单分子测序技术带给基因组学研究及转化医学的革新

革命性的第三代测序系统——PacBio RS

一、概述：

从1980年英国生物化学家Frederick Sanger与美国生物化学家Walter Gilbert建立DNA测序技术并获得诺贝尔化学奖至今已有近三十年了。在这三十年，DNA测序技术取得了令人瞩目的进展，为生命科学研究领域开辟了新的视角，使研究水平上升到新的高度。回顾过去，第一代测序技术帮助人们完成了包括人类基因组计划在内的多种生物基因组序列测定及数据库的建立。第二代测序又以更高的通量，更简易，标准化，自动化的操作以及不断降低的测序费用加速拓展着测序应用的广度和深度，不仅大大促进了基因组学，转录组学，核酸蛋白相互作用等相关生物学研究的发展，还让人们看到了其在医学，临床诊断，药学等等领域广阔的应用前景，让人们对个人基因组时代充满期待。在这里，我们将给大家介绍最新的第三代测序技术及其将带来的革新。

Pacific Biosciences 公司研发的PACBIO RS 单分子实时测序系统，革命性地推出了单分子实时（Single Molecule Real Time, SMRT）DNA测序技术，在测序历史上首次实现了人类观测单个DNA聚合酶合成过程的梦想。使研究者第一次能够利用单个 DNA 聚合酶对天然DNA 的合成进行大规模平行的、连续的实时观察。

一系列的文章也证明了PACBIO RS系统的核心技术——SMRT技术的原理和应用。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2010年6月的《Nature Methods》上。

PACBIO RS系统的一个显著特点在于从样本制备到获得测序结果，所需的时间还不到一天，且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并能实时观察，这对于传染病监控和分子病理学尤为重要。目前的平均读长超过1000bp，其中5%最长的读长甚至可高达数千bp，比二代测序的读长要长很多。可为海量的二代测序数据拼接提供帮助。此外，试剂消耗和样本制备量极少，且不需要常规的PCR扩增步骤，使得假阳性率也大大减少，聚合酶动力学的直接观察也可赋予研究者以测序之外的更多应用。凭借这一系列的优势，Pacific Biosciences公司于2010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。

二、PACBIO RS系统测序原理

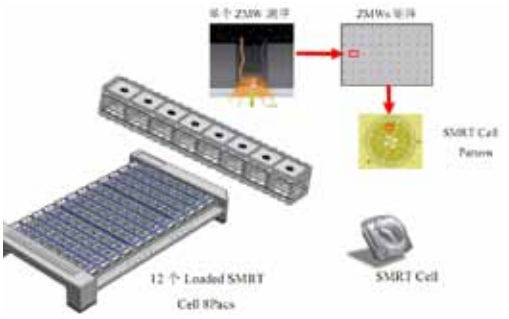

图1 SMRT Cell外观及构造原理

PACBIO RS系统专利的 SMRT Cells (图1)，其含有数以万计纳米级的零模波导孔（zero-mode waveguides, ZMWs）。测序反应是在专利的SMRT cell中进行的，每个SMRT cell中有150,000个ZMW。每个ZMW都能够包含一个DNA聚合酶及一条DNA样品链。这样，SMRT Cell能够平行检测大约75,000个单分子测序反应。

当测序进行时，专利的包被技术保证DNA 聚合酶和模板形成的复合体被锚定在 ZMW 的底部，反应溶液中带有标记着不同荧光磷酸基团的高浓度核苷酸可有效保证聚合酶合成的速度、精确性和持续合成的能力，检测装置则透过ZMW 底部的基层来实时观察 DNA 的合成过程：当 DNA 聚合酶检测到正确的核苷酸并将其插入模板时，这些碱基会在插入位置保留几十毫秒，而游离状态的碱基扩散的速度大约是几十微秒。在这个时间差中，检测器就可以检测到插入碱基的荧光信号，并将其转换成相对应的碱基类型。然后，DNA 聚合酶将碱基所带的标记着荧光集团的磷酸基团剪掉以形成天然DNA 链。此后，荧光信

号迅速衰减至基线并开始下一轮的合成。PacBio RS使用一套优化的算法，将光学系统所捕获的信息翻译成对应的ACGT碱基信息。一旦测序开始，实时的数据就被传送到初步分析系统中，以实时生成碱基组成和质量值等信息。

