

## 五、更多阅读

### 1. 核糖体印记与深度测序技术

将核糖体图谱（ribosome profiling）和深度测序（deep sequencing）相结合，研究人员可以从基因组水平监测蛋白质的翻译状况。

深度测序的强大功能对生物学的各个领域都产生了极大的影响。在诸如全基因组测序等方面，新技术的高效性和经济性使人们得以以一种以前无法想象的方式进行试验研究。而在另一些情况下，例如RNA测序时，借助深度测序可以进行更多的定量分析，获得更大的动态范围。在另一些研究中，例如最近由美国加州大学（University of California）的Jonathan Weissman小组发表的有关翻译图谱（translational profiling）的研究中报道的那样，深度测序不仅是一个有效的定量手段，同时还能提供很多有用的新信息。

使用核酸酶消化mRNA时，在翻译过程中发挥作用的核糖体结合并保护了大约30bp的mRNA片段。Weissman等人将细胞中这些被保护的mRNA片段构建成DNA文库，再使用Illumina公司的测序仪对文库中所有的片段进行测序，最终得到了一幅有关细胞中蛋白质翻译情况的完整“画卷”。

这种方法可以应用于很多方面。首先，它能广泛地用于蛋白质组研究当中。正如Weissman报道的那样，“对于像人类一样复杂的基因组，你真的无法解释清楚细胞表达出来的多肽是什么。而这种新方法刚好给了你一个客观的、全面的机会去弄清楚这些多肽。”现在，Weissman等人正在使用这种新方法研究酵母，因为酵母比较简单，同时也被研究得比较透彻，因此相对来说比较容易研究。但是从理论上来说，该方法是可以应用到其它任何一种物种中的。另外，将该技术与标记有抗原表位的核糖体（epitope-tagged ribosomes）结合使用，还有可能用于研究组织特异性的蛋白质翻译（tissue-specific translation）。Weissman说道：“我认为该技术会将分子神经解剖学（molecular neuroanatomy）一类的学科引向新的纪元。”

其次，在检测蛋白质表达情况时，使用核糖体图谱技术相比检测mRNA丰度来说更准确。研究人员借助核糖体图谱技术为胞内数千种mRNA构建了核糖体印记密度图谱，并通过这些数据获得了蛋白质翻译表达速度方面的数据。据这些研究人员报道，使用蛋白质翻译表达速度方面的数据来判断蛋白质丰度要比用mRNA丰度来预测准确得多。Weissman说道：“对我们来说，定量蛋白质组学（quantitative proteomics）最大的好处就是能客观评价人们的工作究竟做得好不好。”实际上，如果对结合在mRNA链5'端的核糖体数目进行进一步的修正，就能更准确地预测出蛋白质的丰度。

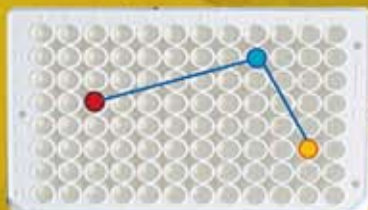
核糖体图谱还可以用于翻译控制（translational control）分析。Weissman等人正在使用该技术对饥饿酵母胞内的翻译反应（translational response）进行研究。毫无疑问，该方法也可以用于高等生物应激或疾病状态下的蛋白质合成反应控制情况。

核糖体图谱技术还具有很高的空间准确性（spatial precision），能准确地反映出究竟是哪一个阅读框被翻译了。因此，可以使用该技术研究程序性框移（programmed frameshift）和终止密码子通读（stop-codon readthrough）等现象。Weissman等人最近在酵母中的工作还发现，该技术可以发现mRNA 5'端非编码区的异常翻译情况。

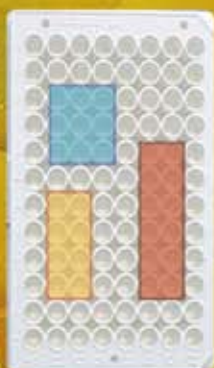
正如Weissman对核糖体图谱技术的总结一样，“我们现在能直接得到全面的、高质量的蛋白质翻译速度方面的数据。通过这些数据我们可以知道哪种蛋白质表达了以及表达了多少。同时，我们还能很方便地对翻译过程本身进行研究。”

原文检索：Natalie de Souza. (2009) Deep sequencing of ribosome footprints. *Nature Methods* 6(4): 244-245.

