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

HIDEKI Nagasaki¹, TAKAKO Mochizuki¹, YUICHI Kodama¹, SATOSHI Saruhashi¹, SHOTA Morizaki², HIDEAKI Sugawara¹, HAJIME Ohyanagi³, NORI Kurata³, KOUSAKU Okubo^{1,4}, TOSHIHISA Takagi^{1,4}, ELI Kaminuma¹, and YASUKAZU Nakamura^{1,*}

Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8510, Japan¹; Fujisoft Incorporated, 3 Kanda-neribeicho, Chiyoda-ku, Tokyo 101-0022, Japan²; Plant Genetics Laboratory, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8510, Japan³ and Database Center for Life Science, 2-11-16 Yayoi, Bunkyo, Tokyo 113-0032, Japan⁴

*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 Author 2013. Published by Oxford University Press on behalf of Kazusa DNA Research Institute.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

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 USA⁹ 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 DDBI Read Annotation Pipeline (DDBI 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 DDB 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 FASTQ¹⁶ 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 lowquality 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 <i>de novo</i> assembly	BLAT ¹⁷ BWA program ^{18,19} SOAP2 ²⁰ Bowtie ²¹ Bowtie 2 ²² TopHat ²³ SAMtools ²⁴ Velvet ^{25,26} SOAPdenovo ²⁷ ABySS ²⁸ Trinity ²⁹			
High-level analysis	Web application Annotation of mapping results Annotation of <i>de novo</i> assembled contigs	Galaxy ³⁰ ANNOVAR ³¹ Cufflinks ³² MACS ³³ 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

