DNA Research **20**, 383–390, (2013)
Advance Access publication on 8 May 2013
doi:10.1093/dnares/dst017

DDBJ Read Annotation Pipeline: A Cloud Computing-Based Pipeline for High-Throughput Analysis of Next-Generation Sequencing Data

Hideki Nagasaki[1], Takako Mochizuki[1], Yuichi Kodama[1], Satoshi Saruhashi[1], Shota Morizaki[2], Hideaki Sugawara[1], Hajime Ohyanagi[3], Nori Kurata[3], Kousaku Okubo[1,4], Toshihisa Takagi[1,4], Eli Kaminuma[1], and Yasukazu Nakamura[1,*]

Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8510, Japan[1]; Fujisoft Incorporated, 3 Kanda-neribeicho, Chiyoda-ku, Tokyo 101-0022, Japan[2]; Plant Genetics Laboratory, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8510, Japan[3] and Database Center for Life Science, 2-11-16 Yayoi, Bunkyo, Tokyo 113-0032, Japan[4]

*To whom correspondence should be addressed. Tel. +81-55-981-6859. Fax. +81-55-981-6889.
Email: yn@nig.ac.jp

Edited by Prof. Masahira Hattori
(Received 10 January 2013; accepted 9 April 2013)

Abstract

High-performance next-generation sequencing (NGS) technologies are advancing genomics and molecular biological research. However, the immense amount of sequence data requires computational skills and suitable hardware resources that are a challenge to molecular biologists. The DNA Data Bank of Japan (DDBJ) of the National Institute of Genetics (NIG) has initiated a cloud computing-based analytical pipeline, the DDBJ Read Annotation Pipeline (DDBJ Pipeline), for a high-throughput annotation of NGS reads. The DDBJ Pipeline offers a user-friendly graphical web interface and processes massive NGS datasets using decentralized processing by NIG supercomputers currently free of charge. The proposed pipeline consists of two analysis components: basic analysis for reference genome mapping and *de novo* assembly and subsequent high-level analysis of structural and functional annotations. Users may smoothly switch between the two components in the pipeline, facilitating web-based operations on a supercomputer for high-throughput data analysis. Moreover, public NGS reads of the DDBJ Sequence Read Archive located on the same supercomputer can be imported into the pipeline through the input of only an accession number. This proposed pipeline will facilitate research by utilizing unified analytical workflows applied to the NGS data. The DDBJ Pipeline is accessible at http://p.ddbj.nig.ac.jp/.

Key words: next-generation sequencing; sequence read archive; cloud computing; analytical pipeline; genome analysis

1. Introduction

Next-generation sequencing (NGS) is an increasingly important technology in genome and molecular biology research, partly because of its rapidity, precision, and cost effectiveness.[1–4] NGS technology allows several analyses such as resequencing, *de novo* assembly of genomes, transcriptome analysis, Chromatin Immunoprecipitation (ChIP) sequencing, and exome analysis.[5] With ever-decreasing sequencing costs, NGS read datasets can now reach terabase sizes. These massive sequencing datasets demand high-performance computational resources, rapid data transfer, large-scale data storage, and competent data analysts. This increase in scale appears to impede data mining and analysis by researchers.

The DDBJ Sequence Read Archive (DRA), released in 2009, is a data archive for NGS raw reads that has been maintained at the DNA Data Bank of Japan (DDBJ) of the National Institute of Genetics (NIG).[6,7]

The DRA is a global provider of public nucleotide sequences in partnership with the International Nucleotide Sequence Database Collaboration (INSDC)[8] consisting of the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) in the USA[9] and the European Read Archive (ERA) of the European Bioinformatics Institute (EBI) in Europe.[10] Researchers may wish to reuse massive read datasets in DRA; however, their DRA file size tends to be too large to be downloaded to a local computer.

A computational system known as the 'cloud', consisting of data service provided via the Internet, was recently developed. Cloud computing allows users to avail services provided by data centres without building their own infrastructure. The infrastructure of the data centre is shared by a large number of users, reducing the cost to each user. To manage the flood of NGS data, several large-scale computing platforms have been recommended.[11–13] Cluster computing is performed by multiple computers typically linked through a fast local area network and functioning effectively as a single computer. Grid computing is performed by loosely coupled networked computers from different administrative centres that work together on common computing tasks. Cloud computing is the computing ability that abstracts away the underlying hardware architecture and enables convenient on-demand network access to a shared pool of computing resources that can be readily provisioned and released. In particular, a model of cloud computing, Software as a service (SaaS), is referred to as 'on-demand software' and is available via a web browser. Cloud computing is a system uniting clusters of personal computers linked together similar to grid computing. The hallmark of cloud computing is that the users can perform computation across the Internet, without the necessity of understanding the underlying architecture.

