SRA 数据库帮助

2008 年 11 月 17 日

1. 简介

近年来，随着科技的进步，新一代大规模平行测序技术诞生了，如 454、Solexa 和 HeliScope 等。这些测序技术可以同时对大量的短片段测序，由于其数据的复杂性及结果的高通量性，使原有的数据库不能适应新的测序结果。而一些小型实验室自身也不具备处理和管理这些复杂数据的能力。因此生命数据中心创建了 SRA 数据库，帮助用户管理这些测序数据，同时有助于科研界共享数据。

SRA 与 Trace 最大的区别是将实验数据与元数据分离。元数据现在可以划分为以下几类。

• Study--study 包含了项目的所有 metadata，并有一个 NCBI 和 EBI 共同承认的项目编号（universal project id），一个 study 可以包含多个实验（experiment）。

• Experiment--一个实验记载实验设计（Design），实验平台（Platform）和结果处理（processing）三部分信息，并同时包含多个结果集（run）。

• Run--一个结果集包括一批测序数据。

• Submission--一个 study 的数据，可以分多次递交至 SRA 数据库。比如在一个项目启动前期，就可以把 study，experiment 的数据递交上去，随着项目的进展，逐批递交 run 数据。LSBI 采用了“项目”和“批次”的数据递交管理单位，study 等同于项目，submission 等同于批次的概念。

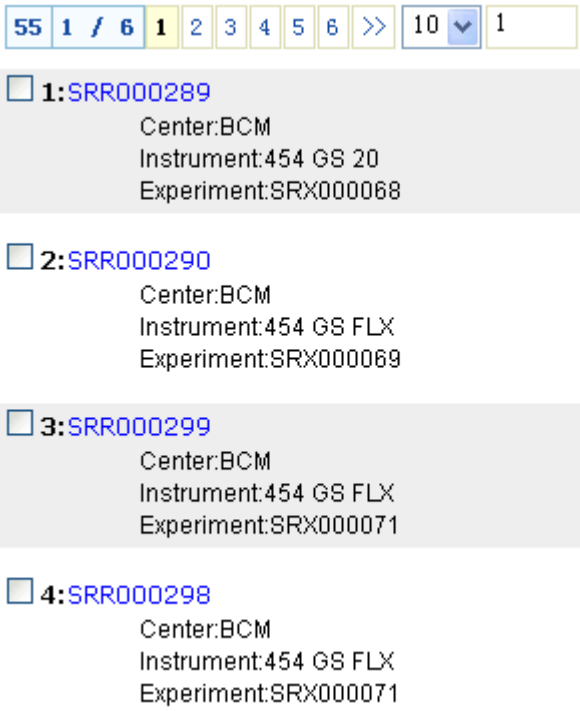
2. 数据库查询与结果展示

左侧菜单栏中，点击 Run、Sample、Experiment、Study 可以进行相关内容的高级检索。SRA 数据库的高级检索可以最多使用三个限定词来进行更精确的检索，三个限定词之间可以用“AND”和“OR”相连接，其中“AND”表示查询的结果中必须包含它所连接的两个关键词，“OR”表示查询的结果中至少包含它所连接的关键词中的一个。

