CSI NGS Portal: An online platform for automated NGS data analysis and sharing

Omer An[1], Kar-Tong Tan[1,2], Ying Li[1], Jia Li[1], Chan-Shuo Wu[1], Bin Zhang[1], Chen Leilei[1], Henry Yang[1]

1 Cancer Science Institute of Singapore, National University of Singapore, Singapore

2 Biological and Biomedical Sciences Program, Division of Medical Sciences, Harvard Medical School, Boston, Massachusetts, USA

Correspondence:

Dr. Omer An: csioan@nus.edu.sg

A/Prof. Henry Yang: csiyangh@nus.edu.sg

Abstract

Next-generation sequencing (NGS) has been a widely-used technology in biomedical research for understanding the role of molecular genetics of cells in health and disease. A variety of computational tools have been developed to analyse the vastly growing NGS data, however, they often require bioinformatics skills and tedious work to handle with. Moreover, processing raw data such as genome alignment and expression quantification consume a significant amount of time before having the data ready for downstream analyses. To facilitate data processing steps minding the gap between biologists and bioinformaticians, we developed CSI NGS Portal, an online platform which gathers established bioinformatics pipelines to provide fully automated NGS data analysis and sharing in a user-friendly website, developed in PHP and JavaScript with an integrated MariaDB database running on a dedicated Linux server. The portal currently provides 14 standard pipelines for analysing NGS data from DNA, RNA, smallRNA, ChIP, RIP, 4C, SHAPE, circRNA, eCLIP and Bisulfite sequencing, and is flexible to expand with new and customised pipelines. The users can upload raw data in fastq format and submit jobs for the desired analyses in a few clicks, and the results will be self-accessible via the portal to view/download/share in real-time. The output can be readily used as the final report or as an input for other tools depending on the pipeline. Overall, CSI NGS Portal helps researchers rapidly analyse their NGS data, share with colleagues and keep it organised without the aid of a bioinformatician. The website is freely available at: https://csibioinfo.nus.edu.sg/csingsportal

Keywords: NGS data analysis; bioinformatics pipelines

Introduction

Next-generation sequencing (NGS) has become a routine in biomedical research thanks to its proven significance and rapidly decreasing cost. Today, an overwhelming number of sequencing protocols are available by various providers, and more of them are to be developed in the near future as the underlying technology advances. In parallel, bioinformatics tools and packages to analyse the growing NGS data are also expanding, at the expense of increasing redundancy, technicality and complexity, which often alienates the wet lab biologists from understanding the data that they have generated. On the other hand, emerging technologies such as supercomputers (e.g. National Supercomputing Centre Singapore, NSCC, https://www.nscc.sg/) and cloud computing (e.g. Amazon Web Services, AWS, https://aws.amazon.com/) offer large-scale parallel computations with high speed, memory and storage, to efficiently deal with the big data generated by the NGS platforms. These technologies, however, are still offering high-cost services, which may sometimes even exceed the cost of the sequencing itself. Yet these options do not eliminate the necessity for a local bioinformatician to perform the downstream analysis and to interpret the results - unless paid for additional bioinformatics analysis service - whose task is to render the computer-generated data to the biological knowledge to address the research questions in query. Despite these facts, surprisingly, the attempts to build up comprehensive NGS data analysis platforms utilising the available tools and the existing technologies for the benefit of the community at free of charge are only a handful (Galaxy [1] https://usegalaxy.org/, Maser [2] https://cell-innovation.nig.ac.jp/maser/).

Addressing these issues, in order to facilitate NGS data analysis and sharing, aiming to bridge the gap between biologists and bioinformaticians, we have developed CSI NGS Portal as a freely

accessible, easy-to-use and comprehensive online platform, offering well-established and fully automated bioinformatics pipelines at the service of the community. Currently, the portal covers more than 10 frequently used NGS data types, committed to expand, and offers one-click data analysis and sharing. A simple and intuitive interface with tabular structure across the website greatly enhances user experience by keeping the data well-organised, easily accessible and queryable. The portal has proven to be successful and useful during its internal uptime in the last 3 years, commencing to extend its scope to the globe.

Portal Implementation

CSI NGS Portal has been developed in a Linux environment by using a mix of several programming languages and employing a vast number of bioinformatics tools and packages (Supplementary Table S1). Specifically, the website has been built in PHP v7 and HTML5, and the dynamic features have been implemented in JavaScript. The website runs on an Apache v2.0 server under a Linux machine with Ubuntu v16.04.6 installation, with an integrated database built with MariaDB v10. The majority of the bioinformatics tools and packages as well as their dependencies are maintained in a Conda environment (https://anaconda.com), and all the software behind the portal are regularly updated to the latest stable versions available. The website interface heavily utilises Bootstrap (https://getbootstrap.com/), a popular front-end component library and open source toolkit for developing with HTML, CSS, and JS. The user experience has been greatly enhanced by the interactive tables with search, filter, sort, edit, export, share and other functionalities owing to the Bootstrap plugins, as well as by displaying the web pages properly on different devices and platforms owing to the Bootstrap's responsive design feature.

