文章编号: 1000-1336(2010)06-0959-05

NCBI 高通量测序数据库 SRA 介绍

熊筱晶

上海工程技术大学化学化工学院, 上海 201620

摘要：随着新一代测序技术的发展，高通量测序技术的应用越来越广泛，其产生的海量数据的存储、查询需要专门的数据库辅助，NCBI 的 SRA(Sequence Read Archive)数据库是高通量测序存储的代表，本文对 SRA 数据库的组织架构，数据形态作了综述分析，并对其存贮的数据进行了总结。

关键词：高通量测序；SRA 数据库

中图分类号：Q819

以Illumina/Solexa技术、Roche/LS454技术、ABI/SOLID 技术及 HELICOS 单分子测序技术为代表的新一代测序技术(next-generation sequencing technology)真正实现了高通量测序(massively parallel sequencing/high through-put sequencing)[1]。随之产生了海量的实验数据，单个 run 产生的数据以 GB 乃至数十 GB 计，除此之外，实验样本等 meta 信息也需要与序列数据整合。高效率的数据存储、提取乃至共享成为高通量测序数据分析必不可少的环节。

在美国国立生物技术信息中心(NCBI)的诸多数据库中，传统测序数据(如毛细管电泳产生的测序数据)的存储有 Trace Archives 数据库，但不适合存储高通量测序数据；GEO 数据库用于存储高通量的芯片实验数据，在 SRA 未建立之前，GEO 数据库也用于存储高通量测序数据，但随着高通量测序数据的累积，专门用于存储此类数据的需求越来越迫切，NCBI 在 2007 年底推出了 SRA 数据库，用于存储、显示、提取和分析高通量测序数据。SRA 数据库，最初的命名为Short Read Archive,现已改为Sequence Read Archive，自建立之初，序列数据迅速累积，涉及多平台，多物种，多种应用的，分层次的 SRA 数据库已初具规模。

收稿日期：2010-06-20

上海工程技术大学科技发展基金项目(A-2500-10-01005)资助

作者简介：熊筱晶(1979-)，女，硕士，讲师，通讯作者，E-mail:luzifer.limm@gmail.com

1. SRA 数据库的组织架构

1.1 meta数据

meta 数据是指与测序实验及其实验样品相关的数据，如实验目的、实验设计、测序平台、样本数据(物种，菌株，个体表型等)，在SRA数据库中，meta 数据分如下层次来存储：

(1)研究课题(study)。在 SRA 数据库中，研究课题的检索号(accession number)以前缀 DRP，ERP 或 SRP 开头。一个研究课题致力于一个特定的研究目的，由一个或多个测序中心来完成，往往是某个基因组计划(genome project)的项目,有特定的研究类型(如全基因组测序，转录组分析，宏基因组学分析等)。包含一个或多个实验。研究课题的详细信息可以通过http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?study=xxx(xxx为研究课题的检索号)来查询。(2)样本信息(sample)。样本的检索号以前缀 DRS，ERS 或 SRS 开头。样本信息可以包括物种信息、菌株(品系)信息、家系信息、表型数据、临床数据，组织类型等。样本信息可以通过 http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?sample=xxx(xxx为样本的检索号)来查询。(3)实验信息(experiment)。实验的检索号以前缀 DRX，ERX 或 SRX 开头。实验是 SRA 数据库的最基本单元，就像 PubMed 数据库的每一篇文献是 PubMed 数据库的基本单元一样。一个实验隶属于某个研究课题，对一个或多个样本进行测序，产生的测序数据以 runs 的形式存储于 SRA。实验信息可以通过

http://www.ncbi.nlm.nih.gov/sites/entrez?db=sra&report=full&term=xxx(xxx为样本的检索号)来查询。

1.2 序列数据

包括序列及其质量信息等，在 SRA 数据库中以 run 为单元存储。run 的检索号以前缀 DRR，ERR 或 SRR 开头。一个实验可以包含一个或多个 runs。run 的信息可以通过http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=viewer&m=data&s=viewer&run=xxx(xxx为 run的检索号)。

2．SRA 数据库的现有数据

截至 2010 年 4 月 13 日，SRA数据库中共包括 16,795 个实验，隶属于 1866 个研究课题，涉及的测序样本共有7913个,共包含41,179个runs,包含了3.23×10^{13}数目的碱基，约为人类基因组的大小的一万倍。图 1 描述了自 SRA 数据库自开始接受提交测序数据以来实验数目的增长情况。

