

IV 功能性基因组学的序列普查方法

现在的新一代测序技术使全基因组测序变得容易。除此之外，新的应用和新的测定法则已经出现，大大提高了我们了解基因功能的能力。

正如爱迪生没想到发明的电唱机应用于玩音乐，许多发生在基因组学的事件与此相类似。新一代大规模高通量DNA测序平台已经出现。它们的作用是取代繁杂的毛细管测序系统，为我们带来高质量的人类基因组序列。这些仪器，主要出自454和illumina（前身solexa）公司以及最近加盟的ABI公司^[1]，并于最近两年里出现在DNA测序中心，这预示将比标准毛细管为基础的测序技术生产更多的序列（一次运转可测1Gb）。还有一些新的仪器正处于研发阶段。美国国立卫生研究院（National Institutes of Health）和国家人类基因组研究所（National Human Genome Research Institute, NHGRI）号召把个性化基因分析费用从1000万美元降到大约1000美元，且许诺“如果你们研发新技术，我们会购买”，以及许多如微生物环境基因组学等研究的基因序列需求，大大刺激了这些生物公司发展新技术。正如所希望的，世界基因组中心正致力于研发高效和便宜的测序仪。阅读长度的限制，错误率和组合算法问题，包括其他方面的问题，至少在目前还没有完全阻碍全基因组测序的进行。

正如音乐是爱迪生发明电唱机计划外的功能，在研发费用低的基因组测序技术的实验中，一个新“序列普查”计数法出现了。如果你需要测量一个RNA或DNA‘组’，微阵列技术已不再是唯一的手段。目前采用大规模高通量高产量的DNA序列分析系统，测量mRNA、小RNAs、转录因子多聚体、染色体结构，DNA核酸酶酶切反应及DNA甲基化状态的基因组解析的新方法，正在快速的成熟。

这些全球性“序列普查”方法原理十分简单：要了解复杂的核酸，就测序。直接测序——在无细菌克隆的条件下——获得足够的序列，然后给每段基因指定起始位点，而不是试图确定其全序列。一个短序列（或有时是一对，即一端一个）能确定一个生物样本中的数百万的核酸分子。你无须从分子的起始位置排序，因为一些25-35bp的小片段提示你可用信息学方法分析每个片段在参照基因组中的位置。定位后，计算这些小片段的个数和分析其在整个基因组中的分布。使用这些方法的样本必须含有非常多的独特片段，每个片段对应着样本中的一个不同的分子。在新检测方法费用降低，能提供更丰富的信息后，这些概念在SAGE和大规模平行测序技术（massively parallel signature sequencing, MPSS）^[2,3,4]等早期方法理论的基础

上逐步形成。

这些新一代“基于序列（Seq-based）”方法自然地适合功能性基因组学的应用，因为它们迅速而且以较低的成本产生了大量的短序列片段，并严格的遵守“降低成本”宗旨。也就是说，他们并不需要测序整个基因组，而只是其中的小部分，例如mRNA、甲基化或未甲基化片段、特殊蛋白质结合的DNA或RNA以及核酸酶作用的DNA区域。该系统获取序列数据来源于单个DNA扩增片段，而不是质粒克隆。降低成本是一个长期追求的目标——目前似乎已有一些成本较低的测序仪器出现——现在追求的是以与微点阵杂交（microarray hybridization）一样或者比其更低的成本尽量测得一种核酸的所有序列。“Seq-based”方法，绕过了一些微阵列长期存在的技术问题，其中包括以百万计的DNA探针合成微阵列，但其比例却与大型基因组规模不相符；大量的交叉杂交；持续性固有杂交信号造成的定量干扰。然而，这些新的方法并不能解决一切。举例来说，应用“Seq-based”方法比当代微点阵杂交能获得更多的基因组序列，但仍有15-20%的人类基因组内容是不能定位的，因为它们在基因组中不止出现一次。

序列普查方法最古老的原理是染色质免疫沉

淀，或者芯片，在最近几个月里曾发表了几篇文章介绍ChIP-Seq研究^[5,6,7,8]（图1）。这些实验是为了查找体内由重要的DNA结合蛋白质占据的DNA位点。要做到这一点，需用识别DNA结合蛋白的特异性抗体免疫沉淀活细胞的目的蛋白质。转录过程中有许多蛋白质发挥着重要作用，包括RNA聚合酶和一些辅助因子，特定的修饰组蛋白，序列特异性转录因子和其它参与DNA复制与修复蛋白质。