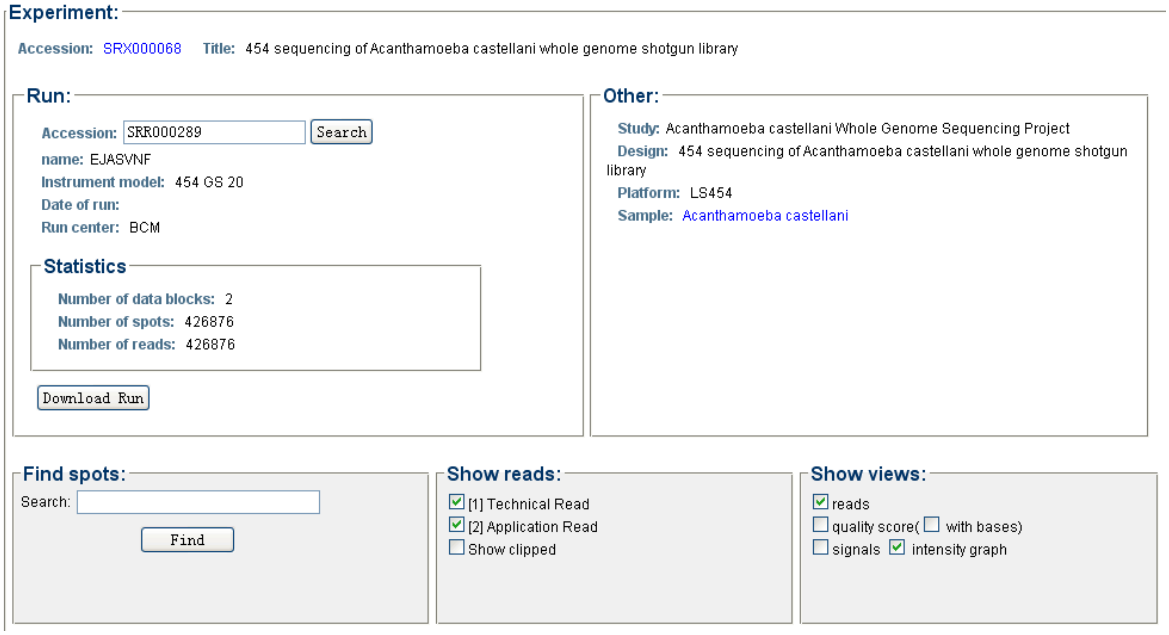
2.1. Run 数据搜索与结果展示

点击左侧菜单栏进入 Run 界面后，再点击下拉菜单，可以看到 4 种限定词，CAC、INSTRUMENT MODEL、CENTER 和 EXPERIMENT CAC，即 Run 的 AC 编号、仪器型号、测序中心缩写和实验的 AC 编号。

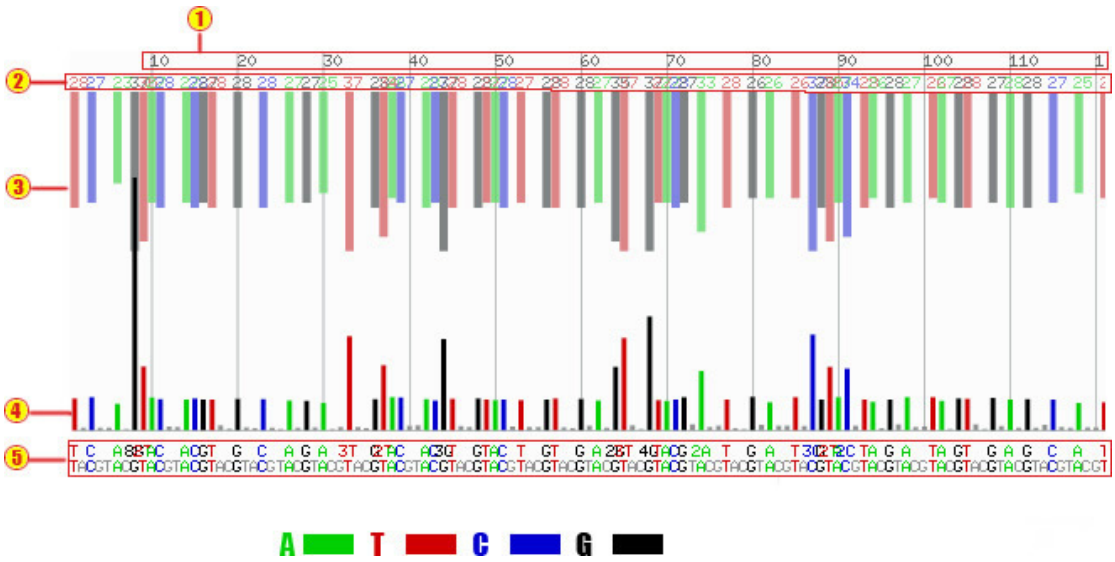
选择 CAC，输入"SRR000001"查询后，可见如下界面，显示了 Run 的摘要信息，如 AC 编号、测序中心、仪器型号和实验的 AC 编号。



单击 SRR000001 可以看到详细的测序信息，包括基本信息、互交区和测序结果图谱



测序结果图谱说明



1）碱基位置序号刻度，以 1 0 为单位显示
2）对应位置碱基测序质量打分的值，只显示被读出的碱基对应的数值
3）质量得分直方图，对应 2）中所示碱基
4）色谱值直方图，同样对应 2）中所示碱基
5）分为上下两行。上行为被读出碱基，下行显示所有有信号碱基