三、PACBIO RS系统配置

PACBIO RS 系统内置的机械臂能够自动将SMRT Cell在储存区、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 bell模板的制备过程，测序过程，碱基判定结果以及 QV 分值等，并实时播放测序过程中记录的影像信息。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。

四、PACBIO RS系统的测序方案

PACBIO RS 系统可提供多种SMRT测序方案，如标准测序、环形比对测序、频闪测序，每种方法都利用了长片段读取的优势并结合SMRT bell模板形式（模板处理后形成的一个类似哑铃的结构）和单DNA 聚合酶原理。

1）标准测序：从插入片段的一端测到另一端，只测一次。适合插入片段在 1-6Kb间。主要应用于再测序和重头测序。

2）环形比对测序：从插入片段的一端测到另一端后，由于模板是哑铃状结构的，因此，继续测序将得到互补链的序列信息，从而可通过对插入片段的反复测序来获得高准确度的单分子测序结果。适合插入片段在 250-500bp。主要用于发现和确定稀有变异。

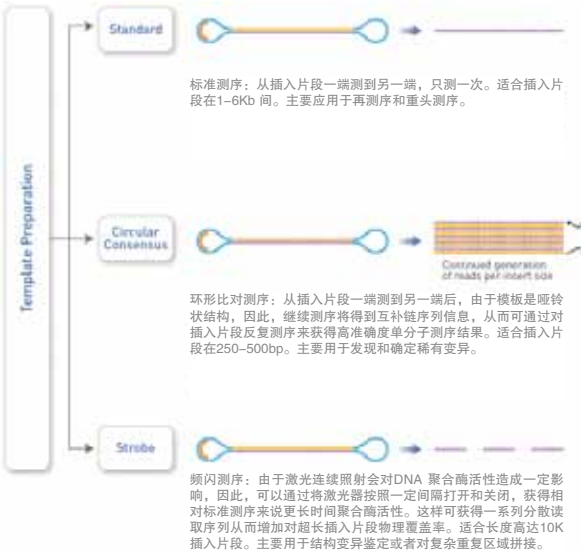

图2 SMRT测序的三种方案（标准、环形比对、频闪测序）

3）频闪测序：由于激光连续照射会对 DNA聚合酶活性造成一定影响，因此，可以通过将激光器按照一定间隔打开和关闭，获得相对标准测序来说更长时间的聚合酶活性。这样可获得一系列分散的读取序列从而增加对超长插入片段的物理覆盖率。适合长度高达 10K的插入片段。主要用于结构变异的鉴定或者对复杂重复区域的拼接。

在每一种测序方案中，用户都可以调整参数，适合多种应用和项目类型。这种灵活性也提供了分多个步骤解决一个问题的能力。比如，使用者可以展开一个序列，使用频闪测序产生数千个碱基对的物理读长，然后用标准测序实现长读长的单分子结果读取，最终，使用环状测序进行多次测序实现前所未有的准确率并发现和确定稀有突变。另外，短的运行时间和灵活的耗材包装，提供了对各种规模的研究课题都经济有效的配套方案。

五、PACBIO RS系统的优势

相对于二代测序，突破性的三代测序有如下优势：

1）单分子实时检测，每条碱基链的数据都得到评估，更易发现稀有变异

2）最低DNA样品量仅500ng，无需PCR扩增，测序覆盖深度不受序列中 GC含量差异影响。可以对 100%A+T区域测序

3）除碱基序列信息外，可同时获得动力学信息，因此可对诸如 DNA甲基化之类的碱基变化情况直接进行分析

4）无与伦比的长序列读长

5）测序速度快，聚合酶反应速度（平均）> 1 base/sec，从样本制备到出结果，需时<1天

6）灵活性极强，有一系列数据管理和分析软件使二代三代测序流程实现无缝整合

7）测序流程简单

六、PACBIO RS系统在基因组学应用

1. 同 NGS平台联合使用

1）基因组序列拼接

沼泽红假单胞菌（Rhodopseudomonas palustris）DX-1菌株基因组GC含量高达60%，二代测序获得了大量短序列数据。为了有效拼接二代测序获得的1320个序列重叠群，研究者首先用PACBIO RS系统的SMRT测序方案中的长读长结果来增加序列重叠群的长度，然后用频闪读取结果将大的序列重叠群链接起来，最终将序列叠群的数