# Primer Array (qPCR 芯片)



筛选您感兴趣的  
表达基因和表达模式



## 原理

- Q-PCR Primer Array是基于SYBR Green染料法的Real-Time PCR检测技术，是对不同样品中具有相似生物功能的基因家族或自选的特异pathway基因进行表达量差异分析的系统。
- 该产品是把优化并通过实验验证的各对All-in-One™ Q-PCR Primer按使用浓度交联到96-Well-qPCR反应板中。
- 该产品使用时仅需加入All-in-One™ Q-PCR Mix，样品cDNA及PCR级H<sub>2</sub>O即可上机反应。
- 通过Q-PCR Primer Array，研究人员可以快速、可靠、精确的筛选出感兴趣的特异基因和基因表达间的关系。

**GeneCopoeia™**  
Expressway to Discovery

GeneCopoeia, Inc.  
19520 Amaranth Drive  
Germantown, Maryland 20874  
USA  
Tel: 301-515-6982  
Fax: 301-515-6983  
Web: <http://www.genecopoeia.com>

 复能基因  
FulengGen

广州复能基因有限公司  
地址：广州科学城国际企业孵化器D座8楼 510663  
电话：(020)32052376、32052410、32290874  
传真：(020)32052877  
网址：<http://www.fulengen.com>  
服务热线：020-32068595 020-32052376-666  
电子信箱：[sales@fulengen.com](mailto:sales@fulengen.com)

## 2. 如何将数十亿的短片段测序结果定位到庞大的基因组序列当中

随着新一代测序仪的出现，人们获得了大量的短片段序列，如何对这些短片段作图就成了一个大问题。现在有什么办法可以解决这个问题呢？上述办法又是基于何种原理工作的呢？

新一代测序仪可以以极快的速度以及极其低廉的价格获得大量的序列，这已经改变了基因组学的面貌。这些新测序仪一经出现，马上就成为了全基因组测序的主力军，广泛应用于各种测序相关的实验检测，包括基因表达谱检测、DNA与蛋白质相互作用检测和RNA剪切研究等。例如，它们可用于对RNA进行测序，即先通过逆转录将其变成cDNA，然后再对cDNA进行测序，这样就能发现一些未知的基因，并据此发现新的RNA剪切方式。也可以将测序技术应用于ChIP，弄清楚与蛋白质共沉淀的DNA片段的序列。这种方法能用于研究转录因子与DNA调控元件之间的相互作用。此外，对肿瘤细胞全基因组测序也能发现一些新的致癌突变。

但在新一代测序仪带来方便的同时也带来了问题，即被称为“阅读片段作图（‘read mapping’）”的问题。美国Illumina公司、Applied Biosystems（ABI）公司和Helicos公司等开发的测序仪在测序时产生的都是长约25bp~100bp左右的小片段序列，即“read”。这些小片段都是待测样品大片段的某一部分。与对未知的全基因组进行测序，即与将所有小片段组装成一个完整基

因组的工作相比，人们现在大部分的工作实际都可以参照“参考基因组”（也称“模式基因组”，小词典1）进行。因此，要了解小片段“read”的作用，首先要知道它们在参考基因组中的确切位置，而对这些小片段进行定位的过程就称作“作图”（mapping），或“定位”（aligning）到参考基因组中。在作图中，有一个问题需要注意，那就是进行定位（本文将在后面的“短片段作图软件”一节中对此做详细介绍）时不能出现大的“间隙”。而在对RNA进行测序时，因为存在内含子的缘故，这一点就显得尤为突出。因此，对RNA进行测序时就允许有较大的间隙出现（这将在下文“剪切后的短片段作图软件包”一节进行详细讨论）。

