G&I
Genomics & Informatics

Original article

eISSN 2234-0742
Genomics Inform 2020;18(1):e8
https://doi.org/10.5808/GI.2020.18.1.e8

Received: January 9, 2020
Revised: February 29, 2020
Accepted: March 11, 2020

*Corresponding author:
E-mail: bulee@kribb.re.kr

Bioinformatics services for analyzing massive genomic datasets

Gunhwan Ko[1], Pan-Gyu Kim[1], Youngbum Cho[2], Seongmun Jeong[2], Jae-Yoon Kim[2], Kyoung Hyoun Kim[2], Ho-Yeon Lee[2], Jiyeon Han[3], Namhee Yu[3], Seokjin Ham[4], Insoon Jang[4], Byunghee Kang[4], Sunguk Shin[5], Lian Kim[6], Seung-Won Lee[7], Dougu Nam[8], Jihyun F. Kim[5,9], Namshin Kim[2], Seon-Young Kim[10], Sanghyuk Lee[3], Tae-Young Roh[4,11], Byungwook Lee[1*]

[1]Korea Bioinformation Center (KOBIC), KRIBB, Daejeon 34141, Korea
[2]Genome Editing Research Center, KRIBB, Daejeon 34141, Korea
[3]Department of BioInformation Science, Ewha Womans University, Seoul 03760, Korea
[4]Department of Life Sciences and Division of Integrative Biosciences & Biotechnology, Pohang University of Science & Technology (POSTECH), Pohang 37673, Korea
[5]Department of Systems, Biology Division of Life Sciences, and Institute for Life Science and Biotechnology, Yonsei University, Seoul 03722, Korea
[6]Bioposh Inc., Daejeon 34016, Korea
[7]SeqGenesis, Daejeon 34016, Korea
[8]School of Life Sciences, Ulsan National Institute of Science and Technology, Ulsan 44919, Korea
[9]Strategic Initiative for Microbiomes in Agriculture and Food, Yonsei University, Seoul 03722, Korea
[10]Genome Structure Research Center, KRIBB, Daejeon 34141, Korea
[11]SysGenLab Inc., Pohang 37613, Korea

The explosive growth of next-generation sequencing data has resulted in ultra-large-scale datasets and ensuing computational problems. In Korea, the amount of genomic data has been increasing rapidly in the recent years. Leveraging these big data requires researchers to use large-scale computational resources and analysis pipelines. A promising solution for addressing this computational challenge is cloud computing, where CPUs, memory, storage, and programs are accessible in the form of virtual machines. Here, we present a cloud computing-based system, Bio-Express, that provides user-friendly, cost-effective analysis of massive genomic datasets. Bio-Express is loaded with predefined multi-omics data analysis pipelines, which are divided into genome, transcriptome, epigenome, and metagenome pipelines. Users can employ predefined pipelines or create a new pipeline for analyzing their own omics data. We also developed several web-based services for facilitating downstream analysis of genome data. Bio-Express web service is freely available at https://www.bioexpress.re.kr/.

Keywords: analysis pipeline, cloud computing, genomic data, web server, workflow system

Introduction

Next-generation sequencing (NGS) technology has revolutionized the researches in biology and medicine during the last decade. It is routinely used in genomics field, and explosive growth of NGS data has resulted in ultra-large-scale datasets and various computational problems [1]. Public archives for sequencing data such as the Sequence Read Archive have grown rapidly and now exhibit a doubling time of 10–18 months [2]. In Korea, genomic data have been increasing rapidly in recent years. As of February 2020, approximately 277 TB of genomic data have been deposited in Korea Bioinformation Center (KOBIC) database.

It is not easy for typical researchers to analyze these massive genomic datasets. To obtain results from the data, researchers need to use high-performance computing (HPC) environments with sufficient storage space and CPU cores. In addition, the difficulties in creating complicated computational pipelines and maintaining software packages tend to overwhelm bench biologists and prevent them from attempting to analyze their own genomic data [3]. Despite the availability of a vast set of computational tools and methods for genomic data analysis in public, it is still challenging for a genomic researcher to organize these tools, integrate them into workable pipelines, find accessible computational platforms, configure the computing environment, and perform the actual analysis.

A promising solution to address this computational challenge is cloud computing, where CPUs, memory, and storage are accessible in the form of virtual machines [4]. The cloud computing, by definition, refers to the on-demand delivery of IT resources and applications via the Internet [5]. The Software as a Service (SaaS) cloud service for applications provides the perfect solution for the analysis of massive genomic datasets. SaaS is a method of software delivery in the IT field that allows data to be accessed from any device with an Internet connection and web browser. In recent years, cloud computing has rapidly emerged as a viable option for quickly and easily acquiring computational resources and pipelines for large-scale NGS data analyses [6].

The parallelism techniques in HPC infrastructure are used to process all the produced data in a feasible time [7]. Parallel computing is a type of computation in which many calculations or the execution of processes are carried out simultaneously. However, it is still challenging to integrate bioinformatics experiments with parallel techniques in the HPC environments. Many applications developed for the analysis of genomic data are either tools running only on a parallel platform, such as a MapReduce platform, or general-purpose (mainly Linux-based) programs. It is crucial to integrate these two types of platform-based applications on a single pipeline.

In this study, we present Bio-Express, a software package for deploying an on-demand computing cloud with minimal user intervention. The goal of Bio-Express is to provide a web-based analysis environment in which all genomic researchers, including those with limited or no programming knowledge, can easily analyze their own genomic data. The Bio-Express Graphic User Interface (GUI) provides a workflow editor in which users can simply use a predefined analysis pipeline or create a multistep analysis pipeline using preinstalled programs. The analysis pipelines on Bio-Express are exactly reproducible, and all analysis parameters and inputs are permanently recorded. Bio-Express makes it simple to perform a multistep analysis using simple drag and drop functionality. We also developed several web-based services for facilitating downstream analysis of genome data such as gene-set enrichment analysis.

Methods

Hardware

All runs of analysis pipelines on Bio-Express are performed on a cluster of five master nodes and 33 data nodes. The hardware system of Bio-Express consists of 800 core CPUs, 2 TB of memory, and 800 TB of disk storage in total. Each node has an Intel Xeon E502690 v2 3.0 GHz CPU, 96 GB of memory, and 28 TB of disk storage. The data node HDD configuration consists of the Hadoop Distributed File System and a solid-state drive (SSD) cache. The node manager handles the individual data nodes in a Hadoop cluster.