分布于世界各地的研究机构向 SRA 提交数据，图 2 统计了高通量测序实验的主要的完成机构，我国的北京华大基因研究中心(BGI)在其中也占有一席之地。

截至目前，SRA 数据库中的测序数据来自四个测序平台，分别为：Roche_LS454(4965 个实验)，Illumina(11,549个实验)，ABI_SOLID(272个实验)和 HELICOS(9 个实验)。值得注意的是 HELICOS 单分子测序技术的应用，从早期的基因组大小较小的酿酒酵母的转录组分析，到家鼠(*Mus musculus*)的转录组分析，乃至 2009 年 9 月用单分子测序技术低成本地实现人类个体的全基因组测序。

目前 SRA 数据库已涵盖了超过 1500 种物种的测序数据，表 1 列举了排名前 15 位的物种。人、果蝇、小鼠、线虫、酵母、拟南芥和水稻等传统模式生物仍然是研究的热点，此外，宏基因组学研究也占了测序研究的一大部分，如海洋微生物宏基因组学研究和人体肠道微生物元基因组学研究。

3．从 SRA 看高通量测序的应用

统计现有的 1866 个研究课题的研究类型，结果如表 2 所示。

全基因组测序方面，最著名的全基因组测序的项目莫过于国际千人基因组计划(1000 Genomes Project)[3],如1000 Genomes Project Pilot 1(对180个Hapmap个体的低覆盖率测序),1000 Genomes Project Pilot 2(对两个三体家系,共六个个体的高覆盖率测序)以及对诸如Han Chinese, Japanese, Yoruba, Toscan, Luhya 等不同种族的个体的测序项目。由美国国家人类基因组研究院(NCI/NHGRI)主导的国际肿瘤基因组计划(The Cancer Genome Atlas Project, TCGA)[4]则是癌基因组学研究方面的范例。宏基因组学方面，受到广泛关注的是人体肠道微生物元基因组研究(Human Gut Microbiome Initiative)[5],国际人类微生物组(Human Microbiome Project)[6]以及国际海洋微生物普查计划(International Census of Marine Microbes, ICoMM)。