2.2. Sample 数据搜索与结果展示

点击左侧菜单栏进入 Sample 界面后，再点击下拉菜单，可以看到 2 种限定词，CAC 和 TAXON ID，即 Sample 的 AC 编号和分类编号。

选择 CAC，输入"SRS000001"查询后，可以看到 Sample 的摘要信息，如 AC 编号、通用名、分类编号号。

单击 SRS000001 可以看到详细的样本信息

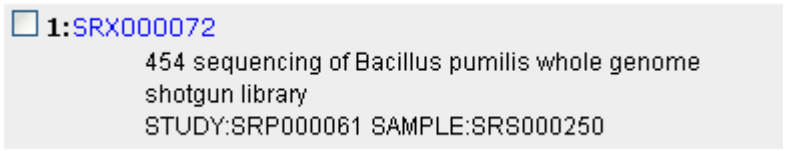
1: SRS000249

Accession	SRS000249
Sample Name	Apis mellifera
Tax ID	7460
Experiments	SRX000071

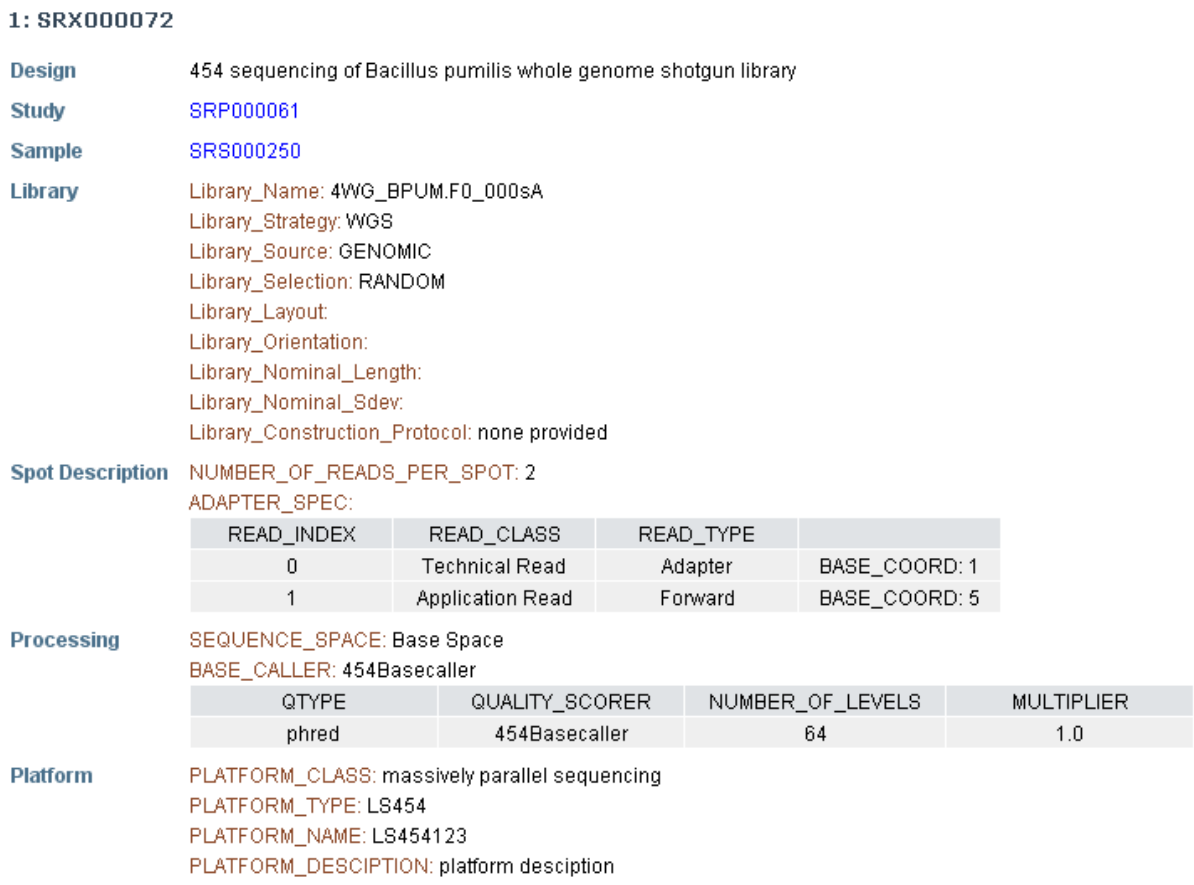
2.3. Experiment 数据搜索与结果展示

点击左侧菜单栏进入 Experiment 界面后，再点击下拉菜单，可以看到 4 种限定词，CAC、DESIGN DESCRIPTION、STUDY CAC 和 SAMPLE CAC，即 Experiment 的 AC 编号、实验设计、Study 的 AC 编号和 Sample 的 AC 编号。

选择 CAC，输入"SRX000001"查询后，可见如下界面，显示了 Experiment 的摘要信息，如 AC 编号、实验设计、Study 的 AC 编号和 Sample 的 AC 编号。



单击 SRX000001 可以看到详细的实验信息

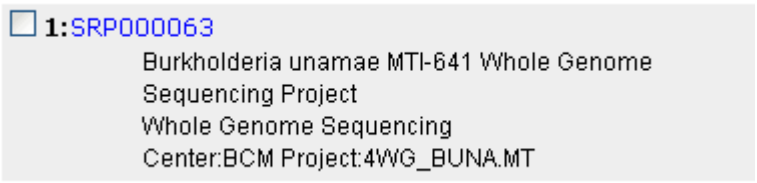


Processing 图标框中从左至右依次为打分类型、打分器、分级数量和放大倍数

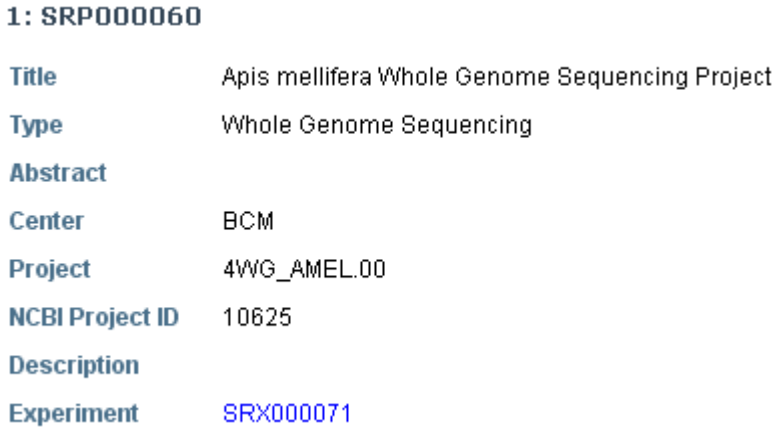
2.4. Study 数据搜索与结果展示

点击左侧菜单栏进入 Study 界面后，再点击下拉菜单，可以看到 5 种限定词，CAC、STUDY TITLE、STUDY TYPE、CENTER NAME 和 PROJECT ID，即 Study 的 AC 编号、课题名称、课题类型、测序中心名称和项目编号查询。