量减少到15。为了检测重叠部分，研究者开发了一种基于suffix tree的有效算法，可在较大范围允许特异性的匹配和错误，为二代测序的短读取结果与PacBio的长读取结果的重叠读取提供很大的灵活性和灵敏性。这套算法的使用增加了SMRT测序对基因组上长的、低复杂性的区域扫描的能力，使得二代测序与PacBio的读取结果相结合的读长比单独使用二代测序的读长要更长很多，N50（即覆盖50% 所有核苷酸的最大序列重叠群长度）数值也更大（图3）。最后，我们使用PACBIO RS系统的频闪测序技术将大的重叠群序列以一种类似mate-pair测序的方式将多个跨越长距离的读取结果连接起来。另外，PACBIO RS系统读取结果也可与Sanger测序结合，来增加在DX1基因组上观察到的重叠群的数量（图3）。总之，PACBIO RS系统标准长读取结果和频闪读取结果的组合方便了细菌基因组测序的完成，使海量的二代测序短序列结果的拼接更加便利。通过二代和三代数据的混合拼接，可以大幅提高基因组拼接效率。 NGS平台所获得的海量数据量，可在PacBio测序步骤得到解读，从中获得更多有益的信息。

Assembly	Number Contigs	Total Contig Length	Max Contig Length	Mean Contig Length	N50
Abyss PE Illumina contigs	1,320	5,042,272	46,839	3,819	4,821
PacBio long reads + Abyss PE contigs	103	5,396,171	369,642	52,390	122,926
+ strobe scaffolding	19	5,235,513	982,175	348,094	595,229
Sanger reference	53	5,372,418	689,257	101,366	202,894
+ strobe scaffolding	2	5,307,230	2,973,544	2,653,615	2,973,544

图3 采用不同方法的基因组序列拼接结果对比。对于大于1kb的重叠群选择性拼接的统计学分析。数据表示碱基对。Scaffolding通过Bambus Pop M.程序完成 (2004)。

2）寻找结构变异的特征

频闪测序，是SMRT测序对传统paired-end方法的扩展，已被应用于检测结构变异。频闪方法对结构变异分析的应用，最初是在已知结构变异的人类基因组DNA的两个F黏粒上被证实的：这两个F黏粒中的一个在克罗恩病相关基因IRGM，AC207974，的上游有两个缺失。研究者准备了一个目标大小为7kb的单个SMRT bell文库。跨度范围从3kb到7kb的频闪读取结果通过使用相同的试剂而仅改变读取时间产生。结果PACBIO RS系统收集到类似mate-pair的2个亚读取结果及3个频闪亚读取结果。AC207974上的两个缺失可以通过频闪测序轻易地检测到(图4A)。另外，在SMRT测序时，每一个合成事件的精确时间都被记录下来，并可被用于在物理覆盖率下比较碱基间隔的时间。通过比较一个未排列的序列和参考序列，合成时间和覆盖率之间应该是接近1：1的关系。对于一个缺失区域，对比参考序列，这部分区域测序的时间就会减少，减少的程度与

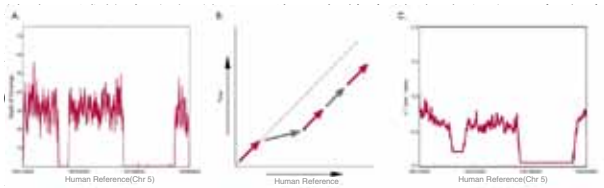

图4 对AC207974进行频闪测序

15.1kb的限制性片段上含有3个亚读取结果和4个亚读取结果的频闪数据。这个片段结果是在测试大于10kb的SMRT bells的测序质量和产量时的部分结果时产生的。该结果可分为三个类型。紫色代表跨越并锚定于相对参考序列的7kb插入片段的频闪测序结果，橙色代表定位于7kb掺入序列与参考序列连接处的频闪测序结果，而蓝色代表将两个插入序列彼此锚定及插入序列与参考序列之间锚定的频闪测序结果(图5)。与配对末端策略相比，这种革新的、多个亚读取结果的方法，可以使研究者获得更丰富的有关结构变异的关联信息。使用3个或者更多的亚读取结果，可同时跨越插入片段，揭示插入序列，连接插入片段并将其定位到基因组参考序列，以及更准确地解决边界重叠部分的断点问题。