当然，上述问题都不是伴随新一代测序仪的出现而出现的新问题，即使在经典的Sanger毛细电泳测序法中也有与之相应的专门用来处理定位问题的程序。不过，这些程序既不能处理短片段测序仪获得的大量序列数据，也不能定位长度较短的短片段序列。使用传统的BLAST或BLAT软件分析ChIP或RNA测序结果，可能会花上几百甚至几千个小时。幸运的是，人们现在有了新的分析软件。在选择一款分析软件之前，要先弄清楚，为什么用计算机处理作图问题会出现问题？人们现在已经解决了其中的哪些问题？还存在哪些问题？还有没有其它机遇？

### 2.1 短片段作图

#### 2.1.1 对短小片段作图存在哪些问题？

问题1：实际操作。如果参考基因组很大，而我们手上又有数十亿计的短片段序列，那么该如何处理这么庞大的数据呢？如何将每一条短片段定位到参考基因组中相应的位置上？序列比对是生物信息学中的一个传统问题，有大量的文献著作介绍了各种不同的比对方法，既有精确严格的方法也有不那么严格的方法。不过，从实际应用的角度出发，要将数十亿的短小片段定位到哺乳动物基因组大小级别的参考基因组中需要借助效率非常高的算法进行处理才有可能办到。

问题2：处理策略。如果某个短小片段属于参考基因组里的一个重复元件，那么就应该弄清楚它来自重

复元件中的哪一个拷贝。但这是不太可能实现的，所以分析程序一般都只能给出该短片段可能属于参考基因组中哪几个位点。同时，由于测序错误或者检测样品间以及检测样品和参考基因组间出现变异等情况，使上述问题变得更加严重。同样，在RNA剪切体作图中也存在上述问题，而且由于内含子的问题使得情况更为复杂。

Illumina、ABI、Roche、Helicos以及其它众多测序仪生产厂家开发的测序仪每一轮测序都能获得百万计的短片段序列，不过要对一个基因组进行完全测序则需要进行好几轮检测，这也就意味着要想获得一份完整的全基因组图谱必须对数百万甚至是数十亿的短小片段进行作图、定位和拼接。比如，最近由Ley小组做出的癌症基因组序列就是通过132轮测序，对80亿条短小片段进行作图后得到的结果。使用BLAST或BLAT比对法，借助大型的超级计算机只需要几天就能获得这个癌症的基因组序列结果，但这并非人人都能享有。为了能让更多的人用更廉价的计算机也能进行类似的作图分析，人们开发了一套新的比对定位程序，使用这种新程序即使在普通的台式机上也能对数亿计的短小片段进行作图分析。测序仪器生产厂商也会提供一些专门的作图软件，例如Illumina公司开发的ELAND程序等。本文将着重探讨第三方开发的软件，这些软件中很大一部分都是开放源代码的免费程序。这些软件主要都是建立在这样一种算法之上，即充分利用短小DNA序列的特点来作图，而不需要依靠计算机强大的处理能力、内存容量等条件。

2.1.2 短片段作图软件

Maq和Bowtie（见表16）都属于上述提及的程序。它们使用的是一种称作“建立索引（indexing）”的策略。同时，人们也对大量的DNA序列建立了一份索引，借助这份索引就能快速地找到其中的短DNA片段了。Maq软件是基于一种直接的但是很有效的策略——空位种子片段索引法（spaced seed indexing）（图12a）。它将一个短片段（read）分成了4条长度相等的更短的片段——种子片段（seed）。如果整段短小片段（read）可以与参考基因组序列完全配对，那么很显然所有的种子片段（seed）也理所应当应该与参考基因组序列完全配对。但如果其中有一处错配，例如SNP，那么肯定有一条种子片段无法与参考基因组序列完全匹配。依次类推，如果出现了两处错配就会导致一条或两条种子片段无法与参考基因组序列完全匹配。因此，对所有种子片段两两组合后的片段（共有6种组合方式）进行比对，就有可能找出该短小片段在基因组中最有可能的位点。Maq软件采用的这种“空位种子片段索引法”（spaced seed indexing）作图时的效率非常高。

Bowtie软件采用的则是另一种完全不同的策略，该策略借鉴了Burrows-Wheeler转换（Burrows-Wheeler transform）这种数据压缩算法技术，将完整的人类基因组序列索引压缩到不到2GB大小（这是当前主流台式机甚至是笔记本电脑都能达到的水平），而空位种子片段索引法至少需要50GB。Bowtie每次都只

