DDBJ Read Archive and DDBJ Read Annotation Pipeline: An archive database and an analytical tool for next-generation sequence data

Introduction

DNA Data Bank of Japan (DDBJ) is one of three databanks that constitute the DDBJ/EMBL-Bank/GenBank International Nucleotide Sequence Database (INSD), which was established through cooperative work with the European Bioinformatics Institute (EBI) in Europe and the National Center for Biotechnology Information (NCBI) in the USA[1]. We started a novel data archive database, DDBJ Read Archive (DRA), for massive amounts of raw sequencing reads from next-generation sequencers. The expert annotators of DRA issue original accession numbers for submitted data. Concurrently, we developed the preliminary version of raw read annotation pipeline tool called DDBJ Read Annotation Pipeline. This analytical pipeline tool supports reference genome mapping, de novo assembly and further annotation analyses, such as Single Nucleotide Polymorphism (SNP) detection. These new services will aid users’ research and provide easier access to DDBJ databases.

Method and Results

2.1 DDBJ Read Archive

Since May 2009 we have operated a new repository of sequencing reads, DDBJ Read Archive (DRA) (http://trace.ddbj.nig.ac.jp/dra/index_e.shtml), to archive raw output data from new sequencing platforms [2]. In 2007, NCBI set out Short Read Archive (SRA) to accommodate the data from next-generation sequencing platforms. Early in 2008, EBI began operating European Read Archive (ERA) and late in the same year DDBJ started to accept sequencing data from next-generation technologies such as Roche-454 Life Sciences GS FLX, Illumina Genome Analyzer and Applied Biosystems SOLiD. Initially, we prepared submission files at DDBJ and uploaded them to SRA. In June 2009, we started to issue our own internationally recognized accession numbers with prefix ‘DR’. Most submissions are from Japan. DRA has released 12 submissions by FTP and these data can also be retrieved from SRA. Considering the number of next-generation machines running in Japan and other Asian countries, the number of submissions to DRA is expected to increase. DRA uses the same metadata formats as SRA and ERA, and accessions of the Submission (DRA), Study (DRP), Experiment (DRX), Sample (DRS) and Run (DRR) metadata objects with the prefix indicated in parentheses followed by a six-digit number (e.g. DRA000001).

We are developing a submission system for DRA to improve submission throughput. As a first step, we have developed a web system, DRA Meta Checker, to validate metadata in XML file format (http://trace.ddbj.nig.ac.jp /DRAMetaChecker). This checker first validates uploaded XML files against an SRA XML schema, and then validates what cannot be validated by the schema, such as reference integrity among the XML documents, and correspondence between taxonomy ID and organism name. Detailed error, warning and usage messages are displayed after the validation process to help users create their metadata by themselves. In addition, we developed EXCEL-based file, ‘DRA sheets’, which contains an Excel macro to generate the metadata XML files(Fig.1A). Submitters can submit their metadata either in Excel file format or in XML file format (they can be validated by...
the DRA Meta Checker) as they prefer. For data transfer, submitters can use the FTP service of DDBJ or send a disk by a return-paid courier service. Once files have been received, the DRA team validates, issues accessions and uploads the data to SRA.

2.2 DDBJ Read Annotation Pipeline

Automatic tools for the analysis of raw sequencing reads registered in DRA may be convenient and valuable for experimental biologists. We have developed a read annotation pipeline tool to annotate DRA-registered raw sequencing reads with high throughput. The proposed annotation pipeline uses input data from FASTQ formatted files in the DRA databases. The pipeline consists of two subprocesses: basic analysis for reference genome mapping and de novo assembly, and high-level analysis of structural or functional annotations for combining automatic and manual operations, such as SNP detection and expression tag counts (Fig.1B).

The DDBJ Read Annotation Pipeline has the following three features. First, there is a shortcut for the submission of analytical results to DDBJ databases, which means that map/assembly outputs are converted to DRA formats, and similarly the results of annotation are converted to DDBJ-based INSD formats. The second feature is high throughput, achieved by the use of a cluster computing system in DDBJ. The third feature is flexibility to select appropriate analytical tools from multiple candidates.

Analytical results of de novo assembly are converted to the submission format of DDBJ Mass Submission System (MSS). Then analytical results of genome mapping are converted to one of submission formats for DDBJ DRA. Users can access these output files at FTP server. As a preliminary step for high-level annotation, analytical tools for SNP detection and RNA-seq analysis have been implemented. The RNA-seq annotation at high level analysis computes tag frequency corresponding to expression within gene regions on human, mouse, or plant genomes.

In general, to analyse massive amounts of raw reads requires high-level bioinformatics expertise. On the other hand, the DDBJ read annotation pipeline enables experimental biologists to obtain results of automatic annotations by simply manipulating a graphical user interface.

3 Future Works

DRA will develop original automatic submission and data retrieval systems. Thus in future all data to be deposited to SRA will be processed in DDBJ. As for the DDBJ Read Annotation Pipeline, currently, the pipeline only has the function of automatic annotation. To screen automatically annotated results, manual curation is indispensable (e.g. [3]). Therefore, a user support function for further manual curation will be added to the pipeline tool.

References