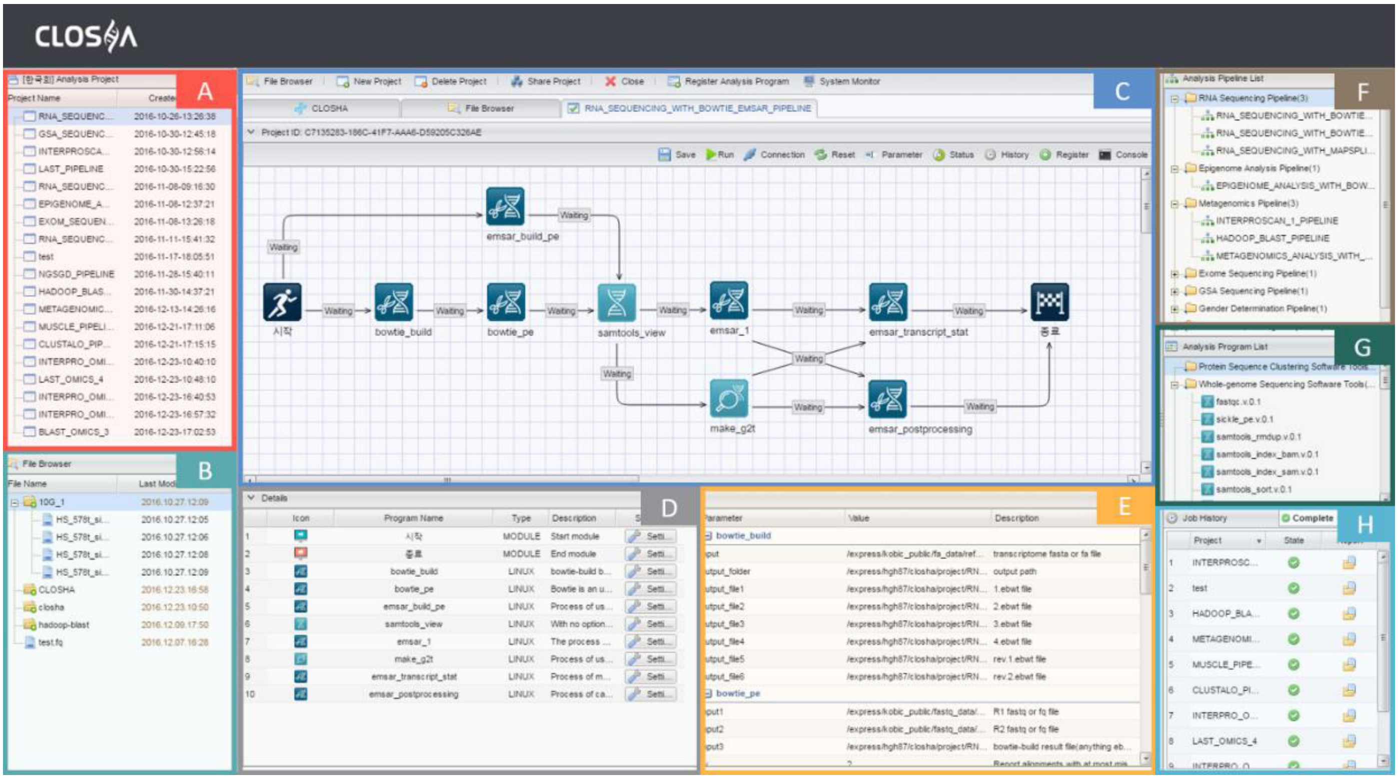
Graphic User Interface (GUI)

The GUI workspace of Bio-Express consists of eight panels: the user's projects, the file explorer, the canvas, the analysis programs of the current pipeline, the program parameter settings, the pipeline panel, the program panel, and the job execution history (Fig. 1). Among these panels, the canvas is the most important panel and is used for creating and modifying workflows by arranging and connecting activities to drive processes. The canvas provides the working surface for creating new workflows or editing predefined ones. The canvas makes it simple to perform multistep analyses using drag and drop functionality.

Pipelines

The workflows, or analysis pipelines, in the canvas are commonly depicted as directed acyclical graphs, in which each of the vertices has a unique identifier and represents a task to be performed. Additionally, each of the tasks in a workflow can receive inputs and produce outputs. The outputs of a task can be directed through another task as an input. An edge between two vertices represents the channeling of an output from one task into another. The edges determine the logical sequence. A task can be executed once all of its inputs can be resolved. If one of user pipeline programs fails, users can select the program of the pipeline to view more detailed information on errors, and resume the whole pipeline from the failed program after fixing the errors.

https://doi.org/10.5808/GI.2020.18.1.e8



Project Name	Create
RNA_SEQUENC...	2016-10-26-13:26:38
GSA_SEQUENC...	2016-10-30-12:45:18
INTERPROSCA...	2016-10-30-12:56:14
LAST_PIPELINE	2016-10-30-15:22:56
RNA_SEQUENC...	2016-11-08-09:16:30
EPIGENOME_A...	2016-11-08-12:37:21
EXOM_SEQUEN...	2016-11-08-13:26:18
RNA_SEQUENC...	2016-11-11-15:41:32
test	2016-11-17-18:05:51
NGSGD_PIPELINE	2016-11-28-15:40:11
HADOOP_BLAS...	2016-11-30-14:37:21
METAGENOMIC...	2016-12-13-14:26:16
MUSCLE_PIPELI...	2016-12-21-17:11:06
CLUSTALO_PIP...	2016-12-21-17:15:15
INTERPRO_OMI...	2016-12-23-10:40:10
LAST_OMICS_4	2016-12-23-10:48:10
INTERPRO_OMI...	2016-12-23-16:40:53
INTERPRO_OMI...	2016-12-23-16:57:32
BLAST_OMICS_3	2016-12-23-17:02:53

File Name	Last Mod
10G_1	2016.10.27 12:09
HS_578t_si...	2016.10.27 12:05
HS_578t_si...	2016.10.27 12:06
HS_578t_si...	2016.10.27 12:08
HS_578t_si...	2016.10.27 12:09
CLOSHA	2016.12.23 16:58
closha	2016.12.23 10:50
hadoop-blast	2016.12.09 17:50
test.fq	2016.12.07 16:28

	Icon	Program Name	Type	Description	S
1		시작	MODULE	Start module	Setti...
2		종료	MODULE	End module	Setti...
3		bowtie_build	LINUX	bowtie-build b...	Setti...
4		bowtie_pe	LINUX	Bowtie is an u...	Setti...
5		emsar_build_pe	LINUX	Process of us...	Setti...
6		samtools_view	LINUX	With no option...	Setti...
7		emsar_1	LINUX	The process ...	Setti...
8		make_g2t	LINUX	Process of us...	Setti...
9		emsar_transcript_stat	LINUX	Process of m...	Setti...
10		emsar_postprocessing	LINUX	Process of ca...	Setti...

