RUbioSeq+: a multiplatform application that executes parallelized pipelines to analyse next-generation sequencing data

Miriam Rubio-Camarillo[1]*, Hugo López-Fernández[2,3]*[$], Gonzalo Gómez-López[1], Ángel Carro[1], José María Fernández[4], Coral Fustero Torre[1], Florentino Fdez-Riverola[2,3], Daniel Glez-Peña[2,3].

[1]Bioinformatics Unit (UBio), Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain; [2]Higher Technical School of Computer Engineering, University of Vigo, Ourense, Spain; [3]Instituto de Investigación Biomédica de Vigo (IBIV), Vigo, Spain; [4]Spanish National Bioinformatics Institute (INB), INB Node 2, Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), 28029 Madrid, Spain.

* These authors contributed equally to this work

[$] Corresponding author

Abstract

Background and Objective: To facilitate routine analysis and to improve the reproducibility of the results, next-generation sequencing (NGS) analysis requires intuitive, efficient and integrated data processing pipelines. **Methods**: We have selected well-established software to construct a suite of automated and parallelized workflows to analyse NGS data for DNA-seq (single-~~-~~nucleotide variants (SNVs) and indels), CNA-seq, bisulfite-seq and ChIP-seq experiments. **Results:** Here, we present RUbioSeq+, an updated and extended version of RUbioSeq~~ (ref)~~, a multiplatform application that incorporates a suite of automated and parallelized workflows to analyse NGS data. This new version ~~, that now ~~includes: (i) an interactive graphical user interface (GUI) that facilitates its use by both biomedical researchers and bioinformaticians, (ii) a new pipeline for ChIP-seq experiments, (iii) pair-wise comparisons (case-control analyses) for DNA-seq experiments, (iv) and improvements in the parallelized and multithreaded execution options. ~~novel and multiplatform application that incorporates a suite of automated and parallelized workflows to analyse NGS data. The software supports DNA-seq (single-nucleotide and copy number alteration analyses) for panels, and for exome and whole genome sequencing experiments, as well as for bisulfite-seq and ChIP-seq workflows. RUbioSeq+ supports parallelized and multithreaded execution, and its interactive graphical user interface~~

~~(GUI) facilitates its use by both biomedical researchers and bioinformaticians.~~ Results generated by our software have been experimentally validated and accepted for publication. **Conclusions:** RUbioSeq+ is free and open to all users at http://rubioseq.bioinfo.cnio.es/.

Keywords: NGS analysis, parallelized workflows, whole-genome, variant calling, ChIPSeq, Bisulfite-Seq, CNV, HPC, SGE.

1. Introduction

The increasing use of next-generation sequencing (NGS) studies has revealed the need for integrated and reliable pipelines to analyse deep-sequencing experiments in a reproducible way. This issue is especially relevant in hospitals and research institutes where regular analyses accentuates the demand for intuitive and automated workflows that accelerate the delivery of final results, minimizing human technical error, and ensuring the reproducibility and fidelity of the data obtained (1).

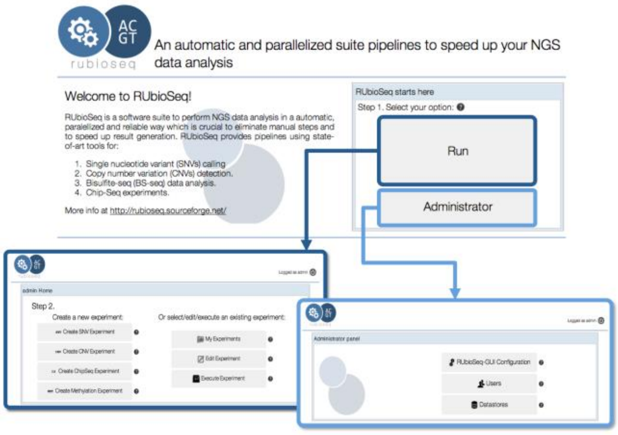
NGS data is usually analysed in a set of successive stages that are executed routinely. Highly specific software exists to carry out each of these particular steps, constituting a growing and diverse catalogue that includes quality control utilities, read aligners, variant callers, peak finders, functional annotators, mutational impact predictors, etc. (2, 3). In such a diverse scenario, the quality of those NGS tools is very heterogeneous in terms of implementation and documentation, and in many cases the applications are complicated for non-specialist users to employ, requiring a solid expertise in bioinformatics to manage them properly.