The bioinformatics pipelines (Table 1, Supplementary Data) consist of in-house generated scripts (written in Bash, Perl, R) and/or integrated tools mostly from freely available open source projects (written in various programming languages), and all of the pipelines have been integrated into the portal with wrapper Bash scripts. The pipelines are implemented in a modular structure, which makes it easy to add new ones in the future. To submit a job, the required inputs from the users are raw fastq files, basic sample annotations and simple analysis design. The job submission is subject to queue system, which handles the jobs at the background by managing the available resources on the server. In order to utilise the resources efficiently and to obtain the results rapidly, the jobs are parallelised via multi-threading per sample wherever supported, and via simultaneous run of multiple samples per job wherever applicable.

To monitor overall job progress and to debug in case of failure, a job log has been implemented. The job log displays real-time information for each step of the pipeline with timestamp, which makes it easier to identify the source of the error upon failure. The keyword "ERROR" highlighted in red is reserved to denote the failure for this purpose. It is possible for the pipeline to continue running for the next steps even though a step has failed if they are independent from each other. Moreover, log files of the individual steps are written to a dedicated "LOGS" folder.

The portal has been designed in mind with minimalist approach for the user input and exhaustive approach for the pipelines run. To achieve this, the processes are automated as much as possible allowing users to focus on the analysis results rather than the procedure. Provided that the user inputs have no error - to ensure which a number of control functions are implemented at the backend - the pipelines are guaranteed to run successfully to output the expected results. In case

of unexpected failure due to technical reasons such as syntax change upon software update, the

job can be easily rerun by the admin with the same parameters after fixing the problem, for which

no action is required from the user.

Table 1: Bioinformatics pipelines implemented on CSI NGS Portal

Bioinformatics Pipeline	Analysis Steps	Tools & Packages	Sequencing Types	Normal/Control /Reference Samples	Replicate Samples[1]	Overall Runtime
1. **DNA-Seq**	Genome alignment	BWA (mem) [3]	Single/Paired end	Optional[2]	NA	~ 1 day
	Mutation calling	GATK4 Mutect2 [4, 5]				
	Mutation annotation	ANNOVAR [6]				
2. **RNA-Seq**	Genome alignment	STAR [7]	Single/Paired end	NA	NA	~ 2 hours
	Gene expression	HTSeq-count [8]				
	Isoform expression	Salmon [9]				
	Alternative splicing	in-house Perl				
3. **Diff-Exp**	Genes table	Bioconductor DESeq2 [10]	Single/Paired end[3]	Required	Required (min 2 samples)	~ 10 minutes
	Interactive report	Bioconductor regionReport [11]				
	Heatmap	ggplot2 (Wickham 2016)				
	Pathway enrichment	Bioconductor ReactomePA [12]				
	Gene set enrichment analysis	GSEA [13]				
4. **Pathway-Enrichment**	Enrichment plots	Bioconductor ReactomePA [12], enrichplot [14]	NA	NA	NA	~ 1 minute
5. **RNA-Editing**	Genome alignment	BWA (mem) [3]	Single/Paired end	NA	NA	~ 20 hours
	Variant calling	Samtools mpileup [15]				
	Candidates selection	adapted from [16]				
6. **smallRNA**	Genome alignment	NovoAlign	Single/Paired end	NA	NA	~ 1 hour
	smallRNA expression	in-house Perl				

7. 4C-Seq	Genome alignment	BWA (mem) [3]	Single/Paired end	Optional	Optional (2 samples)	~ 10 minutes
	Interactions	Bioconductor r3Cseq (Thongjuea et al. 2013)				
	Report	Bioconductor r3Cseq [17]				
8. ChIP-Seq	Genome alignment	Bowtie2 [18]	Single/Paired end	Required	NA	~ 6 hours
	Peak calling	MACS2 [19]				
	Motif enrichment	HOMER [20]				
	UCSC track hub	in-house Bash				
9. RIP-Seq	Genome alignment	STAR [7]	Paired end	Required	Optional (2-10 samples)	~ 8 hours
	Peak calling	in-house Bash				
	UCSC track hub	in-house Bash				
10. SHAPE-Seq	Transcriptome alignment	Bowtie2 [18]	Single/Paired end	Required	NA	~ 10 hours
	Reactivity calculation	icSHAPE [21]				
	Structure prediction	RNAfold [22, 23]				
11. rMATS	Genome alignment	STAR [7]	Single/Paired end	Required	Required (2-10 samples)	~ 2 hours
	Alternative splicing	rMATS [24]				
12. circRNA	Genome alignment	STAR [7]	Single/Paired end	NA	NA	~ 1 hour
	circRNA expression	in-house Perl				
13. eCLIP-Seq	Demultiplexing	eclipdemux [25, 26]	Single/Paired end	Required	NA	~ 1 day
	Mapping	STAR [7]				
	Peak calling	clipper [27]				
	Peak normalisation	eCLIP [25, 26]				
14. Bisulfite-Seq	Genome alignment	bowtie2 [18]	Single/Paired end	NA	NA	~ 3 days
	Methylation calling	Bismark [28]				
	UCSC track hub	in-house Bash				

Usage of the tools and packages in CSI NGS Portal and website links to their original sources are given in Supplementary Table S1.