arameter	Value	Description
bowtie_build		
put	/express/kobic_public/fa_data/ref...	transcriptome fasta or fa file
utput_folder	/express/hgh87/closha/project/RN...	output path
utput_file1	/express/hgh87/closha/project/RN...	1.ebwt file
utput_file2	/express/hgh87/closha/project/RN...	2.ebwt file
utput_file3	/express/hgh87/closha/project/RN...	3.ebwt file
utput_file4	/express/hgh87/closha/project/RN...	4.ebwt file
utput_file5	/express/hgh87/closha/project/RN...	rev.1.ebwt file
utput_file6	/express/hgh87/closha/project/RN...	rev.2.ebwt file
bowtie_pe		
put1	/express/kobic_public/fastq_data/...	R1 fastq or fq file
put2	/express/kobic_public/fastq_data/...	R2 fastq or fq file
put3	/express/hgh87/closha/project/RN...	bowtie-build result file(anything eb...
	2	Report alignments with at most mis

	Project	State
1	INTERPROSC...	✓
2	test	✓
3	HADOOP_BLA...	✓
4	METAGENOMI...	✓
5	MUSCLE_PIPE...	✓
6	CLUSTALO_PI...	✓
7	INTERPRO_O...	✓
8	LAST_OMICS_4	✓
9	INTERPRO_O...	✓

Fig. 1. The interface of the Bio-Express workspace. The Bio-Express workflow editor has eight panels: the user's projects (A), the file explorer (B), the canvas (C), the analysis programs of the current pipeline (D), the program parameter settings (E), the pipeline list (F), the program list (G), and the job execution history (H).

The transfer of data

The bottleneck of cloud computing is the transfer of data into clouds. Therefore, we developed a fast file transfer tool, Gbox, for uploading massive genomic datasets to the cloud server from the user's local computer and for downloading the resulting files to the local. The client program of Gbox can be downloaded from the website and be installed on the user's computer. Gbox has a file transfer at a rate of approximately 10 Gigabits per second, capable of dealing with big data over the web. Currently, Gbox has no file size limitations and storage limit on the Bio-Express cloud server.

Scalability

Scalability is one of the most attractive prospects of cloud computing and provides a useful safety net when a user's needs or demands change. The resource and job manager of Bio-Express distributes computing resources to user jobs within a parallel computing infrastructure. Its aim is to satisfy user's demands for computation and achieve a good performance in overall system's utilization by efficiently assigning jobs to resources. The resource and job manager analyzes the application performance during runtime and predicts the demand for load balancing, i.e., when to add/remove resources or redistribute workload. Thus the scalability of Bio-Express improves the execution speed of job by efficient assignment of computing resources.

Results

The analysis pipelines can be divided into two types: predefined and user-created. As of February 2020, Bio-Express contains approximately 170 analysis tools and 57 predefined analysis pipelines for genome, transcriptome, epigenome, and metagenome data. Users can employ a predefined pipeline suitable for their data by selecting a pipeline in the pipeline panel. If users want to create a new analysis pipeline, they can build their own pipeline either from scratch or by modifying a predefined pipeline. The following sections describe representative predefined analysis pipelines in the pipeline panel.

Genome pipeline

For the analysis of genome data and high-density single nucleotide polymorphism (SNP)-arrays, we developed 25 pipelines and programs, which can be grouped into seven categories: (1) discovery of variants for human, animal and plant data, (2) discovery of candidate genes from whole-exome sequencing data of rare diseases, (3) identification of somatic mutations, SNPs and short INDELs from cancer genomes, (4) clonality and evolutionary analysis of cancer genomes, (5) structural variation and copy-number analysis of whole-genome sequencing data, (6) population genomic analysis of whole-genome sequencing data, and (7) association studies and genomic predictions from the high-density SNP-arrays. We also developed two dockerized workflows that can be used for the discovery of SNPs, short INDELs, or copy-number variations from germline and somatic sample data, and for population genomics analysis in evolutionary studies. The two dockerized workflows were developed using the Workflow Description Language, developed on the Data Sciences Platform at the Broad Institute.

We developed two tools using genome data: GenoCore [8] and SEXCMD [9]. GenoCore is a new method for selecting a core collection using modified statistical measures related to genetic allele coverage and diversity. It can be used to select core subsets from plant genotype datasets, which is important for increasing cost-effectiveness and shortening the time required for the analyses of genome-wide association studies (GWAS), genomics-assisted breeding of crop species, etc. SEXCMD is a pipeline that can extract sex marker sequences from reference sex chromosomes and rapidly identify the sex of individuals from whole-exome/genome and RNA sequencing (RNA-Seq) data.

Transcriptome pipeline

The analyzing an organism's transcriptome is important for understanding the functional elements of a genome [10]. RNA-Seq is a deep-sequencing technique that can be used to explore and profile the entire transcriptome of any organism [11]. Fig. 2 shows a typical schematic overview of the RNA-Seq analysis pipeline on the canvas. The pipeline, often referred to as the tuxedo pipeline, includes five analysis tools: TopHat 2.1.1 [12], Cufflinks 2.1.1 [13], Cuffmerge 2.1.1, Cuffdiff 2.1.1, and limma voom 1.0 [14]. TopHat is a fast splice junction mapper that is used to align RNA-Seq reads to large genomes and analyze the mapping results to identify splicing junctions between exons. Cufflinks is used to assemble these alignments into a parsimonious set of transcripts and then estimate the relative abundances of these transcripts. The main purpose of Cuffmerge is to merge several Cufflinks assemblies, making it easier to produce an assembly GTF file suitable for use with Cuffdiff. Cuffdiff is then used to identify significant changes in transcript expression, splicing, and promoter use. Finally, voom robustly estimates the mean-variance relationship and generates a precision weight for each individual normalized observation, which can be used to calculate differentially expressed genes from transcript ex-