In the light of this situation, a number of initiatives have been proposed to provide effective solutions that facilitate the systematic analysis of NGS experiments. For example, HugeSeq (4), Bcbio-nextgen (5) and Omics-pipe (6) are recent open-source pipelines available to analyse NGS data in an automated and proficient manner. These reflect the remarkable effort to provide powerful specialized computational frameworks to analyse NGS data. However, they usually lack an adequate Graphic User Interface (GUI) and thus, they are complicated for researchers without computational or bioinformatics skills to use. Besides, the current versions of HugeSeq and Bcbio-nextgen do not support the analysis of several seq-based experiments, such as copy-number alteration sequencing (CNA-seq), bisulfite-seq or ChIPseq. Galaxy (7) represents a large and flexible web-based platform that provides support for NGS experiments (e.g. ChIPseq, variant analysis, etc.) and although it may be installed as a local server, its set up requires advanced technical skills. Despite its potential, Galaxy's NGS toolbox is still in beta state and it does not support either CNA-seq or bisulfite-seq analysis. Moreover, using it online demands high bandwidth capacities due to the size of the NGS files. Other interesting proposals such as GeneProf (8) or CisGenome (9) provide ChIPseq analysis through an interactive GUI, but they do not provide support for other types of NGS techniques.

Here we present RUbioSeq+, a multi-platform application for the integrated analysis of NGS data. This software uses well-established tools to implement pipelines for DNA-seq, CNA-seq, bisulfite-seq and ChIP-seq experiments. RUbioSeq+ is free and it includes the entire core functionalities implemented in the original release of RUbioSeq (10), while expanding the capability of RUbioSeq by supporting parallelized analysis of full genomes in computing farms. Moreover, we have included two novel pipelines for ChIPseq analysis covering sample quality control, read alignments, assessment of biological replicates, and peak calling for the detection of histone marks and transcription factors binding sites. RUbioSeq+ also incorporates a new and accessible GUI, designed for interdisciplinary research groups where bioinformaticians and biomedical researchers work together. Thus, the GUI supports two types of profiles: a) a basic user profile through which users with limited skills in bioinformatics can execute all the NGS analysis tasks offered by the software; and b) an administration mode where bioinformaticians and advanced users can manage and configure all the technical parameters in the application (Figure 1). To our knowledge there is currently no other application with similar characteristics, as well as such a large number of automated workflows and utilities for NGS data analysis (Table 1).

Table 1. RUbioSeq+ functionalities compared with state-of-the-art NGS tools.

	Quality Control	Alignment	SNVs	CNVs	ChIP-Seq	Bisulfite-Seq	Annotation	HPC	GUI
RubioSeq+	●	●	●	●	●	●	●	●	●
Omics-pipe	●	●	●		●			●	
Bcbio-nextgen (2013)	●	●	●	●	●		●	●	
HugeSeq (2012)		●	●	●			●	●	
MutationTaster (2014)		●	●						
Ngs-backbone (2011)	●	●	●					●	
SeqGene (2011)			●	●			●		
SHORE (2008)		●	●	●	●		●		
GeneProf (2014)	●	●			●			●	●
HiChIP (2014)	●	●			●			●	
Cisgenome (2011)					●		●		●
Haemcode (2013)	●	●			●				
MethylSig (2014)						●	●		
SVmerge (2010)		●		●					
m-HMM (2014)				●					
HMMCopy (2012)				●					
SVseq (2011)		●		●					
WaveCNV (2014)				●					
TOGGLE (2015)	●	●	●				●	●	
NEAT (2015)	●	●			●			●	●
Churchil (2015)	●	●	●	●				●	
GotCloud (2015)	●	●	●	●				●	



An automatic and parallelized suite pipelines to speed up your NGS data analysis

Welcome to RUbioSeq!