And Larger Account gin Diguesti Logout NALVSIS Account Set optional parameters of the paired-end analysis Strippidad TPP upload Perprocessing Set (age out) Northow Step 20 Map Besping and 14 Consensing Processing Step 20 Map		Setting for Reference Genome Mapping c.ddbj.nig.ac.jp/pipeline/SelectGenomeNext.do							
ACCOUNT grin ID (guest) Logout Setting for Reference Genome Mapping AALYSIS BACK NEXT ANALYSIS Step1 for vert references sequence Deprocessing Marping / dropp 2 Back (ais (for small-size reference) • refgenome.fasta query1.fastq.(fasta) > out.sai Coltons Lisacy (click) Step2) Map Back (ais (for small-size reference) • refgenome.fasta query1.fastq.(fasta) > out.sai Coltons Lisacy (click) Step2) Map Back (ais (for small-size reference) • refgenome.fasta query1.fastq.(fasta) > out.sai Step3 (fire count) Step3 (ais (fire small-size reference) • refgenome.fasta query1.fastq.(fasta) = out.sai Step3 (fire count) Step3 (onvert meread alignment to .BAM format samtools wire -S		www.img.ac.gp/pipenne/selecteenonnerex.coo							
Running Status Setting for Reference Genome Mapping ANALYSIS Atta setup DRA Start FTP upload DRA Start DRA Start FTP upload Step1) Convert reference sequence Dwa in ord Proprocessing Options. Usage (cick) ae nove Assembly sep2 Workflow Context tase Options. Usage (cick) ae nove Assembly Step2) Map bwa ain -14 Dwa ain -14 Descriptions Step3) vinig*: Remove reads, which are mapped multiple on the genome fasta query1.fastq(fasta) > out1.asi Step3) Vinig*: Remove reads, which are mapped multiple on the genome fasta outbam jot.tasm step1: Step3) Options.usam step3: Step3) Detect DNA polymorphism step3: Step3) Options.usam step3: Step3: Detect DNA polymorphism step3: Step3: Detect DNA polymorphism step3: Step3: Detect DNA polymorphism step3: Step4: Convert the read alignment to .BAM format sambols pileup. Co-	DDBJ	Select Query Files - Select Tools - Set QuerySet - Set GenomeSet - Set Map Options - Confirmation -							
Setting for Reference Genome Mapping ANALYSIS ANALYSIS Ata setup DRA Start FP upload DRA Start FP upload Set optional parameters of the paired-end analysis Set optional parameters of the paired-end analysis Preprocessing Genome (SNP/Shott inde) Index (Tag count) Chilbesed (Tag count) Chilbesed (Tag count) Chilbesed (Tag count) Chilbesed (Tag count) Step3) Lonvert references explanation Step3) Lonvert references Step3 (Calck) step4 Decision usage (Calck) step3 Step3 (Calck) step4 Dob Read Finds Step3 (Calck) step1 Step3 (Calck) step1 Step3 (Calck) step4 Step3 (Calck) step4 Step3 (Calck) step5 Step5 (Stept (Tag count) Step5 (Calck) step3 step4 (Calck) <		Running Status							
Setting for Reference Genome Mapping BACK NEXT AMALYSIS Data satup DATA Start FTP upload HTTP upload DRA import Preprocessing Mapping / de noro Assembly tsp-2 Workflow Genome (RNPShot Inde) Control (Rg court) Chip Seal Step 1 Convext reference sequence bwa and r.4 Control (Rg court) Chip Seal Control (Rg court) Chip Seal Step 2 Mapping step 1 Preprocessing step 1 Mapping step 1 de noro Assembly step 1 Preprocessing step 1 Contact US, PDEA Mad Annotaton Bewelopenet Team.	ACCOUNT								
ANALVSIS BACK: NEXT. Deta stutp Deve Set optional parameters of the paired-end analysis BACK: NEXT. DPA inport Step1) Convert reference sequence Imaging / Step1) Convert reference sequence Workflow Step2) Map Back: Step2) Map Workflow Step2) Map Was aln -14 refgenome.fasta query1.fastq(.fasta) > out1.sai DBS startus Step3) vinid": Remove reads, which are mapped multiple on the genome.fasta in 1.sai in2.ai query1.fastq(.fasta) query2.fastq(.fasta) Step3) vinid": Remove reads, which are mapped multiple on the genome.fasta out.sam. Step3) Convert the read alignment to .BAM format Step3) Step4) Convert the read alignment to .BAM format samtools view -US -0 out.bam out.sam Step3) Detect DNA polymorphism Step4) Convert the read alignment to .BAM format Step3) Detect DNA polymorphism samtools pieup -cfreference.fasta out.bam bcfools view -bvcg -> OCONTACL Us. Dob Read Annotation DDB faed Annotation Step6) Analysis for Depth, Coverage Dob Read	ogin ID [guest]								
