS 已成型，目前正在早期试用阶段。今年 2 月份，PacBio 宣布了 10 位北美的早期试用客户，包括贝勒医学院、冷泉港实验室、美国能源部联合基因组研究所、华盛顿大学基因组中心和斯坦福大学等等，并与第二季度开始发货。8 月份，PacBio 又迎来了它的首位欧洲试用客户，大名鼎鼎的 Wellcome Trust Sanger 研究院。

整个测序系统包括 PacBio RS 仪器以及专利的耗材。耗材部分包括 SMRT Cell 和试剂盒。PacBio RS 上的每个反应消耗一个 SMRT Cell。8 个 SMRT Cell 包装成 8Pac 的形式。另外，他们还提供了三种试剂盒，用途各不同。Template Preparation Kit 将 DNA 转化成 SMRTbell 文库形式。Binding Kit 则将这个文库与聚合酶结合，为测序做准备。Sequencing Kit 包含了实时测序所需的试剂。每个样本可在单个 SMRT Cell 上测序，也可根据项目需求在多个 SMRT Cell 中测序。因此，每个反应的价格取决于实验设计。

PacBio 预计，到 2013 年，个人基因组的测序能在 15 分钟内完成，费用低于 1000 美元，人人都可以消费得起。说实话，2008 年我第一次听到这个消息时，心中饱存疑惑，看来我实在是低估了 PacBio 的实力，也低估了测序技术发展的脚步。

（生物通 余亮）

单分子测序灵敏检测 5-hmC

甲基化研究如今开展得如火如荼。除了人们熟知的 5-mC, 很多研究人员也将目光转向了另一种 DNA 修饰——5-羟甲基胞嘧啶 (5-hmC)。5-hmC 是 5-mC 的羟基化形式, 1952 年在噬菌体 DNA 中首次发现。几十年来, 羟甲基胞嘧啶的研究一直围绕着它对于噬菌体 DNA 的保护作用。直到最近, 人们发现它与多个关键的细胞功能有关, 包括胚胎干细胞发育、正常的骨髓细胞生成和受精卵发育。

若想深入了解 5-hmC 的生物学功能, 必须开发出一种灵敏的测序方法, 以揭示它在基因组中的位置。而现有的测序方法, 如亚硫酸氢盐测序, 无法区分 5-mC 和 5-hmC。为此, 美国芝加哥大学与 Pacific Biosciences 公司的研究人员合作, 利用第三代单分子测序技术和 5-hmC 的选择性化学标记方法来高通量检测 5-hmC。该研究成果于 11 月 20 日发表在《Nature Methods》在线版上。

Pacific Biosciences 公司开发的单分子实时 (SMRT) DNA 测序是一种第三代的测序技术, 它利用单个 DNA 聚合酶分子来开展 DNA 合成, 并监控核苷酸的不断掺入。核苷酸掺入的实时记录, 不仅产生了序列读取, 还带来了关于聚合酶动力学的宝贵信息。通过这些信息, 研究人员能鉴别 DNA 碱基的修饰。

一般来说, 聚合酶在修饰碱基周围的速率要比未修饰 DNA 慢。通过比较每个模板位点的掺入时间、脉冲间持续时间 (IPD), 可定量比较这一差异。利用这一方法, 研究人员可直接检测 DNA 甲基化, 包括 N6-甲基腺嘌呤、5-mC 和 5-hmC。聚合酶动力学的测定是 SMRT 测序的内在组成, 不会对 DNA 一级序列的测定有任何不良影响。

研究人员还对 5-hmC 进行了选择性化学标记。首先利用 β -葡萄糖苷转移酶为 5-hmC 引入叠氮修饰, 之后通过点击化学 (click chemistry) 反应安装一个生物素标签。通过这种方法, 研究人员富集了基因组中的 5-hmC, 减少所需的测序量。此外, 5-hmC 上的大体积标签在 SMRT 测序中产生了更大的动力学信号, 以便在较低覆盖度时更自信地指定修饰碱基。

为了检验这种方法的可行性, 研究人员检测了小鼠胚胎干细胞 (mESC) 中的 5-hmC。在选择性标记和富集之后, 他们制备了 SMRTbell DNA 文库, 用于鸟枪法 SMRT 测序, 并分析动力学信号。他们检测到清晰的 HS-N3-5-gmC 信号。由于动力学信号的大幅增加, 在单分子读取中可轻松指定 5-hmC 的位点, 而无需与未修饰的对照样品进行比较。

作者认为, 这种技术将是研究 5-hmC 生物学功能的有力方法。(生物通 余亮)

[欢迎索取PacBio RS测序仪的更多资料!](#)

原文检索

Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine

Nature Methods (2011)
doi:10.1038/nmeth.1779

We describe strand-specific, base-resolution detection of 5-hydroxymethylcytosine (5-hmC) in genomic DNA with single-molecule sensitivity, combining a bioorthogonal, selective chemical labeling method of 5-hmC with single-molecule, real-time (SMRT) DNA sequencing. The chemical labeling not only allows affinity enrichment of 5-hmC-containing DNA fragments but also enhances the kinetic signal of 5-hmC during SMRT sequencing. We applied the approach to sequence 5-hmC in a genomic DNA sample with high confidence.

单分子测序助力脆性 X 综合征的研究

生物通报 美国加州大学戴维斯分校（UC Davis）医学院及 Pacific Biosciences 公司的研究人员利用单分子实时测序技术，对脆性 X 综合征的关键基因 FMR1 中的 CGG 片段扩增进行了测序。他们在第 15 届脆性 X 和早发性认知缺损国际研讨会上公布了他们的研究成果。

脆性 X 综合征（又称 Martin-Bell 综合征）是一种遗传性综合征。它导致一系列特征性症状，包括生理、智力、情绪、以及行为上的异常。症状的轻重各有不同。该疾病伴随着 X 染色体上一段三核苷酸序列（CGG）的扩增。这种扩增导致了一种称为 FMR1 的蛋白质无法在病人体内表达，而该蛋白质是神经正常发育所必需的。

脆性 X 智力低下（FMR1）基因位于 X 染色体上，其 5'非翻译区（5'UTR）包含了 CGG 三核苷酸的重复序列。大部分人的 CGG 片段重复 5-44 次。CGG 片段的扩增与认知力和生殖功能受损相关。如果一个人含有 45-54 个重复序列，则被认为是“灰色地带”，而 55-200 个重复序列被称为前突变，200 个以上重复序列则被称为全突变。

中度 CGG 重复序列扩增（55-200 个 CGG 重复）可导致脆性 X 相关性震颤/共济失调综合征（FXTAS），一种以意向性震颤、步态共济失调和其他症状为特征的迟发性神经变性疾病 1。而高度 CGG 重复序列扩增（>200 个 CGG 重复）可引起转录沉默和 FMR1 编码蛋白（FMRP）的功能丧失，从而导致脆性 X 染色体综合征 1。

由于目前的测序方法都依赖于大量相同拷贝的读取，因而大量扩增的 CGG 重复（>100）只能生成不连贯的信号（图 1）。这阻碍了研究人员获得单碱基分辨率的测序数据。

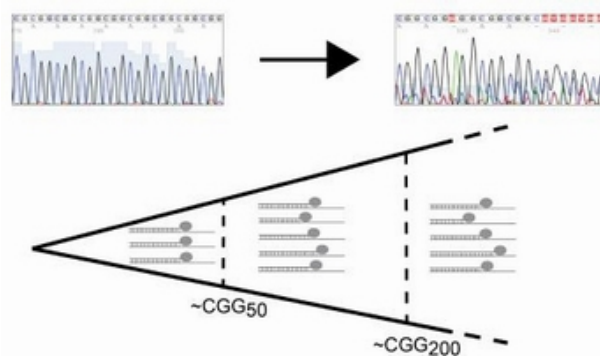
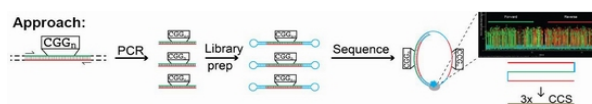


图 1：随着重复扩增数量的增多，生成大量相同的拷贝变得越来越困难。这阻碍了对大部分疾病范围(>100 次重复)内重复扩增序列的准确测序。

于是，研究人员利用 Pacific Biosciences 公司的单分子实时（SMRT）测序技术，试图从大量扩增的 CGG 重复序列中产生高质量的测序数据（图 2）。他们首先利用 PCR 从人类基因组 DNA 中扩增出带不同 CGG 重复的 FMR1 序列，制备出待测文库。随后加上发夹测序接头，生成闭合的环状测序模板。之后利用 Pacific Biosciences RS 对样品进行测序。最后采用 BLASR 算法将长读取装配成环状共有序列（CCS），而目标序列的覆盖度至少为 3 倍。



结果表明，环状共有序列的读取产生了生理相关范围的重复扩增。而对长的重复等位基因的测序能力似乎仅受读长限制。原始读长超过 10 kb，使

得环状共有序列覆盖了超过 750 个 CGG 重复片段。

[索取PacBioRS单分子实时测序系统的详细资料](#)

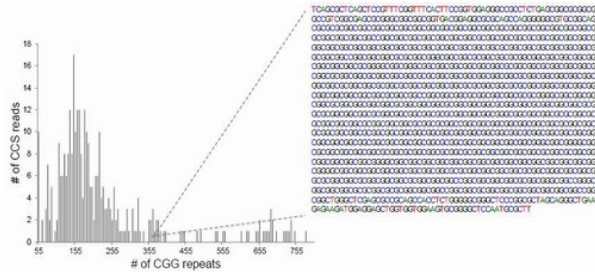


图 3：一段 CGG 重复片段范围内的环状共有序列读取图。接近 155 个重复的峰是 SMRT cell 被动上样偏向（passive loading bias）与较小插入

片段优势的综合结果。示例：环状共有序列读取包含 372 个 CGG 重复。

研究人员认为，通过这种方法，如今能够以单碱基分辨率完整鉴定所有与疾病相关的等位基因，从而有利于疾病的准确诊断。（生物通 余亮）

参考文献

1. Oostra, B.A. and R. Willemsen, FMR1: a gene with three faces. *Biochim Biophys Acta*, 2009. 1790(6): p. 467-77.
2. Eid, J., et al., Real-time DNA sequencing from single polymerase molecules. *Science*, 2009. 323(5910): p. 133-8.