选择 CAC，输入"SRP000001"查询后，可见如下界面，显示了 Study 的摘要信息，如 AC 编号、课题名称、测序中心、课题类型、项目和链接。



单击 SRP000001 可以看到详细的课题信息



3. 数据下载

LSBI 的 SRA 数据库提供如下两种数据下载方式：

1）下载打包的核酸数据

点击“链接”，你可以 FTP 下载我们发布的所有 SRA 数据。

2）下载查询到的核酸序列数据

在查询结果的页面，点击“下载”按钮，即可下载你所查询到的 SRA 数据。

4. 数据提交

一个 SRA study 所包含的内容，应该在一个 LSBI 的项目中提交。即 SRA study 和 LSBI 项目为 1 对 1 关系。一个 study 的内容可以在一个项目下，分成几个批次提交，每次提交不同的内容。

一个批次的 SRA 数据，包括一个.info 文件和一个名为 DATA，装有提交原始文件的子文件夹。子文件夹中内容为描述 metadata 的 xml 文件或者 sff 等格式的数据文件。一个完整的 study，包括一个或多个 study.xml, experiment.xml, sample.xml 和 run.xml，以及一个或多个数据文件。但是一个批次的提交数据不一定包括所有的文件。
Run.xml 和该 xml 中包括的所有数据文件，必须要在一个批次中提交。

1)请先确认您已是数据中心网站注册的用户，否则请登陆中心网站,注册。
2)登陆中心网站后,点击左侧菜单的 mydata,选择已有项目或创建新项目。
3)选择已有批次或创建新批次。在创建批次时，选择要提交的数据类型为“SRA”。
4)在点击批次下的 submit data 按钮后，进入提交页面。
5)首先下载离线提交附件（subdesc.bch），作为离线提交的标识文件，是离线提交必须的附件之一。
6)按照 SRA 的数据格式标准，处理生成数据文件，连同标识文件一起，通过提交页面上显示的路径（为 ftp://lifecenter.sgst.cn:2121/SRA/****/）上传至服务器指定目录。

5. 数据格式

5.1. Info 格式

Tag 名称	内容描述	必填	备注
NAME:	提交人的姓名	是	
EMAIL:	E-mail 地址	是	
LAB:	实验室		
INST:	机构名		
FAX:	传真		
TEL:	电话	是	
ADDR:	地址	是	
COUNTRY:	国家	是	
FILENAME	DATA 子文件夹下的文件名称	是	
FILETYPE	该文件的类型	是	可取值: study | experiment | sample | run | data
CHECKSUM_MET		是	可取值:MD5

HOD			
CHECKSUM		是	
COMMENTS	用户的说明		

5.2. Sff 格式

sff 文件格式是专门设计用来记录 454 原始数据的，以下以一个具体的 sff 文件中开头的一部分为例，对此格式进行简要的说明，红色文字为说明文字。

以下为 Common Header Section:
magic_number: 2e736666　　固定：2e736666，表示 sff 文件
version　　:　　固定：0001，为 sff 目前的版本
index_offset :216448320
index_length :2540212　　文件索引的 length 和 offset
number_of_reads :127010　　read 的数目
header_length :440　　header 的长度
key_length :4　　key 长度，见 key_sequence 字段
number_of_flows_per_read :400
每个 reads 的 flow 数，一个 flow 测 ATGC 四种碱基的一种

flowgram_format_code :1
目前只有一种，所以固定为 1。这种 code 把每个 flowgram_value 写成一个整数，实际数值为 flowgram_value/100
flow_chars :TACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACG
TACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTAC
GTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTA
CGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACG
TACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTAC
GTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACG
每个 flow 测的碱基，该字段长度应等于 number_of_flows_per_read 的值

key_sequence :TCAG　　key 序列，即每个 reads 开始固定的序列
以下为 Read Section，每个 read 一个
read0info----------　　标示 0 号 read 的开始
read_header_length: 32　　Read header section 的长度
name_length: 14　　该 read 的 Name 的长度，见 name 字段
number_of_bases: 158　　碱基的数目
clip_qual_left: 5
clip_qual_right: 111
clip_adapter_left: 0
clip_adapter_right: 0