bala setup Uwa Set optional parameters of the paired-end analysis PRA import Preprocessing de novo Assembly step1 Step2 Map vorticow Genome (SNP/Short drd) RNA import Step2 Map vorticow Step2 Map vorticow Back (Tista) Step2 Map vorticow Back (Tista) Step2 Map vorticow Back (Tista) Vorticow Back (Tista) Step3 vorticities Vorticow Step3 vorticities Step3 vorticities Vorticow Step3 vorticities Step3 vorticities Vorticow Step3 vorticities Step3 vorticities Step3 vorticities Vorticow Step3 vorticities Step3 vorticities Step3 vorticities Step3 vorticities Step3 vorticities Vorticow Step4 Convert the read	Logout	Setting for Reference Genome Mapping							
DRA Start bwa FTP upload Set optional parameters of the paired-end analysis Step1 Convert reference sequence Preprocessing bwa index <= is (for small-size reference) ? refgenome.fasta	ANALYSIS	BACK NEXT							
FTP upload Set optional parameters of the paired-end analysis HTTP upload Set optional parameters of the paired-end analysis HTTP upload Set optional parameters of the paired-end analysis Preprocessing Start Set optional parameters of the paired-end analysis Preprocessing Start Set optional parameters of the paired-end analysis Options usage (dick) de novo Assembly de novo Assembly Step2) Map bwa ain -14 refgenome.fasta query1.fastq(.fasta) > out1.sai Dots read reg count) > out.sam JOB STATUS Step3) fungit: Remove reads, which are mapped multiple on the genome from out.sam. step1. Step3) fungit: Remove reads, which are mapped multiple on the genome from out.sam. step1. Step3) fungit: Remove reads, which are mapped multiple on the genome from out.sam. step1. Step3) fungit: Remove reads, which are mapped multiple on the genome from out.sam. step1. Step3) fungit: Remove reads, which are mapped multiple on the genome from out.sam. step1. Step3) fungit: Remove reads, which are mapped multiple on the genome from out.sam. step1. Step3) fungit: Remove reads, which are mapped multiple on the genome.fasta out.sam step2. All status Please choose one of the following. Please									
HTTP upload Step1) Convert references sequence DRA Import Preprocessing Mapping / de novo Assembly Step2) Map Bre-2 Step2) Map Workflow Genome (Asta query1.fastq(.fasta) > out1.sai Genome (ASP/Short Inde)) Step2) Map bwa aim 14 refgenome.fasta query1.fastq(.fasta) > out1.sai Bow aim 14 refgenome.fasta query1.fastq(.fasta) > out1.sai Bow aim 14 refgenome.fasta query1.fastq(.fasta) > out2.sai bwa aim 14 refgenome.fasta query1.fastq(.fasta) > out2.sai bwa aim 14 refgenome.fasta query1.fastq(.fasta) > out2.sai bwa aim 14 refgenome.fasta in1.sai in2.sai query1.fastq(.fasta) > out2.sai bwa aim 25 Step3) 'uniq': Remove reads, which are mapped multiple on the genome from out.sam. step1. Step3) 'uniq': Remove reads, which are mapped multiple on the genome from out.sam. step1. Step3) betect DNA polymorphism step5. Step5.0 betect DNA polymorphism step5.0 betect DNA polymorphism Step5.0 betect DNA polymorphism Step6) Analysis for Depth, Coverage samtools wiew -bxrcg >> DDM Read Annotation Peering. Step6) Analysis for Depth, Coverage samtools wiew out.var.raw.bcf <t< td=""><td></td><td>bwa</td></t<>		bwa							
HTP upload PRA import Preprocessing Preprocessing Options usage (click) de novo Assembly isp-1 Options usage (click) de novo Assembly isp-2 Workflow Genome (SNPShort Indel) RN-seq (Fig count) ChiP-seq Step1) (and all + 4 Preprocessing Workflow Genome (SNPShort Indel) RN-seq (Fig count) ChiP-seq Step2) viniqt: Remove reads, which are mapped multiple on the genome frasta in1.sai in2.sai query1.fastq(.fasta) > out2.sai JOB STATUS step1. Preprocessing de novo Assembly step1. Preprocessing de novo Assembly step1. Ge novo Assembly step2. step3. Step3) Convert the read alignment to .BAM format samtools view -bS - o out.bam out.sam de novo Assembly step5. Detect DNA polymorphism Please choose one of the following. HELP © TUTORIAL © out.var.raw.bcf voltalis, Double dad Annotation Ppene. Dobl Kead Annotation		Set optional parameters of the paired-end analysis							
Preprocessing Start atter if convent research equations ap-1 p-1 ap-1 bwe index <= is (for small-size reference)									