表16 短片段分析软件一览表

| 软件名称      | 站点                                           | 是否开放源代码 | 能否处理ABI的彩色空隙测序结果 | 能分析的最长序列长度bp |
|-----------|----------------------------------------------|---------|------------------|--------------|
| Maq       | http://maq.sourceforge.net                   | 是       | 能                | 127          |
| Bowtie    | http://bowtie.cbcb.umd.edu                   | 是       | 不能               | 没有相关信息       |
| BWA       | http://maq.sourceforge.net/bwa-man.shtml     | 是       | 能                | 没有相关信息       |
| Mosaik    | http://bioinformatics.bc.edu/marthlab/Mosaik | 不是      | 能                | 没有相关信息       |
| Novoalign | http://www.novocraft.com                     | 不是      | 不能               | 没有相关信息       |
| SOAP2     | http://soap.genomics.org.cn                  | 不是      | 不能               | 60           |
| ZOOM      | http://www.bioinfor.com                      | 不是      | 能                | 240          |

把一段短片段序列中的一个碱基与经Burrows-Wheeler转换压缩过的参考基因组序列进行比对（图12b）。经过这种连续的比对，最终也能找出这段短片段在参考基因组中的定位。如果Bowtie软件发现短片段中的某个碱基在参考基因组中没有很好地配对，那么软件就会退回到上一个碱基重新进行比对。实际上，Burrows-Wheeler转换使得Bowtie软件通过碱基逐个比对，直至完成全长短序列比对的方法解决了短序列作图的问题。从本质上来说，Bowtie软件使用的算法要比Maq采用的复杂得多，但Bowtie软件却比Maq软件分析的速度快30倍。