name: EM7LHVR02GV4ED　　该 read 的名字

flowgram_values: 101 11 108 19 11 92 13 916 34 16 90 121 16 294 18 102 185 22 11 104 27 101 15 15 96 20 182 13 15 98 18 13 94 17 111 102 189 21 15 192 18 293 16 15 182 204 16 11 186 202 104 11 95 20 93 349 25 16 178 110 16 95 15 99 13 19 187 97 15 11 188 16 365 17 99 12 21 295 206 17 104 17 14 102 18 104 17 11 185 99 15 14 371 18 15 97 102 15 104 101 190 17 16 180 103 20 11 97 21 99 18 16 111 17 188 109 19 92 19 11 100 16 190 195 16 15 195 187 98 15 15 92 188 18 101 103 19 12 100 10 98 17 14 100 20 106 13 100 15 107 104 17 11 100 100 106 13 9 95 12 13 228 104 18 13 84 11 409 20 107 16 10 104 104 17 222 17 10 107 5 11 254 16 210 19 11 106 5 14 109 102 18 10 104 99 17 11 95 12 420 18 108 15 5 111 108 13 217 21 12 11 30 13 17 11 18 12 15 9 14 10 14 10 13 9 16 11 13 11 10 9 12 13 13 9 12 13 12 9 10 12 11 9 8 12 12 9 11 11 13 9 8 11 10 11 9 11 10 11 7 12 11 11 8 10 10 8 9 8 14 8 9 9 12 8 8 8 14 8 9 9 12 7 9 9 13 8 8 11 9 9 7 10 12 8 7 11 13 9 8 9 13 8 7 9 11 8 7 10 13 8 7 9 12 7 8 10 16 8 7 10 19 8 8 10 12 9 8 9 11 8 6 10 13 8 7 12 13 8 8 15 15 8 8 10 12 8 8 10 11 8 7 10 12 8 8 9 12 9 8 9 14 9 7 10 12 9 9 8 13 8 9 7 13 10 8 9 11 8 10 8 10 8 8 10 12 9 8 9 10 9 8 9 12 9 8 9 10 9 8

每个 flow 的读数，该字段应该包含 number_of_flows_per_read 个数字

flow_index_per_base: 1 2 3 2 0 0 0 0 0 0 0 0 3 1 2 0 0 2 1 0 3 2 3 2 0 3 3 2 1 1 0 3 0 2 0 0 3 0 1 0 3 0 1 0 1 2 2 1 0 0 3 0 1 2 2 3 0 1 3 0 2 0 0 0 2 3 0 0 1 0 2 3 2 3 0 1 3 0 0 0 3 1 2 1 1 0 3 0 1 3 2 3 2 0 1 2 3 2 0 1 0 3 0 1 0 1 3 1 0 2 1 3 2 3 2 2 2 1 3 1 1 3 3 0 1 3 2 0 0 0 2 3 1 2 0 3 3 0 0 2 0 3 3 1 3 1 3 2 0 0 0 2 3 1 2 0 2 5

测出序列每个 base 在 flowchart 上对应的位置，比如该条数据的 flow-chars 为：TACGTACGTAC…

flow_index_per_base 为：1 2 3 2 0 0 0 0 0 0 0 0 3…

则序列为第一位的 T，第 1+2=3 位的 C，第 3+3=6 位的 A，第 6+2=8 位的 G，第 8+0=8 位的 G…即序列为 TCAGG…

bases:TCAGGGGGGGGGCGAAAGTTGATCCATCGTTGGAAATTAATTAACTCGGGCCGA
GCCGCCTTTTCAAACCTGATTATTTTGTCGTTGGTGATCCGATCCGGCCGGTGTTCGCT
GAGACACGCAACAGGGGATAGGCAAAGGCACACAGGGGATAGGNN

测出的序列，长度应等于 number_of_bases 的数值

quality_scores: 28 27 26 36 29 20 15 11 8 6 3 1 25 21 36 32 13 28 31 23 28 27 27 30 22 27 26 26 28 32 24 33 25 36 32 13 30 21 34 28 31 23 34 27 28 27 26 36 32 17 29 20 26 27 27 31 24 27 32 24 36 32 18 2 27 36 32 14 35 28 28 28 28 31 23 27 36 32 19 3 27 28 28 28 32 25 29 21 28 27 27 26 32 24 26 26 27 32 25 33 26 33 26 31 24 27 26 32 24 28 28 27 27 27 27 27 27 28 27 27 27 27 33 27 28 23 39 35 22 9 27 28 28 34 28 27 30 26 2 35 28 27 26 28 28 27 26 39 35 23 10 27 26 27 34 28 0 0

每个碱基的 quality score，长度应等于 number_of_bases 的数值

Read1info----------　　标示另一个 read，1 号 read 的开始

6. 常见问题

1.有其它存储 SRA 数据的地方吗？

有，NCBI 和 Ensembl 都存放了 SRA 数据。我们的数据库正在建设中，目前只存储了部分数据信息。

2.ftp 多久更新一次？

每周我们都会把所有 SRA 的数据放到 ftp 上，供用户下载。

3.什么是 RCF 格式？

RCF 是 Relieved Compress Format 的缩写，和存放在服务器上的数据内容一致。为了减少占用的磁盘空间和计算时间，该文件格式经过了一系列详细的测试后，投入使用。服务器在载入数据的时候会将它重新处理和压缩，所有的色谱数据都存放在这种 RCF 文件中。RCF 包含两种算法：derivation 和 Huffman，压缩比较高，同时比较简单，不需要占用大量的系统资源。