The detailed descriptions, expected input and output of the pipelines are given in Supplementary Data and on the website Docs page.

Overall runtime is the approximate time elapsed for 1 sample to finish all the analysis steps once the job starts running, and may vary

depending on the data size, pipeline parameters and server load. However, runtime for additional samples under the same job do not

multiply proportionally due to the parallelisation.

[1]: Ideally technical replicates rather than biological replicates. Numbers in parenthesis denote the samples in total.

[2]: For somatic mutation calling, matched normal DNA sample is highly recommended. Use of "tumor-only mode" is useful only for

specific purposes.

[3]: Not directly applicable to "Diff-Exp" pipeline, instead refers to the samples from the "RNA-Seq" where this pipeline starts from.

Website Usage

The website is publicly accessible and fully functional via major web browsers. The portal has no sign up or login requirement, hence it is open to all without any authentication, however, authorisation to data is provided via a browser cookie. Upon access to the website on the browser, a random cookie is created on the user's local computer and associated with a stable user id and a generic username, where the latter can be changed to a personal one any time by the user, which is solely used for data sharing among users. The returning users are recognised via the browser cookie, which allows user-specific data to be available until it expires (10 days after the job finish date and 30 days since the last access to the website). Upon expiry, the user accounts and the associated data are removed from the server to restore disk space, except the data annotation, which is permanently stored in the database to avoid re-annotating the same data upon re-uploading in the future, provided that the filename is identical. For safety reasons, the cookie is computer and browser specific, which means that the users can access their data next time only by using the same browser from the same computer, otherwise regarded as a new user. Alternatively, for the sample annotation and the job results, the users may still access to their data on a different computer/browser by sharing them with themselves under a different username by using the "Shared With" field. For example, a user may upload data and run jobs from his computer at work with the username "user86_work", and he can access the results from his computer at home sharing them with his other username "user86_home". However, to delete or to modify the data, he still needs to use the original account "user86_work" from which the data are uploaded, or they will be automatically removed after account expiry. Such a system ensures data privacy while keeping sharing results easy and simple. In addition to the data analysis, the portal also contains detailed

documentation of the pipelines ("Docs" page) and answers to the frequently asked questions ("FAQ" page).

Website Framework

The website framework of CSI NGS Portal consists of 5 major steps ("Upload", "Annotate", "Submit", "Jobs", "Browse"), each of which is built as an individual webpage (Figure 1). Although each page is independently accessible on the website menu, they are interconnected via automated data transitions, i.e. the output of each step is reflected as the input of the next step. More specifically, successfully uploaded files via the "Upload" page are inserted into the annotation table on the "Annotate" page, and properly annotated samples become available to the job submission on the "Submit" page, and status/progress of the submitted jobs can be monitored on the "Jobs" page, and finally, text output of the finished jobs can be queried at the "Browse" page. All the pages with a table structure ("Annotate", "Jobs", "Browse" pages) are equipped with advanced features (search, filter, sort, edit, export and share options) to enhance user experience and data organisation. Each page has a "README" section to guide to the usage and the expected input/output data, and explained further as follows:
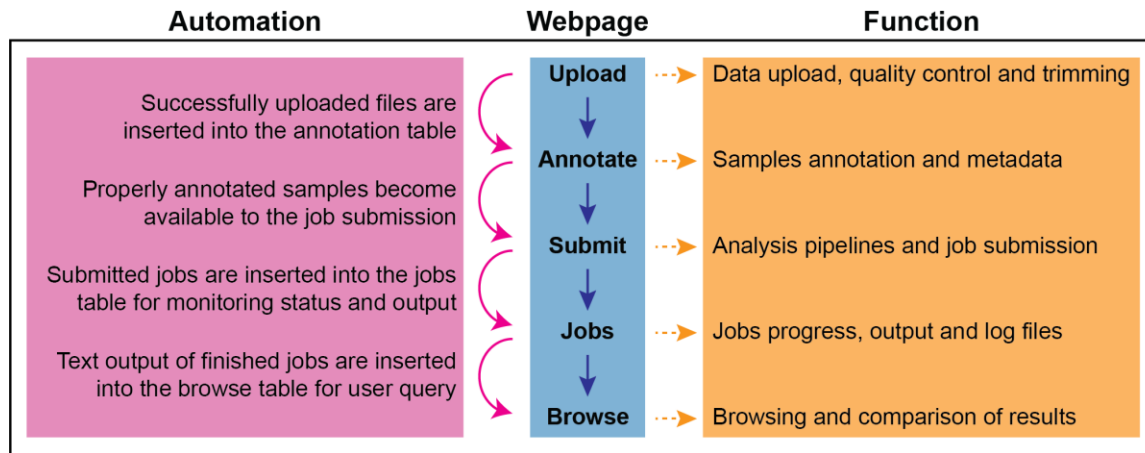
Figure 1: Website Framework of CSI NGS Portal. The middle panel shows the logical flow of the website from top to the bottom, where the automated steps of the data transitions from one page to another are described on the left panel, and the function of each page expecting user input/action is given on the right panel.