重测序方面的应用包括发现 SNP 或大的基因组变异(structural variant)等,如:Parkhomchuk等[7]利用

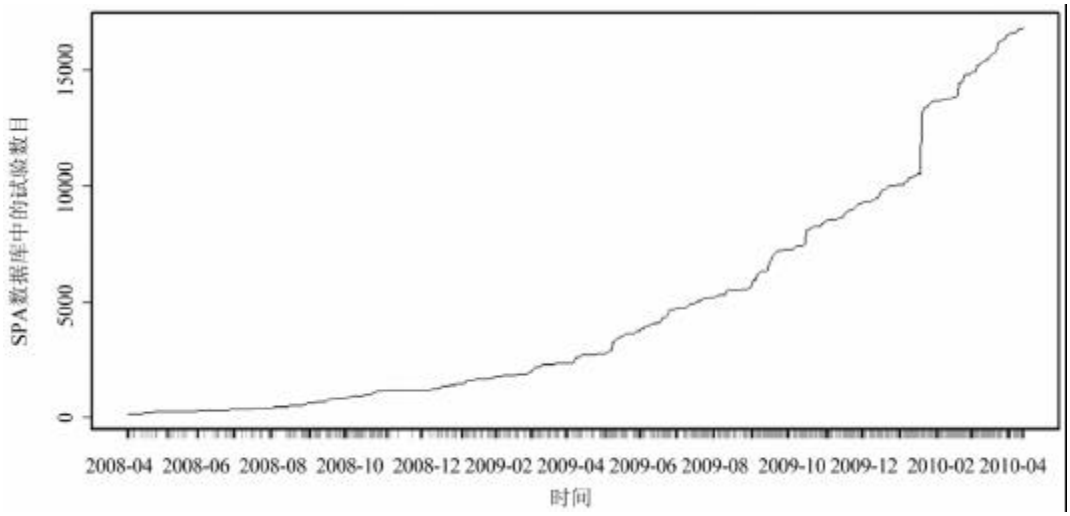

图 1　SRA 数据库中的实验数目

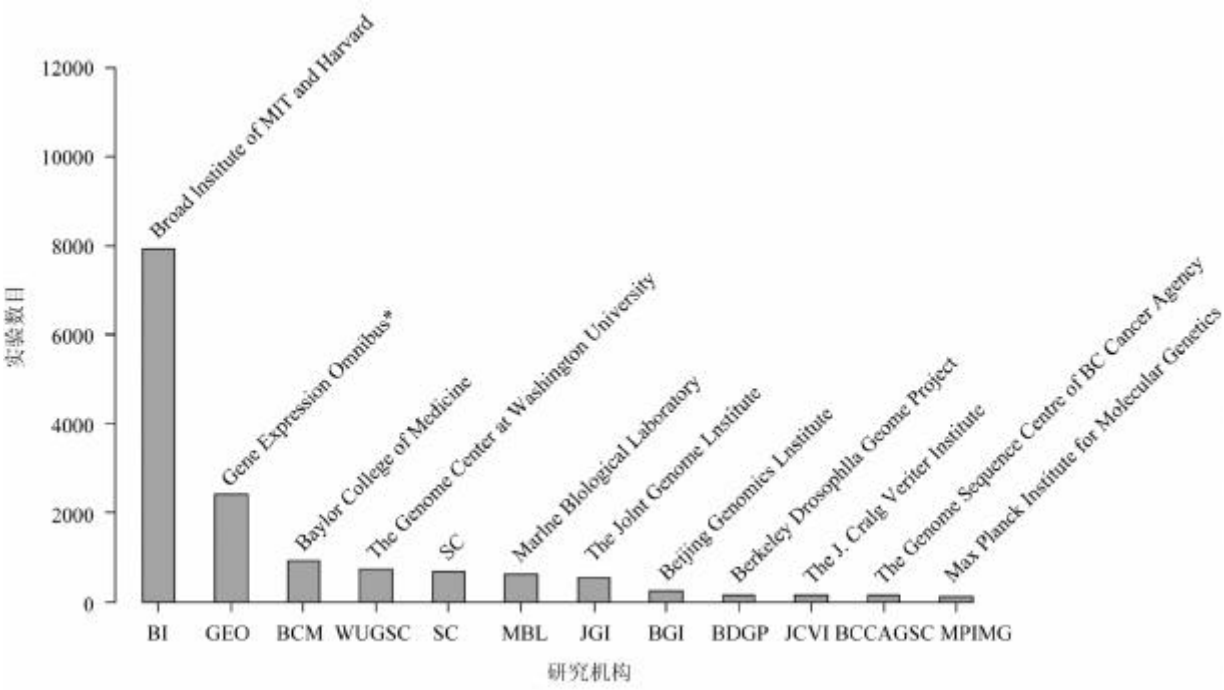

图 2　SRA 数据库中的实验完成机构

*GEO 为 NCBI 的另一数据库。

表 1　SRA 数据库中的测序数据所对应的物种

Taxonomy_id	物种	实验数目	样本数目
9606	人(*Homo sapiens*)	8118	2595
7227	果蝇(*Drosophila melanogaster*)	768	548
408172	海洋微生物宏基因组(marine metagenome)	701	682
10090	小鼠(*Mus musculus*)	498	441
32630	构建物(synthetic construct)	280	12
69293	刺鱼(*Gasterosteus aculeatus*)	252	12
30611	婴猴(*Otolemur garnettii*)	186	16
1280	金黄色葡萄球菌(*Staphylococcus aureus*)	180	72
6239	线虫(*Caenorhabditis elegans*)	175	167
4932	酿酒酵母(*Saccharomyces cerevisiae*)	133	112
3702	拟南芥(*Arabidopsis thaliana*)	124	124
1773	结合杆菌(*Mycobacterium tuberculosis*)	115	82
367110	链孢霉(*Neurospora crassa* OR74A)	106	2
4530	水稻(*Oryza sativa*)	77	77
408170	人体肠道微生物元基因组(human gut metagenome)	76	38

化学诱变剂处理大肠杆菌并进行重测序，从全基因组范围内研究变异的发生情况，发现变异在基因组中并非随机分布。Ng 等[8]对极少的个体(总数为 12 个，8 个不同种族的 Hapmap 个体和 4 个无亲缘关系的 FSS 疾病——一种孟德尔显示遗传疾病——的受累患者)的300 Mb基因编码外显子组(exomes)进行重

表 2　SRA 数据库中的研究课题

研究类型	研究课题的数目
全基因组测序(Whole Genome Sequencing)	1032
转录组分析(Transcriptome Analysis/RNASeq)	344
宏基因组学研究(Metagenomics)	186
表观遗传学研究(Epigenetics)	144
重测序(Resequencing)	38
群体遗传学(Population Genomics)	14
基因调控表达研究(Gene Regulation Study)	12
癌基因组学研究(Cancer Genomics)	7
合成生物基因组学(Synthetic Genomics)	1
其他(Other)	88
总计	1866