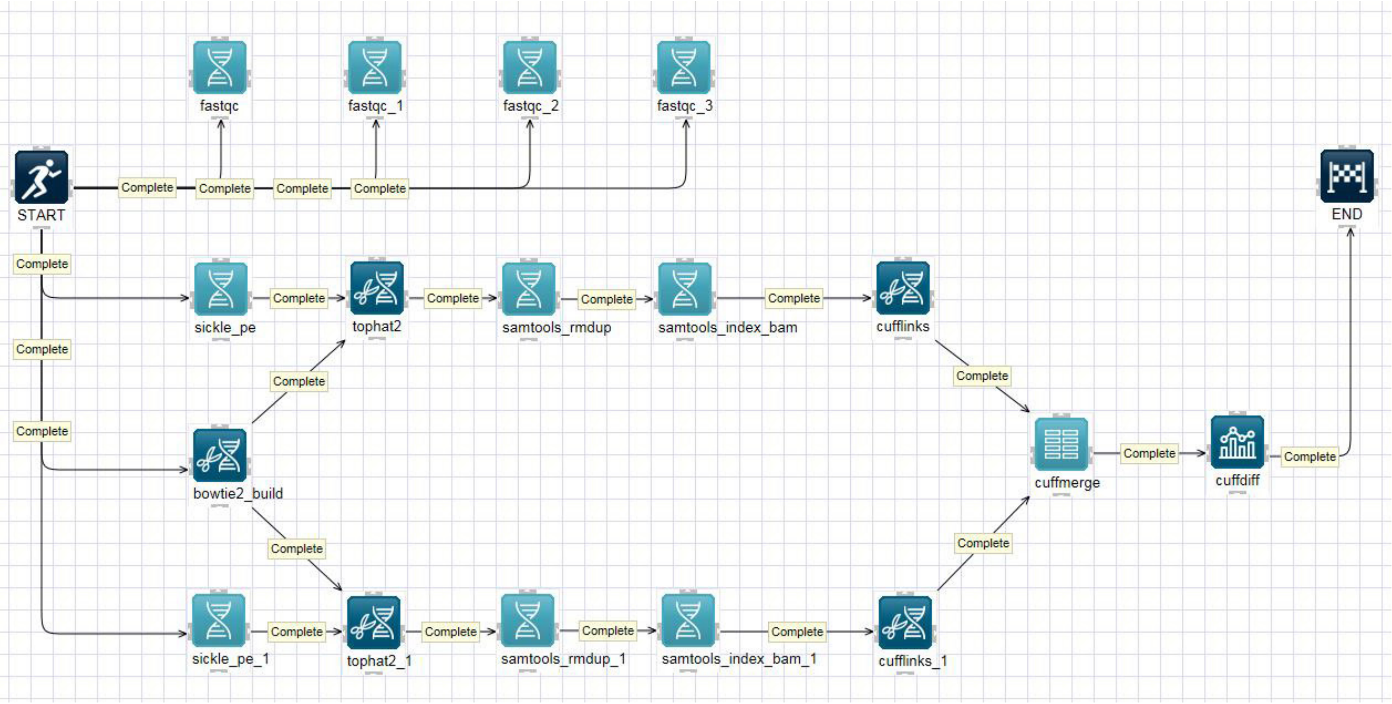


Fig. 2. Screenshot of the RNA-sequencing (RNA-Seq) schematic diagram and its pipeline. The RNA-Seq pipeline was implemented on the canvas.

https://doi.org/10.5808/GI.2020.18.1.e8

pression levels. Several other pipelines for RNA-Seq data analysis are available at Bio-Express, including MapSplice2-RSEM [15], Bowtie-EMSAR [16], STAR-HTSeq [17], and STAR-RSEM [18].

Epigenome pipeline

Epigenetic changes, including histone modifications and DNA methylation, provide a differential gene regulatory mechanism without altering DNA sequences [19]. Histone modifications occur mostly at histone tails by acetylation, methylation, phosphorylation, and ubiquitination. The accurate mapping of the called peaks of these modification sites is a critical step for understanding epigenetic transcriptional regulation. A popular, fast applicable pipeline for histone modification mapping was established by comparing various peak calling programs such as CisGenome [20], MACS1 and MACS2 [21], PeakSeq [22] and SISSRs [23], RSEG [24], SICER [25], hiddenDomains [26], BroadPeak [27], PeakRanger-CCAT, and PeakRanger-BCP [28]. For the best performance to define the exact binding sites of proteins in DNA, we tested 12 histone modifications using different peak calling programs, and we suggest the MAC2 program for narrow peak identification and PeakRanger-BCP for broad peak identification. The analysis pipeline for histone modifications is summarized in Fig. 3; the input files in fastq format are preprocessed by cudapt, fastq_quality_filter, and paired_sequence_match.py and then read quality is tested with FastQC. After mapping reads onto the reference genome, peak calling or domain calling is followed by application of MACS2 or PeakRanger-BCP. The final output is produced with annotation information. This simple pipeline is open to the public under the Bio-Express portal provided by KOBIC.

Metagenome pipelines

The analysis of metagenome data can be categorized into three parts (Fig. 4): whole metagenome shotgun sequence data analysis (shotgun metagenomics), whole transcriptome shotgun sequence data analysis (RNA-Seq), and 16S rRNA gene amplicon sequence data analysis (16S sequencing). In shotgun metagenomics, there are three pipelines: the assembly-based gene profiling, scaffold-binning, and reference-guided analysis pipelines. In the assembly-based gene profiling pipeline, sequence reads are assembled using SOAPdenovo-63mer [29]; gene regions in the assem-