RUbioSeq is a software suite to perform NGS data analysis in a automatic, paralellized and reliable way which is crucial to eliminate manual steps and to speed up result generation. RUbioSeq provides pipelines using state-of-art tools for:

1. Single nucleotide variant (SNVs) calling
2. Copy number variation (CNVs) detection.
3. Bisulfite-seq (BS-seq) data analysis.
4. Chip-Seq experiments.

More info at http://rubioseq.sourceforge.net/

Figure 1. RubioSeq+'s main interface. Standard users may easily access all the analyses supported by RUbioSeq+ through its main interface. Alternatively, the advanced administrator profile provides access to the technical configuration and user's administration.

2. Materials and methods

RUbioSeq+ is written in Perl language. It is designed to run on ordinary UNIX workstations and it has been successfully tested in Linux and Mac OS X, as well as on HPC systems with SGE or PBS as cluster job schedulers. Windows users may also run RUbioSeq using the Docker client (http://www.docker.com). Its modular programming design provides a high degree of flexibility to facilitate the creation of shared libraries and functions through the distinct execution branches (e.g., the cluster job management module) that stabilizes the code, helps with software maintenance and avoids code redundancy. This facet will also facilitate the inclusion of additional functionalities and extensions in future versions of the tool.

2.1 Third-party software integration

In light of the overwhelming list of applications currently available for NGS data analysis, we have selected well-established software to construct pipelines for DNA-seq, CNA-seq, bisulfite-seq and ChIP-seq experiments (Table 2).

Table 2. Software included in RUbioSeq+. All the tools included in RUbioSeq have been upgraded in RUbioSeq+.

		RUbioSeq+	RUbioSeq
Shared Software	Java	1.7	1.6
	Samtools	0.1.19	0.1.16
	PicardTools	1.107	1.6
	FastQC	0.10.1	0.10.1
	BWA	0.7.10	0.6.2
	Bfast-bwa	0.7.0b	0.7.0a
	VCFtools	Not used	0.1.19
	BEDTools	2.16.2	2.16.1
	GATK	3.1-1	2.3.9
SNV	VEP	73	66
CNV	CONTRA	2.0.3	2.0.3
ChipSeq	MACS2	2.0.10	Not used
	CCAT	3	Not used
	IDR	1.0	Not used
Methylation	Bismark	0.10.1	0.7.7
	Bowtie	0.12.7	0.12.7
	Fastx Toolkit	0.0.13.2	0.0.13.2
	FiloTools	1.1.0	Not used

The current release of RUbioSeq+ integrates more than 20 tools and utilities that can be used for NGS data analysis. Each tool included follows the open-source initiative and has been described previously in the scientific literature. RUbioSeq+ saves user's time by providing a quick and easy installation toolkit. The full set of tools integrated in RUbioSeq+ may be installed in 3 ways. First, through a LiveCD that contains both the whole RUbioSeq+ suite and its GUI, along with all the documentation and dependencies. This LiveCD has been built from scratch, taking as a base system the latest stable Ubuntu x86_64 release (14.04.1 at the time of writing), and it also contains the LXDE graphics desktop (from the Lubuntu release). The RUbioSeq+ LiveCD can also be used to install RUbioSeq+ plus Ubuntu on any x86_64 machine that fulfils the RUbioSeq+ memory requirements, either real or virtual. Second, a Docker image (ubio/rubioseq:latest) stored in the public Docker Hub. This option permits users to deploy the container on Unix-based or Windows systems, on desktops, physical servers, virtual machines, in data centres, and even via public and private clouds. Third, a simple script provided along with the software allows the user to easily install the software and the dependencies (only for UNIX-based systems). Additionally, our software admits multithreading and parallelized processing in computational farms, which allows multiple samples to be analysed simultaneously, allowing the rapid and efficient generation of results.

RUbioSeq+ has two branches at the level of the source code: (a) the RUbioSeq+ source code itself, which is publicly available in SourceForge (http://rubioseq.sourceforge.net); and (b) RUbioSeq+GUI, a project for the logic web application and a further development for the persistence storage, which is publicly available in GitHub (https://github.com/hlfernandez/rubioseq-gui-webapp).

2.2 Level-based design

Workflows involve operative modules acting at different functional levels, each level representing an independent stage of the analysis (e.g. alignment level, calling level). The level-based design allows users to launch the workflows in their entirety or partially, depending on the laboratory requirements, thereby providing the greatest flexibility and speed when running the analyses using different parameters.