利用基因组学来对付疾病爆发

这个夏天，德国爆发了严重的大肠杆菌疫情，3500 人感染，50 人死亡。全世界多个研究小组迅速调动他们的实验室，对这种细菌基因组进行测序和分析，在几天之内找到有效的治疗方案。马里兰大学医学院的 David Rasko 正是其中的一位，他与来自丹麦国立血清研究所的小组合作，对细菌进行测序，并研究了此次爆发的流行病学。在本周举行的美国分子病理学会年会上，Rasko 谈了他的经验，他认为全基因组测序能在快速响应疾病爆发上占据一席之地。

Rasko 谈到，来自华大基因、德国哥廷根大学、美国/丹麦小组的研究人员及其他都利用了不同的技术对细菌进行测序，而每一种技术都有其自身的优势。利用 Life Technologies 的 Ion Torrent 个人型操作基因组测序仪（PGM），德国和中国的研究人员能够在 3 小时内测序，并在 5 天内开发出诊断检测。随后研究小组发布了多名研究人员分析出的原始数据。

Rasko 的小组则利用 Pacific Biosciences 的技术来对细菌测序。尽管 PacBio 的仪器耗时略长，但 Rasko 表示他的研究小组产生了更长的读取。由于所有技术和快速分析是由专家完成的，所以研

究小组能够将此次爆发菌株的基因组与其他大肠杆菌菌株进行比较，构建了一个系统发育树，并确定此菌株（O104:H4）是从未出现过的，携带了志贺氏毒素。

功能基因组学分析告诉研究人员，就像平时细菌感染一样使用环丙沙星进行治疗，会使病人感觉更糟，这正是因为志贺氏毒素。而比较基因组学分析表明，爆发菌株中存在的一个质粒赋予了它耐药性的基因。Rasko 认为，所有这些表明全基因组测序及随后的基因组学分析可协助医生更及时、彻底、高效地对付疾病爆发。（生物通编译）

[欢迎索取PacBio RS测序仪的更多资料！](#)

致命大肠杆菌研究成果登上 《新英格兰医学杂志》

大肠杆菌疫情总算告一段落。这种突如其来的大肠杆菌最初在德国出现，然后迅速蔓延至欧洲其他国家、甚至美国，共夺走了 52 人的生命。它为何有如此大的威力？全世界研究人员都在努力破解。近日，一个国际小组利用 Pacific Biosciences 的单分子测序技术，深入研究了此次暴发菌株的致病性和进化起源，该研究成果发表在《新英格兰医学杂志》上。

研究人员对暴发的菌株以及 11 种相关的大肠杆菌菌株进行了测序，确定暴发菌株 C227-11 是肠聚集性大肠杆菌（EAEC）中的一员，但它与其他 O104:H4 菌株不同，因为它包含了编码志贺毒素 2 的噬菌体以及其他毒力和耐药性因子。

他们确定，最近与肠出血性大肠杆菌（EHEC）菌株的水平基因交换引发了这种高毒力暴发菌株的出现。同时，他们还发现，某些抗生素（包括环丙沙星）提高了志贺毒素 2 基因的表达，这表明应谨慎使用某些类别的抗生素，以抵抗这种新出现的病原体。

这项研究的作者来自第三代测序公司 PacBio、马里兰大学医学院、世界卫生组织、丹麦 Statens Serum 研究所等机构。他们提到，在此次暴发之前，只有三个 EAEC 基因组经过了测序，因此对这种大肠杆菌的系统发育的基因组规模了解很有限。

为了进一步深入了解这种菌株的进化历史，他们使用 PacBio RS 对暴发菌株 C227-11、7 种其他的 EAEC O104:H4 菌株以及 4 种参考菌株进行了测序。平行利用三台 PacBio RS，他们获得了每个分离株的约 75 倍覆盖度。平均读长达 2067 个碱基。之后，他们比较了 11 个菌株的数据，包括 Ion Torrent 和 454 的测序结果，以便找出所有分离株的拷贝数变异和单核苷酸变异。

研究人员还利用了 53 个大肠杆菌和痢疾杆菌的基因组数据，生成了系统发育树，描绘暴发菌株的进化。尽管 EAEC 分离株分布在整个树上，但所有的 EAEC O104:H4 菌株形成了一个不同的分支，有着非常保守的核心基因组。

此次德国暴发的编码志贺毒素的菌株与缺乏编码志贺毒素的噬菌体的 EAEC O104:H4 菌株之间的相似度说明噬菌体整合到 EAEC 基因组是一个最近发生的事件。此外，暴发菌株位于 EAEC O104:H4 分支证实了暴发菌株不是典型的肠出血性大肠杆菌。

文章的作者之一，马里兰大学医学院的助理教授 David Rasko 谈到：“这种多菌株的测序数据和分析显著增加了关于致命大肠杆菌的科学信息量，并对其致病媒介有了关键的了解。我们的结果提供了迄今为止最完整的基因组数据，并强调了 DNA 测序的重要性，它帮助我们了解细菌基因组的可塑性如何促进新病原体的出现。”（生物通 薄荷）

原文检索

Origins of the E. coli Strain Causing an Outbreak of Hemolytic-Uremic Syndrome in Germany

JGI 购入 PacBio 取代 454 明年测序 47Tb

美国能源部联合基因组研究所（简称 JGI）正在调整其测序技术，因其准备在 2012 财年产生超过 47 Tb 的 DNA 序列，其中一半以上的测序产量将用于其最大的用户计划，群体测序计划（Community Sequencing Program，简称 CSP）。

据 JGI 发言人 David Gilbert 透露，JGI 目前拥有 8 台 Illumina 的 HiSeq, 2 台 Illumina 的 MiSeq 和 1 台 PacBio RS 测序仪。该研究所刚刚淘汰了 2 台罗氏 454 的仪器和 5 台 Illumina 的 GAIIx 测序仪，并购进了第二台 RS 系统，Gilbert 表示，RS 系统本周到货，应该在下个月开始运行。

JGI 的主管 Eddy Rubin 告诉《In Sequence》，由于 JGI 扩充了 Illumina 测序仪的队伍，并利用 PacBio RS 加强其产生长读取的能力，因此减少了对罗氏 454 仪器的依赖。他解释道：“我们能够用 Illumina 更经济地开展多个应用，并且我们希望能够用 PacBio 做之前罗氏（454 测序仪）所做的许多事情。”

长读取对许多项目而言仍然有优势，包括那些过去从未鉴定的基因组或宏基因组的测序工作。Rubin 表示，JGI 计划在多个长读取应用中使用 PacBio 平台，在这之前他们是使用罗氏 454 的测序技术。

Rubin 谈到：“我们的重点之一是 de novo 测序。对于 de novo 测序，我们之前不了解其基因组——如宏基因组环境、真菌、植物等——长读取确实是一个优势，因此 PacBio 的确有帮助。”

尽管目前利用 PacBio 测序的成本仍相对较贵，但他表示，PacBio RS 的长读长和所产生的错误的特点很吸引人。

“PacBio 从通量角度而言，成本不便宜，这是它目前的一个缺点，但它的长读长确实很棒，”Rubin 解释道。“另外它产生的错误是随机错误的——看起来可以很好的覆盖不同 GC 含量的基因组区域（高 GC 和低 GC 区域）。”

不过 JGI 的大多数测序工作仍将使用 Illumina 的机器进行（相对 PacBio 来说，Illumina 现阶段的测序成本更低），Rubin 称之为“干了大部分活的主力军”。

41 个新计划

JGI 上周宣布，它已经从参与其 2012 群体测序计划的 152 名申请者中挑选出 41 个研究项目。

这家位于加州 Walnut Creek 的基因组学机构为 CSP 选中的项目提供测序服务和分析协助，Rubin 解释道。对他们而言，参与研究者必须能够证明，他们具有金融手段使项目的其他方面得以实现，包括采集样本和 JGI 所提供的数据处理的资金。

Rubin 告诉《In Sequence》，符合资格的项目是那些涉及生物圈几乎每个方面的项目，除生物医学外，包括过去从未鉴定的基因组、植物-微生物相互作用，及微生物群落研究。

批准在 2012 年开展的 41 个项目涉及方方面面，从旨在充实生命之树各分支的基因组测序研究，到与生物修复和/或生物能源生产相关的宏基因组研究。

除了 DNA 测序,一些 CSP 计划还将包括 RNA 测序,它们旨在评估基因表达或宏转录组。

2012 CSP 计划的一部分将涉及特定物种的全基因组测序,包括一项由俄勒冈州立大学和加州大学河滨分校的研究人员领导的研究,它将对大约 1,000 种真菌基因组进行测序,以便为 577 个真菌家族中的每一个建立两个或更多的参考基因组。

除此之外,2012 CSP 计划中的许多是旨在了解群落基因组学以及特定环境中物种之间的相互作用,而不是单个基因组,Rubin 解释道。

例如,CSP 大型计划之一,由北卡罗莱纳大学研究人员 Jeff Dangl 领导的研究小组正在寻找植物之间的相互作用,如拟南芥、玉米或生物能源作物芒草,以及围绕这些植物根部的跟际微生物群落。

Rubin 表示:“在未来某个时候,我们希望通过引进某个微生物群落来改善植物生长。过去,基因组学在很大程度上是了解一种微生物或一棵树。现在是要了解它们的相互作用。”

总体而言,预计 JGI 在 2012 财年产生的序列中的~55%是用于 CSP 计划。剩余的测序能力将投入到美国能源部的生物能源研究中心、JGI Director 的科学计划、低剂量辐射研究计划,及国际合作生物多样性小组(ICBG)计划。