The DDBJ Read Annotation Pipeline (DDBJ Pipeline) was released in 2009 with the aim of supporting users wishing to submit NGS data analysis results to the DDBJ database, a cloud computing-based analysis pipeline for DRA NGS data. This pipeline comprises two analytical components: a basic analytical process of reference mapping and *de novo* assembly and a process of multiple high-level analytical workflows. The main workflows of the high-level analysis offer structural and functional annotations. The DDBJ Pipeline, which is a web application based on the SaaS model of cloud computing, assists in the submission of analysed results to DDBJ databases by automatically formatting data files and facilitates the web-based operation of NIG supercomputers for high-throughput data analysis. Although conventional web-based genome-analysis pipelines, such as NCBI Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP)[14] and Rice Genome Automated Annotation System (RiceGAAS),[15] perform genomic annotation of a draft sequence, their main target is Sanger-based sequence reads in small datasets. In contrast, the DDBJ Pipeline processes multiple datasets of terabase size using the computational resources of NIG supercomputers (the system is introduced in http://sc.ddbj.nig.ac.jp/index.php/en/).

In this report, we introduce the DDBJ Pipeline system with respect to its hardware and software configuration and outline its usage statistics since 2009. At present, the NIG supercomputer time is provided free of charge. We believe that provision of computational services for NGS data analysis without cost to users will increase the use of public data and accelerate data submission to public databases.

2. Materials and methods

2.1. Basic analysis

The pipeline accepts single- or paired-end reads in FASTQ[16] format and simple metadata describing the organism and experimental conditions associated with the reads. The type of sequencer is immaterial, providing that the data format of the reads is followed. Users submit their NGS data and XML-formatted metadata to the DRA. They may also submit the NGS data to a DDBJ Pipeline directory. Users subsequently analyse the NGS data in the pipeline using the accession numbers. The FASTQ-formatted sequences and metadata are loaded from the DRA databases. The DDBJ Pipeline allows pre-processing by trimming low-quality bases from both ends of the reads. The pre-processing function returns statistics and figures describing read qualities by read position, and these outputs enable users to set trimming parameters. The FASTQ files are used either for genome mapping or for *de novo* assembly. The basic analysis supports various mapping and *de novo* assembly tools for the NGS data according to the user's preference. (Analytical programme tools hosted in the DDBJ Pipeline[17–36] are listed in Table 1). Optional analytical parameters can be selected. Sequential commands from pre-processing to outputting analytical results are preset for easy operation. Reference data, such as the relevant genome sequence, can be retrieved from DDBJ databases by the Simple Object Access Protocol (SOAP).[37] Users can confirm the error rate by read position and can trim low-quality bases from the reads. The numbers of mapped reads, genome coverage, depth, and maximum contig length are reported. Output files from all processing stages including SAM-formatted files,[24] if supported by the tool, can be downloaded from an FTP server. A multiple FASTA file, which is convenient for subsequent submissions to the whole-genome shotgun (WGS)

Table 1. Analysis programmes hosted in the DDBJ Pipeline

Analysis type	Usage	Analysis tools
Basic analysis	Mapping	BLAT[17]
		BWA program[18,19]
		SOAP2[20]
		Bowtie[21]
		Bowtie 2[22]
		TopHat[23]
		SAMtools[24]
	de novo assembly	Velvet[25,26]
		SOAPdenovo[27]
		ABySS[28]
		Trinity[29]
High-level analysis	Web application	Galaxy[30]
	Annotation of mapping results	ANNOVAR[31]
		Cufflinks[32]
		MACS[33]
	Annotation of *de novo* assembled contigs	GENSCAN[34]
		GeneMark.hmm[35]
		BLAST[36]

Numbers in brackets following analysis tools are citations.

section of DDBJ, is built on the basis of consensus sequences from mapping or contig files from *de novo* assembly. The basic analysis system is built mainly using Perl 5.8, Java 6, PostgreSQL 8.3, and gnuplot (http://www.gnuplot.info/). Mapping and *de novo* assembly are performed on NIG supercomputers using 704 8-core 2.60-GHz Intel Sandy Bridge CPUs with 64 GB RAM and 1.6 TB storage, and 96 8-core 2.66-GHz Intel Xenon CPUs with 10 TB RAM, respectively.