For example, the ChIPseq workflow could be executed at four levels depending on the input stage: 1) alignment phase and FastQC analysis if the user starts with unaligned reads; 2) duplicate markings if the reads have been aligned but the user wants to check the level of library duplication; 3) normalization steps if the libraries to be compared have to be balanced; and/or 4) straightforward ChIPseq calling, peak annotation and IDR control if allowed by the input data. Once the workflow has been fully executed, the users might want to modify the parameters of the peak caller and rerun the analysis from step 3, without having to execute the full workflow. Each round of execution will generate the associated result files, together with their corresponding process logs for identification.

Moreover, to adapt RUbioSeq+ to the analysis of whole genomes, we reimplemented the alignment step using a parallelized design. Thus, starting from the fragmented raw data files supplied by sequencers, RUbioSeq+ processes the fragments side-by-side to generate a single alignment file that can be used as an input in the next stage. This implementation saves time and computational demands, performing the analysis of whole genomes in an efficient manner.

2.3 Graphical User Interface

The RUbioSeq+ GUI is implemented as an AJAX-enabled web application programmed in Java 1.7. The ZK development framework was used to construct a rich web user interface with many features of a desktop application, and a HSQLDB (HyperSQL DataBase) is used to store the application data (users, configurations, experiments, etc.). Since the user interface has been developed using web technologies, the RUbioSeq+GUI can be used in two different modes: (a) as stand-alone application using Jetty Runner 8.1; or (b) as a web front-end that is installed in a dedicated web server. The second option is especially useful for users of a bioinformatics unit that have at their disposal a computer cluster or dedicated server. In such a scenario, RUbioSeq+ can be installed into the head node of the cluster so that users can access RubioSeq+GUI via the server, configuring their experiments there and launching analyses that will be executed by the cluster machines.

The RUbioSeq pipeline and RUbioSeq+ GUI has been integrated as follows. First, RUbioSeq+ GUI provides an interface to visually configure experiments, avoiding the need for users to manually write the XML RUbioSeq experiment configuration files. The XML files generated then feed a new external background process in RUbioSeq that has been installed and configured previously. Finally, the progress of the experiments running is monitored by reading the standard RUbioSeq log files. In summary, the integration of the two modules is achieved in a file-based and loosely coupled architecture (Figure 2).

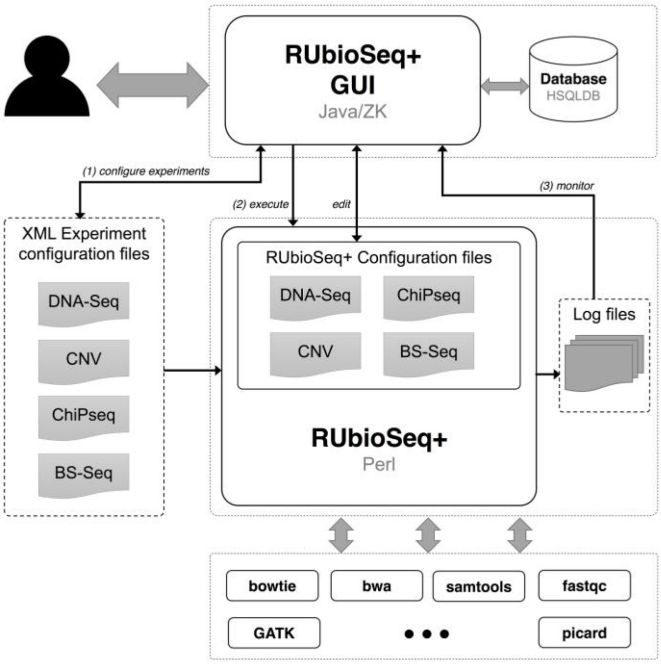


Figure 2. Integration between the RUbioSeq pipeline and RUbioSeq+GUI. (1) The GUI allows the user to visually create experiments that are then represented in XML files and passed to the pipeline for execution (2). The pipeline is launched when requested by the user via the GUI (3) and as the experiment is executed, log files are created by the pipeline. The GUI monitors these files in real time in order to present the user a progress meter. Finally, the GUI can also be used to configure the pipeline.