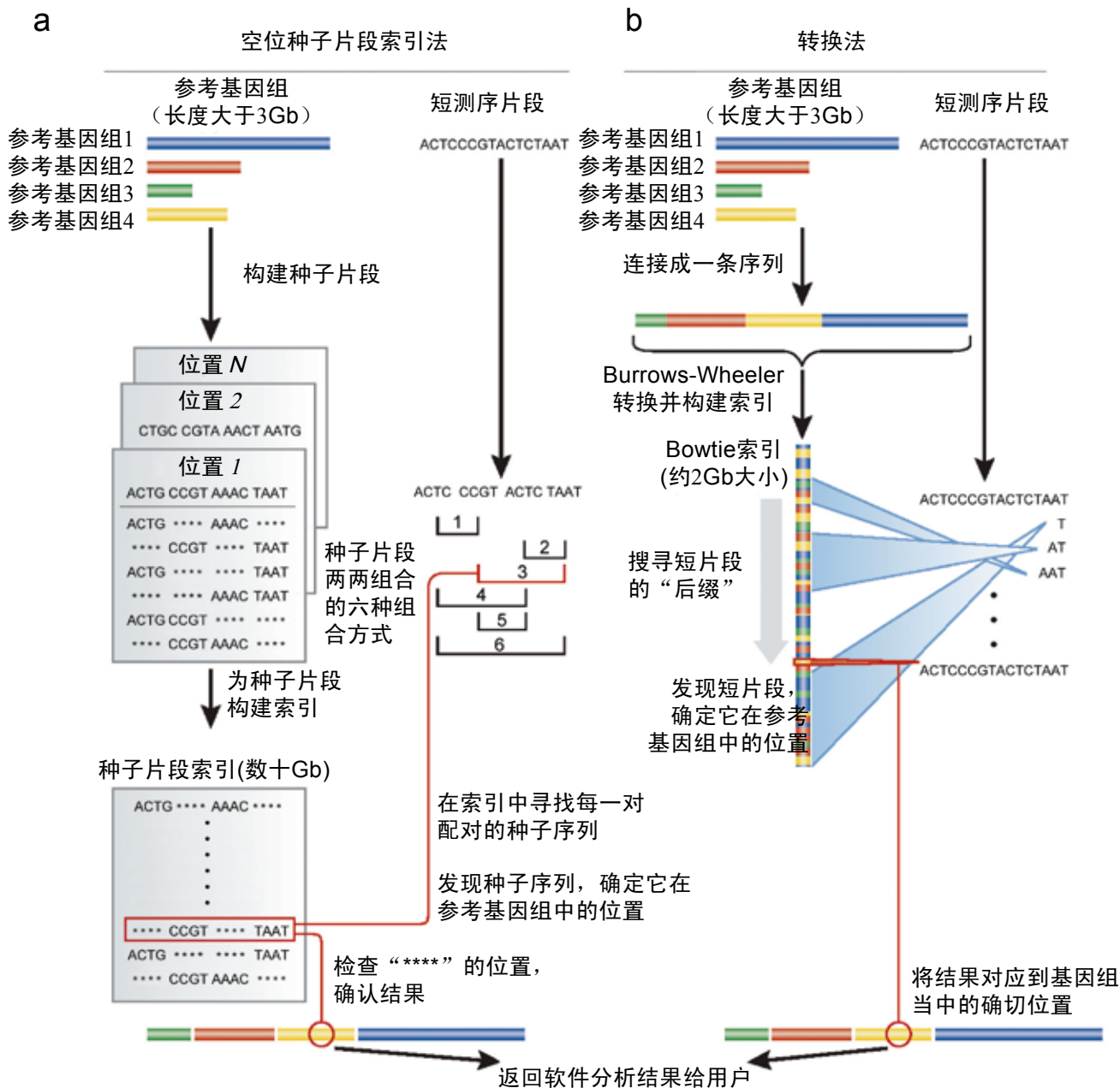


图12 最近出现的两种将20bp~200bp长度的短片段测序结果定位到参考基因组当中的方法。

(a) 基于空位种子片段的Maq软件索引法。首先，将参考序列分成长度相同的片段，称为“种子片段”，然后将这些种子片段组合起来形成一个索引表。再将短测序片段等分成4条更短的“种子片段”，然后将4条种子片段两两配对，形成6对配对种子序列，按照这些配对序列到种子片段索引表中寻找相应的序列。由于种子片段的数量非常大，因此在有些算法中（比如Maq采用的算法）就将这些片段打包处理了。(b) Bowtie软件采用的是基于Burrows-Wheeler转换法的方法，该方法能以一种非常节约存储空间的方式储存参考基因组序列。在该方法中，短片段中的碱基会按照从右至左的方向逐个和参考基因组序列进行比对，随着被比对过的片段长度的增加，它们在参考基因组中对应的范围就会缩小，如图中蓝色“光束”所示。当全部短片段中的碱基都被逐个比对完毕之后，就能将该片段定位到基因组当中了。Burrows-Wheeler转换法比空位种子片段索引法处理的速度快得多，这可能是因为对转换后的参考基因组序列进行搜索的效率要高所致。

Bowtie软件和Maq软件的默认模式中至多都只会允许两个错配位点，不过有时有些用户需要允许更多的错配位点存在。Bowtie软件和Maq软件能够分析的短序列长度范围在20bp~40bp之间，它们都经过优化设计以使其适合用于人类基因组再测序计划（human resequencing project）。不过，现在Illumina公司最新的测序仪已经能够获得长约100bp的“短”片段序列，还有一些测序项目，例如细菌或真菌基因组测序项目等获得的片段序列与目前已经测得的类似物种全基因组序列之间存在着较大的差异。再加之随着新测序仪的不断涌现，测序结果的质量也在不断提高，但这些测序结果却极易受到各种因素的影响，例如样品文库的准备、测序操作步骤、甚至是放置测序仪器实验室的温度等等。鉴于此，面对上述这些新出现的“问题”，人们也应该采取相应的措施，调整Maq软件和Bowtie软件的各种参数使之适应这些新情况。

表16列出的是几种新的开放源代码的短片段序列作图软件，它们的安装和使用都很简单。

Bowtie软件包中包括预置的大肠杆菌基因组索引和部分大肠杆菌短片段序列。要使用该软件分析数据只需输入下面的命令就会生成一个表格式的报告，给出每一个匹配短序列的编号、在参考基因组中的位置、以及发生错配的位点个数和具体位置。

```
bowtie e_coli reads/e_coli_1000.fq
```