↓ : Logical flow   ∫ : Automated step   ⤑ : Page function

1. Upload

This page (Figure 2a) allows to upload raw data files in FASTQ [29] format (http://maq.sourceforge.net/fastq.shtml). All the pipelines on the portal start from the fastq file, followed by genome/transcriptome alignment, in order to standardise the data processing with a suitable mapper for the specific task and refrain user from the tedious alignment step. The file format requirements and the restrictions are given on the website. Successfully uploaded files are displayed as downloadable links including full file name, file size, file owner and action buttons for sequence quality check and processing, whereas failed uploads display an appropriate error message. A FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) report for each file is auto-generated in the background upon completion of the upload, allowing users to check the sequencing quality via a set of quantitative and visual metrics. A failed FastQC run is also displayed, indicating that the fastq file is corrupted, and such a file should not be proceeded to submit a job. In this case, user is expected to delete the corrupted file, fix the issue and re-upload the corrected file until the FastQC report is successfully generated. If needed, a trimming interface is also available employing Trimmomatic [30] to trim the adapter/primer/barcode or other custom sequences and remove the low-quality reads with a variety of options. Dragging and dropping of multiple files/folders are supported for the file upload interface, and batch upload/cancel/delete of files are available with a single click. A progress bar on top of the page displays the overall upload status showing upload speed, estimated time remaining, upload percentage and upload size out of overall size. For fair usage, there are quota restrictions on the file number and the file size per user. Uploaded files may be automatically renamed to comply with the file naming rules, e.g. spaces are replaced with underscores. Closing browser tab or internet disconnection interrupt active file uploads. The uploaded files are private to the user and may be deleted only by the file owner or

upon data expiry, whichever is earlier. Further instructions are given in the "README" section to avoid any possible problems in the next steps.

2. Annotate

Successfully uploaded files are auto-inserted into the annotation table on this page (Figure 2b) awaiting user action. A single entry is inserted per sample based on the filename regardless whether the sample comes from single-end (1 file) or paired-end (2 files) library, provided that the filenames follow the naming rules explained on the website. An incremental, unique and stable id is assigned to each sample, in addition to the unique filenames and custom sample names. The sample annotation consists of "required" and "optional" sections with predefined fields, where the former must be fully filled before job submission as it contains relevant information to the pipeline. All the editable fields offer "in-place" editing, i.e. editing by clicking directly on the html element rather than using a separate panel or dialog box and without reloading the page, which makes sample annotation a quick and easy task. The annotation data may be modified and shared only by the file owner, i.e. user who has uploaded the files, and remain visible as long as the raw file exists on the "Upload" page. For RNA-Seq samples, "Diff-Exp Group" must be specified to perform a "Diff-Exp" analysis to determine the sample groups to be compared ("contrast" parameter in DESeq2, explained further in the "Pipelines" section). The sample annotations are permanently kept in the database, and restored upon re-uploading of the same data with identical filenames, saving time for the users who need to re-analyse the same samples in the future.

3. Submit

This page (Figure 2c) allows job submission to a comprehensive number of bioinformatics pipelines for NGS data analysis (Table 1). Each pipeline has a user-friendly modal window with a simple interface for setting up the analysis design for the job. The required inputs may vary among the pipelines, although certain inputs such as sample files are common to all major pipelines. Wherever needed, information boxes are available to explain the pipeline options and parameters in further detail, as well as cross links to help direct users to the original documentation of the integrated tool. A number of control functions are also implemented at the back-end to prevent possible user mistakes before job submission. Even though the alignment step is required for all the pipelines starting from fastq file, the user may choose to opt out/in for certain analysis steps simply by un/selecting the check box, e.g. to skip/perform alternative splicing under RNA-Seq pipeline. Opting out for analyses not needed will benefit to the users by reducing the overall runtime for the job as well as the use of the server resources. The interface allows to add multiple samples per pipeline, and multiple pipelines per job within a single submission, which will run in parallel as needed. Currently, hg19 is the default human reference genome available to all pipelines, and other reference genomes (hg38, mm10, mm9) are additionally available to certain pipelines. A description file such as datasheet or metadata can be optionally attached to the job submission for future reference, which can be in any format. Also, an optional e-mail address field is available to receive notification upon job completion. Finally, a name and description of the project must be inserted before job submission for reference. The accuracy of the user inputs on this page is crucial to ensure the pipelines to run without failure, which can be further checked in the job details explained in the next section.