3. Results and discussion

We present here RUbioSeq+, a novel and improved version of the RUbioSeq suite for the analysis of NGS data. Our application consists of a multiplatform collection of automated and parallelized pipelines to analyse DNA-seq, CNA-seq and bisulfite-seq experiments. Additionally, RUbioSeq+ includes two new ChIP-seq workflows based on MACS2 and CCAT tools for peak calling. The software can provide an intuitive GUI that can be implemented to support integrated workflows for bioinformaticians and biomedical researchers, and it is being used extensively at our institute for both research and clinical purposes.

3.1 Workflows

RUbioSeq+ workflows are divided into distinct analytical branches that may be executed independently to analyse distinct NGS experiments. Moreover, the default parameters have been tweaked based on developer's recommendations and on our daily experience performing routine NGS analyses for biomedical researchers.

Our application automates the NGS analysis, reducing human errors and improving the reproducibility of deep-sequencing studies. To this aim, the software generates an XML configuration file that is associated to every execution and may be imported to reproduce the same technical conditions of any particular implementation. Additionally, RUbioSeq+ provides an intuitive framework in which each pipeline can be checked and controlled. Thus, quality control of the sample is addressed in every workflow, and all the processes are monitored through log files to trail the progress and errors at both the sample and data analysis levels. The complete results of RUbioSeq+ are saved in a project directory tree that maintains a structured organization for the output files.

Our software employs standard formats of input and output files, supporting both single and paired-end experiments. The execution of each pipeline with our software starts from either raw data files (FASTQ and ABI SOLiD formats) or aligments (BAM, see Figure 3), after which RUbioSeq+ is very flexible and adjustable, allowing the users to customize the parameters employed at each stage of the workflow.

In order to assure the reproducibility of the workflows employed, the software automatically exports the full technical configuration of the pipeline into a .XML file. When required, this configuration file can be reloaded into RUbioSeq+, replicating the technical conditions used in the original execution. By reusing configuration files, every analysis performed by RUbioSeq+ may be controlled, evaluated and reproduced systematically.

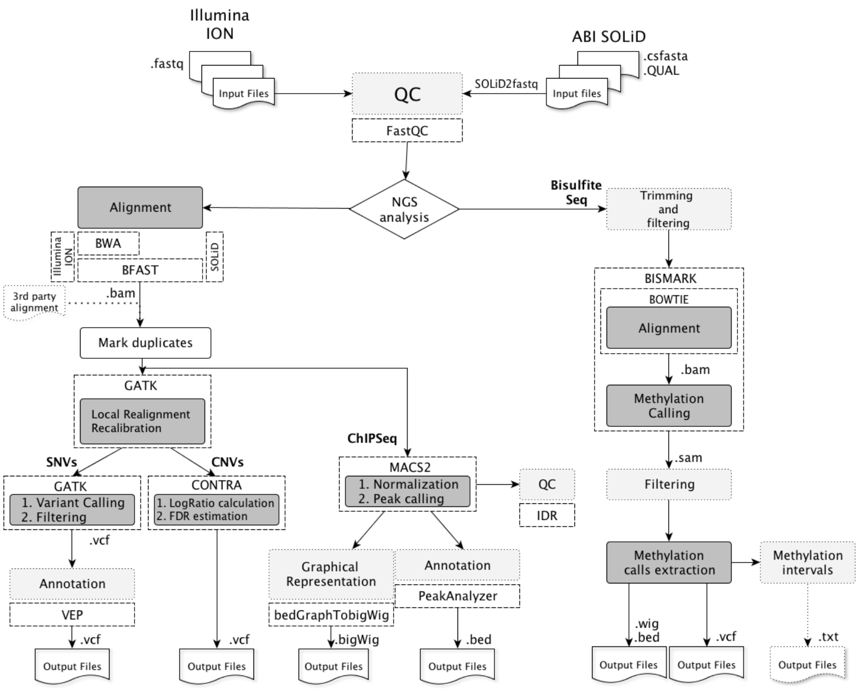


Figure 3. Schematic RUbioSeq+ flow diagram. The dark gray boxes correspond to the main steps in the pipelines, while the light gray boxes indicate the optional steps.