在Maq软件中输入以下命令也会得到同样的结果。

```
maq.pl easyrun -d outdir  
reference.fasta reads.fastq
```

对于一次实验来说，短序列片段能否与参考基因组相匹配实际上取决于很多因素。假设被测序的DNA片段中几乎没有错配位点，大多数作图软件也只能定位出70%~75%的短片段序列。这个结果和使用Sanger测序法获得的80%的结果比起来低得令人吃惊，说明现在新一代测序技术还不成熟。这提示人们，很多短片段都需要与参考基因组中的多个位点进行比对，而大部分的作图软件都只会给出短片段在参考基因组中的一个匹配位点。

有了序列定位的软件，接下来就可以了解这些短片段具体在参考基因组中的什么位置了，同时也可知道SNP都位于基因组中的什么地方。SAM软件包能满足这些要求。SAM软件包（<http://samtools.sourceforge.net>）包括一体化的碱基调用和浏览器（base caller and viewer），它能使用Maq和Bowtie两种分析软件。

实际上，大部分短片段作图软件设计的初衷都是为了服务于人类全基因组再测序工作，但是调整软件参数之后，它们也能应用于其它方面。Maq和Bowtie这两种分析软件的操作手册都写得非常详细，它们给出的备选方案多到“吓人”的程度。现在还出现了越来越多的短片段作图软件（表16），不过每一款软件都无法达到十全十美的境界，而且各有偏重，这就给人们选择软件及其配置参数带来了麻烦。幸运的是，人们能够得到帮助。SeqAnswers message board（<http://www.seqanswers.com>，图13）就是一个非常好的论坛，它是一个短片段作图软件开发人员经常光顾的论坛。最流行的SeqAnswers线程一般都包括目前用于初步分析的软件和短序列数据可视化处理的软件一览表。



图13 SeqAnswers message board。图片来源：SeqAnswers。

### 2.1.3 剪切后的短片段作图软件包

要将RNA的逆转录片段cDNA重新定位到基因组当中需要更加复杂的专业化算法。要将不同外显子经过剪切拼接之后生成的RNA短片段重新定位到基因组中和将一个外显子生成的RNA短片段重新定位到基因组中是完全不一样的（图14）。

在RNA逆转录产物cDNA的定位操作中用到的诸如ERANGE（<http://woldlab.caltech.edu/rnaseq>）这类软件包都会用到已知基因的外显子位置和内含子位置信息作为参考。这样，ERANGE软件包就能“横跨”多个外显子构建新的参考序列，然后再调用Maq程序或者Bowtie程序将剪切后的RNA片段定位到参考序列中了。因为这种方法不能发现新的（人们未知的）剪切模式，所以有些科研人员就使用了一种“机器学习法”（machine learning method）来预测新的剪切模式。该方法借助现有的参考序列注释信息在统计模型（statistical model）上进行过演练。与此相反，TopHat软件包（<http://tophat.cbcb.umd.edu>）则不需要借助任何注释信息，它使用的是Bowtie软件来发现包含有短片段的外显子，然后再将余下的短片段定位到前面发现的各种外显子连接体当中。还有一款程序G-Mo.R-Se（<http://www.genoscope.cns.fr/externe/gmorse>）使用的也是这种策略，不过它是借助RNA测序数据而不是通过Bowtie软件来发现外显子的。