Mapping benchmarks in the basic analysis were calculated using the whole-genome NGS data from the Japanese rice cultivar 'Omachi' (paired-end reads of accession number DRR000719) and using the complete genome sequences of the Japanese rice cultivar 'Nipponbare' as reference (accession numbers NC_008394–NC_008405). *De novo* assembly was performed using whole-genome NGS reads of *Escherichia coli* NDM1Dok01 (paired-end reads of accession number DRR001003).

2.2. High-level analysis

Because advanced analysis after mapping and *de novo* assembly requires several workflows with variable functions, the high-level analysis system was mainly designed to use the Galaxy interface,[30] a genomic workbench with a graphical user interface. To date, single-nucleotide polymorphism (SNP) analysis, transcriptome analysis (RNA-seq), and ChIP-sequencing have been implemented using SAMtools,[24] Cufflinks,[32] and MACS,[33] respectively (Table 1). These analyses are performed using mapped results in the SAM format generated by the basic analysis. Users can modify parameter settings through the graphical user interface and execute the analysis flows repeatedly using Galaxy's Workflow and History methods. For SNP analysis, a figure showing the frequency distribution of SNPs over the entire genome can be produced. For transcriptome analysis, mapped results are sent to Cufflinks to quantify gene structures and expression values. In addition, these results can be visualized by genomic regions linked to the UCSC genome browser site (http://genome.ucsc.edu/).

The high-level analysis system requires the Galaxy environment, Cairo (http://www.cairographics.org/), and Perl modules from CPAN (http://www.cpan.org/) for graphical output. The analysis is performed on the same nodes as the mapping, using the 704 8-core 2.60-GHz Intel SandyBridge CPUs with 64 GB RAM and 1.6 TB storage.

3. Results and discussion

3.1. System configuration of the proposed pipeline

The system outline of the DDBJ Pipeline is summarized in Supplementary Fig. S1. Apache Tomcat and DB servers for the DDBJ Pipeline run on an NIG supercomputer, and an FTP server that handles data import and export and resides outside the supercomputer. Reads imported from the DRA are sent via its built-in FTP server.

The pipeline is built as a cloud computing-based web application, and its flow follows two steps. The basic analysis receives transferred reads and maps them to reference genomes or assembles them. The high-level analysis generates results closer to the research goals, such as genome contig construction, SNP detection, or expression analysis.

NGS data are transferred either to an analysis server for basic analysis or to Galaxy interface servers for high-level analysis, both residing within an NIG supercomputer. Classified on the basis of purpose, the data are analysed by the supercomputer nodes using the qsub command of the UNIVA grid engine.

3.2. A pipeline for high-throughput analysis of NGS data

In the basic analysis, the DDBJ Pipeline provides the following useful functions: (i) data transfer: at the start of analysis, users can specify three methods for query data: FTP uploading, secure copy from DRA if the data have been pre-registered to DRA, or HTTP uploading. If users wish to use public data as query data, they may choose directory upload from the DRA, whose data are shared with SRA and ERA. Public data may be used not only as query data, but also as reference sequences for mapping. (ii) Pre-processing in the form of trimming off low-quality parts of sequence reads: basecalling quality is not uniform and may influence mapping or assembly quality. Although trimming off less accurately identified bases is effective to maintain the quality, several analysis tools can be used as optional functions.[38–41] The DDBJ Pipeline outputs trimmed

reads and figures showing the distributions of read quality scores. (iii) Parameter changes to software components: the DDBJ Pipeline allows modification of some options and parameters of the software and allows users to limit reads to uniquely mapped reads in the SAM file by removing multiread sets (Fig. 1). (iv) Confirming job status and ensuring confidentiality: the DDBJ Pipeline communicates with web applications to analyses the NGS data using DDBJ supercomputers and currently supports 11 mapping or *de novo* assembly software packages (Table 1). During pre-processing, mapping, or *de novo* assembly on the supercomputer, users can confirm the status of their operation through a web browser (Fig. 2). The user's jobs are listed along with their status ('running', 'complete', 'error', etc.) and elapsed times. When the jobs are completed, the DDBJ Pipeline notifies users by e-mail. Output is not limited to SAM-formatted files as mapping results or FASTA files as assembly results, but includes intermediate files, work logs, and statistical data, including mapping coverage and depth or N50 contig size of assemblies. The DDBJ Pipeline and the NIG supercomputer system, which execute the pipeline jobs, may be accessed by an unspecified number of users. To protect user confidentiality, the DDBJ Pipeline does not allow users to identify any other user except for the demo user, and users may never access each other's queries except for public data and results.