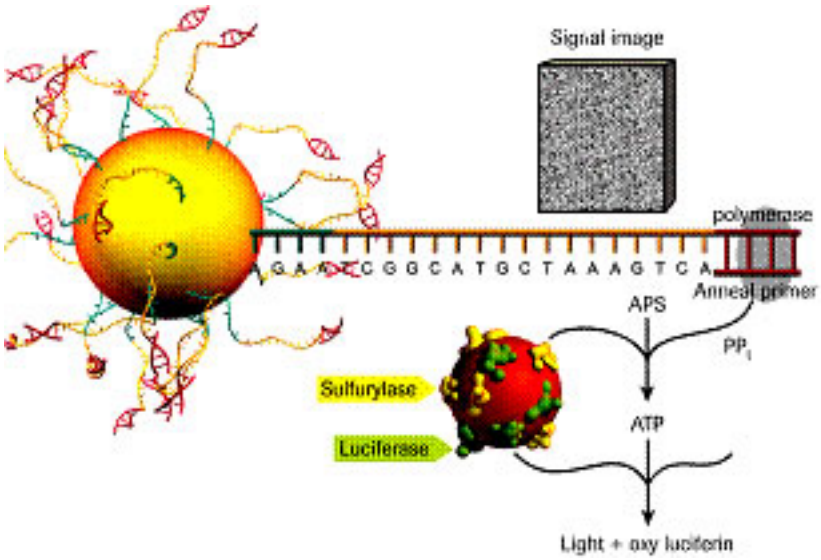
7. 附录

7.1. GS FLX 系统超高通量测序

2005 年底，罗氏诊断公司和 454 公司推出了基于焦磷酸测序法的超高通量基因组测序系统——Genome Sequencer 20 System (GS 20)。自问世以来，已经被世界上几乎所有从事基因组测序和相关结构功能研究的顶级实验室配备使用，Nature，Science，PNAS，Cell 等世界知名期刊上发表的文章就已经近五十篇。在已有 GS 20 的技术基础上，罗氏诊断公司和 454 公司不断加大创新投入，又于 2007 年推出了通量更大，读长更长，准确性更高的第二代基因组测序系统——Genome Sequencer FLX System (GS FLX)，相信它的出现，必将掀起测序技术的进一步革命；对整个基因组学的研究将产生巨大的推动作用。

GS FLX 系统超高通量测序技术原理

GS FLX 系统的测序原理和 GS 20 一样，也是一种依靠生物发光进行 DNA 序列分析的新技术；在 DNA 聚合酶，ATP 硫酸化酶，荧光素酶和双磷酸酶的协同作用下，将引物上每一个 dNTP 的聚合与一次荧光信号释放偶联起来 (图 16)。通过检测荧光信号释放的有无和强度，就可以达到实时测定 DNA 序列的目的。此技术不需要荧光标记的引物或核酸探针，也不需要进行电泳；具有分析结果快速、准确、灵敏度高和自动化的特点。



GS FLX 系统的操作过程

GS FLX 系统提供了完整的从样品制备到后续的生物信息学分析解决方案。

1）样品种类：GS FLX 系统支持各种不同来源的样品序列测定：包括基因组 DNA，PCR 产物，BAC，cDNA，小分子 RNA 等等。

2）样品 DNA 打断：样品例如基因组 DNA 或者 BAC 等被打断成 300－800 bp 的片段；对于小分子的非编码 RNA，这一步骤则不需要。短的 PCR 产物则可以利用 GS 融合引物进行扩增后直接进行步骤 4)的工作。

3）衔接子连接：借助一系列标准的分子生物学技术，将 A 和 B 接头（3’和 5’端具有特异性）连接到 DNA 片段上。接头也将在后继的纯化，扩增和测序步骤中用到。图中仅仅显示了后续步骤中要用到的单链的 DNA 片段。

4）一条 DNA 片段＝一个磁珠：接头使成百上千条 DNA 片段结合到它们自己唯一的磁珠上，此磁珠被单个油水混合小滴包被后，在这个小滴里进行独立的扩增，而没有其他的竞争性或者污染性序列的影响；整个 DNA 片段进行平行扩增。