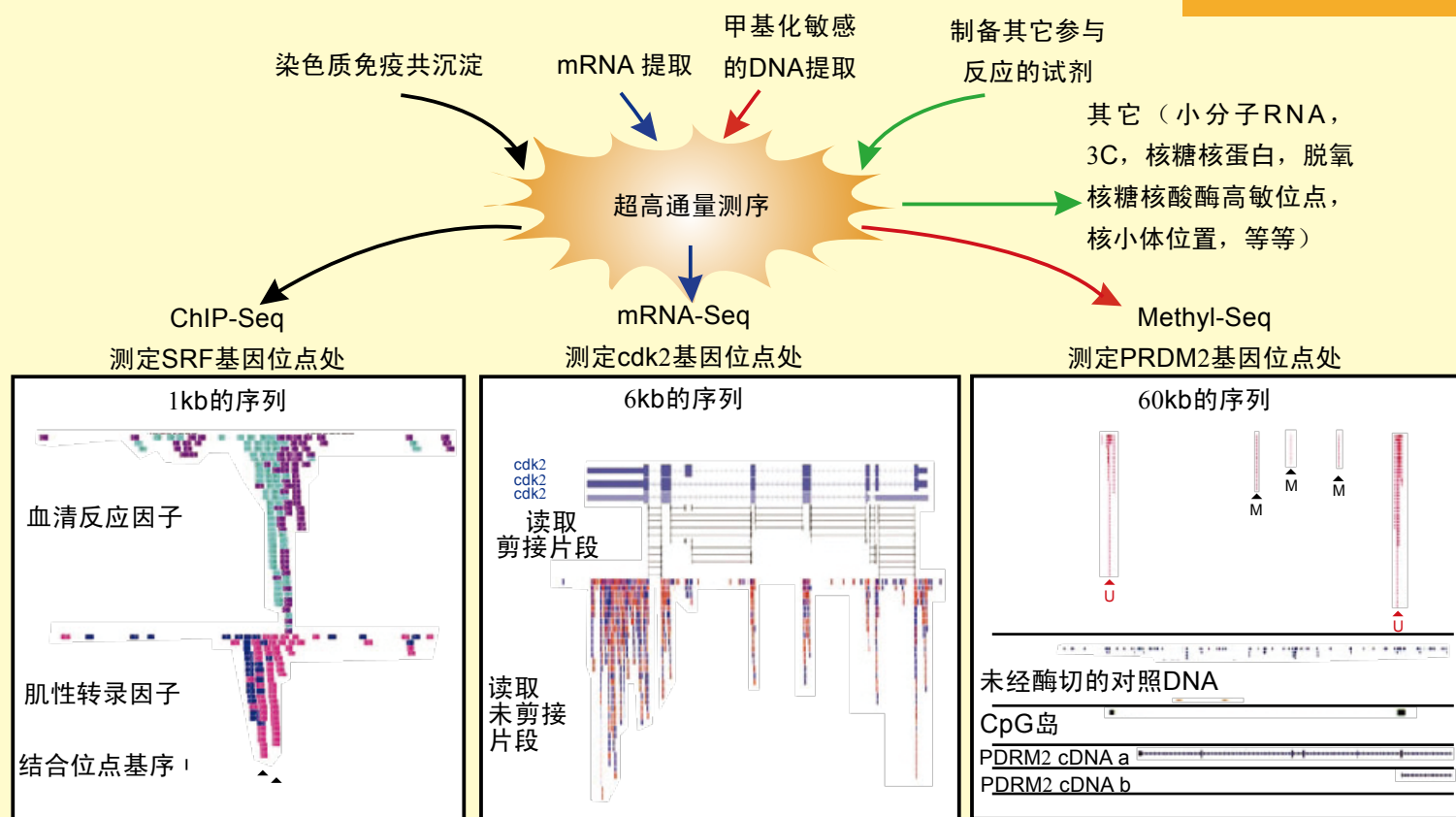


图1 序列普查试验 每个彩色矩形对应一个illumina (solexa) 所读的序列，且参照UCSC上的序列顺序将每个序列映射到直线中相应位置。

ChIP-Seq (左)，以c2c12细胞的血清反应因子 (serum response factor, SRF) 和肌性转录因子 (myogenic transcription factor, MyoD) (B. Williams, G. Kwan, A. Mortazavi, S. Sharp and B.W; 该成果未发表) 为例，蓝色和粉色 (MyoD)，绿色和紫色 (SRF) 指示着序列的方向，三角形指示着各个转录因子的结合位点基序。ChIP-Seq中，矩形集中分布在结合位点基序，颜色显示结合位点的方向性。

mRNA-Seq转录图谱 (中)，以肌原性细胞为例，由多聚 (A) +完成的 (A.Mortazavi, B.Williams and B.W.;该成果未发表)，RNA片段的间隙显示为黑色，随机用红色和蓝色显示序列方向性。两个外显子与交替的剪接亚型符合。Refseq上的所有剪接片段和其余的基因模型已被发现。

Methyl-Seq图 (右)，显示的是甲基化以及未甲基化片段被限制性内切酶HpaII或者MspI酶切的位点 (D. Johnson and R.M.M.; 该成果未发表)。