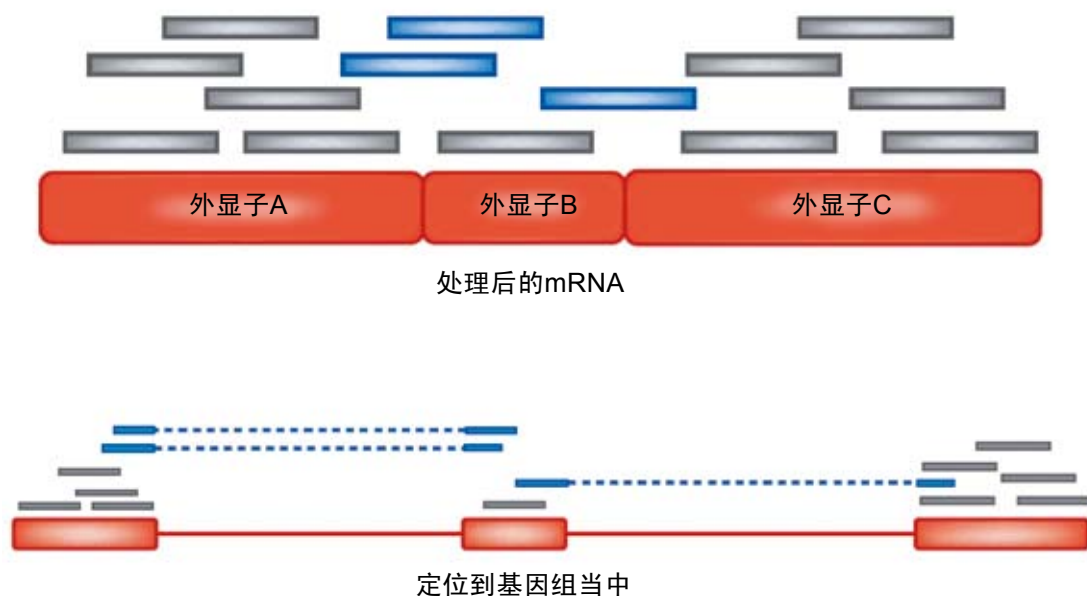


图14 RNA测序会产生大量的短片段，使用Bowtie软件或Maq软件将这些短片段定位到基因组当中只能处理图中同一外显子内的黑色片段样结果，而不能处理跨外显子的蓝色片段结果。但使用TopHat软件或ERANGE软件就能处理这些跨外显子的蓝色片段结果。

## 2.2 局限性及存在的问题

现有的用于短片段作图的方法都有其各自的局限性。比如，Maq和Bowtie软件在处理插入或缺失片段时就几乎不起作用。

有些软件，例如SHRiMP（<http://compbio.cs.toronto.edu/shrimp>，图15）就能支持ABI公司的“彩色空隙（color space）”测序结果，但大部分软件都是不支持该结果的。剪切后短片段作图软件同样存在类似问题，而且它们还有自己的特殊问

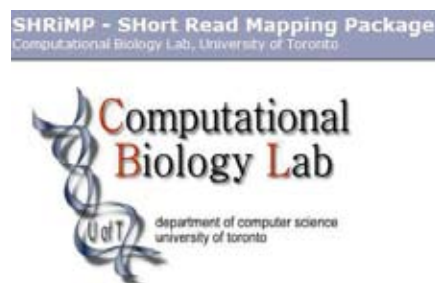


图15 SHRiMP。  
图片来源：Computational Biology Lab.

题。例如，基于注释信息的软件当然最多只能获得和注释信息相当的结果，但很多物种的全基因组注释信息都仅仅是同源预测信息或计算机预测信息。如果“机器学习方法”受到错误的注释信息“操纵”的话，也不会得出好结果。

因此，对于短片段作图软件的开发设计人员来说，还有很多问题需要去解决。所有的测序仪器生产厂家都在努力得到更长的测序片段结果，现有的短片段作图软件能应付这些“大家伙”吗？Maq、Bowtie以及其它几种短片段作图软件都可以处理长度超过100bp的测序片段结果，但这只是在特定的情况下，而且只有原本就是针对长片段设计的软件，例如BLAT才能更好地处理这类测序结果。另外，如果测序的样品物种序列和现有的参考序列差异很大，那该如何调整作图软件参数呢？软件能够自动调整参数吗？这样做出来的图质量又如何呢？上述这些问题的解决方案都依赖于采用的检测方法和分析范围。不过，随着技术的进步，相信所有这些问题很快都会被克服的。

原文检索：Cole Trapnell & Steven L Salzberg. (2009) How to map billions of short reads onto genomes. *Nature Biotechnology*, 27(5): 455-457.