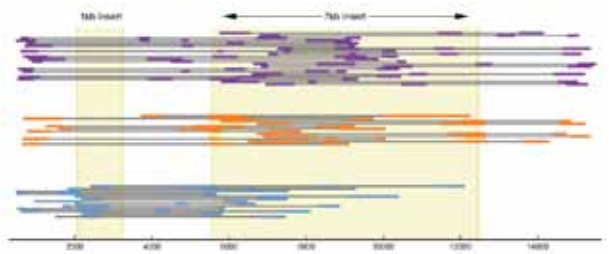

图5 对AC223433的15.1kb的限制性酶切片段进行频闪测序

3) 模糊配对拼接数据的更正

频闪测序也被应用于更正含有复杂重复结构的基因组拼接数据，解决模糊配对的问题。比如在沼泽红假单胞菌（Rhodopseudomonas palustris）的杂交拼接上，应用PACBIO RS系统的频闪测序对于分辨不同的重叠群的配对之间的大量模糊的连接是十分重要的，这可以使得来自二代Illumina测序和标准SMRT测序读取结果的杂交拼接的重叠群的数目显著减少。图6A显示的是频闪测序解决重复区域之间连接的重要性的一个具体示例，六对重叠群和一个1.5kb的重复序列之间的模糊配对。在这个例子中，1.5kb的插入片段（contig57）导致其他6个重叠群的模糊配对。而使用3个亚读取结果的频闪测序可解决模糊配对的问题(图 6B)。

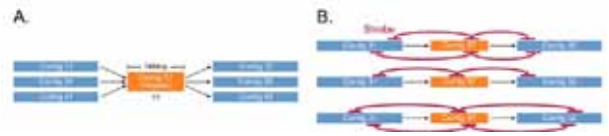

图6 沼泽红假单胞菌的模糊配对拼接的解决方案

2. 表观遗传学

除碱基序列信息外，PacBio可同时获得动力学信息，不同的碱基变化类型，其动力学特征具有特异性，从而可以根据标准动力学图谱来确认未知的碱基变化情况。

1）无需重亚硫酸盐转换检测 DNA甲基化：

SMRT测序系统独特的检测方式可以无需重亚硫酸盐转换而直接检测DNA甲基化。在SMRT测序时，磷酸连接的核苷酸掺入的动力学特征的变化会引起被修饰碱基特异

的动力学合成特征的改变。在最初的研究中，这些变化的动力学特征被用于研究基因组样品中不同CpG甲基化并鉴定dam甲基化。而且，将这种方法与单个DNA分子的环状比对测序结合可以达到碱基对水平的分辨率，并能检测表观遗传学修饰。现在的重亚硫酸盐测序技术受到短读长和基因组序列复杂性降低的限制，而这个新提出的方法可以对更高重复的基因组区域进行甲基化模式作图。

2）检测表观遗传修饰

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，其它的表观遗传学修饰，以及各种形式的DNA伤害，也能用这种方法检测。

3. 稀有突变检测

二代测序检测的是经过 PCR 扩增后的大量分子的总信号，一方面 PCR 扩增过程中的错配等，会变成系统误差影响最终的测序结果。另一方面，稀有变异在扩增过程中，很容易被淹没在群体中而无法被检出。 PacBio无需扩增、对每一个单分子模板都进行评估，对于混合群体中低至 1/100的稀有突变都可以检出。

由于目前二代测序技术的低灵敏度等问题，使得检测混合群体中诸如单核苷酸多态性（SNP）等稀有突变的过程非常繁琐且难度较大，结果使得稀有突变被淹没在大量数据中。而且，这些测序方法只会产生一个多分子且依赖于整体的一致序列，而不是从一个单一的DNA模板分子产生高可信度的序列，而高可信度序列对于稀有突变的高度敏感性往往是必需的。相反的，SMRT测序本质上来自于单个聚合酶-模板复合物，其SMRT bell模板和DNA聚合酶进行高速的链替换，使单分子测序具有很高的准确率。如前所述，SMRT bell模板由一个双链DNA插入片段连接到两个末端的发卡结构上，形成了一个结构上线性、拓扑学上环状的DNA模板。因此，单个DNA聚合酶可在单一DNA模板上产生多个正向和反向的读取结果。结合SMRT环形比对测序的随机误差模式，可实现对稀有突变的高准确率检测。