直到最近，正在研究的ChIPed DNA的一部分或者大部分片段的分析结果几乎都是通过微阵列杂交方式得到的 (即所谓ChIP-chip)。虽然这种方式已经产生了许多宝贵的数据，但是ChIP-chip具有上述的微阵列相关的局限性。然而，ChIP-Seq解决了其中最大的问题，因为它没有微点阵的设计与构造，所以就可以为任何生物测序。即使已经检测了测序过程中可能存在的序列误差，但由于ChIP-Seq不依赖实验者选择的DNA探针，其结果是无法预测的。ChIP-Seq没有受到交叉杂交信号所致的错误以及模糊信号的干扰，定量可能更为准确，因为计数序列是“数字化的”，而不是连续的。最后，与ChIP-chip 相比，ChIP-Seq能以更高的分辨

率定位结合位点。按大小来进行选择的步骤和计算法则（图1）通常决定了结合位点集中在低于40bp的区域。

RNA-Seq是超高通量测序技术的另一个应用，目前正被多个实验室开发和测试，而且可能比ChIP-Seq有更广泛的用途。由于454平台致力于加快发现小RNAs的速度，RNA-Seq等其它一些测序仪从而被开发出来^[9,10]。十多年来，研究mRNA一直都采用微阵列技术。我们知道的转录子系统，从酵母和细菌到小鼠、人类和芥菜种子，其中大部分RNA序列都来源于微阵列技术。然而，RNA-Seq可以提供更多信息。举例来说，即使用更密集排列阵列也很少得到RNA剪接序列，但使用RNA-Seq就可以解决这点（图1）。此外，RNA-Seq可通过高采样，从低丰度的样本（或罕见的样本亚细胞株）检测到RNAs，且准确率很高。我们已经发现RNA-Seq存在许多缺陷，而且还有一些问题仍未发现。举例来说，标测大规模的转录基因给各种剪接分类配对，需要尾部牢固结合的配对标签和/或更长的标签。所有RNA-Seq数据，面临着一系列新的信息学方面的难题，其中包括如何最好地诠释和标记多个形似的基因位点。某一转录子图谱的质量，依赖于其RNA是否常见，因为罕见的转录子没有足够的标签来标测其剪接，甚至不能验证存在的特异外显子，除非RNA样本经过事先标准化。

虽然没有公布结果，但有几个科研小组已开始使用超高通量测序仪来帮助测定人类基因组中CpG岛的甲基化状态（图1）。他们先用硫酸氢钠处理CpG岛样品，接着对样品碎片测序，这样大大减少了样品

量。虽然没有得到人类基因组中所有CpG岛位点，但是相对于目前最有效检测CpG的试验来说，他们能够用一个更简单的实验得到更多的数据。类似做法正被用来检测DNA酶-酶切位点，这些位点一般存在于活细胞基因组中的规律性区域。

为了避免被当前的成绩冲昏头脑，认识到Seq-based方法的局限性以及超高通量测序仪的应用还有待开发的事实尤为重要。即使有高价值的用途，这个新方法也需要稳定支持生产的协议，目前一项支持该方法的国家卫生研究院的编码计划正在起草和完善。虽然在排序的公式和检测阳性信号的灵敏度方面取得了进步，但这两项仍有待提高。另外，仪器本身是完全不能防弹的，而且有价值的标签生产量也低于人们的期望值，Illumina（Solexa）和ABI的（SOLiD）系统还只是处于生产成对端标签的初步阶段，而该检签对于检测RNA的剪接异构体非常重要。

虽然有许多不确定性因素，但是这个令人振奋的前沿研究正在开展，以后生物学就建立在研究细胞株的种类和状态的基因规律基础上。破译基因组不是我们唯一的目的，尽管这是非常有益的一步。我们更大的目的是要了解基因如何决定细胞的类型和它们的功能。在正常和实验或基因多样化条件下，能定期和全面地分析基因的投入产出，有可能改变我们了解基因组合的规律。在纯粹的技术层面上，利用Seq-based方法测定许多不同种类的功能性基因组将使整合和合成基因电路和网络变得容易。

原文检索：<http://www.nature.com/nmeth/journal/v5/n1/full/nmeth1157.html>

参考文献

- [1] Shendure, J. et al. Science 309, 1728–1732 (2005) .
- [2] Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Science 270, 484–487 (1995) .
- [3] Brenner, S. et al. Nat. Biotechnol. 18, 630–634 (2000) .
- [4] Margulies, M. et al. Nature 437, 376–380 (2005) .
- [5] Barski, A. et al. Cell 129, 823–837 (2007) .
- [6] Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Science 316, 1497–1502 (2007) .
- [7] Mikkelsen, T.S. et al. Nature 448, 553–560 (2007) .
- [8] Robertson, G. et al. Nat. Methods 4, 651–657 (2007) .
- [9] Lu, C. et al. Genome Res. 16, 1276–1288 (2006) .
- [10] Ruby, J.G. et al. Cell 127, 1193–1207 (2006) .



小会 编译