具体来说,JGI 明年将产生 47.1 Tb 碱基,其中大约 30%将用在生物能源研究中心,这是美国能源部资助的研究中心,致力于将生物质转化成生物能源。JGI Director 的科学计划已分配 10%,而低剂量辐射研究计划和 ICBG 计划将占据剩余的 5%。

30 倍的增加

在 2009 年之前,JGI 主要依赖 Sanger 测序来开展 CSP 及其他计划(详见 IS 7/11/2008)。从那时起,该研究所已逐步转向新一代测序平台。

Gilbert 提到,近几年来,通过采用高通量测序技术,JGI 已实现测序产量的巨大增长:在 2009-2011 年间,研究所的测序产量已提高近 30 倍,而经费仍保持在约 7,000 万美元。

特别是 CSP,2012 年的测序产量有望达到 2011 年的 4-5 倍。

Rubin 表示:“所有宏基因组计划正在越变越大。”

“在 Sanger(测序)时,宏基因组计划是数千万个碱基对。如今我们的宏基因组计划有万亿个碱基。”他补充道。

为了处理它所产生的海量序列数据,JGI 开始利用伯克利超级计算机构来储存和分析序列数据,此机构也被称为国家能源研究科学计算中心(NERSC)。Rubin 谈到,为此,JGI 正将基因组算法转化成一种能与 NERSC 的高性能计算硬件兼容的格式。

他提到,除了 DNA 和 RNA 测序能力,该中心还在开发其他测序相关的能力,包括从那些无法培养的单个微生物细胞中扩增 DNA 并测序的技术—此技术将实现一些项目,在对指定微生物群落进行宏基因组测序的同时,对同一环境中的微生物进行单细胞测序。

“JGI 过去主要提供序列分析,而我们将逐步提供其他东西:我们提供单细胞基因组学,我们有一个小 DNA 合成计划,”Rubin 谈到。“我们正逐步提供其他各种能力,开始成为一个新一代基因组中心。”

德国总理默克尔启动第三代测序仪

2011年9月13日，来自柏林的消息——德国总理默克尔在访问柏林 Max Delbrück 分子医学中心时，启动了该中心新安装的一台单分子实时 DNA 测序系统——PacBio RS。该系统是由美国加利福尼亚州的 Pacific Biosciences 公司开发并制造的。Jonas Korlach 博士，PacBio 技术的共同发明者也出席了该仪式，同时他也是一位土生土长的柏林人。

PacBio RS 是一台革命性的第三代 DNA 测序系统，它融合了新颖的单分子测序技术和高级的分析技术来实时揭秘生物学。此系统有着其他系统无以伦比的序列读长，平均达 1000 个 DNA 碱基。与第二代测序系统相比，它能更快获得结果——不到一天，且测序前无需扩增 DNA，避免了潜在的错误。PacBio RS 平台被业界普遍看好，在未正式推出之前，已有多家机构在排队购买。它于今年 4 月正式推出。



德国总理默克尔按下了 PacBio RS 系统的启动按钮。（图片版权归 MDC 所有）



PacBio 技术的共同发明者 Jonas Korlach 博士（右）向默克尔总理解释这台仪器如何工作。（图片版权归 MDC 所有）

Max Delbrück 分子医学中心（简称 MDC）研究细胞内疾病的发展，关注基因和蛋白。该中心利

用非常先进的技术，深入了解基因和蛋白的分子网络，并研究它们的调控和相互作用，以及对疾病发展的重要性。MDC 的 Nikolaus Rajewsky 教授向默克尔总理解释道：“为了实现这一点，必须处理巨大量的数据。这种海量的数据可通过数学和统计学的方法以及高性能的计算来分析。”

Rajewsky 教授对这一技术发表评论认为：“SMRT 技术的突出特点是我們不仅能够了解 DNA 如何合成，还能通过它产生的数据定量分析基因调控、RNA 功能、表观遗传调控、DNA 修饰和基因组结构。这项技术让我们更深入地了解基因调控网络，并为个性化医疗带来新方法。”

Max Delbrück 分子医学中心的 PacBio RS 测序系统是欧洲安装的第三台系统。除了 MDC，目前只有英国的 Sanger 研究院安装了这种测序仪。Pacific Biosciences 欧洲副总裁 Terry Pizzie 表示：“自从我们 4 月份正式推出以来，欧洲市场表现出了极大的兴趣。看到产品被我们的客户以及总理认可，感觉很棒，而总理也是一名物理学家。”

Korlach 博士也认为这一事件具有象征意义：“柏林墙倒塌时，我就在这里，这为我带来了出国留学的机会，我才能够康奈尔大学与 Steve Turner 共同发明了这项技术。回到我开始的地方，与大家分享我们多年艰苦努力的成果，是一种荣誉，如今这台产品将为德国的科学发现和医学创新作出贡献。”（生物通 余亮）

[索取 PacBio RS 系统的最新资料或询价](#)

PacBio 正式发售第三代测序仪

生物通报道，第三代测序公司 Pacific Biosciences 上周宣布已开始正式发售 PacBio RS 系统。当天它发售了两台 RS 测序仪，一台是发往美国国家生化防卫分析与对策中心（NBACC），另一台是发往冷泉港实验室。

PacBio RS 是一台革命性的 DNA 测序系统，它融合了新颖的单分子测序技术和高级的分析技术来实时揭秘生物学。此系统有着其他系统无以伦比的序列读长，平均达 1000 个 DNA 碱基。与第二代测序系统相比，它能更快获得结果——不到一天。RS 测序仪的发售有望扩展 DNA 测序在更多领域的应用，诸如癌症研究、病原体检测和农业。

NBACC 基因组学的小组负责人 Nick Bergman 博士评价道：“PacBio RS 系统是 NBACC 测序项目的重大扩展，其长读长和通量灵活性为我们鉴定微生物病原体提供了很多新的选择。我们非常激动能在多个应用中率先使用它。”

冷泉港实验室曾是 RS 测序仪的早期试用客户之一。W. Richard McCombie 博士表示，他们非常看重长读长在多个项目中的价值，这些项目包括了解人类基因组中的结构变异，以及植物基因组的 de novo 测序。

除了这两个机构，第一批仪器还将发往北美和欧洲的多个机构，包括生物技术公司、商业化服务供应商、政府机构和科研实验室。Wellcome Trust Sanger 研究院也是 PacBio RS 的早期试用客户之一，最近，他们也将升级到商业化的硬件配置。

Sanger 研究院测序技术的主管 Harold Swerdlow 博士表示：“为了维持基因组学的前沿地位，我们探索基因组测序中的新机会。我们计划使用 PacBio RS 来改善病原体的 de novo 拼接，并提高一些物种的序列信息覆盖度，在未来，我们将通过甲基化位点的直接检测来探索表观遗传学。”

Pacific Biosciences 的 CEO Hugh Martin 在一次电话会议中表示，该公司计划在迟些时候用最终发布的软件升级测试系统。他还谈到，目前 PacBio 有 44 台仪器的积压订单，计划在未来几个季度陆续发售。

他表示：“我们计划以非常可控的方式发送并安装积压的系统。我们的目标是成功培养每个客户。因此，我们分阶段发送系统，以便我们的安装和支持资源能够有足够时间支持每一位客户，然后再转去下一位。”

Pacific Biosciences 上周还宣布，其第一季度净亏损较去年同期增长 15%，因为该公司正在准备 PacBio RS 系统的正式发售。第一季度净亏损达 3480 万美元，平均每股亏 0.66 美元。PacBio 于去年十月上市，筹集资金 2 亿美元。

（生物通 余亮）

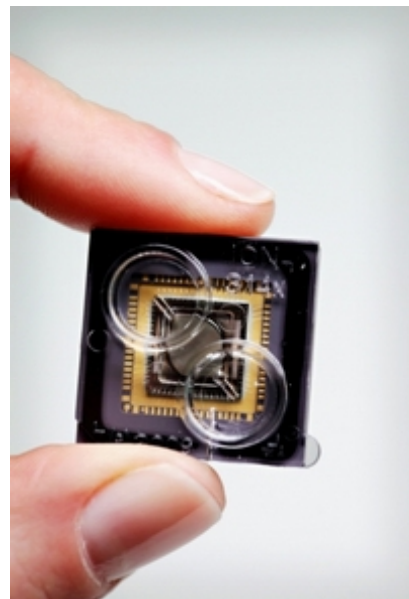
半导体测序仪 每个实验室都触手可及

半导体技术已经改变了它所触及的每一个行业，从计算到摄影再到音乐。如今，它正在改变生命科学。也许，你很难想象，一块半导体芯片就是一台测序仪，但事实上，Life Technologies 的 Ion 个人化操作基因组测序仪 (PGM™) 正是如此。难怪在 2010 年美国基因组生物学技术进展年会上甫一亮相时，就立刻惊艳全场。

当时，这台测序仪还不属于 Life Technologies，它的制造商是没啥名气的 Ion Torrent 公司。这家公司总部位于美国康涅狄格州吉尔福德，是由 Jonathan Rothberg 创办的。是不是觉得这个名字有些耳熟？没错，他正是 454 生命科学的创始人。2007 年他成功将 454 卖给了罗氏公司，之后创办了 Ion Torrent。去年 8 月，Life Technologies 收购了这家公司。

Rothberg 博士的梦想是打造测序领域的“数码相机”，既实惠又易用，还能快速得到结果。他确实做到了。与其他测序平台相比，Ion PGM™ 测序仪快速、简捷，扩展性好，更关键的一点是，价格实惠。它让测序平民化，让每个实验室都触手可及。因此，Ion PGM™ 测序仪在推出以来的 6 个月内创下了多项纪录。它的通量已提高了 10 倍，它的流程成为行业中最快的，并且它是世界上销售最快的测序仪，已发往 40 多个国家。