ep-1 bwa index [a is (for small-size reference) \$) refgenome.fasta Preprocessing Apping / de novo Assembly step2 Map bwa aln -t.4 refgenome.fasta query1.fastq(.fasta) > out1.sai bwa aln -t.4 refgenome.fasta query2.fastq(.fasta) > out2.sai hords bwa sampe refgenome.fasta query2.fastq(.fasta) > out2.sai bwa sampe refgenome.fasta query2.fastq(.fasta) > out2.sai bwa sampe refgenome.fasta in1.sai in2.sai query1.fastq(.fasta) > out2.sai bwa sampe refgenome.fasta query2.fastq(.fasta) > out2.sai bwa sampe refgenome.fasta query2.fastq(.fasta) > out2.sai bwa sampe refgenome.fasta query2.fastq(.fasta) > out2.sai bwa sampe refgenome.fasta step1. refgenome.fasta fep1. gf Uniq (optional) step1. Step3) Uniq!: Remove reads, which are mapped multiple on the genome from out.sam. step1. samtools view -b5 - o ub bam out.sam step2.Ail status Please choose one of the following. ettP.P samtools pileup -c f refgenome.fasta out.bam bcftools view -b - out.bam out.var.raw.bcf samtools pileup -c. f refgenome.fasta out.bam bcftools view -> out.sam <t< td=""><td></td><td>Step1) Convert reference sequence</td></t<>		Step1) Convert reference sequence							
Options usage (dick) de novo Assembly ep-2 Workflow Genome (SMP/Short Indel) RNA-seq (Tag count) ChIP-seq Void (MP/Short Indel) DBS STATUS Step2) Map Perprocessing Preprocessing Rep2-All status Itep1. Mapping Isep1. Mapping Isep1. Mapping Isep1. Mapping Isep2-All status PEPerocessing Preprocessing Contact Us. DBK Reed Annotation Pipeline. Dewelopment Team. Step3) Analysis for Depth, Coverage samtools sort - o out.bam out.gorted.bam samtools sort - o out.bam out.gorted.bam samtools sort - o out.bam out.gorted.bam samtools sort - o out.pam out.gorted.bam Step6) Analysis for Depth, Coverage samtools sort - o out.pam out.gorted.bam samtools sort - o out.pam out.gorted.bam Step6) Analysis for Depth, Coverage samtools sort - o out.pam out.gorted.bam Step7) Create assembled sequences in FASTA file from pileupped reads to submit WGS division of DDBJ.									
Mapping / Options Ussige (Click) Verkflow Step2) Map bwa alm + 14 refgenome.fasta query1.fastq(.fasta) > out1.sai Genome (SNP/Short Inde) bwa alm + 14 refgenome.fasta query2.fastq(.fasta) > out2.sai MoxActiow sampe refgenome.fasta query1.fastq(.fasta) > out2.sai Inde refgenome.fasta query1.fastq(.fasta) > out2.sai Inde refgenome.fasta query1.fastq(.fasta) > out2.sai Inde sampe out.sam Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Image: samtools view -bS - out.bam Step1. Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Image: samtools view -bS - out.bam out.sam Step2.All status Step5) Detect DNA polymorphism Please choose one of the following. FELP @ Please choose one of the following. Image: samtools pileup -c -frefgenome.fasta out.bam bcftools view -bvcg > > ODRI Read Annotation Pipeline. Development Team. Step6) Analysis for Depth, Coverage samtools pileup -c / reference.fa Samtools pileup -c - freference.fa out_sorted.bam > out.pileup per pileup. for CoverageDepth.pl out_sorted.bam > out_sorted.bam > out.pileup per pileup.c. This command does not appear in the list. Step7) Create assembled sequences i		uwa iliuza (-a is (io) sinali-siza relaterica) 😜 reigenome.tasta							
Mapping / de novo Assembly sp-2 Step2) Map Norkflow Genome (SNP/Short Inde) bwa ain -t 4 refgenome.fasta query1.fastq(.fasta) > out2.sai Bow a lan -t 4 refgenome.fasta query2.fastq(.fasta) > out2.sai Bow a lan -t 4 refgenome.fasta query1.fastq(.fasta) > out2.sai Bow a sampe refgenome.fasta in1.sai in2.sai query1.fastq(.fasta) query2.fastq(.fasta) > out.sam Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Proprocessing step1. ✓ Uniq (optional) Step2) Alary is exp3. Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. #LP ✓ Uniq (optional) Step2. All status Step5. Detect DNA polymorphism NorkIAL Samtools pileup -c ✓ Contact Us. DBX Read Annotation Pipeline. Please choose one of the following. ✓ ELP [@] UTURIAL Samtools pileup -u - C50 - BQ0 -d100000000 -frefgenome.fasta out.bam bcftools view -bvcg >> Ø out.var.raw.bcf volt.sam out_sorted.bam samtools pileup -c - frefgenome.fasta out.bam bcftools view -bvcg >> Uut.var.raw.bcf volt.sorted.bam samtools pileup -c - freference.fa - out.var.fit.vcf Step6) Analysis for Depth, Coverage samtools pileup -c - freference.fa </td <td></td> <td>Options usage (click)</td>		Options usage (click)							
ap-2 bwa ain -14 refgenome.fasta query1.fastq(.fasta) > out1.sai Genome (SNP/Short Indai) bwa ain -14 refgenome.fasta query2.fastq(.fasta) > out2.sai Bowa Grag count) bwa ain -14 refgenome.fasta query2.fastq(.fasta) > out2.sai IntB-seq bwa ain -14 refgenome.fasta query2.fastq(.fasta) > out2.sai bwa ain -14 refgenome.fasta query2.fastq(.fasta) > out2.sai bwa ain -14 refgenome.fasta query2.fastq(.fasta) > out2.sai bwa ain -14 refgenome.fasta query1.fastq(.fasta) > out2.sai bwa ain -14 refgenome.fasta query2.fastq(.fasta) > out2.sai bitpli Step5) Detect DNA polymorphism samtools pileup - C -1 refgenome.fasta out.bam bcftools view DDBL Read Annetation Pipeline. bcftools view out.v									