~~2~~3.~~2~~1.1 ChIPseq

RUbioSeq+ provides two novel workflows to analyze ChIPseq experiments. Using FASTQ or BAM files as primary input, RUbioSeq+ can apply MACS2 (11) and CCAT (12) to detect sharp or broad peaks, respectively.

The main steps in the workflows available include: a) Quality control through FastQC (13); b) short-read alignment with BWA (14); c) duplicate marking with Picard tools (15); d) normalization performed with SAMtools (16) in which the files are equilibrated in terms of size to make the experiments comparable; e) peak calling with MACS2 or CCAT; f) an optional step to assess the reproducibility of biological replicates using IDR (17); and f) peak annotation with PeakAnnotator (18).

In both the MACS2 and CCAT workflows, peak calling steps are performed on each sample. This strategy allows experiments to be analysed in parallel, making the protocol adaptable to support a variety of experimental ChIPseq designs (e.g. case+input, case vs control, case+input vs control+input, etc.).

The output of this workflow includes a standard BED file containing the coordinates and annotation of the significant peaks detected, and a bigWig file to upload and display the results in a genome browser.

~~2~~3.~~2~~1.2 DNA-seq

This workflow accepts FASTQ, ABI SOLID raw data or BAM files as the primary input to analyse ~~single nucleotide variants~~SNVs and indels in full genomes and exomes. The pipeline is divided into four main modules: (a) short-read alignment with a combination of BWA~~ and ~~+BFAST aligners(19), and a quality control analysis using FastQC; (b) duplicate marking using Picard tools, realignment and recalibration using GATK (20, 21); (c) GATK variant calling (variant standard database annotation) and advanced filtering permitting the user to choose between Hard Filtering using GATK's VariantFiltration walker or GATK's variant quality score recalibrator (VQSR); and (d) Pair-wise comparisons to detect variants (e.g. case vs control). Variant impact is calculated using Ensembl Variant Effect Predictor (VEP, 22). All the output files are generated in standard formats, such as BAM and VCF.

~~2~~3.~~2~~1.3 CNA-seq

RUbioSeq+ performs CNV detection for exome sequencing experiments using FASTQ and BAM as the input files. This workflow uses: (a) BWA and BFAST aligners combination to generate BAM files; (b) GATK to mark duplicates and for alignment recalibration; and (c) CONTRA software (23) to estimate case-control log ratios for each region targeted, and to evaluate the false discovery rate (FDR) of the gains and losses detected, generating the output files in standard VCF format.