4. Jobs

This is the main page (Figure 2d) of the portal where the users have full control over their jobs. Specifically, users can:

1. check submitted job details to make sure everything is correct and as intended,

2. delete the entire job or the individual samples anytime,

3. monitor the job status if it is queued, running or finished,

4. monitor the job progress via real-time log with timestamp,

5. access to the output files in real-time for view/download,

6. share/unshare job results with other users anytime.

The jobs table by default displays only the most important fields due to the space constraint, which can be expanded by selecting more columns in the "Columns" action button. Alternatively, all the fields can be viewed with the "Details" button (+ sign) including overall job progress log. Importantly, the number of days to expiry is indicated in the job status column, after which the job and the associated data are automatically removed. The jobs can be shared with other users anytime, by simply inserting their usernames into the "Shared With" field separated by space(s) and/or comma: e.g. *topuser1,sevgi55 mike_86, john*. Likewise, usernames can be removed to unshare the job. Finally, the results column displays the output of the pipelines, as being the most important part of the page and possibly of the portal. The results become available stepwise in real-time, e.g. as the job keeps running before full completion, and all the output are available to directly view on the browser or to download to the local computer with single click. Alignment data (in .bam, .bigwig, .vcf formats) can be directly viewed in Integrative Genomics Viewer (IGV) [31] installed on the user's local computer without downloading the original files, allowing comparison

of samples between different pipelines and even different jobs. Due to the high volume of the NGS data, users are strongly encouraged to download the necessary output files and delete the job as soon as possible, which will help efficient usage of the storage and faster browsing experience on the portal for everyone.

5. Browse

This page (Figure 2e) provides interactive tables to browse the job results for the pipelines with reasonable size of text output ("DNA-Seq", "RNA-Seq", "RNA-Editing", "smallRNA", "4C-Seq", "circRNA"). As soon as such a job is finished, the results are automatically inserted into the database as a part of the pipeline, so that they become available for browsing and comparison. The results can be queried by either genomic feature (e.g. gene symbol, smallRNA identifier etc.) or genomic interval (e.g. chromosome, start, end) or both depending on the pipelines used. The query returns matching results from all the jobs under the same category belonging to the user as well as those that are shared with the user. For example, assuming a user has 3 DNA-Seq jobs and 2 other DNA-Seq jobs shared with him by his colleague, he will be able to compare the mutations in *gene X* across all the samples under the 5 jobs in a single query. The query results can be further tuned by using the action buttons on the table.

a) Upload

b) Annotate

c) Submit

Figure 2: Website Interface and Usage of CSI NGS Portal. a) Upload page **b)** Annotation page **c)** Submit page **d)** Jobs page **e)** Browse page. Key features and usage information are highlighted in text boxes. Further details on the quotas, rules and usage instructions are given on the "README" sections on the website.

Portal Features

Among many features of the portal (Table 2), the most powerful of them are its usability, modularity and flexibility. It is usable because of its simple design yet powerful functionality, and automation of not only the pipelines but also the data transitions between the pages. It is modular so that a new pipeline can be readily integrated to the portal complying with the existing website framework. It is flexible so that there is no restriction for the script language used in the background or the pipeline parameters collected from the user, making virtually any tool that is compatible with Linux environment also work on the portal.

Table 2: Features of CSI NGS Portal

	Full-automation	All the pipelines run from input to output without intervention with minimal user input.
	Usability	User-friendly and simple design with interactive tables having search, filter, sort, edit, export and share options.
	Modularity	Repertoire of pipelines is easy to expand complying with the existing website framework.
	Flexibility	Pipelines written in virtually any script language can be integrated independently of the website code.
	Transparency	The pipelines documentation are available online with the descriptions and the code.
	Responsive design	The website can be functionally displayed on multiple devices and platforms with different window/screen sizes.
	Quality control	FastQC report is auto-generated upon file upload, sequence and quality trimming are optionally available with options.
	User privacy	No personal information is collected, secure, random cookies for authorisation and dynamic usernames for data sharing are used.
	Data privacy	Data can be edited, deleted or shared only by the owner, expired data are completely removed from the server.
	Data sharing	Uploaded raw fastq files are private to the user, results can be optionally shared/unshared with other users any time.
	Data availability	Data is fully accessible via the portal until expiry (10 days, subject to revision upon usage and server capacity).
	Data download	All the data can be downloaded to local computer with a few clicks via browser and command line.
	IGV-integrated	Alignment (.bam, .bigwig) and mutation (.vcf) data can be viewed in local IGV without downloading the original files.
	Real-time logging	Real-time overall job progress log and individual tool log files are generated useful for tracking and debugging.
	E-mail notification	User is notified upon job completion if e-mail address is provided during job submission (optional).
	Parallelisation	Jobs are parallelised by multi-threading and by simultaneous run of multiple samples wherever possible.
	New pipelines	Popular and established bioinformatics pipelines for new data types are continuously added.
	Up-to-date	All the tools and packages are regularly updated to the latest stable versions available.