研究者已经用带有200bp（含‘T’或‘C’的SNP）插入片段的SMRT bell证明环形比对测序可以用于混合群体中稀有突变的检测。研究者将每个等位基因中获得的SMRT bell模板按照T/C比为0:100, 2.5:97.5, 5:95, 10:90, 25:75, 50:50, 及100:0的比例进行混合（图7A），然后使之形成聚合酶－模板复合体并进行环形比对测序。随后通过SMRT测序对混合比进行定量，并被标绘出‘T’等位基因的观察值和预期值的比例。将‘T’等位基因中获得的检测频度与起始混合物进行作图，可以很容易检测到低至2.5%的频度（图7B）。

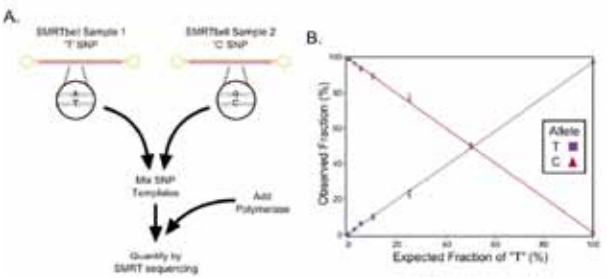

图7 SNP检测

4. 全转录本测序

相对于芯片而言，二代测序对于基因表达的研究，其灵活性和改进的数据丰度均已提高。然而，由于mRNA选择性剪接的频率和多样性，很多情况下，检测及定量都不足以完整描述一个特定基因的表达图谱。为了更好地研究某一特定基因的表达情况，一般需要单个转录本的更长的读长。利用PACBIO RS系统标准的SMRT测序方法，长的单分子读取结果产生于标准的SMRT bell模板，可跨越整个转录本的长度。

MCF7细胞是一个稳定的并已被深入研究的乳腺癌细胞系，含有大量差异基因表达和突变相关的可用数据。研究者采用MCF7细胞总RNA制成的cDNA文库，来评估一个标准SMRT测序和样品制备过程。在准备SMRT bell模板之前，起始mRNA样品首先被富集。标准SMRT测序的结果检测到大量不同的转录本及相关的剪接变体。研究者在MCF7细胞及扩散性导管乳腺癌中监测到上调的ARL6IP1基因的3个剪接变体（图8）。剪接变体包括标准形式及两种不常见的异形体。由图可知，用SMRT技术对MCF7的转录本测序时，在一个读长中即可跨越5’非翻译

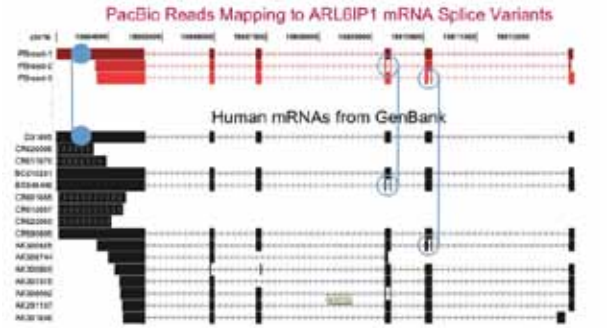

图8 全长转录本

区到3’非翻译区。常见的转录本(PRread-1 to D31885)和两种少见的异形体(PBread-2 to BX648445 and PBread-3 to AK300825)均能被检测到。由此可见，PacBio的长读长特点令其可以跨越整个转录本的长度，从而可以对特定基因

的表达详情进行完整描述，这些特点都体现其在基因组学研究中的应用价值。

七、PACBIO RS系统在转化医学应用

将测序信息整合到诊断和其他临床应用中，通过更好地检测并描述传染病和其他疾病的状态，来实现对病人护理方案的及时调整和改善。为达到最大功效，诊断测序需要快速地产生分析后的测序数据，最好在一天内完成。通过加速样品制备，减少仪器运行时间和快速分析测序数据，PACBIO RS系统的SMRT测序可在一天内从带有病毒的拭子样品中得到测序结果，可完全适合快速的临床样品检测。