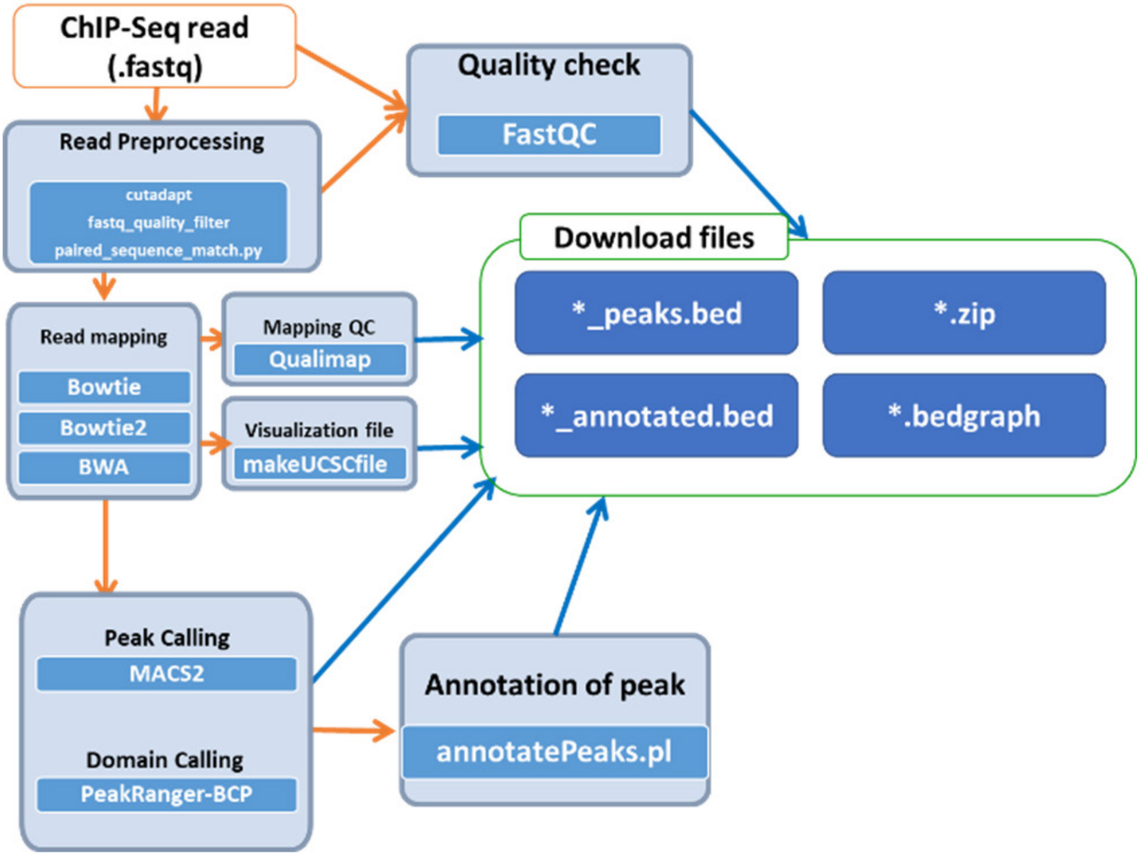


Fig. 3. Workflow for the histone modification analysis pipeline. ChIP-Seq, chromatin immunoprecipitation sequencing.

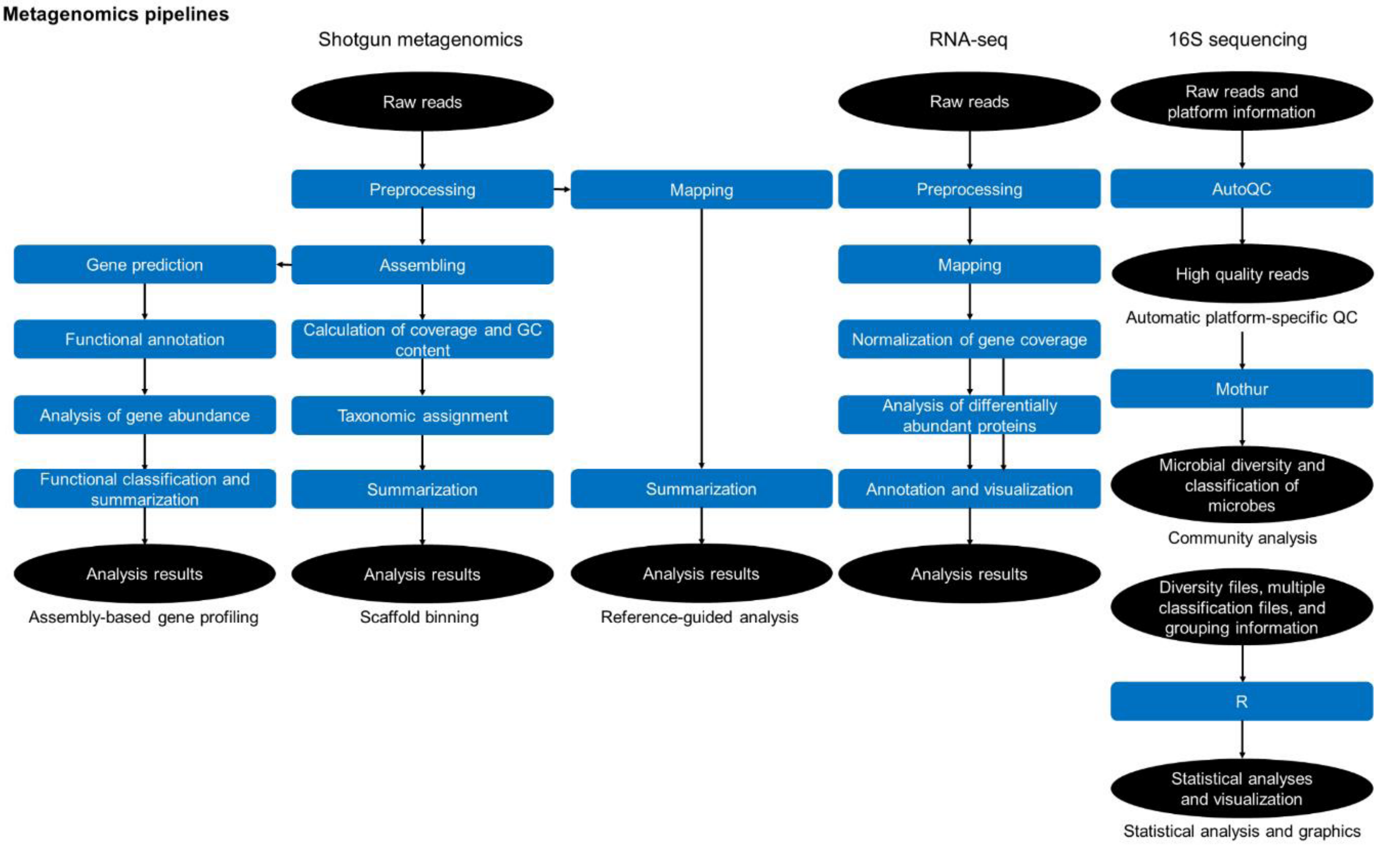

Metagenomics pipelines
Shotgun metagenomics
RNA-seq
16S sequencing
Raw reads
Preprocessing
Mapping
Gene prediction
Assembling
Functional annotation
Calculation of coverage and GC content
Analysis of gene abundance
Taxonomic assignment
Functional classification and summarization
Summarization
Summarization
Analysis results
Analysis results
Analysis results
Assembly-based gene profiling
Scaffold binning
Reference-guided analysis
Raw reads
Preprocessing
Mapping
Normalization of gene coverage
Analysis of differentially abundant proteins
Annotation and visualization
Analysis results
Raw reads and platform information
AutoQC
High quality reads
Automatic platform-specific QC
Mothur
Microbial diversity and classification of microbes
Community analysis
Diversity files, multiple classification files, and grouping information
R
Statistical analyses and visualization
Statistical analysis and graphics

Fig. 4. Simplified workflow diagram of the metagenomics pipelines.

bled sequences are predicted using MetaGeneMark [18], and the functions of the gene regions are assigned by the BLAST program with the COG and GenBank nr databases.

In the scaffold-binning pipeline, the coverage and GC content of the scaffolds are calculated, and taxonomic identifiers are assigned to the scaffolds using MEGAN [30] and HMMER 3.0 [31]. In the reference-guided analysis pipeline, sequence reads are mapped with the BWA program with reference genes or genomes. In the RNA-Seq category, sequence reads are mapped and normalized, statistical analyses are performed to identify differentially abundant genes, and finally, the results are annotated.

The 16S sequencing category is composed of three modules in sequential order: automatic platform-specific quality control (QC), community analysis, and statistical analysis and graphics. We developed a program, AutoQC, for the automatic platform-specific QC module. AutoQC uses platform-specific conditions to efficiently remove erroneous reads. AutoQC is freely available at https://sourceforge.net/projects/autoqc/. The community analysis module mainly reveals the microbial diversity and classification of microbes using Mothur. In the statistical analysis and graphical statistical analyses like pMANOVA test [32] are performed and the analysis results are visualized.