Icons are made by Freepik and obtained from www.flaticon.com.

Comparison to similar Platforms

CSI NGS Portal stands as a unique platform for fully automated analysis of NGS data. There is no other freely available and publicly accessible platform which offers identical service in terms of completeness, comprehensiveness and simplicity. There are few other resources, however, performing a similar job in providing NGS data analysis online which are compared to CSI NGS Portal next (Table 3).

Galaxy [1] (https://usegalaxy.org/) is currently the most popular project among these resources, standing as a freely available web server and open-source software for NGS data analysis, which expanded over 10 years as a scientific workflow management system towards data intensive biomedical research. However, Galaxy is primarily designed for users who have knowledge on how to build up a bioinformatics analysis pipeline (workflow) as it provides a stepwise (or tool-based) usage rather than complete pipelines, sometimes with redundant options for the individual steps and exhaustive parameters for each step. Therefore, it may not be user-friendly for a pure biologist without any prior bioinformatics skills. In comparison, CSI NGS Portal requires minimal user inputs, assuming no advanced bioinformatics knowledge from the users, who are focused on the results rather than the procedure, thereby targeting a wider user profile standing as a real user-friendly public tool for NGS data analysis. Specifically, the actions required from the users to submit a job are 1) upload raw data, 2) annotate samples, and 3) design analysis, each of which can be easily done with a few clicks. After the job submission, monitoring its progress, viewing/downloading/comparing the results and sharing them with other users are available in real-time on the portal. In addition, the portal provides detailed documentation of the pipelines and help on interface usage, and flexible to expand with new and customised pipelines from other labs

in the future. The existing Galaxy users can still manually export the job results from CSI NGS Portal to Galaxy in supported formats (.bam, .bigwig, .vcf, .bed, .txt) for further analysis. Thus, CSI NGS Portal adds a new asset to collaboratively support biomedical research with the existing platforms.

Another online platform developed for NGS big data analysis and sharing is Maser [2] (https://cell-innovation.nig.ac.jp/maser/), offering built-in bioinformatics pipelines and genome browser for data visualisation. Although the features of the two platforms are comparable in overall, CSI NGS Portal covers two times more NGS data types compared to Maser yet with a high overlap, and require no sign up allowing quicker access for the users. On the other hand, CSI NGS Portal provides UCSC track hubs [32] rather than an embedded browser for the supported pipelines (ChIP-Seq, RIP-Seq, Bisulfite-Seq).

It is noteworthy to mention many other efforts to analyse growing NGS data, which have been serving to the community as useful resources for years. However, these tools are not directly comparable to CSI NGS Portal as they either present pre-analysed datasets (databases), focus on a single domain (e.g. functional genomics by GenePattern [33], ZENBU [34], RNA-Seq by RAP [35], miRNA analysis by miRMaster [36]), or offer distributed bioinformatics software/framework/workflow systems consuming user's owned resources, rather than a publicly accessible online service, which require bioinformatics capability for the installation and large resources for the usage (GobyWeb [37], Eoulsan [38], Sequanix [39], Taverna [40], Arvados (https://doc.arvados.org/), Anduril [41], BioQueue [42], DolphinNext [43].

Table 3: Comparison of CSI NGS Portal to other NGS data analysis platforms

Platform Name	Number of pipelines/ NGS data types	Full pipelines	Data visualisation	Data sharing	Custom workflow building	Code availability	Local installation	Registration/ Login
CSI NGS Portal	14[a]	Yes	Static[1]	Yes	No	Pipeline level	In progress	Not required
Galaxy	Multiple[b]	No	Dynamic	Yes	Yes	Source level	Yes	Required
Maser	7[c]	Yes	Dynamic	Yes	Limited	No	No	Required
RAP	1[d]	Yes	Static	No	No	No	No	Required
miRMaster	1[e]	Yes	Static	No	No	No	No	Not required

Features common to all platforms such as user data upload, results download, job log, pipelines documentation etc. are omitted, and commercially available or paid platforms are excluded from the comparison.

[a]: DNA-Seq, RNA-Seq, Diff-Exp, Pathway-Enrichment, RNA-Editing, smallRNA, 4C-Seq, ChIP-Seq, RIP-Seq, SHAPE-Seq , rMATS, circRNA, eCLIP-Seq, Bisulfite-Seq [b]: Pipelines are available as workflows for different data types [c]: RNA-Seq, ChIP-Seq, Bisulfite-Seq, Exome-Seq, De novo genome sequencing, Metagenome, CAGE/SAGE-Seq [d]: RNA-Seq [e]: miRNA-Seq

[1]: Diverse plots in pdf format, comprehensive html reports, links to IGV and UCSC track hubs are provided depending on the pipeline and accessible via single clicks on the browser.