Workflow bwa aln -t 4 refgenome.fasta query1.fastq(.fasta) > out1.sai Genome (SNP/Shott Indel) bwa aln -t 4 refgenome.fasta query2.fastq(.fasta) > out2.sai BRVA-seq (Tag count) bwa asmpe refgenome.fasta in1.sai in2.sai query1.fastq(.fasta) query2.fastq(.fasta) ChIP-seq refgenome.fasta in1.sai in2.sai query1.fastq(.fasta) query2.fastq(.fasta) Proprocessing step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Image: Convert the read alignment to .BAM format Samtools view -bS -o out.bam out.sam Step5) Detect DNA polymorphism Please choose one of the following. Image: Convert the read alignment to .BAM format Samtools view -bS -o out.bam out.sam Step5) Detect DNA polymorphism Please choose one of the following. Image: Convert the read alignment to .BAM format Samtools pileup -c -frefgenome.fasta out.bam bcftools view Step5) Detect DNA polymorphism Please choose one of the following. Step6) Analysis for Depth, Coverage samtools not wort.var.raw.bcf vcfutils.pl varFilter -D10000 > out.var.fit.vcf Step6) Analysis for Depth, Coverage samtools sort -o out.bam out_sorted.bam samtools pileup -c - freference.fa * This command does not appear in the list.	and the second	Step2) Map							
Control Service Genome (SNP/Short Inde) bwa aln +t 4 refgenome.fasta query2.fastq(.fasta) > out2.sai Bwa sampe refgenome.fasta query2.fastq(.fasta) > out2.sai bwa sampe refgenome.fasta in1.sai in2.sai query1.fastq(.fasta) > out2.sai bwa sampe > out.sam Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Image: Control time in the init init init init init init init ini		bwa aln -t 4 reference fasta query1.fasta(.fasta) > out1.sai							
Indel) Trefgenome.fasta in 1. sai in2. sai query1.fastq(.fasta) query2.fastq(.fasta) PRNA-seq (Tag count) Step3)'uniq': Remove reads, which are mapped multiple on the genome.fasta in 1. sai in2. sai query1.fastq(.fasta) query2.fastq(.fasta) IOB STATUS Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Identity Treprocessing itep1. Uniq (optional) Step4) Convert the read alignment to .BAM format samtools view -bS -o out.bam out.sam Step5) Detect DNA polymorphism Please choose one of the following. Itel P ∅ UTUTORIAL © samtools pileup -u - C50 - BQ0 -d100000000 -f refgenome.fasta out.bam bcftools view -bvcg - >> © out.var.raw.bcf out.var.raw.bcf © bcftools view out.var.raw.bcf vofutils.pl varFilter -D10000 > out.var.fit.vcf Step6) Analysis for Depth, Coverage samtools pileup -c - freference.fa * * This command does not appear in the list. Step7) Create assembled sequences in FASTA file from pileupped reads to submit WGS division of DDBJ.									
RNA-seq (Tag count) bwa sampe refgenome.fasta in1.sai in2.sai query1.fastq(.fasta) query2.fastq(.fasta) IOB STATUS > out.sam IOB STATUS Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Idep1. Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Idep1. Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Idep1. Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Idep1. Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Idep1. Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Idep1. Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Idep1. Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Idep1. Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Idep1. Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Idep1. Step3)'uniq': Remove reads, which are mapped multiple on the genome from out.sam. Idep2.All status Step5) Detect DNA polymorphism Idep2.All status Step5) Detect DNA polymorphism. Please choose one of the following. out.var.r		bwa aln -t 4 retgenome.fasta query2.fastq(.fasta) > out2.sai							