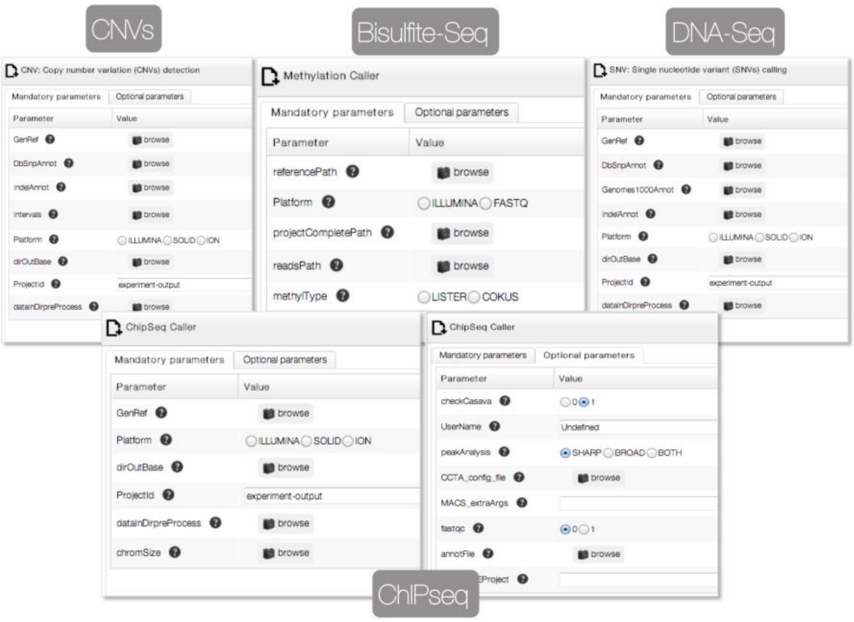
~~2~~3.~~2~~1.4 Bisulfite-seq

RUbioSeq+ integrates a bisulfite-seq workflow that supports the analysis of full genomes from FASTQ files. The workflow extracts methylation calls for each independent sample included in a particular experiment, but it can also construct a unique file with the whole calling information from the full set of samples included in such experiments. The pipeline includes: (a) quality control, sequence alignment and methylation calling with Bismark (24); (b) methylation call extraction in standard VCF format; and (c) an optional calculation of the percentage methylation in specific intervals.

3.2 Using RUbioSeq+

Since legal requirements impede many laboratory computers holding sensitive data from being connected to the internet, RUbioSeq+ can also be run on a computer that is offline. In this particular case, some of the online functionalities required by RUbioSeq+ (e.g., VEP) must be installed locally onto the computer. Accordingly, RUbioSeq+'s administrator would have to set up VEP to run on the local installation.

The RUbioSeq+ GUI also facilitates the administration of the tasks it carries out and thus, managing the technical configuration of RUbioSeq+ is straightforward when handled through the administrator's profile. Similarly, the specific parameters of each pipeline may be easily set up by standard users using intuitive menus (Figure 4).



CNV: Copy number variation (CNVs) detection

Mandatory parameters | Optional parameters

Parameter	Value
GenRef	browse
DbSnpAnnot	browse
indelAnnot	browse
Intervals	browse
Platform	○ ILLUMINA ○ SOLID ○ ION
dirOutBase	browse
ProjectId	experiment-output
datainDirpreProcess	browse

Methylation Caller

Mandatory parameters | Optional parameters

Parameter	Value
referencePath	browse
Platform	○ ILLUMINA ○ FASTQ
projectCompletePath	browse
readsPath	browse
methylType	○ LISTER ○ COKUS

SNV: Single nucleotide variant (SNVs) calling

Mandatory parameters | Optional parameters

Parameter	Value
GenRef	browse
DbSnpAnnot	browse
Genomes1000Annot	browse
IndelAnnot	browse
Platform	○ ILLUMINA ○ SOLID ○ ION
dirOutBase	browse
ProjectId	experiment-output
datainDirpreProcess	browse

ChipSeq Caller

Mandatory parameters | Optional parameters

Parameter	Value
GenRef	browse
Platform	○ ILLUMINA ○ SOLID ○ ION
dirOutBase	browse
ProjectId	experiment-output
datainDirpreProcess	browse
chromSize	browse

ChipSeq Caller

Mandatory parameters | Optional parameters

Parameter	Value
checkCasava	○ 0 ● 1
UserName	Undefined
peakAnalysis	● SHARP ○ BROAD ○ BOTH
CCTA_config_file	browse
MACS_extraArgs	
fastqc	● 0 ○ 1
annotFile	browse
EProject	

Figure 4. RUbioSeq+ uses automated pipelines to detect genomic ~~single nucleotide variants~~SNVs, indels, copy number variations, methylated positions and ChIPseq peak signals. Each workflow is displayed in a graphic menu, where the mandatory and optional parameters are accessed, along with a comprehensive help section.

Finally, despite depending on more than 20 different software packages, some of them difficult to install and setup, the installation of RUbioSeq+ is a straightforward process. The software is easy to configure and examples of its use are provided as step-by-step video tutorials. In addition, three installation options are provided: 1) a customized 64-bit LiveDVD based on Ubuntu 14.04.1 on which RUbioSeq+ and all its dependencies are bundled, ready to be used on any computer; 2) A Docker image (ubio/rubioseq:latest) stored in the public Docker Hub; and 3) Manual installation through a simple script provided with the software.

4. Conclusions

RUbioSeq+ is a multiplatform application for the integrated analysis of NGS data. Our software implements pipelines for the analysis of SNVs~~single nucleotide~~, indels, copy-number variation, bisulfite-seq and ChIP-seq experiments using well-established tools to perform these common tasks. The results obtained by RUbioSeq+ have already been validated and published (25, 26). Moreover, RUbioSeq+ comes with a new and accessible GUI enabling researchers without advanced bioinformatics skills to straightforwardly manage complex NGS workflows.