Ion Torrent 技术使用专有的大规模并行半导体传感器芯片，对 DNA 复制期间产生的氢离子开展实时测定。离子半导体芯片上高密度的孔阵列提供了数百万个单独的反应器，而集成的流体通路让试剂流过整个传感器芯片。这种流体体系、微体系机械设计和半导体技术的独特组合，让我们能够将遗传信息 (DNA) 直接翻译成数码信息 (DNA 序列)，快速生成大量的高质量数据。[点击索取 Ion PGM 测序仪的更多资料](#)



扩展

英特尔的创始人之一戈登·摩尔曾提出过著名的摩尔定律，即集成电路中可容纳的晶体管数量将每两年翻一番。而 Life Tech 和 Rothberg 似乎并不满足于此，他们许下每六个月将 PGM 产量提高 10 倍的承诺。如今，新的 Ion 316™ 芯片有着 100 Mb 的通量，比原先的 Ion 314™ 芯片高了 10 倍。而第四季度上市的 Ion 318™ 芯片将带来 1 Gb 的通量。这三种芯片都能用于 Ion PGM™ 测序仪，为用户带来了灵活选择通量的灵活性，从 10 Mb 到 1 Gb。

简捷

Ion Torrent 使用半导体技术和简单的试剂对 DNA 进行测序，而并非使用光作为中介体。Ion

Torrent 将化学信号转换成数码信息，就像一块 CMOS 芯片将光子转换成像素。无修饰的核酸，无需酶化学级联，无需任何荧光和化学发光。仅适用经济的标准试剂。

快速

Ion PGM™ 测序仪拥有目前最快的新一代测序流程，从构建文库到数据分析，不到 8 小时。就测序运行本身来说，对于 100 bp 读取，全新的 Ion 316™ 芯片需要 2 小时运行时间，与 Ion 314™ 芯片相同。在测序短读取时，运行时间缩短至 1 小时内。单日流程加快了研究周期和文章发表，可实现更多的研究和市场应用。

基于此，在德国暴发大肠杆菌疫情时，多个研究机构都第一时间采用了 Ion PGM™ 测序仪。德国明斯特大学医院的 Dag Harmsen 博士认为：“从我这个公共卫生官员的角度来看，（PGM 的）最大优势在于其快速，而速度也是此刻最需要的。”

开山之作

也许有些人质疑 Ion PGM™ 测序仪是否真的如广告单页上宣传的那样好，但同行评议的文章应该会让他们信服。上个月，Life Technologies 的研究人员在《Nature》在线版上发表了[第一篇大作](#)，介绍他们如何利用半导体测序的策略对细菌基因组和人类基因组进行测序。

研究人员首先利用此平台对费氏弧菌和沼泽红假单胞菌进行了 de novo 基因组测序，覆盖度分别为 6.2 倍和 6.9 倍。之后，他们利用更大的离子芯片，对大肠杆菌的基因组进行了 3 次测序，覆盖度在 11.3-58.4 之间。

研究人员报道，对于每次读取的前 50 个碱基，每个碱基的准确率超过 99.6%，对于前 100 个碱基，大约在 98.9%。在包含均聚物的部分基因组，

准确率有所下降，但 5 碱基均聚物的准确率仍达到 97.3%。读长基本在 100 bp 左右，但有些也达到了 200 bp 甚至更长。

之后，他们对利用大约 1000 块离子芯片，对戈登·摩尔先生的核基因组进行了测序，平均覆盖度达 10.6 倍。研究小组发现了大约 260 万个 SNP。其中 99.95% 的杂合基因型和 99.97% 的纯合基因型通过 ABI SOLiD 产生的序列得到了验证。

应用

Ion PGM™ 系统实现了所有可能的测序应用。以创纪录的时间开展多重扩增子测序。此外，200 bp 的读取让长扩增子测序成为可能。在微生物测序项目中，Ion PGM™ 系统能带来最优的结果，极其均一的覆盖和长读长。316 和即将上市的 318 芯片则更适合转录组、小 RNA 测序以及 ChIP-Seq。

美国国立癌症研究院翻译基因组学实验室的主任 Stephen Chanock 也很看好这一平台。他认为该技术不仅便宜，测序准确度也很高。当然，就准确度而言，Ion PGM™ 测序仪还是比不上目前最大、最昂贵的测序仪。这些大家伙花费几十万美元就可以在一次测序反应里得到数百 Gb 的碱基序列，它们是目前进行高精度人类基因组测序工作的首选。而 PGM 测序仪则比较适合进行快速、小规模测序工作，比如对细菌的基因组进行测序或者对很多人基因组里的某一个基因区域进行测序，以帮助确诊疾病等。

据 Rothberg 介绍，Ion PGM™ 测序仪的读长很快将达到 400 bp，而更高通量的 318™ 芯片也将上市，也许，未来的天地更加广阔。（生物通 余亮）

[请为 Ion PGM 测序仪投上一票！](#)

我的样本，我的研究。MiSeq 个人测序系统

人类已经无法阻止 Illumina 了。最新的 HiSeq 2000 测序仪已经实现了每次运行产生 600 Gb 的数据量。据说，在 1-2 年内，通量还将进一步提高，每次运行将产生 1000 Gb 的数据。此外，在高通量测序市场上，Illumina 占据了近 7 成的市场份额，成为近几年生物科技领域的最大赢家。

从最初的 Genome Analyzer，到 Genome Analyzer IIx，再到最近的 HiSeq 1000 和 2000，测序仪的通量在不断攀升，发表文献数量也屡创新高。然而，并非每个实验室都需要如此高通量的测序仪。对于很多实验室来说，即使买了，也没有足够的样品让仪器满负荷运行。Illumina 当然也考虑到这个问题，于是在今年初推出了个人型测序平台——MiSeq 系统。



MiSeq 个人测序系统采用 Illumina 久经考验的 TruSeq™ 边合成边测序技术，但售价要低得多。它是唯一一台在单个仪器上整合了扩增、测序和数据分析的新一代测序仪，每次运行能产生超过 1 Gb 的数据，且占地面积仅为 2 平方英尺。MiSeq 有着革命性的流程和无可比拟的准确性，这让它成为快速且经济高效的遗传分析的理想平台，适合广泛的应用。

[点击索取MiSeq系统的详细资料](#)

久经考验的新一代测序试剂

MiSeq 系统特有全新的射流结构，能使试剂循环时间缩短 5 倍。以 Illumina 久经考验的边合成边测序技术为基础，通过专利的可逆终止试剂方法对数百万个片段进行大规模平行测序，在单个碱基掺入增长的 DNA 链时检测它们。当每个 dNTP 加入时对荧光标记的终止子成像，随后切割，允许下一个碱基的掺入。由于每个测序循环中四种可逆终止子结合的 dNTP 都存在，所以自然竞争让掺入偏差最小化。根据每个循环的信号强度测定直接检出碱基，与其他技术相比大大降低了原始错误率。最终结果是高度准确的 base-by-base 测序，避免了序列内容特异的误差，实现了可靠的碱基检出，即使是那些重复序列区和均聚物。由 TruSeq 技术推动的 Illumina 测序带来了最高的数据完整性，以及最高产量的无偏差读取和最多的>Q30 碱基检出。

革命性的测序流程

MiSeq 提供了最简单的新一代测序流程。以低至 50 ng 的起始 DNA 制备文库，时间不到 2 小时。通过直观的触摸屏界面开展简单的仪器操作，即插即用的试剂带有 RFID 追踪，具有自动化的便利。小巧、一体化的 MiSeq 平台整合了簇生成，末端配对测序和完整的数据分析，不再需要辅助性设备，节省了宝贵的实验室台面空间，也降低了初期投入的成本。

快速的周转时间

有了 MiSeq, 您能在几小时而不是几天内获得结果。两小时内制备您的测序文库, 然后放入 MiSeq, 开展 1 小时的克隆扩增和最快 3 小时的测序。最后, 在内置的仪器计算机上开展完整的数据分析, 从质量打分的碱基检出到变异体检出和比对, 不到 2 小时即可完成。

最广的应用范围

让 MiSeq 带您傲游广阔的测序天地吧。有了它, 您既可以开展快速、经济高效的毛细管电泳 (CE) 测序应用, 包括扩增子测序、克隆检查和小基因组测序; 也有机会实现高效的新一代测序应用, 如高度多重的 PCR 扩增子测序、定向重测序、ChIP-Seq、小 RNA 测序, 以及更多。对于目前的新一代测序用户, 可利用 MiSeq 系统以一小部分的时间和费用完成小型项目。Illumina 近期也发布了一些 MiSeq 的应用指南。

扩增子测序 1

研究人员从福尔马林固定、石蜡包埋 (FFPE) 的配对肿瘤/正常样本的基因组 DNA 中扩增得到 KRAS 癌基因外显子 2 区域的扩增子, 并利用 MiSeq 对其测序。在单个 MiSeq 运行中, 观察到每个扩增子的覆盖度超过 15 万倍, 足以检测即使是高度降解 DNA 中接近检测极限 (~0.5%) 的变异。

在本研究中, 他们对五个扩增子进行了多重分析, 并以极深的覆盖度测序。MiSeq 仪器上的测序运行能够生成最多 500 万个簇, 最多同时容纳 48 个样本的 48 个带索引的扩增子, 产生平均 2,500× 的覆盖度。本研究中的 KRAS 扩增子长度为 76 bp。MiSeq 平台能够产生重叠的末端配对读取, 以创建融合读取, 覆盖最长 300 bp 的扩增子。然而, 由于 FFPE DNA 的高度降解性质, Illumina 建议使用 200 bp 以下的 FFPE 扩增子。