As an example of a benchmark of the system, the DDBJ Pipeline enables the mapping of 34.7 million 75-base NGS reads to a 383-Mb reference genome using BWA program[18,19] in 6.5 h and can assemble 24.4 million 80-base paired-end WGS reads in 10.5 min with SOAPdenovo,[27] using the cloud computing system. NGS technologies reduce the cost and time required for sequencing, and the resulting data increase

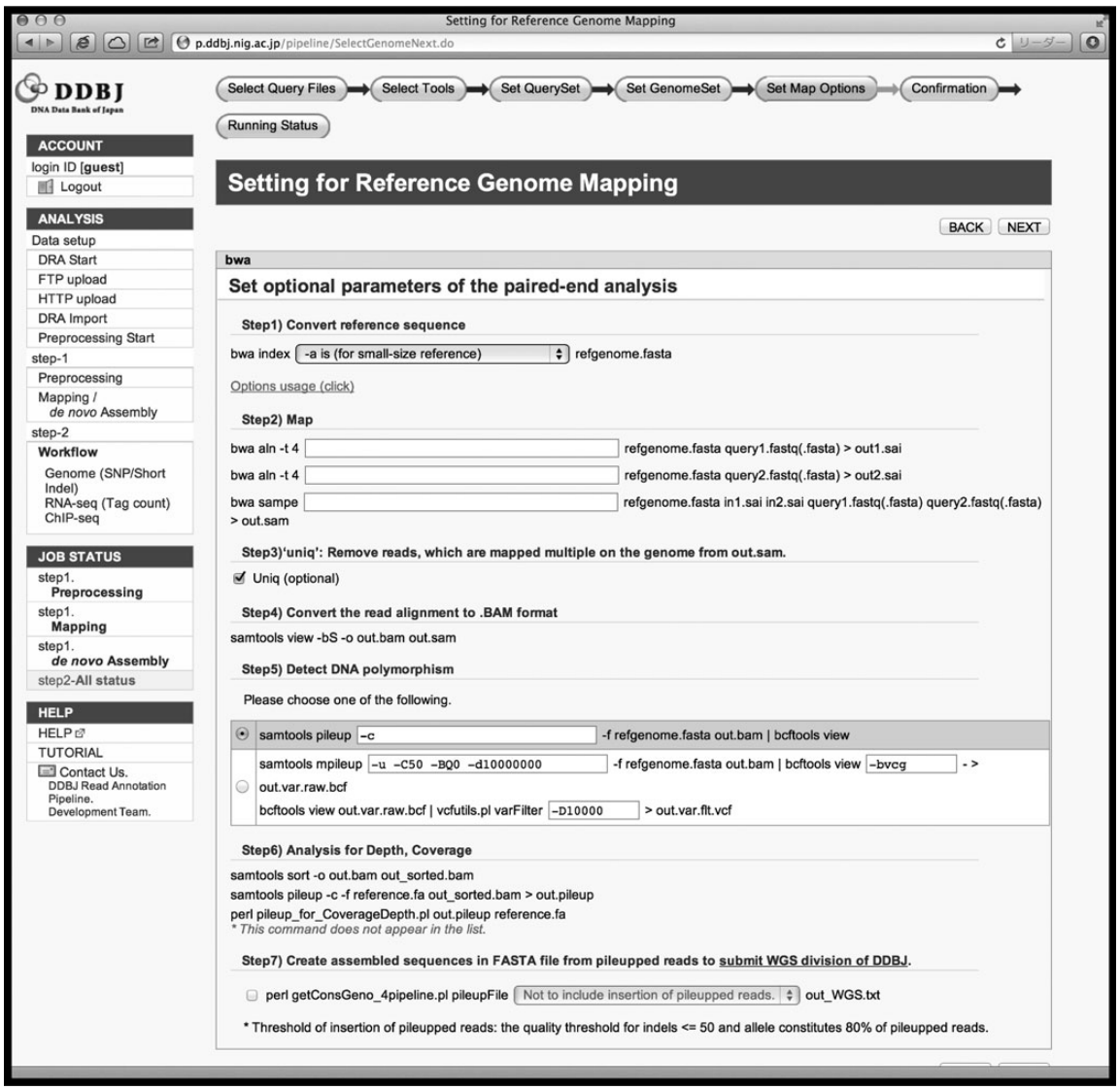

Figure 1. Interface for modifying the settings of analysis tools in basic analysis of the DDBJ Pipeline.