Conclusion

CSI NGS Portal is a unique platform which provides a free online service for fully automated NGS data analysis from raw data to final output with minimal user input. The portal covers most of the popular NGS data types and allows to share the results with colleagues. The website has a simple and user-friendly interface primarily designed for non-bioinformaticians, with detailed documentation of the pipelines and "README" information on the usage. The website is freely accessible and publicly available for academic use. With its comprehensive coverage and expanding potential, CSI NGS Portal stands as a promising and long-lasting resource fostering biomedical research.

Maintenance

All the tools and packages used in the pipelines are regularly updated to the latest stable versions available by using a Conda environment which ensures that there are no compatibility issues. The modified or new pipelines are carefully tested before release for public. Similarly, the data sources such as annotation databases are also updated and kept in the most comprehensive version available as a rule of thumb. Nevertheless, the users are encouraged to report any bugs or make suggestions to facilitate user experience via given contact e-mail on the portal.

In addition, the website is constantly monitored for usage in order to:

1. improve overall user experience,

2. improve portal performance,

3. reduce common user mistakes,

4. fix potential bugs,

5. prevent abuse.

User Privacy and Data Security

The portal ensures user privacy and data security by taking a set of precautions on the website and the server including:

- No record of real user information, e.g. no sign up or password requirement, usage of dynamic usernames for data sharing, optional e-mail address used only for job notification etc.,

- Cryptographically secure and randomly generated cookies for user recognition and data authorisation,

- Encrypted internet connection via https protocol,

- Server protection by a strict firewall,

- User-restricted data access and full control upon sharing, i.e. unshare and delete,

- Restriction of sensitive functions to data owner, such as delete, edit and share,

- Back-end control functions to prevent potential user mistakes,

- Backup of non-physical data i.e. sample annotations,

- Constant monitoring of website usage to prevent abuse.

To further strengthen data security, users are encouraged to pay attention to additional points including but not limited to:

- Avoid leaving computer unattended to prevent cookie theft,

- Download the results as soon as the job is finished and delete from the website,

- Share data with trusted people and with caution, e.g. a simple typo may cause sharing data with another user not intended,

- Report bugs as soon as encountered.

Limitations

Despite the expanding applications of automated systems in the global context, not every system can be automated in perfection without human intervention, so is true for the bioinformatics systems. Automation on CSI NGS Portal inherently bears a certain level of trade-off between speed and complexity in the NGS data analysis. For the sake of speed and ease of use, the portal is designed as to require minimal user input without advanced options, which may not always provide the power of fine-tuning the analysis for the individual tools and functions in the pipelines. However, for the majority of cases, standard pipelines with default parameters result in the desired output by the user and are sufficient for the downstream analysis, particularly in comparison studies.

Future Work

The portal will be expanded with new NGS data analysis pipelines such as neoepitope prediction for personalised medicine and single cell RNA-Seq in the near future. Other popular pipelines will be added to the portal on demand. We also encourage the users to submit their own pipelines to share with the community. There is also room for improvement of the existing pipelines, e.g. cross-pipelines analysis, availability of more reference genomes, as well as of the website interface, e.g. built-in visualisation tool. Building a docker container for the local installation of the web application is in progress.

Funding

This research was supported by The National Research Foundation Singapore; The Singapore Ministry of Education under its Research Centres of Excellence initiative; and RNA Biology Centre at CSI Singapore NUS from funding by the Singapore Ministry of Education's Tier 3 grants [MOE2014-T3-1-006].

Acknowledgements

We greatly thank to all the lab members of CSI Singapore NUS for their usage, feedback and support of CSI NGS Portal.

Authors' Contributions

H.Y. conceived the project. O.A. designed, developed and maintains CSI NGS Portal. K.T.T. wrote the initial code for DNA-Seq and RNA-Editing pipelines, Y.L. wrote the code for smallRNA pipeline, J.L. wrote the code for alternative splicing part of RNA-Seq pipeline, C.S.W. wrote the code for ChIP-Seq, RIP-Seq and Bisulfite-Seq pipelines, B.Z. wrote the code for circRNA pipeline, O.A. wrote the code for the remaining pipelines and integrated all the pipelines to the portal. C.L. and H.Y. supervised the project. O.A. wrote the manuscript, and all authors read and approved the final version of the manuscript.

Disclosure Declaration

The authors declare that they have no competing interests.

References

1.   Afgan E, Baker D, Batut B et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update, Nucleic Acids Res 2018;46:W537-W544.

2.   Kinjo S, Monma N, Misu S et al. Maser: one-stop platform for NGS big data from analysis to visualization, Database (Oxford) 2018;2018.

3.   Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 2009;25:1754-1760.

4.   McKenna A, Hanna M, Banks E et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res 2010;20:1297-1303.

5.   Cibulskis K, Lawrence MS, Carter SL et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples, Nat Biotechnol 2013;31:213-219.

6.   Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, Nucleic Acids Res 2010;38:e164.

7.   Dobin A, Davis CA, Schlesinger F et al. STAR: ultrafast universal RNA-seq aligner, Bioinformatics 2013;29:15-21.

8.   Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data, Bioinformatics 2015;31:166-169.