研究人员认为, MiSeq 仪器上的测序结果是极其准确的, 因为过滤掉了预测准确率低于 Q30 的数据。对于此数据组, 原始读取准确率为 99.8%, 且平均背景噪音极低 (<0.25%)。扩增子测序的结果清晰表明了低丰度变异的可靠检测, 这是癌症突变研究的关键要求。

de novo 细菌基因组测序 2

研究人员还在 HiSeq 2000 和 MiSeq 平台上对从大肠杆菌参考菌株制备而来的同一个文库进行测序, 以及 de novo 装配。

利用 Illumina 的 TruSeq 样品制备试剂, 将基因组 DNA 制备成测序文库。对于 MiSeq 仪器上的测序, 将样品放在试剂槽内, 并与流动槽一起上样至仪器中。所有后续步骤在 MiSeq 仪器上自动进行, 包括簇生成和 2×150 bp 的末端配对测序。在 HiSeq 流动槽上, 利用 cBot 和 TruSeq v3 clustering 试剂将 K12 MG1655 库制备成簇, 接着在 HiSeq 仪器上开展 2×100 bp 的末端配对测序。

对于 HiSeq 和 MiSeq 测序仪上的数据比较, 利用 CASAVA 1.8a5 离线分析两组碱基检出文件, 并利用 Velvet 完成 de novo 装配。结果显示, 两个平台都产生了高质量数据 (85% 以上的碱基高于 Q30) 和均衡的 GC 覆盖。用这些数据进行 de novo 装配也产生了相似的结果, 极好地覆盖了参考基因组。

他们认为, MiSeq 系统上产生的测序结果很好地预测了高通量 HiSeq 2000 测序平台所带来的结果, 让 MiSeq 成为试行大规模研究或开展需要速度和准确性的独立实验的理想选择。(生物通 余亮)

MiSeq 系统的初步性能参数 3

读长	扩增和测序的总时间*	产量
1×35 bp	4 小时	>120 Mb
2×100 bp	19 小时	>680 Mb
2×150 bp	27 小时	>1 Gb
读取	>340 万个通过过滤的簇, 以及>680 万个末端配对读取	
性能	对于 1×35 bp 的运行, 90%以上的碱基高于 Q30 对于 2×100 bp 的运行, 80%以上的碱基高于 Q30 对于 2×150 bp 的运行, 75%以上的碱基高于 Q30	

* 包含末端配对读取, 如果适用。

注: 所有性能参数都是初步的, 有可能更改。

以 Illumina 官方网站公布的为准。

参考资料

1. <http://www.illumina.com.cn/systems/miseg/amplicon.asp>
2. <http://www.illumina.com.cn/systems/miseg/ecoli.asp>
3. <http://www.illumina.com.cn/systems/miseg/q.asp>

三款热门个人型测序仪比拼

2009 年，生物通抢先发布《新一代测序技术之三国时代》，聚焦了 Illumina、罗氏和 ABI 的新一代测序仪。如今两年过去了，测序市场上的竞争愈加激烈，第二代测序仪的升级版频出，而第三代测序仪也加入了激战。各厂商的你追我赶，使得测序通量不断攀升，而测序成本也一降再降。然而，除了一些大实验室和基因组中心，更多的实验室只能是笑看风云。无奈，门槛太高，而且也没有足够的样品让仪器满负荷运行。为了让更多的实验室能够有机会使用新一代测序技术，各大公司也争相推出了个人型测序仪。

生物通近期对市场上三款热门的个人型测序仪进行了介绍，详见：

- 罗氏 454 的 GS Junior: [《读长vs.通量 二代测序中哪个更重要？》](#)

- Life Technologies 的 Ion PGM: [《半导体测序仪 每个实验室都触手可及》](#)

- Illumina 的 MiSeq: [《我的样本，我的研究。MiSeq 个人测序系统》](#)

在此，生物通小编将对这三款测序仪的性能参数进行比较，主要是通量、读长、质量和费用。我想，这四个参数也是大家在购买测序仪前最先考虑的。文中的数据主要来自官方网站、官方发布的应用指南以及发表的文章，当然，数据仅供参考。

通量

在这一参数上，MiSeq 无疑具有最大优势。对于 2×150 bp 的末端配对运行，MiSeq 的通量已超过 1 Gb。GS Junior 目前每次运行可得到超过 35 Mb 高质量过滤后数据。Ion PGM 的 314 芯片有着 >10 Mb 的通量，而 316 芯片则提高了 10 倍，达到 100 Mb 以上。据 Life Technologies 介绍，第四季度上市的 318 芯片将带来 1 Gb 的通量。

读长

谈到读长，罗氏一定是二代测序仪制造商中笑得最开心的。GS Junior 的平均读长达到 400 bp。在通量上，GS Junior 虽稍逊一筹。然而，它的长读长为后续数据处理所带来的优势也是其他平台无法比拟的。以不久前的德国大肠杆菌事件为例，短读长平台多日未解的数据结果，GS Junior 数据 5 分钟内即给出了相应的生物学结论。

根据最近公布的数据，Ion PGM 测序仪的读长已经超过 100 bp。据 Ion Torrent 的创始人 Rothberg 介绍，读长很快将超过 200 bp，并达到 400 bp。如果真是那样，将可以与 GS Junior 相媲美。

表 1. 三个平台的读长和通量比较

仪器	运行时间	读长 (bp)	通量 (Mb)
Ion PGM – 314 芯片	2 小时	100	>10
Ion PGM – 316 芯片	2 小时	>100	>100
MiSeq	27 小时*	2×150	> 1000
GS Junior	10 小时	400	>35

* 27 小时为扩增和测序的总时间。

准确性

自 Ion PGM 测序仪推出以来的 6 个月内，它的准确性已提高了一个数量级。Ion Torrent 充分利用其简单测序试剂的固有准确性，实现了更多的信号，更好的信号处理以及改善的碱基检出。根据其

发表在《Nature》上的论文，利用 Ion PGM 测序仪对大肠杆菌测序时，对于每次读取的前 50 个碱基，每个碱基的准确率超过 99.6%，对于前 100 个碱基，大约在 98.9%。在包含均聚物的部分基因组，准确率有所下降，但 5 碱基均聚物的准确率仍达到 97.3%。¹

而同样是对大肠杆菌测序，MiSeq 平台也产生了高质量的数据，其中 85% 以上的碱基高于 Q30，且有着均匀的 GC 覆盖。² 此结果与 HiSeq 平台的结果相似，表明 MiSeq 可很好地预测了高通量 HiSeq 2000 测序平台所带来的结果。

费用

想必大家都对这三款仪器的价格很感兴趣。但仪器售卖时涉及的因素较多，可能还需要配套仪器、试剂和耗材，因此生物通很难以一个准确的数字来概括每台仪器的价格。就国外公布的价格来说，Ion PGM 是最低的，而 MiSeq 和 GS Junior 差不多。至于每次运行的费用，大家肯定也很想知道。下面是国外学者整理的单次运行的试剂费用，仅供大家参考。³

表 2. 三个平台的试剂费用比较（单位：美元）

仪器	试剂费用/运行*	试剂费用/Mb
Ion PGM – 314 芯片	500	< 50
Ion PGM – 316 芯片	750	< 7.5
MiSeq	750	0.74
GS Junior	1100	22

* 包括单个样品所有阶段的样品制备

应用

这三款个人型测序仪的价格比高端仪器要低得多，且有着快速的周转时间，因此适合快速、小规模测序工作，比如对细菌的基因组进行测序或者对很多人基因组里的某一个基因区域进行测序，以帮助确诊疾病等。

对于微生物等基因组的 de novo 测序，GS Junior 无疑是最合适的，MiSeq 和 Ion PGM 也适合，但装配可能比 454 更具挑战性。而对于定向位点的重测序，MiSeq 的性价比最高，Ion PGM 的数据则少于 MiSeq。

另外，个人型测序仪也在不断升级。Ion PGM 的 318 芯片即将上市，如果真如它所预计的那样，通量达到 1 Gb，读长达到 400 bp，那市场恐怕又是另一番格局。当然，GS Junior 和 MiSeq 也不会固步自封，也许不久之后的性能也会大大提升。如果您对个人型测序仪感兴趣，请关注生物通的最新报道。（生物通 余亮）

参考资料

- Jonathan MR et al. (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature. 475, 348–352.
- <http://www.illumina.com.cn/systems/miseq/ecoli.asp>
- Glenn, TC (2011) Field Guide to Next Generation DNA Sequencers. Molecular Ecology Resources. doi: 10.1111/j.1755-0998.2011.03024.x

面对越来越多的突发疫情，我们该如何应对？

今年 5 月源于德国的大肠杆菌疫情，目前该疫情已导致 35 人死亡，773 人出现严重并发症及致肾功能衰竭，确诊感染病例高达约 3255 起。深圳华大基因的研究团队，使用 Illumina 公司的新一代高通量测序系统完成了第一个真正意义上的该大肠杆菌基于 de novo 测序的组装结果，比早前发布的版本更加完整，并且修正了之前可能出现的组装错误。

Illumina 即将推出的 MiSeq 系统，非常适合此类突发疫情病原体的确定和监控。



- 一天内可以完成包括文库制备、SBS（边合成边测序）测序、数据分析的整体实验流程，第一时间确认病原体。

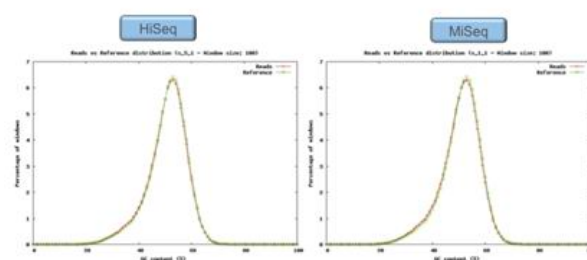
- 对于越来越复杂的细菌和病毒的变异情况，de novo 测序是确认病原体所必须的。MiSeq 一次运行可产生 1—1.5Gb 数据结果，一次运行的数据量即可完成 de novo 组装。

- 具有大量文献验证的 SBS（边合成边测序）技术原理，适合处理复杂的基因组测序，即使是连续的重复碱基的测序也不会出现任何问题。

- 使用双端对读测序，可产生 2 X 150 bp 测序读长，便于 de novo 组装。

- 操作便捷的台式仪器，从簇生成、测序反应到数据分析，都可以在一台仪器上完成，只需要 20 分钟左右的手动操作，适合任何实验室应用。

MiSeq vs HiSeq 2000 *E. coli* Data – 2x100bp
Even Coverage Across Range of GC Content