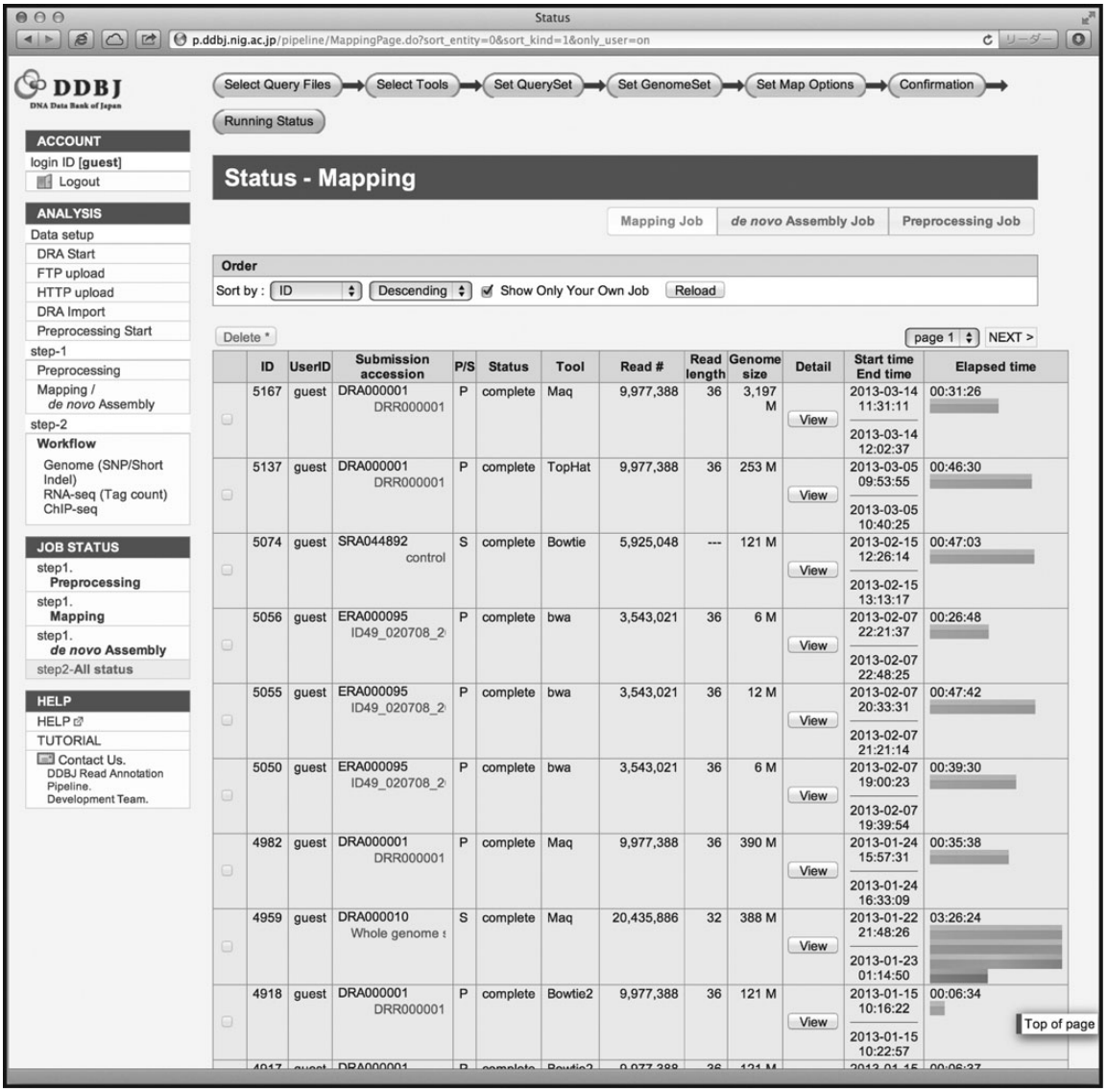

Figure 2. Job status list in basic analysis of the DDBJ Pipeline. Jobs executed in the DDBJ Pipeline are shown in lists, and users may manage the jobs, for example, by downloading results or by halting the jobs. The bars at the right end of the list indicate elapsed times.