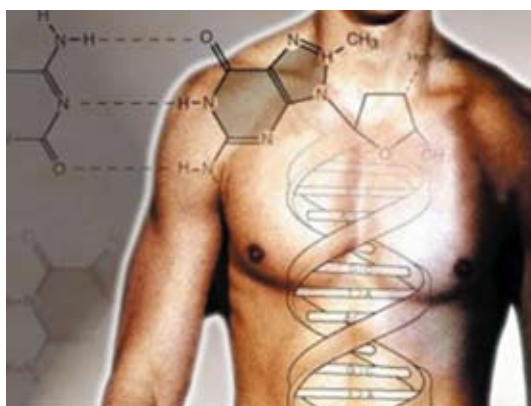


## 小词典

### “参考”基因组（‘reference’ genome）

每一个种内的物种的基因组都有数量一定且相对恒定的基因和基因排列方式，但由于某些基因或基因片段的突变形成了不同的物种。通常可以在种内以某一代表性物种的基因组作为模式基因组，以利于对其它物种的研究。这种“某一代表性物种的基因组”就是“参考”基因组了。

## 3. 更快——只需15分钟、更便宜——只需100美元的人类基因组测序技术即将面世



很快，我们就会看到最新的人类基因组测序技术问世。有了这种新技术，人类基因组测序的费用将大为降低，平均每个样品只需要花费100美元。同时，该技术的测序速度要比目前市场上广泛应用的第二代测序技术快2万倍，而且我们可以借助该技术实时的观测到人类基因组DNA被扩增的过程。

Stephen Turner是太平洋生物科技公司（Pacific Biosciences）的首席技术官（Chief Technology Officer），他表示，最新的商业化单分子实时测序仪（Single Molecule Real-Time sequencing, SMRT）将于2010年上市。

十年前，塞莱拉基因公司（Celera Genomics）和人类基因组计划组（Human Genome Project）都花费了

数年的时间才得到完整的人类基因组序列图。

但到了2008年，由于有了新一代的测序仪，所以我们只用了几个月的时间就获得了James Watson的个人完整基因组序列。

现在，有了SMRT测序仪，Pacific Biosciences公司希望可以用这款测序仪在几分钟之内完成人体基因组测序的工作。

我们在进行人类基因组计划工作时采用的研究策略，就是利用了细胞复制DNA的天然机制。

使用DNA聚合酶复制DNA链的方法获得了数十亿计，各种长度的DNA片段分子。然后在每一个片段末端都加上一小段荧光标记分子，该荧光标记分子只能对DNA链末端最后一个碱基进行标记，然后根据DNA片段的长度将这些分子排列整齐，我们就可以像读书一样，按照长短顺序把DNA序列末端的碱基一个一个给读出来了。

不过SMRT测序仪采用的并不是这种方法——等DNA聚合酶完成了复制工作之后再判读序列的方法，该技术采用的是实时对DNA聚合酶的工作状态进行监测的方法，每一个DNA链分子都被吸附在小孔的底部，在DNA聚合酶复制的同时，SMRT测序仪就会实时读出每一个碱基，这样就能得到完整的序列。

在SMRT测序仪使用的每一个碱基上都带上了特有的荧光标记，一旦某个碱基掺入了新合成的DNA链，就会发出特异性的荧光信号，实时探测仪就可根据该荧光信号判断出该位点是A、C、G、T中的哪一个碱基。

发明SMRT技术的科研人员们希望能够将该技术进一步改进，成为芯片式的多通道并行处理测序仪，这样还可以进一步加快测序速度。

“如果我们能同时处理100万个片段分子，那么我们就能够在15分钟之内获得完整的人类基因组序列图，”Turner说道。

SMRT测序仪在提高测序速度的同时还能够提高测序的准确率。由于用SMRT测序仪进行测序时出错是随机发生的，也就是说每个位点出错的概率都一样，没有哪个位点会更容易出错，因此如果进行多次重复测序应该能够提高准确率。

SMRT测序仪刚刚在第51界美国医学物理学家协会年会(Annual Meeting of American Association of Physicists in Medicine)2009年工业物理学大会(Industrial Physics Forum)上第一次亮相。

原文检索: <http://www.freshnews.in/coming-soon-15-minute-100-human-genome-sequencing-157761>

视频: <http://www.scivee.tv/node/11409>

筱玥/编译

科研综述 研究前沿 热点话题  
技术方法 专题译述 生命百态  
会议展览 教学视频



[www.LifeOmics.com](http://www.LifeOmics.com)