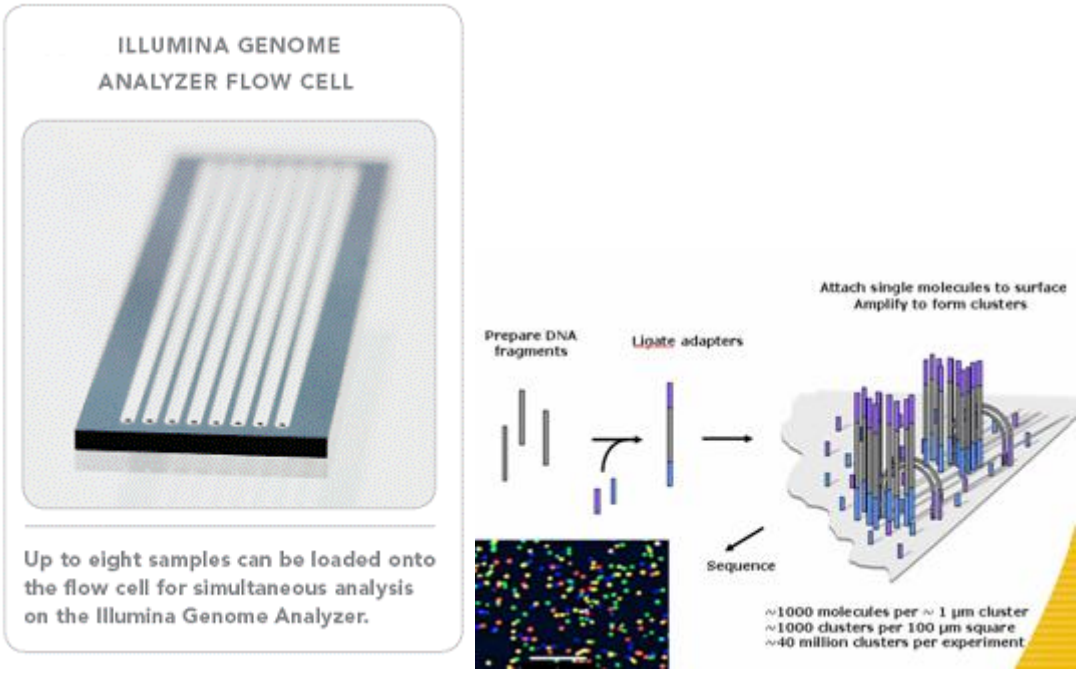
5）一个磁珠＝一条读长：经过 PCR 扩增后，每个磁珠上的 DNA 片段拥有了成千上万个相同的拷贝。经过富集以后，这些片段仍然和磁珠结合在一起，随后就可以放入到 PicoTiterPlate 板中供后继测序使用了。

6）数据读取和分析工具：GS FLX 系统 在 7.5 小时的运行当中可获得 40 多万个读长，读取超过 1 亿个碱基信息。GS FLX 系统提供三种不同的生物信息学工具对测序数据进行分析，适用于不同的应用：例如多达 3 GB 序列的重测序，对比已知参考序列进行的扩增产物差异分析，及 120 MB 的从头测序工作等。

7.2. Solexa 高通量测序法原理

Solexa 方法是利用单分子阵列测试 genotyping ，此种测序法首先是将 DNA 从细胞中提取，然后将其打断到约 100 － 200bp 大小，再将接头连接到片段上，经 PCR 扩增后制成 Library 。随后在含有接头的芯片（ flow cell ）上将已加入接头的 DNA 片段绑定在 flow cell 上，经反应，将不同片段扩增。在下一步反应中，四种荧光标记的染料应用边合成边测序（ Sequencing By Synthesis ）的原理，在每个循环过程里，荧光标记的核苷和聚合酶被加入到单分子阵列中。互补的核苷和核苷酸片断的第一个碱基配对，通过酶加入到引物上。多余的核苷被移走。这样每个单链 DNA 分子通过互补碱基的配对被延伸，利用生物发光蛋白，比如萤火虫的荧光素酶，可通过碱基加到引物后端时所释放出的焦磷酸盐来提供检测信号。针对每种碱基的特定波长的激光激发结合上的核苷的标记，这个标记会释放出荧光。荧光信号被 CCD 采集， CCD 快速扫描整个阵列检测特定的结合到每个片断上的碱基。通过上述的结合，检测可以重复几十个循环，这样就有可能决定核苷酸片断中的几十个碱基。

Solexa 的这种方法，可在一个反应中同时加入 4 种核苷的标签，采用边合成边测序（ SBS － sequencing by synthesis ），可减少因二级结构造成的一段区域的缺失。并具有所需样品量少，高通量，高精确性，拥有简单易操作的自动化平台和功能强大等特点，此反应可以同时检测上亿个核苷酸片断 ，因此在同一个芯片或几个芯片上花费很少（只需常规方法的 1 %）的成本就可测试全基因组。



随着 DNA 测序表达谱产品（ DNA sequencing ,expression profiling ）以及 microRNA 分析平台 -Solexa Genome Analysis System. 相继问世，使得此种方法在更多的领域得到应用。

7.3. HeliScope 测序技术简介

落于美国麻省剑桥（Cambridge）的 Helicos Biosciences，也许是最早致力于单分子测序技术研发的生命科学公司。公司总裁兼首席运营官 Steve Lombardi 说，Helicos 公司的单分子测序系统可以对 DNA 和 RNA 进行分析，对序列突变进行检测，并有可能用于实验胚胎学的检测。该系统计划 在 2008 年投入应用。“我们的目标是建立一个可用于大型实验的商业化遗传学分析系统。”