the submissions to public archives.[8] Current developments in NGS technologies are leading to increases in both read length and the number of reads, and novel biological strategies are being developed to utilize the sequencing systems. The expansion of read numbers requires expanded computational resources, particularly for *de novo* assembly.[42] The use of a computational cluster system allows decentralized processing, resulting in scalability for efficient analysis of NGS data. The DDBJ Pipeline supports not only the use of public domain data, but also the submission of mapping and assembly results to the WGS division of DDBJ in FASTA format. Basic analysis of the DDBJ Pipeline is accessible at http://p.ddbj.nig.ac.jp.

3.3. *Genome-wide annotation by the high-level analysis of NGS data*

The DDBJ Pipeline supports not only basic analyses, such as mapping, but also high-level analyses via the Galaxy interface, which has the advantage of modifiability and easy maintenance.[30] User data such as login accounts, e-mail addresses, and passwords for access to Galaxy are shared with those of basic analysis. Therefore, basic analysis results, which are mapping or assembly jobs identified by job IDs, can be imported into Galaxy.

The high-level analysis has recently been augmented with the following four analysis methods:

(i) *SNP detection*: Users may view the pileup data produced by the basic analysis and identify SNPs using

ANNOVAR.[31] They may also view figures showing SNP distribution on the genome (Supplementary Fig. S2A and 2B).

(ii) *RNA-Seq analysis*: Expression analysis using TopHat[23] produces a SAM file that is subsequently processed using Cufflinks.[32] Downloaded Cufflinks results are sent to the UCSC genome browser site (http://genome.ucsc.edu/cgi-bin/hgGateway),[43] allowing the user to identify read expression patterns within genome-wide images. Cuffcompare, an analysis function of Cufflinks, is also available.

(iii) *ChIP-Seq analysis*: The DDBJ Pipeline supports MACS[33] for ChIP-Seq analysis.

(iv) *Annotation of contigs by de novo assembly*: A length filter is applied to remove short fragments. For gene finding in contigs, gene prediction tools, such as GENSCAN for eukaryote and GeneMark.hmm for prokaryote data,[34,35] are applied. BLASTX is used for similarity searching against known proteins[36] (Supplementary Fig. S2C). Supplementary Figs S2A, S2B, and S2C showing the whole-genome distribution of SNPs or the functional annotation of assembled contigs provide researchers with inspiration for new discoveries. In addition, new strategies for applying NGS technologies to novel biological analyses will be developed in the future. Therefore, the high-level analysis has the flexibility to be modified.

The high-level analysis of the DDBJ Pipeline can be accessed at http://p-galaxy.ddbj.nig.ac.jp.

3.4. *Usage statistics of the DDBJ Pipeline*

As a beta version offering only the basic analysis component, the DDBJ Pipeline has been open to the public via the Internet with updates since August 2009. Some analytical tools have been replaced according to the frequency of their use since then. From the start of recording in June 2010, the number of jobs submitted was around 1800 (Table 2), not considering those used for development and demonstration. The number of mapping jobs (1428) was nearly quadruple that of *de novo* assembly jobs (326).

Table 2. Job numbers for the basic analysis of the DDBJ Pipeline (since June 2010)

Year	Pre-processing	Mapping	*de novo* Assembly	Total
2010	—[a]	674	35	709
2011	11	310	152	473
2012 (January–September)	33	444	139	616
Total	44	1428	326	1798

[a]Pre-processing was still under construction.

3.5. *Building an environment supporting the use of NGS data by biologists*

Deposits of NGS data in public databases (DRA, ERA, and SRA) are rapidly increasing each year.[6] However, the NGS database has been used only as a data repository. Bioinformaticians use the data to test their own computer analysis programmes, whereas general biologists lacking computational skills rarely use the huge and unwieldy datasets. In this report, we present an example of how biologists may use public NGS data to their advantage. Genome sequences used as references for re-sequencing are often updated, resulting in the shifting of mapped positions of reads. Researchers may wish to compare SNPs that have been newly mapped and detected by them to previously detected SNPs in the public database. SNP positions in databases, such as dbSNP,[44] based on older reference genomes will lead to confusion. The DDBJ Pipeline not only provides a computational environment for analysing NGS data, but also permits seamless access to the public domain data, including NGS short reads and complete genomes. Researchers familiar with the DDBJ Pipeline will be able to re-perform reference mapping quickly using public NGS reads with current genome sequences. It may occur that comparing or merging SNP data from their own dataset with the public data using the DDBJ Pipeline allows re-analysis with preferred parameters. We expect the DDBJ Pipeline to analyse users' NGS data, thereby accelerating the submission of NGS data to public databases such as DDBJ.