Control Image: Control Preprocessing Step4) Convert the read alignment to .BAM format Step1. Step4) Convert the read alignment to .BAM format samtools view -bS -o out.bam out.sam Step5) Detect DNA polymorphism Please choose one of the following. Please choose one of the following. IELP © Image: Contact Us. DDBJ Read Annotation Pipeline. samtools pileup -u - C50 -BQ0 -d100000000 -freigenome.fasta out.bam bcftools view -bvcg - >> DBJ Read Annotation Pipeline. out.var.raw.bcf vcfutils.pl varFilter -D10000 > out.var.fit.vcf Step6) Analysis for Depth, Coverage samtools pileup -c -freference.fa out_sorted.bam samtools pileup -c -freference.fa out_sorted.bam samtools pileup _cf- CoverageDepth.pl out.pileup reference.fa 'This command does not appear in the list. Step7) Create assembled sequences in FASTA file from pileupped reads to <u>submit WGS division of DDBJ.</u>	RNA-seq (Tag count)								
Step1. Preprocessing Mapping samtools view -bS -o out.bam out.sam step1. samtools view -bS -o out.bam out.sam step5) Detect DNA polymorphism Please choose one of the following. Image: the transmitter of the top of		Sten3)'unig': Remove reads, which are manned multiple on the genome from out sam							
Preprocessing step1. Step4) Convert the read alignment to .BAM format Mapping samtools view -b5 - o out.bam out.sam step1. samtools view -b5 - o out.bam out.sam gep2.All status Step5) Detect DNA polymorphism Please choose one of the following. Please choose one of the following. Image: Contact Us. DDBJ Read Annotation Pipeline. DDBJ Read Annotation Pipeline. -c Step6) Analysis for Depth, Coverage samtools solt - o out.bam out_sorted.bam samtools pileup _c - freference.fa Step6) Analysis for Depth, Coverage samtools solt - o out.bam out_sorted.bam samtools pileup_cor_CoverageDepth.pl out.pileup reference.fa * This command does not appear in the list. Step7) Create assembled sequences in FASTA file from pileupped reads to <u>submit WGS division of DDBJ</u> .									
step1. Mapping mapping samtools view -bS - o out.bam out.sam samtools view -bS - o out.bam out.sam Step5) Detect DNA polymorphism step2.All status Please choose one of the following. MUTORIAL Samtools pileup -c -f refgenome.fasta out.bam bcftools view Step4) Contract Us. Samtools mileup -u -C50 -BQ0 -d10000000 -f refgenome.fasta out.bam bcftools view DDBJ Read Annotation Pipeline. Development Team. bcftools view out.var.raw.bcf vcfutils.pl varFilter -p10000 > out.var.fit.vcf Step6) Analysis for Depth, Coverage samtools pileup -c - f reference.fa samtools pileup -c - f reference.fa Step6) Analysis for Depth, Coverage samtools poleup.cor CoverageDepth.pl out.pileup reference.fa > out.var.fit.vcf Step7) Create assembled sequences in FASTA file from pileupped reads to submit WGS division of DDBJ. Step7) Create assembled sequences in FASTA file from pileupped reads to submit WGS division of DDBJ.		Uniq (optional)							
Mapping step1. step1. de novo Assembly step5. Detect DNA polymorphism Please choose one of the following. Please choose one of the following. INTUTORIAL OContact Us. DDBJ Read Annotation Pipeline. Development Team. Step6) Analysis for Depth, Coverage samtools pileup_co_for CoverageDepth.pl out_pileup samtools pileup_co_for CoverageDepth.pl out_pileup per pileup_for CoverageDepth.pl out_pileup * This command does not appear in the list. Step7) Create assembled sequences in FASTA file from pileupped reads to submit WGS division of DDBJ.		Sten4) Convert the read alignment to BAM format							
step 1.									
Step2-All status Step5 / Detect DNA polymorphism HELP © Please choose one of the following. Image: Step5 / Detect DNA polymorphism Please choose one of the following. Image: Step5 / Detect DNA polymorphism Please choose one of the following. Image: Step5 / Detect DNA polymorphism Please choose one of the following. Image: Step5 / Detect DNA polymorphism Please choose one of the following. Image: Step5 / Detect DNA polymorphism Please choose one of the following. Image: Step5 / Detect DNA polymorphism Please choose one of the following. Image: Step5 / Detect DNA polymorphism Please choose one of the following. Image: Step5 / Detect DNA polymorphism Please choose one of the following. Image: Step5 / Detect DNA polymorphism Please choose one of the following. Image: Step5 / Detect DNA polymorphism Image: Step5 / Detect DNA polymorphism Image: Development Team. Step6 / Analysis for Depth, Coverage Image: Step5 / Detph. Please choose one of the following on the please on the please on the list. Step6 / Analysis for Depth, Plout please role the list. Image: Step5 / Create assembled sequences in FASTA file from pileupped reads to submit WGS division of DDBJ. Step5 / Create assembled sequences in FASTA file from pileupped reads to submit WGS division of DDBJ.		samtools view -DS -O out.bam out.sam							