测序成功定位了候选基因。

群体遗传学研究方面采用比较基因组学的方法来研究群体或进化遗传学的问题。如，Hohenlohe 等[9]测序比较了分别来自海洋的两个和源自淡水的五个无鳞甲三刺鱼(*Gasterosteus aculeatus*)群体，鉴定了45,000个SNP(single-nucleotide polymorphism)，评估其群体遗传多样性与分化现象。Rubin 等[10]测序比较了家鸡和红色原鸡群体，鉴定了 7 百万个 SNP，推断在家鸡驯化过程中具有选择性清除(selective sweeps)作用的染色体位点。

转录组分析方面，高通量测序可以应用于发现新的转录物、研究可变剪切，定量比较不同条件下个体的转录情况、根据转录组的表达确定多种微生物的系统发生树等。Sharma 等[11]利用原创的 dRNA-seq技术对幽门螺杆菌(*Helicobacter pylori*)的所有转录物的 5′端进行高通量测序，确定转录起始位点和多顺反子 / 操纵子结构，揭示了幽门螺杆菌复杂的转录机制，与此同时，还鉴定了 60 种左右的小 RNA。Tuch 等[12]利用 SOLID 测序研究比较了三个口腔鳞癌患者和正常对照的转录组差异和等位基因特异性表达(allelicspecificexpression)情况。

表观遗传学研究方面，Ruike 等[13]将甲基化免疫共沉淀技术和高通量测序相结合(MeDIP-seq)，用以研究正常的乳腺上皮细胞和乳腺癌细胞系的全基因组甲基化情况。Skene 等[14]利用研究 MeCP2 蛋白(一种与甲基化 DNA 专一性亲和的蛋白质) 在染色质上的结合情况。UCSF-UBC 则发起了人类表观基因组学测序项目(Human Reference Epigenome Mapping Project)，致力于高通量测序在表观遗传学研究上的应用。

4. 结语

SRA 数据库作为高通量测序数据存储的代表，尚处于初步发展阶段，随着高通量测序在生物学上的进一步应用，SRA 及同样类型的用于存储、调用高通量测序数据的专门数据库，如 EBI(European Bioinformatics Institute)的ENA数据库—European Nucleotide Archive;DDBJ(DNA Data Bank of Japan)的DRA 数据库隶 DDBJ Read Archive 必将进一步完善。

参 考 文 献

[1] Metzker ML *et al*. Sequencing technologies ——the next generation. *Nat Rev Genet*, 2010, 11: 31-46
[2] Shumway M *et al*. Archiving next generation sequencing data. *Nucleic Acids Res*, 2010, 38: D870-D871
[3] Via M *et al*. The 1000 Genomes Project: new opportunities for research and social challenges. *Genome Med*, 2010, 2: 3
[4] Barrett IP *et al*. Cancer genome analysis informatics. *Methods Mol Biol*, 2010, 628: 75-102
[5] Tschop MH *et al*. Getting to the core of the gut microbiome. *Nat Biotechnol*, 2009, 27: 344-346
[6] Peterson J *et al*. The NIH Human Microbiome Project. *Genome Res*, 2009, 19: 2317-2323
[7] Parkhomchuk D *et al*. Use of high throughput sequencing to observe genome dynamics at a single cell level. *Proc Natl Acad Sci USA*, 2009, 106: 20830-20835
[8] Ng SB *et al*. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 2009, 461: 272-276
[9] Hohenlohe PA *et al*. Population genomics of parallel adaptation in three spine stickleback using sequenced RAD tags. *PLoS Genet*, 2010, 6: e1000862
[10] Rubin CJ *et al*. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*, 2010, 464: 587-591
[11] Sharma CM *et al*. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*, 2010, 464: 250-255
[12] Tuch BB *et al*. Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PLoS One*, 2010, 5: e9317
[13] Ruike Y *et al*. Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. *BMC Genomics*, 2010, 11: 137
[14] Skene PJ *et al*. Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the chromatin state. *Mol Cell*, 2010, 37: 457-468

Introduction to NCBI' s SRA database

Xiao-Jing Xiong

College of Chemistry and Chemical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China

Abstract Next-generation DNA sequencing technology has become widely available and been applied to biological and biomedical research. Tremendous amount of data had been dramatically generated and accumulated. It becomes indispensable to store, query and manage these kinds of data using specially designed databases, among which SRA (Sequence Read Archive) established by NCBI was widely used and accepted. Here, we described data-organization of this database and summarized currently available data in SRA and their applications.

Key words next-generation DNA sequencing; SRA database