[联系我们，索取Illumina MiSeq系统更多的信息。](#)

Illumina 中国代表处：

北京

Tel: 010-64108530

Fax: 010-84539218

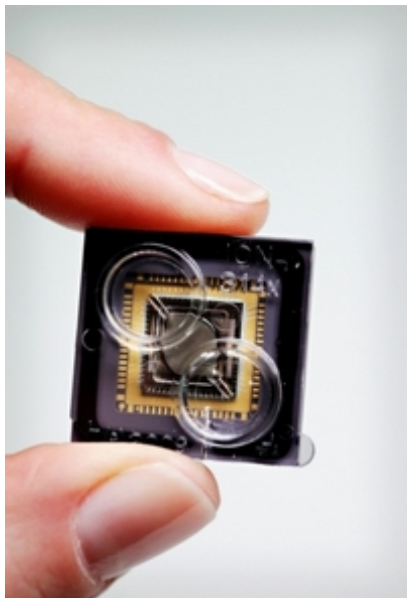
上海

Tel: 021-61134018

Fax: 021-61138320

个人化操作基因组测序仪快速发现德国毒黄瓜致病原因

2011年6月3日，美国加利福尼亚州卡尔斯巴德市——Life Technologies 公司今天宣布，使用 Ion Personal Genome Machine (PGM™)与德国明斯特大学医院合作进行的 DNA 测序得出的初步数据有力地表明，一种新混合型致病大肠杆菌菌株是导致德国致命疫情的原因。



从 DNA 测序仪获得的数据显示，典型的基因存在于两种不同类型的大肠杆菌中：肠聚集性大肠杆菌（EAEC）和肠出血性大肠杆菌（EHEC）。研究人员目前正通过 Ion PGM™对结果进行进一步的数据分析和确认。这些结果有助于了解这种细菌的侵略性并阻止疫情进一步爆发。

“快速全基因组测序结果使我们在几天内发现细菌毒力性状的独特组合……使德国经过此次疫情爆发克隆了一种来自不同大肠杆菌致病变型的独特混合类型，”德国明斯特大学医院卫生研究所德国国家溶血尿毒综合症（HUS）咨询实验室的科学家 Alexander Mellmann 博士说。

Life Technologies 公司位于德国达姆施塔特市的实验室收集的数据将被明斯特大学医院的科学家用于开发更好的测试手段，以积极地识别这种表现为肾衰竭和出血性腹泻的传染性疾病。这是自

疫情爆发以来人们第一次能够详细研究这种菌株致命性如此之强的原因。

Life Technologies 公司负责此次测序工作的 Simone Guenther 博士说：“此次疫情爆发的严重性意味着一定要速战速决。我们破纪录地迅速为明斯特大学医院提供了数据。在先前的疫情爆发中我们肯定需要比现在长得多的时间来达到这一阶段。”

为了防止细菌的进一步传播，Life Technologies 公司本周开始向欧洲的实验室运送其定制的大肠杆菌检测试剂盒，用于筛选疫情爆发中心被认为污染的食物。此次欧洲的疫情爆发已造成 17 人死亡，1000 多人感染。

这些试剂盒是第一道防线，用以检测致病性大肠杆菌是否存在。之后研究人员会使用更细化的试剂盒进行第二轮检测。未来几天内，一旦测序数据分析完毕，Life Technologies 公司便将开发用于检测德国这种混合型菌株的新定制试剂盒。公司可在一周内开发出定制的化验剂。

运往欧洲的试剂盒是大批 Life Technologies 公司研制的分子化验剂的一部分。这些产品可准确检测出绝大多数危害人类的病原体。Ion PGM™是一种采用半导体技术的 DNA 个体化操作基因组测序仪，它可以在短短两个小时内完成高通量和高精确度的 DNA 测序。用于此次测序的 DNA 样本采集自一位患有这种疾病的病人。

Ion Personal Genome Machine测序仪仅可用于研究用途。该仪器测序以快速著称，可以在两个小时内完成相对高通量的测序通量，并得到精准的测序结果。[点击索取Ion PGM测序仪的更多资料](#)

关于Life Technologies
(www.lifetechnologies.com)

Life Technologies公司是一家致力于改善人类生存环境的全球性生物技术公司。该公司的仪器、耗材和服务可协助研究人员加快推进科学和医学的发展，从而让人们的生活变得更加美好。该公司的客户遍及生物学各个领域，包括筛选与转化研究、分子药物、干细胞治疗、食品安全和动物保健以及 21 世纪的法医鉴定等。该公司生产供分子诊断和仅供研究使用的产品。Life Technologies的业界领先品牌，包括创新型Applied Biosystems和Ion Torrent品牌仪器以及Invitrogen、Gibco、Ambion、Molecular Probes和Taqman等被全球各地的生命科学实验室所广泛使用。该公司 2010 年的销售额为 36 亿美元，在全球 160 多个国家和地区拥有员工约 11,000 人，公司持有将近 3900 项知识产权专利及专有许可证，堪称生命科学界最大规模的知识产权资产。如蒙垂询，请访问：

<http://www.lifetechnologies.com>，或浏览Life Technologies的Twitter和Facebook网站：
<http://twitter.com/LIFECorporation>
(@LIFECorporation) 和
<http://www.facebook.com/LifeTechnologies>。

Life Technologies 的安全港声明

本新闻稿包含关于 Life Technologies 预期结果的前瞻性陈述，这些结果涉及多种风险和不确定因素。本新闻稿中的部分内容，包括但不限于行业趋势、Life Technologies 的业务计划、目标、预期和战略的陈述，均包含受各种风险和不确定因素影响的前瞻性陈述，这些风险和不确定因素可能导致实际结果和事件与前瞻性陈述中明示或暗示的有重大差异。任何不属于历史事实的陈述均为前瞻性陈述。可以通过一些特定用词，如“相信”、“计划”、“意图”、“预期”、“目标”、“估计”、“期望”，和/或将来时态、条件句结构（如“将”、“可能”、“可以”、“应当”等）或类似的表达方式，来识别一部分前瞻性陈述。可能引发实际结果与前瞻性陈述有重大差异的一些重要因素，在 Life Technologies 向美国证券交易委员会提交的文件中皆有详述。Life Technologies 不承担任何更新或修改此类前瞻性陈述以反映后续事件或情形的义务。

罗氏 GS Junior 助力致命大肠杆菌的深入分析

截止 16 日，致命的大肠杆菌 O104 菌株目前已经造成德国 38 人及瑞典 1 人死亡。虽说疫情高峰已过，但病菌的源头仍未查清，为欧洲的食品安全蒙上一层阴影。各国的科学家都在加紧研究，希望能更深入地了解这种神秘的大肠杆菌。

近日，英国健康保护局（Health Protection Agency）利用罗氏 454 的 GS Junior 测序系统，对细菌进行了“迄今为止最准确、最详细的遗传分析”，该数据已经上传至 NCBI 网站。

研究表明，该菌株有着独特的特征组合，包含来自两种不同类型大肠杆菌的基因以及它自己的基因。测序拼接表明，该菌株包含了两个染色体外元件或质粒，这些元件可能与它的致病性有关。

454 测序系统所产生的长读长让 HPA 的科学家能够在数小时内将几乎整个基因组拼接成一幅全面的基因组图谱。序列数据将协助后续分析，以鉴定病菌的来源，并了解这种菌株如何进化。

HPA 生物分析部门的主管 Saheer Gharbia 教授表示：“我们很高兴生成了全面的遗传代码数据。我们拼接完成后，立即与科学界共享，以便协助国际组织追踪这种新兴的感染。这项工作显示了 HPA 在监控爆发以及分析数据上的能力。”

454 生命科学的基因组学主管 Chinnappa Kodira 也谈到：“这种长的连续拼接能够更好地鉴定基因组结构、水平基因转移和致病岛，这些对了

解新型大肠杆菌菌株的出现很重要。来自 GS Junior 的长读长序列数据为科学界提供了迄今为止最连续的序列，可在此基础上开展后续分析。”

454 生命科学的总裁 Christopher McLeod 表示：“快速的周转时间以及高质量的长测序读长让 GS Junior 成为理想的研究平台，在新病原体爆发时快速全面地开展全基因组分析。”

[点击索取GS Junior测序系统的更多资料](#)

序列数据可在NCBI的网站上浏览，地址为 <http://www.ncbi.nlm.nih.gov/>。GeneBank ID 为 67929，accession number为afpn00000000。数据也可以在 HPA 的网站上查看，地址为：www.hpa-bioinformatics.org.uk/lgp/genomes。

至此，在人类与大肠杆菌的这场战役中，三大测序公司的测序仪都已登场，发挥着各自的优势。Ion PGM™是一种采用半导体技术的 DNA 个体化操作基因组测序仪，它可以在短短两个小时内完成高通量和高精确度的 DNA 测序。而 Illumina 公司 HiSeq 2000 测序平台助力了第一个真正意义上基于从头测序的组装结果的产生。

（生物通 余亮）

Life Tech 推出取代芯片的 RNA-Seq 方案

生物通报道，Life Technologies 公司近日宣布，它推出了一个适用于 Ion 个人化操作基因组测序仪（PGM）的 Ion Total RNA-Seq Kit。此试剂盒为转录本表达分析提供了一个快速、实惠且易用的方案，可取代芯片。

Ion PGM™ 是一种采用半导体技术的 DNA 个体化操作基因组测序仪，它可以在短短两个小时内完成高通量和高精确度的 DNA 测序。与其它新一代测序仪相比，它不需要激光、照相机或标记，价格当然要便宜很多。因此，它也被世界各地的研究人员所看好。

RNA-Seq 是研究基因表达的理想方法，因为它提供了最完整的信息，如转录本定量以及新颖转录本及突变的检测。然而，新一代测序的费用和复杂度却让很多研究人员望而生畏。