HELP HELP 0° TUTORIAL © samtools pileup -c JDBJ Read Annotation Pipeline. Development Team. Period Step6) Analysis for Depth, Coverage samtools pileup -c - freference.fa out_sorted.bam samtools pileup -c - freference.fa out_sorted.bam samtools pileup -c - freference.fa * This command does not appear in the list. Step7) Create assembled sequences in FASTA file from pileupped reads to submit WGS division of DDBJ.		Step5) Detect DNA polymorphism							
HELP @									
TUTORIAL Samtools mpileup -u -C50 -BQ0 -d10000000 -f refgenome.fasta out.bam bcftools view -bvcg -> DDBJ Read Annotation pipeline. out.var.raw.bcf vcfutils.pl varFilter -D10000 > out.var.flt.vcf Step6) Analysis for Depth, Coverage samtools soft -o out.bam out_sorted.bam samtools pileup -c - f reference.fa out_sorted.bam > out.jeleup reference.fa * This command does not appear in the list. Step7) Create assembled sequences in FASTA file from pileupped reads to <u>submit WGS division of DDBJ</u> .		camtoole alloun a frafaanama fasta out ham Liteftoole view							
Contact Us. DBJ Read Annotation DPBJ Read Annotation pipeline. Development Team. out.var.raw.bcf Step6) Analysis for Depth, Coverage samtools ont_sorted.bam samtools pileup - of reference.fa out_sorted.bam samtools pileup - of reference.fa out_sorted.bam samtools pileup - of reference.fa out_sorted.bam samtools pileup - of reference.fa * This command does not appear in the list. Step7) Create assembled sequences in FASTA file from pileupped reads to <u>submit WGS division of DDBJ</u> .									
DDBJ Read Annotation Pipeline. out.var.raw.bcf Development Team. out.var.raw.bcf Step6) Analysis for Depth, Coverage samtools sort - o out.bam out_sorted.bam samtools pileup - c - f reference.fa out_sorted.bam > out.pileup performent does not appear in the list. Step7) Create assembled sequences in FASTA file from pileupped reads to <u>submit WGS division of DDBJ</u> .		samtools mpileup -u -C50 -BQ0 -d10000000 -f refgenome.fasta out.bam bcftools view -bvcg ->							
Development Team. bcftools view out.var.raw.bcf vcfutils.pl varFilter _p10000 > out.var.fit.vcf Step6) Analysis for Depth, Coverage samtools sort - o out.bam out_sorted.bam samtools pileup - c - f reference.fa out_sorted.bam > out.pileup pert pileup_for_CoverageDepth.pl out.pileup reference.fa * This command does not appear in the list. Step7) Create assembled sequences in FASTA file from pileupped reads to submit WGS division of DDBJ.		out.var.raw.bcf							
samtools sort -o out.bam out_sorted.bam samtools pileup -c -f reference.fa out_sorted.bam > out.pileup perf pileup_for_CoverageDepth.pl out.pileup reference.fa * This command does not appear in the list. Step7) Create assembled sequences in FASTA file from pileupped reads to <u>submit WGS division of DDBJ</u> .		bcftools view out.var.raw.bcf vcfutils.pl varFilterD10000 > out.var.fit.vcf							
samtools pileup -c -f reference.fa out_sorted.bam > out.pileup pert pileup_for_CoverageDepth.pl out.pileup reference.fa * This command does not appear in the list. Step7) Create assembled sequences in FASTA file from pileupped reads to <u>submit WGS division of DDBJ</u> .		Step6) Analysis for Depth, Coverage							
peri pileup_for_CoverageDepth.pl out.pileup reference.fa * This command does not appear in the list. Step7) Create assembled sequences in FASTA file from pileupped reads to <u>submit WGS division of DDBJ</u> .		samtools sort -o out.bam out_sorted.bam							
* This command does not appear in the list. Step7) Create assembled sequences in FASTA file from pileupped reads to <u>submit WGS division of DDBJ</u> .									
per actions Geno Anneline ni pileun File Not to include inegation of nileunned reade A out WCS by		Step7) Create assembled sequences in FASTA file from pileupped reads to submit WGS division of DDBJ.							
Peri gerovnoveno_ryppenne.pr precupi ne [nor to include insertion of pheupped reads.] *] out_roo.td		perl getConsGeno_4pipeline.pl pileupFile Not to include insertion of pileupped reads.							
* Threshold of insertion of pileupped reads: the quality threshold for indels <= 50 and allele constitutes 80% of pileupped reads.		* Threshold of insertion of pileupped reads: the quality threshold for indels <= 50 and allele constitutes 80% of pileupped reads.							