3.6. *The cloud computational system for NGS data analysis*

A recent cloud computing innovation is virtual machines (VMs), which are programmes that perform parallel processing to overcome differences between server platforms. VM technology has been used in bioinformatics.[11–13] CloVR[45] has been developed for analysing bacterial NGS data, and the RseqFlow workflow[46] processes, RNA-Seq data. Cloud computing with VM expands genome informatics, and more tools will appear in future. Although the DDBJ Pipeline accommodates individual users' data, some biologists wish to host NGS analysis packages on their local servers to keep their data private. Therefore, we are studying the future incorporation of a VM package into the DDBJ Pipeline.

Acknowledgements: We thank Yasuhiro Tanizawa, Natsuko Sakakura, Shigeki Watanabe, and Naoko Sakamoto for the support of pipeline users. We particularly thank Toshihisa Okido for helpful discussions. We also thank Natsuko Hama, Naofumi

Sakaya, Tatsuya Nishizawa, and Yukiteru Ono for building an annotation workflow to assemble data in Galaxy.

Supplementary data: Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was partially supported by a grant-in-aid for Scientific Research on Innovative Areas 'Genome Science' and Scientific Research (C) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. The computations for this work were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

References

1. Schuster, S.C. 2008, Next-generation sequencing transforms today's biology, *Nat. Methods*, **5**, 16–8.
2. Chi, K.R. 2008, The year of sequencing, *Nat. Methods*, **5**, 11–4.
3. Mardis, E.R. 2008, The impact of next-generation sequencing technology on genetics, *Trends Genet.*, **24**, 133–41.
4. Lister, R., Gregory, B.D. and Ecker, J.R. 2009, Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond, *Curr. Opin. Plant Biol.*, **12**, 107–18.
5. Metzker, M.L. 2010, Sequencing technologies—the next generation, *Nat. Rev. Genet.*, **11**, 31–46.
6. Kaminuma, E., Mashima, J., Kodama, Y., et al. 2010, DDBJ launches a new archive database with analytical tools for next-generation sequence data, *Nucleic Acids Res.*, **38**, D33–8.
7. Kaminuma, E., Kosuge, T., Kodama, Y., et al. 2011, DDBJ progress report, *Nucleic Acids Res.*, **39**, D22–7.
8. Cochrane, G., Karsch-Mizrachi, I. and Nakamura, Y. 2011, International Nucleotide Sequence Database Collaboration: The International Nucleotide Sequence Database Collaboration, *Nucleic Acids Res.*, **39**, D15–8.
9. Leinonen, R., Sugawara, H. and Shumway, M. 2011, International Nucleotide Sequence Database Collaboration, The sequence read archive. *Nucleic Acids Res.*, **39**, D19–21.
10. Leinonen, R., Akhtar, R., Birney, E., et al. 2011, The European Nucleotide Archive, *Nucleic Acids Res.*, **39**, D28–31.
11. Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L. and Nolan, G.P. 2010, Computational solutions to large-scale data management and analysis, *Nat. Rev. Genet.*, **11**, 647–57.
12. Stein, L.D. 2010, The case for cloud computing in genome informatics, *Genome Biol.*, **11**, 207.
13. Pennisi, E. 2011, Human genome 10th anniversary. Will computers crash genomics? *Science*, **331**, 666–8.
14. http://www.ncbi.nlm.nih.gov/genomes/static/Pipeline.html
15. Sakata, K., Nagamura, Y., Numa, H., et al. 2002, RiceGAAS: an automated annotation system and database for rice genome sequence, *Nucleic Acids Res.*, **30**, 98–102.
16. Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L. and Rice, P.M. 2010, The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants, *Nucleic Acids Res.*, **38**, 1767–71.
17. Kent, W.J. 2002, BLAT—The BLAST-like alignment tool, *Genome Res.*, **12**, 656–64.
18. Li, H. and Durbin, R. 2009, Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics*, **25**, 1754–60.
19. Li, H. and Durbin, R. 2010, Fast and accurate long-read alignment with Burrows–Wheeler transform, *Bioinformatics*, **26**, 589–95.
20. Li, R., Yu, C., Li, Y., et al. 2009, SOAP2: an improved ultrafast tool for short read alignment, *Bioinformatics*, **25**, 1966–67.
21. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. 2009, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biol.*, **10**, R25.
22. Langmead, B. and Salzberg, S. 2012, Fast gapped-read alignment with Bowtie 2, *Nat. Methods*, **9**, 357–9.
23. Trapnell, C., Pachter, L. and Salzberg, S.L. 2009, TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics*, **25**, 1105–11.
24. Li, H., Handsaker, B., Wysoker, A., et al. 2009, The sequence alignment/map format and SAMtools, *Bioinformatics*, **25**, 2078–9.
25. Zerbino, D.R. and Birney, E. 2008, Velvet: algorithms for de novo short read assembly using de Bruijn graphs, *Genome Res.*, **18**, 821–9.
26. Zerbino, D.R., McEwen, G.K., Margulies, E.H. and Birney, E. 2009, Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler, *PLoS One*, **4**, e8407.
27. Li, R., Zhu, H., Ruan, J., et al. 2010, De novo assembly of human genomes with massively parallel short read sequencing, *Genome Res.*, **20**, 265–72.
28. Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J. and Birol, I. 2009, ABySS: a parallel assembler for short read sequence data, *Genome Res.*, **19**, 1117–23.
29. Grabherr, M.G., Haas, B.J., Yassour, M., et al. 2011, Full-length transcriptome assembly from RNA-seq data without a reference genome, *Nat. Biotechnol.*, **29**, 644–52.
30. Goecks, J., Nekrutenko, A. and Taylor, J. 2010, The Galaxy Team: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, *Genome Biol.*, **11**, R86.
31. Wang, K., Li, M. and Hakonarson, H. 2010, ANNOVAR: functional annotation of genetic variants from next-generation sequencing data, *Nucleic Acids Res.*, **38**, e164.
32. Trapnell, C., Williams, B.A., Pertea, G., et al. 2010, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, *Nat. Biotechnol.*, **28**, 511–5.
33. Zhang, Y., Liu, T., Meyer, C.A., et al. 2008, Model-based analysis of ChIP-Seq (MACS), *Genome Biol.*, **9**, R137.
34. Burge, C. and Karlin, S. 1997, Prediction of complete gene structures in human genomic DNA, *J. Mol. Biol.*, **268**, 78–94.