Creating custom (user defined) pipelines

Users can create a new pipeline to analyze their own data on the canvas. To create a new pipeline, users click the 'New Pipeline' button in the top menu and select an analysis pipeline type. Users will have only the [Start] and [End] modules on the canvas immediately upon creating a pipeline after selecting a 'new analysis pipeline design' in the project type. Users can drag and drop their desired analysis programs from the list of analysis programs on the right of the canvas. After the positioning of a desired analysis program on the canvas, when the users place the mouse over the edge of the analysis program icon, a connection mark will be created that can be drawn to the module. Starting from the mark, the connector must be dragged until the icon of the next analysis program to be connected becomes translucent. Users can make connections to the start module, the analysis program and the end module using this method to perform the analysis. The path for the output file is automatically a sub-path of the project in setting the input data. Finally, the analysis pipeline is executed with a message that the analysis has started. The status of the project is displayed on a real-time basis in three modes: Complete, Execute, and Wait.

Users can see the final results by clicking the ‘Results’ icon on the menu and downloading them to the user’s local computer by clicking the ‘Download’ button on the menu bar. Bio-Express also allows users to view files in various formats including text, HTML, and PNG on the screen without having to download the files (Fig. 5).

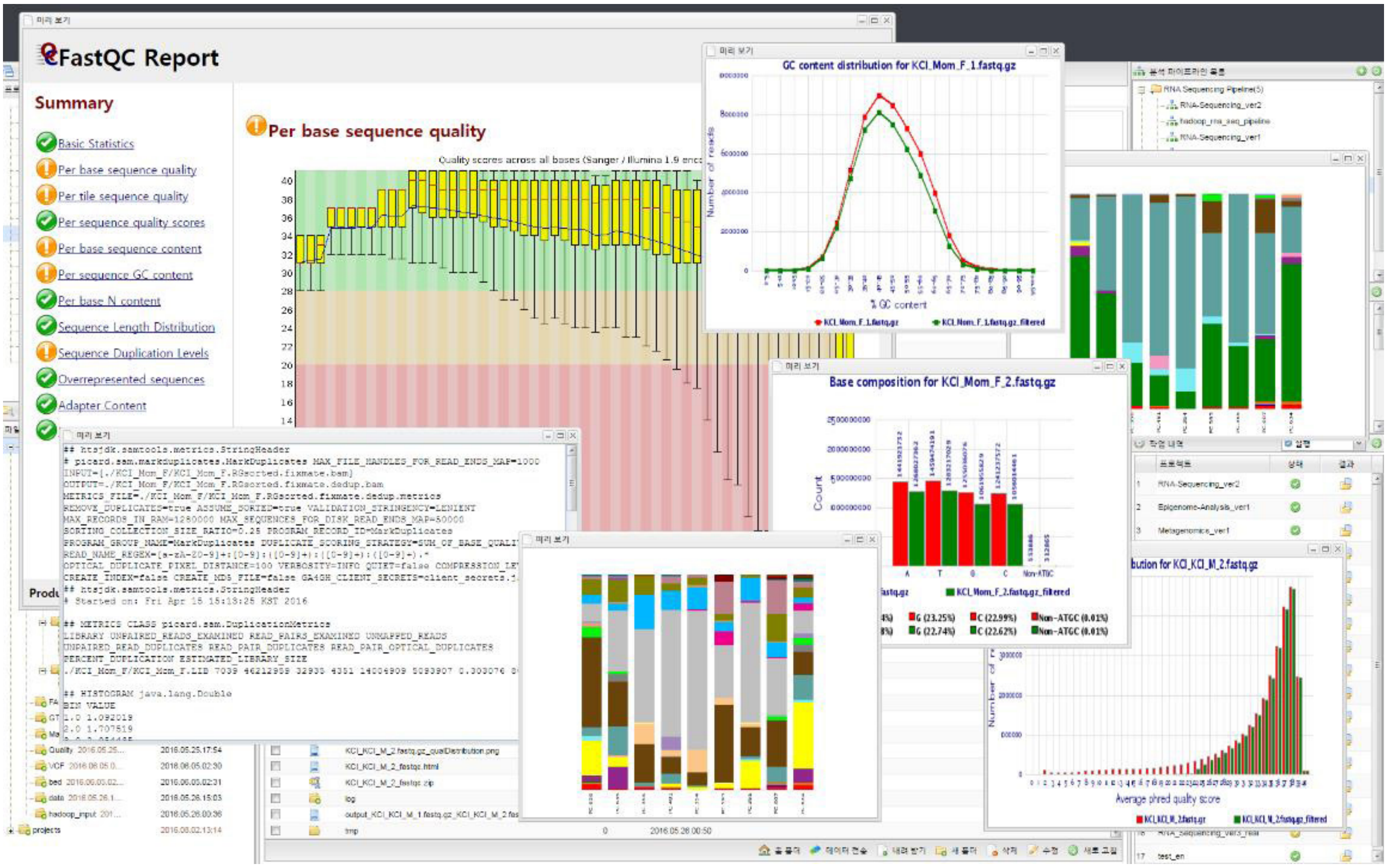
Web servers

We have developed traditional web servers in which the input is a small amount of data such as a gene list. The traditional web servers do not provide automatic scalability to the applications which is the major feature of the cloud server [33]. The developed web servers are ADGO2 [34], ExPathNet [35], GSA-SNP [36], and Barcas [37].

ADGO2 provides biological interpretations of microarray data (gene-set enrichment approach) and a list of genes (gene list over-representation approach) via composite annotation. ADGO2 also supports gene- or sample-permuting gene-set enrichment analysis for RNA-Seq count data. ExPathNet provides network-weighted gene-set clustering that incorporates both gene-set overlap and protein-protein interaction networks. GSA-SNP is standalone software that provides widely used GSA methods for SNP and GWAS data. GSA-SNP2 [38] is an improved version of GSA-SNP that provides fast high-power computation by incorporating the random set model and SNP-count adjusted gene scores. GSA-SNP2 can also visualize protein interaction networks within and across the significant pathways. Barcas is pharmacogenomics data analysis software developed for the mapping and analysis of multiplexed barcode sequencing data. Barcas employs a trie data structure for fast mapping with mismatches allowed and provides many functions, including quality control, data analysis and visualization. Table 1 shows the web servers used for gene-set, pathway, and pharmacogenomic data analysis.