昆士兰医学基因组学中心的主管 Sean Grimmond 博士也这样认为：“尽管 RNA-Seq 提供了理想的灵敏度及分辨力，但费用和速度正在拖它的后腿。而这两个领域正是 Ion Torrent 擅长的，因此，这将是一个非常引人注目的应用。”

Ambion 开发的 Ion Total RNA-Seq Kit 能够分析任何类型的 RNA，包括去除了 rRNA 的总 RNA、poly(A) RNA 及小 RNA。与其他试剂盒不同，它保留了链信息，将所有读取与染色体链比对，提供了关于基因结构的关键信息。

Ion Torrent 为开展 RNA-Seq 研究提供了一个完整的解决方案，从快速的文库制备和简单的测序，再到直观的数据分析工具。

Ion Total RNA-Seq Kit 有着灵活的流程，可生成全转录组或小 RNA 文库。从 200ng 总 RNA 或 5ng miRNA 开始，文库构建低至 6 小时。Ion Total RNA-Seq Kit 利用专利的 Ambion 技术以定向的方式加上接头，保留了链信息。此外，3'和 5'接头同时加上，减少了连接和纯化步骤。文库经克隆扩增后，产生测序即用的模板。

Ion OneTouch™ 自动化流程能够在 3 小时内制备出测序模板，而手工操作只需几分钟。测序可在 1 小时内进行。序列读取由 Torrent Server 处理，并导出一个简单的文件，交由常用的基因组数据分析软件来分析。Partek Genomics Suite 提供了一个直观的用户界面，可对 RNA-Seq 数据进行全面分析。

Life Technologies 将在下半年推出 Ion 318 芯片，它可将 PGM 个人化操作基因组测序仪每次运行的测序通量提高至 1 GB，并可实现包括 RNA 在内的更多测序应用。Ion Torrent 的创始人 Rothberg 预计年底前读长也会达到 400 bp。 [点击索取 Ion PGM 测序仪的更多资料](#)

（生物通 余亮）

罗氏 GS Junior 测序系统与 MagNA Pure Compact 核酸纯化系统联手发现新生物

罗氏的 GS Junior 测序仪通常安装在实验室中用来进行高品质的医疗研究，但是现在它的迷你体积，使得野外实验成为可能！在六月初，一台罗氏 GS Junior 测序仪和罗氏 MagNA Pure Compact 核酸纯化系统一起被安装在在一艘名叫 **Heincke** 的研究船上，并开始了一段不同寻常的旅程。不畏远洋或者 8 到 10 级大风的考验，他们稳定的表现实现了在研究船上工作人员多年来的梦想。



这艘来自 Alfred Wegner 极地海洋研究所（隶属德国不来梅 Helmholtz 协会）的研究船被称作海上移动研究站。海洋生物学家，气候研究和分子生物学家挤在这个极其狭小的空间内共同研究海洋中的那些最小的动物。在这个寸土寸金的实验室里，迷你体积的 GS Junior 测序仪显示了它巨大的优势。“它的小巧体积和易用的数据拼接与分析软件，使得我们可以在船上进行实验并第一时间获得了许多重要的数据。有许多实验等回到岸上几乎是不可能进行。”分子生态学家 Uwe 这样说到。

Alfred Wegner 极地海洋研究所的科学家在世界各地进行极地和海洋的研究调查，分析物种组

成，从而更好地认识环境变化对海洋生态系统以及全球气候变化的影响。除此之外，他们的工作还包括研究那些用传统方法难以检测到的海洋微生物，通过分子标记了解生物多样性，从而认识自然进化过程。John 解释说，“我们试图从各种各样的环境条件下发现各种不同生物。即使是最极端的情况下，比如在 2000 年历史的冰川中我们也找到了仍然存活的生物，获得了许多有价值的结论。”通过现代测序手段，可以在一个给定的环境中确定存活哪些生物，进而研究他们存在的原因。

在本次为期三周的北海海域苏格兰及爱尔兰的沿岸的航行中，GS Junior 测序仪和 MagNA Pure Compact 核酸纯化系统主要用于研究有毒微藻的多样性和生命活动，他们的处女航行已经获得了一个令人振奋的结果：罗氏公司的这些新技术此次至少发现了四个新物种！对于还是学生的 Jan Meyer 来说，本次航行实现了“每一个科学家希望发现新的生命的梦想”。

测序数据分析之专家指南（上）

测序确实是越来越快，也越来越便宜了。随着个人型测序仪的不断上市，许多实验室也跃跃欲试，准备开展这方面的研究。然而，前辈告诉我们，测序并不难，真正困难的工作是数据分析。

目前有不少用于基因组装配和比对的程序和算法，但是该选哪一个呢？许多序列分析的专家认为，这取决于基因组的大小、读取有多长，以及采用的是哪种测序技术。通常，软件还需要优化，以满足每个实验室的独特需求。

为了让大家更好地开展数据分析，《Genome Technology》杂志特邀了一些这方面的专家，与大家分享他们在数据分析方面的经验。通过他们的一问一答，希望您也能从中受益。

Q1: 您使用哪个基因组装配或比对软件，为什么？

Inna Dubchak（美国能源部联合基因组研究所）

我们设计了一个名为 VISTA 的计算系统，它融合了长基因组序列的不同比对算法。不同项目需要不同的算法，如 AVID 适合精确序列或序列草图的整体两两比对，LAGAN 适合精确序列的整体两两比对或多个比对，而 Shuffle-LAGAN 适合精确序列的 glocal 两两比对，因为它检测重排。

Jim Kent（加州大学圣克鲁兹分校）

我们自己比较少做基因组装配。我们主要是从装配好的基因组开始的。关于比对，我们使用 BLAT 进行一个物种内的 RNA/DNA 比对及其他比对，使用 blastz 和 lastz 进行物种之间的两两比对，并用 axtChain 和 chain-Net 进一步处理。关于多个比对，我们正在用 multiz，但也在试一些新软件。

Ian Korf（加州大学戴维斯分校）

因为我们运行 Assemblathon 竞争，所以我们运行多个装配和比对程序。每一个都有自己的优势和劣势。事实上，我们也不知道哪个更好。有时候有的程序更方便。

Li-jun Ma（马萨诸塞大学）

我所做的大部分基因组装配都是利用 Arachne 或 AllPath 完成的，它们是由 Broad 开发的。选择这些装配工具的主要原因是，它们是由 Broad 研究院的软件工程师经过多年艰苦工作而开发的。它们是可靠的工具。它们已经通过很多基因组的多个数据组形式检验过，这些基因组大小不同，性质各异。它们很复杂，但非常可靠。现在，我们也使用 Velvet 来装配 Illumina 的数据组。它使用简单，运行快。

Bud Mishra（纽约大学）

我们是一个生物信息学小组，致力于开发新的装配、碱基检出和比对工具。我们的目标是让这些工具以一种与技术无关的方式共同工作，得到准确的结果，以便更好地进行全基因组关联研究（GWAS）。基于这些目的，我们主要关注一种碱基检出工具 TotalReCaller 和一种装配工具 SUTTA。为了比较，我们也与其他序列装配和比对工具共同使用：SOAPdenovo、Abyss、CABOG、Velvet、TIGR、Arachne 等开放工具以及 Illumina 开发的 Bustrad、Ibis、Rolexa、BayesCall 等。根据我们的经验，对于高覆盖度的 Illumina 序列，可从 Illumina 的 Bustrad 碱基检出开始，然后用 SOAPdenovo 或 Allpaths 来装配读取，之后可能还要确认。今后，一种更加一体化的方法将把

Illumina 的序列和光学图谱与 TotalReCaller 和 SUTTA 结合起来，得到基因型或单体型序列。

Mihai Pop (马里兰大学)

这个问题可没什么标准答案。它取决于实际的应用。对于短读取的快速严格比对，我使用 Bowtie。对于没那么严格的长读取，我使用 MUMmer。这个选择主要是基于习惯和对这些工具的熟悉程度。对于装配，我也依靠多个工具，同样取决于实际应用。对于 Sanger 或 454 读取，我使用 Celera Assembler，而 Newbler 对于 454 数据也非常好，而短序列我会使用 SOAPdenovo 或 Velvet。我通常对那些 unitigging 使用这些装配工具，生成相对保守的 contig。然后，我使用自己的工具 Bambus，掺入 mate-pair 信息或其他相关信息。我使用 Bambus 是因为我能更好控制装配工具实际产生的。大部分现代装配工具只生成一个 FASTA 文件，抛弃了读取在哪里比对的信息。在很多应用（如宏基因组学）中，这个信息才是真正有用的。

Steven Salzberg (约翰霍普金斯大学医学院)

我们使用 6 个主要的基因组装配工具包：CABOG（之前称为 Celera Assembler）、SOAPdenovo、the AMOS package、Allpaths-LG、Velvet，以及我们最近开始用的 SGA。对于比对，我只能回答新一代 DNA 序列读取与参考基因组的比对，我使用 Bowtie（我们小组开发的）。如果读取代表了 RNA-Seq 实验中的 RNA，我们使用 TopHat 来比对，并用 Cufflinks 来装配和定量转录本。

Robert Settlege (弗吉尼亚生物信息学研究所)

我们都使用了多个工具。装配一般使用 Velvet、Newbler、Abyss、SSAKE、CLCBio，而比对也同样使用了多个工具。我们特别喜欢使用 Velvet，它使装配明显加速。CLCBio 也很好，是少数几个能够在单次装配中使用短读取和长读取

的装配工具之一。我们不依赖单个工具的理由是，目前没有一个工具是完美的。

Q2: 您采用哪种方法进行多个序列比对?