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

							itatus						
	o.ddbj.nig	g.ac.jp/p	pipeline/I	MappingPage.do?sort_	entity	/=0&sort_ki	nd=1&only	_user=on	_	_	_	_	C <u>U-Ø-</u>
DDBJ			ery Files	Select Tools		Set Que	rySet	Set Genom	neSet	Set N	Map Optio	ns → Co	nfirmation
ACCOUNT	Rur	nning S	tatus										
gin ID [guest]			-										
Logout	S	tatu	s - N	lapping									
ANALYSIS	_							Mapping	Job	de novo	Assembl	v Job Pr	eprocessing Job
ata setup								mapping				,	oprocessing con
DRA Start		2012											
TP upload	Ord	er											
HTTP upload	Sort	by : 🔳	D	Descending	\$	Show 0	Only Your (Own Job 🕞	Reload				
ORA Import		_			_								
Preprocessing Start	Del	ete *										6	page 1 🛊 NEXT >
tep-1	Der	ele	_		_					-		-	Page I VEAT
Preprocessing		ID	UserID	Submission accession	P/S	Status	Tool	Read #	Read length	Genome size	Detail	Start time End time	Elapsed time
Mapping / de novo Assembly		5167	guest	DRA000001	Ρ	complete	Maq	9,977,388	36	3,197 M		2013-03-14 11:31:11	00:31:26
lep-2	0			DRR000001						IVI	View		
Workflow												2013-03-14	•
												12:02:37	
Genome (SNP/Short Indel)		5137	guest	DRA000001	Ρ	complete	TopHat	9,977,388	36	253 M		2013-03-05	00:46:30
RNA-seq (Tag count)	0			DRR000001							View	09:53:55	
ChIP-seq												2013-03-05	i
JOB STATUS		5074	guest	SRA044892	S	complete	Bowtie	5,925,048		121 M		2013-02-15	00:47:03
				control								12:26:14	
step1. Preprocessing											View	2013-02-15	
step1. Mapping		5056	guest	ERA000095	Ρ	complete	bwa	3,543,021	36	6 M		13:13:17 2013-02-07	00:26:48
step1.	0			ID49_020708_2							View	22:21:37	
de novo Assembly	9										View	2013-02-07	
step2-All status												22:48:25	
HELP		5055	guest	ERA000095	Ρ	complete	bwa	3,543,021	36	12 M		2013-02-07	00:47:42
HELP 🖉				ID49_020708_2							View	20:33:31	
TUTORIAL												2013-02-07	·
Contact Us.												21:21:14	
DDBJ Read Annotation		5050	guest	ERA000095	Ρ	complete	bwa	3,543,021	36	6 M		2013-02-07	00:39:30
Pipeline. Development Team.				ID49_020708_2							View	19:00:23	
Development ream.	1.000											2013-02-07	·
												19:39:54	
		4982	guest	DRA000001	Ρ	complete	Maq	9,977,388	36	390 M		2013-01-24	00:35:38
				DRR000001							View	15:57:31	
												2013-01-24	4
												16:33:09	
		4959	guest	DRA000010	S	complete	Maq	20,435,886	32	388 M		2013-01-22	03:26:24
	0			Whole genome s							View	21:48:26	
												2013-01-23	
												01:14:50	
		4918	guest	DRA000001	Ρ	complete	Bowtie2	9,977,388	36	121 M		2013-01-15	00:06:34
				DRR000001				and direction of the second			View	10:16:22	Top of
												2013-01-15	i i i op or
												10:22:57	
	-	4017	augat	DB4000001	D	annalata	Poutio?	0 077 200	26	101 M		2012 01 15	00.02.27

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 TopHat²³ 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 MACS³³ 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, GENSCAN for eukarvote and such as GeneMark.hmm for prokaryote data,34,35 are applied. BLASTX is used for similarity searching against known proteins³⁶ (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 1 800 (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).

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

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

Year	Pre- processing	Mapping	<i>de novo</i> Assembly	Total	
2010	<u> </u>	674	35	709	
2011	11	310	152	473	
2012 (January– September)	33	444	139	616	
Total	44	1428	326	1798	

^aPre-processing was still under construction.

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 bio-informatics.^{11–13} CloVR⁴⁵ has been developed for analysing bacterial NGS data, and the RseqFlow workflow⁴⁶ 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 No. 4]

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.
- 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.
- Leinonen, R., Akhtar, R., Birney, E., et al. 2011, The European Nucleotide Archive, *Nucleic Acids Res.*, **39**, D28-31.
- Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L. and Nolan, G.P. 2010, Computational solutions to largescale 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.

- 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.
- 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.
- 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, Fulllength 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.
- Wang, K., Li, M. and Hakonarson, H. 2010, ANNOVAR: functional annotation of genetic variants from nextgeneration sequencing data, *Nucleic Acids Res.*, 38, e164.
- 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.
- 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.
- 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.
- Hillier, L.W., Marth, G.T., Quinlan, A.R., et al. 2008, Wholegenome 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.