Comparison between Bio-Express and Galaxy

We compared Bio-Express with Galaxy, an open source system that is the most widely used pipeline system and empowers non-computational users to do computational biology. We performed a comparison experiment between Bio-Express and Galaxy with the same data and the same RNA-Seq pipeline. We used



```
## htsjdk.samtools.metrics.StringHeader
# picard.sam.markduplicates.MarkDuplicates MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000
INPUT=[./KCI_Mom_F/KCI_Mom_F.RGsorted.fixmate.bam]
OUTPUT=./KCI_Mom_F/KCI_Mom_F.RGsorted.fixmate.dedup.bam
METRICS_FILE=./KCI_Mom_F/KCI_Mom_F.RGsorted.fixmate.dedup.metrics
REMOVE_DUPLICATES=true ASSUME_SORTED=true VALIDATION_STRINGENCY=LENIENT
MAX_RECORDS_IN_RAM=1280000 MAX_SEQUENCES_FOR_DISK_READ_ENDS_MAP=50000
SORTING_COLLECTION_SIZE_RATIO=0.25 PROGRAM_RECORD_ID=MarkDuplicates
PROGRAM_GROUP_NAME=MarkDuplicates DUPLICATE_SCORING_STRATEGY=SUM_OF_BASE_QUALI
READ_NAME_REGEX=[a-zA-Z0-9]+:[0-9]:([0-9]+):([0-9]+):([0-9]+).*
OPTICAL_DUPLICATE_PIXEL_DISTANCE=100 VERBOSITY=INFO QUIET=false COMPRESSION_LE
CREATE_INDEX=false CREATE_MD5_FILE=false GA4GH_CLIENT_SECRETS=client_secrets.j
## htsjdk.samtools.metrics.StringHeader
# Started on: Fri Apr 15 15:13:25 KST 2016

## METRICS CLASS picard.sam.DuplicationMetrics
LIBRARY UNPAIRED_READS_EXAMINED READ_PAIRS_EXAMINED UNMAPPED_READS
UNPAIRED_READ_DUPLICATES READ_PAIR_DUPLICATES READ_PAIR_OPTICAL_DUPLICATES
PERCENT_DUPLICATION ESTIMATED_LIBRARY_SIZE
./KCI_Mom_F/KCI_Mom_F.LIB 7039 46212959 32935 4351 14004909 5093907 0.303076 8

## HISTOGRAM java.lang.Double
BIN VALUE
1.0 1.092019
2.0 1.707519
```

Fig. 5. Screenshot of Bio-Express results. Users can view files in various formats, including text, HTML, and PNG on the web.

Table 1. Web servers for gene-set, pathway, and pharmacogenomic data analysis

Tool	Main function	Web address
ADGO2	Gene-set analysis of microarray data	http://www.btool.org/ADGO2
ExPathNet	Gene-set analysis with network-weighted clustering	http://epn.appex.kr/epn/
GSA-SNP2	Gene-set analysis of GWAS summary data	https://sites.google.com/view/gsasnp2
Barcas	Software for analyzing barcode-seq data	http://medical-genome.kribb.re.kr/barseq/

GWAS, genome-wide association studies.

an RNA-Seq case-control sample data set: 42,112,235 paired-end case reads and 40,975,645 paired-end control reads. The total sample size of the case and the control reads is approximately 42 GB. We assigned four CPU cores and 16 GB of memory for a single RNA-Seq job. The same machine was used for the comparison. The execution of the RNA-Seq pipeline on the sample data using Bio-Express takes a total of 3 h 44 min. The execution time using Galaxy was 6 h 11 min, showing Bio-Express has approximately 1.7 times better performance than Galaxy in the execution of the RNA-Seq pipeline. There are two main reasons for the difference in runtime between the two systems. As Galaxy internally processes intermediate data for data conversion after finishing each pipeline program, the execution time is slightly increased due to the internal process of each step. Secondly, Bio-Express has fast access to input and output data by fully utilizing the function of a SSD cache, compared to the Galaxy system.

Discussion

The substantial decrease in the cost of NGS techniques in the past decade has dramatically reshaped the genome research and has led to its rapid adoption in biological research. Nowadays, massive amount of data can be generated quickly using NGS platforms. These data range from the function and regulation of genes, the clinical diagnosis and treatment of diseases, to the omics profiling of individual patients for precision medicine. With the exponential increase in volume and complexity of NGS data, cluster or HPC systems are essential for the analysis of large amounts of NGS data. But the associated costs with the infrastructure itself and the maintenance personnel will likely be prohibitive for small institutions or laboratories.

Cloud-based applications and resources have been developed specifically to address the computational challenges of working with very large volumes of data generated by NGS technology. Cloud computing has changed how we manage computational resources. Increasingly cloud computing is also changing how large computational resources are organized and how scientists in genomics collaborate and deal with vast genome data sets.

We presented a Hadoop based distributed computational framework for large-scale genomic analysis, called Bio-Express, which incorporates a variety of tools and methods. Our system offers a variety of services to researchers. Firstly, Bio-Express allows genomic researchers without informatics or programming expertise to perform complex large-scale analysis with only a web browser using drag and drop functionality. Secondly, Bio-Express is a hybrid system that enables users to use both analysis programs providing traditional tools and MapReduce-based big data analysis programs simultaneously in a single pipeline. Lastly, we also developed a high-speed data transmission solution, Gbox, to transmit a large amount of data at a fast rate.

In the future work, we continuing to add powerful pipelines and programs including the most popular sequence and genome analysis algorithms, and to enable accessible and reproducible genomic science. Secondly, we plan to create a framework with both client-side and server-side components that simplifies the development of web-based visual applications. Visualization and visual analysis are important tools in high-throughput genomics experiments because large datasets do not need to be downloaded. Lastly, we will create a standalone installation package of Bio-Express. The increasingly large size of many datasets and moving the huge datasets is one particularly challenging aspect of current and future genomic science. Hence, local Bio-Express installations near the data are likely to become more prevalent because it makes more sense to run Bio-Express locally as compared to moving the data to a remote Bio-Express server.