要对病毒分离物或其他临床样品进行测序，可以使用两种快速的样本处理策略来产生DNA模板（图8A）。如果目标序列明确，比如是一个特异的病毒，可以使用一个模板富集步骤（比如反转录酶或PCR分别用于RNA或DNA病毒）来产生线性双链DNA模板前体。如果通过RT-PCR从起始病毒RNA中得到cDNA，可以使用标准的SMRT bell制备流程，如果使用特异修饰的引物扩增cDNA的RT-PCR时，可以通过核酸外切酶产生对于SMRT测序引物特异的3’末端突出的线性模板（图9A）。特异修饰的引物，其结构包含特异于目标的序列，即一个已知的位于5’末端的通用序列，和3’区对于病毒靶点是特异的且内部含有硫代磷酸修饰（图9B）。cDNA合成之后，双链DNA被一个5’-3’的核酸外切酶处理，从引物的5’末端降解直到遇到硫代磷酸修饰。然后，单链形成的区域特异的测序引物被复性并被用于定位DNA聚合酶到完整DNA模板的末端。

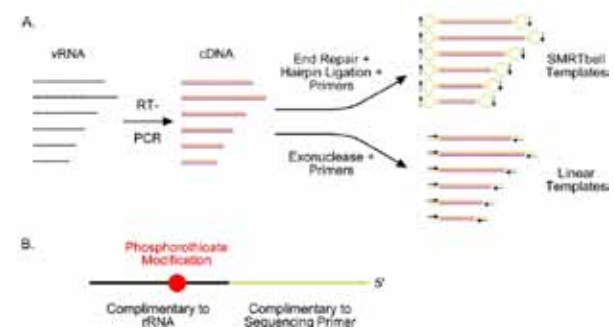

图9 临床检测时获得SMRT测序母板的两种方法

由于无论选择哪一种方案，都可以在9小时内完成测序中的样品提取和准备并产生分析过的测序数据（图10）。图9所示的两种方法现均已被用于检测人类口腔拭子样品上分离的多个病毒株。对于任一个方法，都没有覆盖率或样品大小的偏好，且很容易根据几个SNPs来区分多个病毒株。由此可见，PACBIO RS系统凭借其可在一天内快速完成样本制备工作并获取临床样本的分析后的测序数据的独特优势，实现其在转化医学方面的应用。

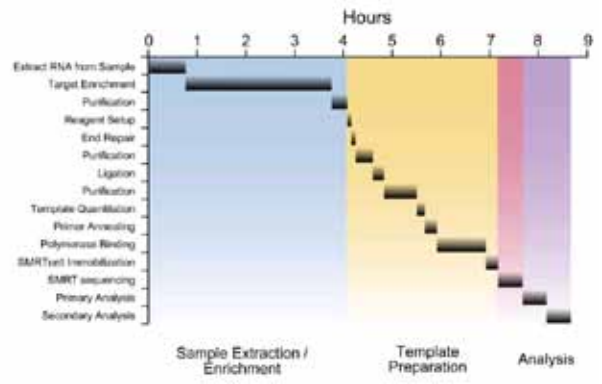

图10 临床检测从SMRT bell制备到测序完成并获得测序数据的时间表

八、展望

除了目前的三种SMRT测序方法（标准，频闪和环形比对）的相关应用，目前PacBio还在积极的进行几种不同的应用研究，比如直接RNA测序的研究。由于SMRT测序技术的核心是黏附于ZMW内的DNA聚合酶的活性位点，利用分子固定的灵活性和荧光检测的本质，通过用一个逆转录酶和RNA模板替换DNA聚合酶和DNA SMRT bell，已经获得了一些前瞻性的初始数据。目前已将逆转录聚合酶加入到PacBio的诱变规划中来改善单分子测序的性能。希望当认证和发售后，直接对RNA模板进行长读长测序的能力将为目前的测序方法提供独特的能力扩展。这种方法因为消除了目前cDNA制备的需求，所以能够降低测序偏向性并减少获得结果所需的成本和时间。相信在不久的将来，PacBio RS可以充分利用其测序灵活性，通过对试剂的不断优化，在不改变仪器硬件的基础上实现更多新的应用，从而进一步提升其在基因组学及转化医学领域的重要性，并可为生物学过程的调控机制研究提供更多新的观点。