9.   Patro R, Duggal G, Love MI et al. Salmon provides fast and bias-aware quantification of transcript expression, Nat Methods 2017;14:417-419.

10.   Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol 2014;15:550.

11.   Collado-Torres L, Jaffe AE, Leek JT. regionReport: Interactive reports for region-level and feature-level genomic analyses, F1000Res 2015;4:105.

12.   Yu G, He QY. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization, Mol Biosyst 2016;12:477-479.

13.   Subramanian A, Tamayo P, Mootha VK et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc Natl Acad Sci U S A 2005;102:15545-15550.

14.   Yu G, Wang LG, Yan GR et al. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis, Bioinformatics 2015;31:608-609.

15.   Li H, Handsaker B, Wysoker A et al. The Sequence Alignment/Map format and SAMtools, Bioinformatics 2009;25:2078-2079.

16.   Ramaswami G, Zhang R, Piskol R et al. Identifying RNA editing sites using RNA sequencing data alone, Nat Methods 2013;10:128-132.

17.   Thongjuea S, Stadhouders R, Grosveld FG et al. r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data, Nucleic Acids Res 2013;41:e132.

18.   Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2, Nat Methods 2012;9:357-359.

19.   Zhang Y, Liu T, Meyer CA et al. Model-based analysis of ChIP-Seq (MACS), Genome Biol 2008;9:R137.

20.   Heinz S, Benner C, Spann N et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities, Mol Cell 2010;38:576-589.

21.   Flynn RA, Zhang QC, Spitale RC et al. Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE, Nat Protoc 2016;11:273-290.

22.   Lorenz R, Hofacker IL, Stadler PF. RNA folding with hard and soft constraints, Algorithms Mol Biol 2016;11:8.

23.   Lorenz R, Bernhart SH, Honer Zu Siederdissen C et al. ViennaRNA Package 2.0, Algorithms Mol Biol 2011;6:26.

24.   Shen S, Park JW, Lu ZX et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data, Proc Natl Acad Sci U S A 2014;111:E5593-5601.

25.   Van Nostrand EL, Pratt GA, Shishkin AA et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP), Nat Methods 2016;13:508-514.

26.   Van Nostrand EL, Nguyen TB, Gelboin-Burkhart C et al. Robust, Cost-Effective Profiling of RNA Binding Protein Targets with Single-end Enhanced Crosslinking and Immunoprecipitation (seCLIP), Methods Mol Biol 2017;1648:177-200.

27.   Lovci MT, Ghanem D, Marr H et al. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges, Nat Struct Mol Biol 2013;20:1434-1442.

28.   Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications, Bioinformatics 2011;27:1571-1572.

29.   Cock PJ, Fields CJ, Goto N et al. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants, Nucleic Acids Res 2010;38:1767-1771.

30.   Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 2014;30:2114-2120.

31.   Robinson JT, Thorvaldsdottir H, Winckler W et al. Integrative genomics viewer, Nat Biotechnol 2011;29:24-26.

32.   Raney BJ, Dreszer TR, Barber GP et al. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser, Bioinformatics 2014;30:1003-1005.

33.   Reich M, Liefeld T, Gould J et al. GenePattern 2.0, Nat Genet 2006;38:500-501.

34.   Severin J, Lizio M, Harshbarger J et al. Interactive visualization and analysis of large-scale sequencing datasets using ZENBU, Nat Biotechnol 2014;32:217-219.

35.   D'Antonio M, D'Onorio De Meo P, Pallocca M et al. RAP: RNA-Seq Analysis Pipeline, a new cloud-based NGS web application, BMC Genomics 2015;16:S3.

36.   Fehlmann T, Backes C, Kahraman M et al. Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs, Nucleic Acids Res 2017;45:8731-8744.

37.   Dorff KC, Chambwe N, Zeno Z et al. GobyWeb: simplified management and analysis of gene expression and DNA methylation sequencing data, PLoS One 2013;8:e69666.

38.   Jourdren L, Bernard M, Dillies MA et al. Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses, Bioinformatics 2012;28:1542-1543.

39.   Desvillechabrol D, Legendre R, Rioualen C et al. Sequanix: a dynamic graphical interface for Snakemake workflows, Bioinformatics 2018;34:1934-1936.

40.   Wolstencroft K, Haines R, Fellows D et al. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud, Nucleic Acids Res 2013;41:W557-561.

41.   Cervera A, Rantanen V, Ovaska K et al. Anduril 2: Upgraded large-scale data integration framework, Bioinformatics 2019.

42.   Yao L, Wang H, Song Y et al. BioQueue: a novel pipeline framework to accelerate bioinformatics analysis, Bioinformatics 2017;33:3286-3288.

43.   Yukselen O, Turkyilmaz O, Ozturk A et al. DolphinNext: A distributed data processing

platform for high throughput genomics. bioRxiv doi: 10.1101/689539.