ORCID

Gunhwan Ko: https://orcid.org/0000-0002-3570-9074
Pan-Gyu Kim: https://orcid.org/0000-0002-6069-1611
Youngbum Cho: https://orcid.org/0000-0002-0335-5961
Seongmun Jeong: https://orcid.org/0000-0002-0038-461X
Jae-Yoon Kim: https://orcid.org/0000-0002-8557-0998
Kyoung Hyoun Kim: https://orcid.org/0000-0002-7652-1879
Ho-Yeon Lee: https://orcid.org/0000-0002-1026-1744
Jiyeon Han: https://orcid.org/0000-0002-6001-1540

Namhee Yu: https://orcid.org/0000-0003-3967-0813
Seokjin Ham: https://orcid.org/0000-0002-6950-2848
Insoon Jang: https://orcid.org/0000-0001-9542-6984
Byunghee Kang: https://orcid.org/0000-0002-6398-5762
Sunguk Shin: https://orcid.org/0000-0002-3812-8320
Lian Kim: https://orcid.org/0000-0002-7417-0366
Seung-Won Lee: https://orcid.org/0000-0002-1878-3766
Dougu Nam: https://orcid.org/0000-0003-0239-2899
Jihyun F. Kim: https://orcid.org/0000-0001-7715-6992
Namshin Kim: https://orcid.org/0000-0001-6361-274X
Seon-Young Kim: https://orcid.org/0000-0002-1030-7730
Sanghyuk Lee: https://orcid.org/0000-0001-9230-7461
Tae-Young Roh: https://orcid.org/0000-0001-5833-0844
Byungwook Lee: https://orcid.org/0000-0003-2083-8508

Authors' Contribution

Conceptualization: JFK, NK, SYK, SL, TYR, BL. Data curation: GK, PGK, BL. Formal analysis: JH, NY, SH, IJ, BK, SS, LK, SWL. Funding acquisition: JFK, NK, SYK, SL, TYR, BL. Methodology: YC, SJ, JYK, KHK, HYL, DN. Writing - original draft: DN, JFK, NK , SYK, SL, TYR, BL. Writing - review & editing: BL.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was supported by the grants from the National Research Foundation of Korea (NRF-2014M3C9A3064552, NRF-2014M3C9A3065221, NRF-2014M3C9A3064548, NRF-2014M3C9A3068554, NRF-2014M3C9A3068822, and NRF-2019M3C9A5069653). A portion of the data used for this study were obtained from the Genome-InfraNet (IDs: 1711041199, 1711057837, 1711031849, 1711042674) at the Korea Bioinformation Center.

References

1. Bansal V, Boucher C. Sequencing technologies and analyses: where have we been and where are we going? iScience 2019; 18:37-41.
2. Kodama Y, Shumway M, Leinonen R; International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res 2012;40:D54-D56.
3. O'Driscoll A, Daugelaite J, Sleator RD. 'Big data', Hadoop and cloud computing in genomics. J Biomed Inform 2013;46:774-781.
4. Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. Nat Rev Genet 2018;19:208-219.
5. Zhou S, Liao R, Guan J. When cloud computing meets bioinformatics: a review. J Bioinform Comput Biol 2013;11:1330002.
6. Navale V, Bourne PE. Cloud computing applications for biomedical science: a perspective. PLoS Comput Biol 2018;14: e1006144.
7. Taylor RC. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics 2010;11 Suppl 12:S1.
8. Jeong S, Kim JY, Jeong SC, Kang ST, Moon JK, Kim N. GenoCore: a simple and fast algorithm for core subset selection from large genotype datasets. PLoS One 2017;12:e0181420.
9. Jeong S, Kim J, Park W, Jeon H, Kim N. SEXCMD: Development and validation of sex marker sequences for whole-exome/genome and RNA sequencing. PLoS One 2017;12:e0184087.
10. Jang YE, Jang I, Kim S, Cho S, Kim D, Kim K, et al. ChimerDB 4.0: an updated and expanded database of fusion genes. Nucleic Acids Res 2020;48:D817-D824.
11. Jeong I, Yu N, Jang I, Jun Y, Kim MS, Choi J, et al. GEMiCCL: mining genotype and expression data of cancer cell lines with elaborate visualization. Database (Oxford) 2018;2018:bay041.
12. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 2012;7:562-578.
13. Ghosh S, Chan CK. Analysis of RNA-Seq data using TopHat and Cufflinks. Methods Mol Biol 2016;1374:339-361.
14. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 2014;15:R29.
15. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011;12:323.
16. Lee S, Seo CH, Alver BH, Lee S, Park PJ. EMSAR: estimation of transcript abundance from RNA-seq data by mappability-based segmentation and reclustering. BMC Bioinformatics 2015; 16:278.
17. Anders S, Pyl PT, Huber W. HTSeq: a Python framework to work with high-throughput sequencing data. Bioinformatics 2015;31:166-169.
18. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res 2010;38:e132.

19. Yang D, Jang I, Choi J, Kim MS, Lee AJ, Kim H, et al. 3DIV: A 3D-genome Interaction Viewer and database. Nucleic Acids Res 2018;46:D52-D57.
20. Jiang H, Wang F, Dyer NP, Wong WH. CisGenome Browser: a flexible tool for genomic data visualization. Bioinformatics 2010;26:1781-1782.
21. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;9:R137.
22. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol 2009;27:66-75.
23. Narlikar L, Jothi R. ChIP-Seq data analysis: identification of protein-DNA binding sites with SISSRs peak-finder. Methods Mol Biol 2012;802:305-322.
24. Lamy P, Wiuf C, Orntoft TF, Andersen CL. Rseg: an R package to optimize segmentation of SNP array data. Bioinformatics 2011;27:419-420.
25. Xu S, Grullon S, Ge K, Peng W. Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. Methods Mol Biol 2014;1150:97-111.
26. Starmer J, Magnuson T. Detecting broad domains and narrow peaks in ChIP-seq data with hiddenDomains. BMC Bioinformatics 2016;17:144.
27. Wang J, Lunyak VV, Jordan IK. BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets. Bioinformatics 2013;29:492-493.
28. Feng X, Grossman R, Stein L. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. BMC Bioinformatics 2011;12:139.
29. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 2010;20:265-272.
30. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res 2007;17:377-386.
31. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14:755-763.
32. Tang ZZ, Chen G, Alekseyenko AV. PERMANOVA-S: association test for microbial community composition that accommodates confounders and multiple distances. Bioinformatics 2016;32:2618-2625.
33. Kim E, Bae D, Yang S, Ko G, Lee S, Lee B, et al. BiomeNet: a database for construction and analysis of functional interaction networks for any species with a sequenced genome. Bioinformatics 2020;36:1584-1589.
34. Chi SM, Kim J, Kim SY, Nam D. ADGO 2.0: interpreting microarray data and list of genes using composite annotations. Nucleic Acids Res 2011;39:W302-W306.
35. Yoon S, Kim J, Kim SK, Baik B, Chi SM, Kim SY, et al. GScluster: network-weighted gene-set clustering analysis. BMC Genomics 2019;20:352.
36. Nam D, Kim J, Kim SY, Kim S. GSA-SNP: a general approach for gene set analysis of polymorphisms. Nucleic Acids Res 2010;38:W749-W754.
37. Mun J, Kim DU, Hoe KL, Kim SY. Genome-wide functional analysis using the barcode sequence alignment and statistical analysis (Barcas) tool. BMC Bioinformatics 2016;17:475.
38. Yoon S, Nguyen HC, Yoo YJ, Kim J, Baik B, Kim S, et al. Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2. Nucleic Acids Res 2018;46:e60.