35. Lukashin, A. and Borodovsky, M. 1998, GeneMark.hmm: new solutions for gene finding, *Nucleic Acids Res.*, **26**, 1107–15.
36. Altschul, S.F., Madden, T.L., Schäffer, A.A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–402.
37. Kwon, Y., Shigemoto, Y., Kuwana, Y. and Sugawara, H. 2009, Web API for biology with a workflow navigation system, *Nucleic Acids Res.*, **37**, W11–6.
38. Hillier, L.W., Marth, G.T., Quinlan, A.R., et al. 2008, Whole-genome sequencing and variant discovery in *C. elegans*. *Nat. Methods*, **5**, 183–8.
39. DiGuistini, S., Liao, N.Y., Platt, D., et al. 2009, De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data, *Genome Biol.*, **10**, R94.
40. Li, R., Li, Y., Kristiansen, K. and Wang, J. 2008, SOAP: short oligonucleotide alignment program, *Bioinformatics*, **24**, 713–4.
41. Narzisi, G. and Mishra, B. 2011, Comparing *de novo* genome assembly: the long and short of it, *PLoS One*, **6**, e19175.
42. Schatz, M.C., Delcher, A.L. and Salzberg, S.L. 2010, Assembly of large genomes using second-generation sequencing, *Genome Res.*, **20**, 1165–73.
43. Kent, W.J., Sugnet, C.W., Furey, T.S., et al. 2002, The human genome browser at UCSC, *Genome Res.*, **12**, 996–1006.
44. Smigielski, E.M., Sirotkin, K., Ward, M., et al. 2000, dbSNP: a database of single nucleotide polymorphisms, *Nucleic Acids Res.*, **28**, 352–5.
45. Angiuoli, S.V., Matalka, M., Gussman, A., et al. 2011, CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing, *BMC Bioinformatics*, **12**, 356.
46. Wang, Y., Mehta, G., Mayani, R., et al. 2011, RseqFlow: workflows for RNA-Seq data analysis, *Bioinformatics*, **27**, 2598–600.