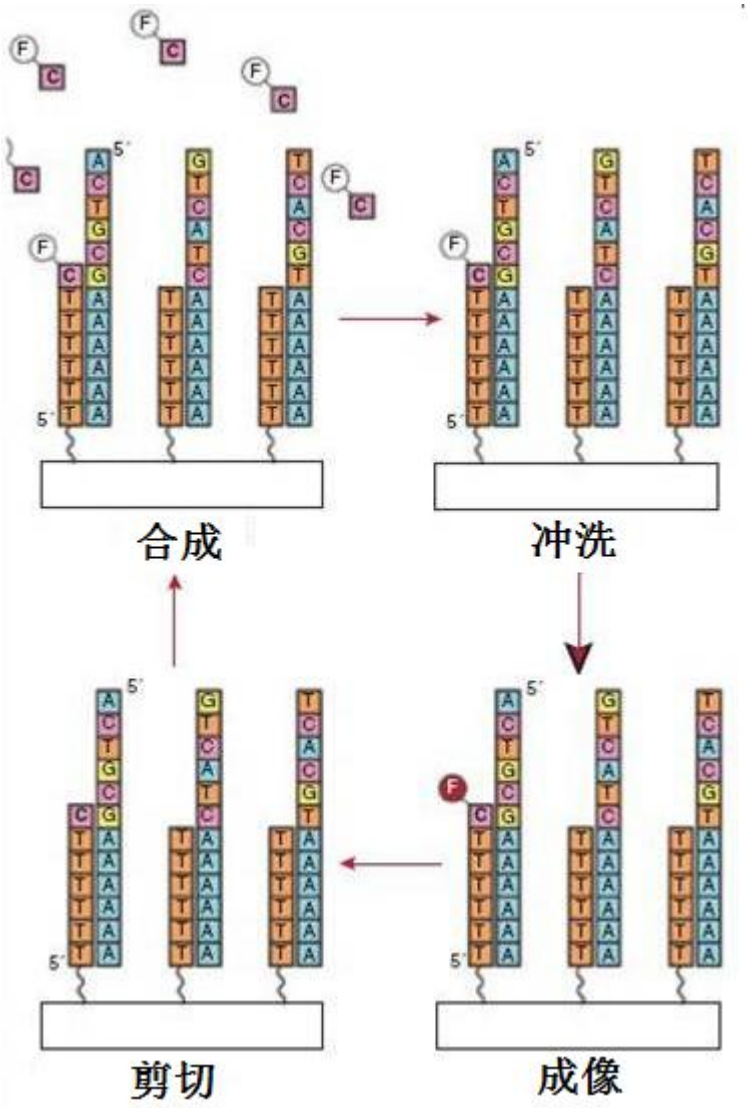
Helicos 的被称作 TRUE 单分子测序技术（true single-molecule sequencing, tSMS）的新型测序方法，是一种基于合成的单分子测序技术，测序过程在 HeliScope single-molecule sequencer（单分子测序仪）中完成。在样品制备中，将待测 DNA 链打断成 100–200 个碱基长度的片段，然后将已知序列的接头－通常为多聚 A （poly（A））尾巴，接到上述 DNA 片段末端。“我们可以在有 poly(T)引物共价相连的界面，检测到 poly(A)尾巴，”Harris 说。所检 测的片段间的间隔对于 tSMS 测序中的成像十分重要。“光学显微镜的分辨率大约在几百纳米左右，因此，待检测分子间的距离至少要达到这个数值，” Harris 解释说。

先将待测片段连接到界面上，然后将已标记的单核苷酸和聚合酶的混合物加至界面。标记核苷酸在聚合酶催化下按照碱基互补原则掺入待测片段中，之后，HeliScope 仪器上的照相机会对整个界面进行成像，并鉴别所有掺入标记核苷酸的片段。

tSMS 关键的一步是在以下过程中：将标记基团切掉，然后在聚合酶作用下再掺入另一个标记核苷酸。通过对这一步骤进行反复循环，加入四种核苷酸，该仪器在单次运行中，即可对上十亿的待测链进行测序，读长在 25 至 45 碱基之间。

对于 Helicos 的研发人员而言，对每一个循环进行同步成像是另一个需要攻克的难关。

“我们无法对整个界面进行一次成像，因此，在每一次循环中，我们需 要每隔几百毫米进行一次成像，”Harris 说。在测序过程中所获得的这一系列高分辨率的图像提供了相当数量的数据点。Heliscope 的每一轮检测需 要计算机 14TB（见文后小词典）的存储空间，但 14BT 的空间足以存储极大量的序列数据了，Harris 说；而成像系统有可能达到每小时处理 1GB 的序 列的速度。尽管目前的这个系统还远没有达到这个标准，但 Harris 和 Lombardi 都表示，该系统在今后化学测序方面的发展拥有极大空间。



Helicos Biosciences 的单分子测序技术示意图