Inna Dubchak (美国能源部联合基因组研究所)

我们对长达 2 Mb 的基因组序列使用 LAGAN，对完整的基因组装配使用 Shuffle-LAGAN 算法的全基因组延伸。

Ian Korf (加州大学戴维斯分校)

我们一般似乎用 T-COFFEE，有时也用 DIALIGN 或 CLUSTALW。

Bud Mishra (纽约大学)

我们的工具是为单体型和宏基因组学设计的，但仍在开发中。

Mihai Pop (马里兰大学)

我的研究中并不常使用多序列比对工具。我发现 CLUSTALW 是一个很好的程序，它灵活、高效且准确，尤其对于 DNA 比对。在装配中，Celera Assembler、Newbler、AMOS 及一些旧的装配算法产生了共装配序列的严格比对。对于包含非常相似序列的大数据组，如 16S 研究，传统的多比对方法不太好，我使用 DNAClust 和 Uclust 等工具来产生序列的“星号”比对。

Steven Salzberg (约翰霍普金斯大学医学院)

对于快速的多序列比对，例如不同物种的 20 个蛋白比对，我个人使用 MUSCLE 程序，我已经用了很多年。如果问题是全基因组多序列比对，如 10 个完整的细菌染色体相互比对，我会使用我之前的博士生 Sam Anguili 开发的 MUGSY 程序。这是个相当好的程序，也是开源的。

下篇请看：

<http://www.ebiotrade.com/newsf/2011-9/2011916174731331.htm>。

测序数据分析之专家指南（下）

测序确实是越来越快，也越来越便宜了。随着个人型测序仪的不断上市，许多实验室也跃跃欲试，准备开展这方面的研究。然而，前辈告诉我们，测序并不难，真正困难的工作是数据分析。

目前有不少用于基因组装配和比对的程序和算法，但是该选哪一个呢？许多序列分析的专家认为，这取决于基因组的大小、读取有多长，以及采用的是哪种测序技术。通常，软件还需要优化，以满足每个实验室的独特需求。

为了让大家更好地开展数据分析，《Genome Technology》杂志特邀了一些这方面的专家，与大家分享他们在数据分析方面的经验。通过他们的一问一答，希望您也能从中受益。

Q1: 您使用哪个基因组装配或比对软件，为什么？

Q2: 您采用哪种方法进行多个序列比对？

详见 [《测序数据分析之专家指南（上）》](#)

Q3: 您如何优化原始数据，以便获得最佳的装配或比对结果？

Inna Dubchak（美国能源部联合基因组研究所）

我们的比对方法最适合装配好的数据。原始数据最好由局部比对程序来处理。

Li-jun Ma（马萨诸塞大学）

质量过滤是关键的一步。我们总是过滤原始读取，除去低质量的读取，修剪接头序列，并除去修剪后非常短的读取。

Bud Mishra（纽约大学）

TotalReCaller 使用原始强度测序数据和参考序列来改善碱基检出，并优化比对结果。既然它使用参考序列，那么似乎不适合 *de novo* 序列装配；然而，在近期 Giuseppe Narzisi 的博士论文中，作者表明通过 *boot-strap* 方法，TotalReCaller 和 SUTTA 联合可显著改善装配质量。

Mihai Pop（马里兰大学）

我主要依靠错误修剪工具，如 *fastx toolkit*。有时我也使用错误纠正工具，但我担心在某些情况下，这些工具可能引入错误。我个人倾向于抛弃可疑的序列，即便它们占据了相当的比例，而不是试图纠正错误。例如，在 16S 研究中，我抛弃那些有一个含糊代码或者太短的序列，通常会抛弃 25-30% 的数据。测序成本正变得足够低，且通量足够高，我们可以承受这些浪费。

Steven Salzberg（约翰霍普金斯大学医学院）

我们经常花费大量精力来修剪载体和低质量序列，这取决于任务。对于全基因组测序项目，我们运行错误纠正软件（如 *Quake*）来修复错误的碱基检出。一些基因组装配工具干得很好，自己能除去低质量的数据或纠正错误，但另一些不行。如果读取是配对的，且片段足够短，配对读取能够重叠，那么我们运行另一个程序，在装配之前将这些配对片段融合成更长的序列。

Robert Settlege（弗吉尼亚生物信息学研究所）

我们的首选方法是猛烈的修剪。如果它看似个接头，去掉。如果它质量有疑问，去掉。通常我们有足够的读取，因此最好猛烈一点。我们之后常将数据补回，看它是否分辨了一些模棱两可。

Q4: 确认装配或比对准确性的最佳方法是什么?

Inna Dubchak (美国能源部联合基因组研究所)

这是个很难的问题。通常我们使用基因组覆盖度统计数字，并与其他确立的比对方法比较。

Jim Kent (加州大学圣克鲁兹分校)

对于装配：与已知参考基因组比较（如果有的话），检查 mRNA/基因组比对，或检查配对读取相对基因组的比对。

Ian Korf (加州大学戴维斯分校)

这是个很难的问题。我们通常不知道正确的答案。对基因组的一部分测序可能非常有用。另一个有用的方法是寻找如高度保守的基因或长转录本。对于序列比对，它取决于你所作的搜索类型。主要有两种搜索，我称之为定位（mapping）和探索（exploring）。在定位序列时，一条序列与另一条序列是相同，或几乎相同的。例如，开展 ChIP-seq 分析，你需要将读取定位回参考序列。你希望比对是相同的，但如果它们有一些错误或多态性也能接受。如果有一些完美比对，你认为比对是准确的。在探索远亲关系时，比如寻找蛋白的同源物，错配和缺口是意料之中的。如果你的序列是已知蛋白家

族的一部分，你的确认策略应当包括，比对与家族的其他成员相符合，也就是说，该家族的保守部分在两两比对中也同样保守。

Li-jun Ma (马萨诸塞大学)

确认装配准确性的方法包括：1) 将装配定位到染色体或连锁图上；2) 将装配与任何已知序列比较，如 PCR 产物、基因、粘粒、BAC 或质粒的序列；3) 如有必要，PCR 扩增你有疑问的基因组区域，以确认装配的准确性。

Bud Mishra (纽约大学)

由于 SUTTA 是为自我确认而设计的，它在装配过程中不断验证。我们也开发了一种新的度量办法，称为 Feature-Response Curve，它能捕获 contig 覆盖之间的交换，以及不同的准确性特征。最近，我们还设计了新的统计学分析工具，能更好地了解各个传统特征之间的关系，并捕获这些特征的核心结构。

Steven Salzberg (约翰霍普金斯大学医学院)

准确性有很多内部和外部的测定。如果使用的話，外部测定非常有用，我指的是与真正的基因组比较。有时这不可能，但对于已知物种如人的重装配，我们还是能够检查。内部测定包括 mate-pair 距离、配对读取的方向，和覆盖深度。我们检查这些参数。我的同事 Adam Phillippy、Mike Schatz，和 Mihai Pop 在 2008 年开发了一个“assembly QC”的工具包，它包含了一些非常有用的软件工具，来检查装配的有效性。（生物通 余亮编译）

测序数据分析之工具大全

测序确实是越来越快，也越来越便宜了。随着个人型测序仪的不断上市，许多实验室也跃跃欲试，准备开展这方面的研究。然而，前辈告诉我们，测序并不难，真正困难的工作是数据分析。

目前有不少用于基因组装配和比对的程序和算法，但是该选哪一个呢？许多序列分析的专家认为，这取决于基因组的大小、读取有多长，以及采用的是哪种测序技术。通常，软件还需要优化，以满足每个实验室的独特需求。

在《[测序数据分析之专家指南](#)》系列文章中，多位专家推荐了他们心水的软件工具。在此，我们整理了这些数据分析工具的信息，供大家参考。

Abyss 1.3.0:

<http://www.bcgsc.ca/platform/bioinfo/software/abyss/releases/1.3.0>

Arachne & AllPath:

<https://www.broadinstitute.org/scientific-community/science/programs/genome-sequencing-and-analysis/computational-rd/computational->

BLAT:

<http://genome.ucsc.edu/cgi-bin/hgBlat?command=start>

Bowtie:

<http://bowtie-bio.sourceforge.net/index.shtml>

CABOG:

http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Main_Page

CLUSTALW:

<http://www.ebi.ac.uk/Tools/msa/clustalw2>

LAGAN & Shuffle-LAGAN:

http://lagan.stanford.edu/lagan_web/index.shtml

MUGSY:

<http://mugsy.sourceforge.net>

MUMmer: <http://mummer.sourceforge.net>

SOAPdenovo:

<http://soap.genomics.org.cn/soapdenovo.html>

TotalReCaller:

<http://bioinformatics.nyu.edu/wordpress/projects/totalrecaller>

Velvet: <http://www.ebi.ac.uk/~zerbino/velvet>

VISTA tools, 包括 AVID:

<http://pipeline.lbl.gov/run5details.shtml>

ABI 数据分析软件:

<http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/ngs-data-analysis-software.html>

Illumina 的软件:

<http://www.illumina.com/software.ilmn>

ION PGM 测序仪的社区:

<http://www.iontorrent.com/forum/>

另外附上一些精选的数据分析工具：

名称	链接	评论
基因组重测序		
Bwa	http://bio-bwa.sourceforge.net	比对工具
Dindel	http://sites.google.com/site/keesalbers/soft/dindel	小的插入/缺失发现
Erds	http://www.duke.edu/~mz34/erds.htm	拷贝数变异发现
Pindel	http://www.ebi.ac.uk/~kyle/pindel/	小的插入/缺失发现
Samtools	http://samtools.sourceforge.net	操控比对后数据的工具
Sequence Variant Analyzer	http://www.svapproject.org	在基因组背景下显示变异
Chip-Seq		
Findpeaks	http://vancouvershorttr.sourceforge.net	
RNA-Seq		
Cufflinks	http://cufflinks.cbc.umd.edu	测定转录本丰度
Tophat	http://tophat.cbc.umd.edu	剪接点定位
De Novo 拼接		
Oases	http://www.ebi.ac.uk/~zerbino/oases/	根据转录组数据拼接
基因组浏览器		
Integrated Genome Browser	http://www.bioviz.org/igb/	
Integrative Genomics Viewer	http://www.broadinstitute.org/software/igv/	