5. Mode of availability

The RUbioSeq+ GUI developed by the SING group is licensed under a GNU GPL 3.0 License (http://www.gnu.org/copyleft/gpl.html). Furthermore, the RUbioSeq+ software and full documentation are free and publicly available under Creative Commons License at http://rubioseq.bioinfo.cnio.es.

Authors' contributions

~~MR-C, GG-L and DGP conceived, coordinate and design RUbioSeq+ project;~~ MR-C developed RUbioSeq+ source code; HL-F and DG-P developed RUbioSeq+ GUI; MR-C and GG-L conceived, coordinate and design RUbioSeq+ project; GG-L, CFT and DG-P wrote the manuscript; ~~CFT implemented the manuscript;~~ FF-R ~~and DG-P~~ drafted the manuscript critically; MR-C, GG-L and AC tested the application; AC and MR-C built and tested RUbioSeq+ docker image and~~; AC~~ implemented RUbioSeq website; JMF built RUbioSeq+ LiveDVD and the installation script.

Acknowledgements

We are thankful to CNIO's Translational Bioinformatics Unit staff for their assistance in beta testing and for their useful suggestions. We also thank O. Graña, E. Carrillo-de-Santa-Pau and F. Al-Shahrour for their advice regarding the manuscript and for fruitful discussions. SING group thanks CITI (*Centro de Investigación, Transferencia e Innovación*) from University of Vigo for hosting its IT infrastructure.

Funding

M.R-C is funded by the BLUEPRINT Consortium (FP7/ 2007-2013) under grant

agreement number 282510. J.M.F is funded by the INB Node 2 - CNIO, a member of Proteored - PRB2-ISCIII and is supported by grant PT13/0001, of the PE I+D+i 2013-2016, funded by ISCIII and FEDER. H.L-F is funded by a postdoctoral fellowship from the Xunta de Galicia. F.F-R and D.G-P are funded by the European Union's Seventh Framework Programme FP7/REGPOT 2012 2013.1 under grant agreement n° 316265 (BIOCAPS) and the "Platform of integration of intelligent techniques for analysis of biomedical information" project (TIN2013-47153-C3-3-R) financed by the Spanish Ministry of Economy and Competitiveness.
C.FT is funded by the "Spanish National Youth Guarantee Implementation Plan" (2013/2016) financed by the Spanish Ministry of Economy and Competitiveness.

Conflict of interest statement

The authors have no conflicts of interest to declare.

References

1) Brown, JR, Dinu V: High performance computing methods for the integration and analysis of biomedical data using SAS. Comput Methods and Programs Biomed. 2013 Dec; 112(3): 553-562.

2) Trapnell C, Salzberg SL: How to map billions of short reads onto genomes. *Nat Biotech*. 2009, 27(5):455-457.

3) Ding L, Wendl MC, McMichael JF, Raphael BJ: Expanding the computational toolbox for mining cancer genomes. *Nat Rev Genet*. 2014, Aug;15(8):556-70.

4) Lam HY, Pan C, Clark MJ, Lacroute P, Chen R, Haraksingh R, O'Huallachain M, Gerstein MB, Kidd JM, Bustamante CD, Snyder M: Detecting and annotating genetic variations using the HugeSeq pipeline. *Nat Biotechnol.* 2012, Mar 7;30 (3):226-9.

5) https://bcbio-nextgen.readthedocs.org

6) https://bitbucket.org/sulab/omics_pipe

7) Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*. 2010, Aug 25;11(8):R86.

8) Halbritter F, Vaidya HJ, Tomlinson SR: GeneProf: analysis of high-throughput sequencing experiments. *Nat Methods* 2011, Dec 28;9(1):7-8.

9) Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH: An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol*. 2008, Nov;26(11):1293-300.

10) Rubio-Camarillo M, Gómez-López G, Fernández JM, Valencia A, Pisano DG: RUbioSeq: a suite of parallelized pipelines to automate exome variation and bisulfite-

seq analyses. *Bioinformatics* 2013*,* Jul 1;29(13):1687-9.

11) Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* 2008, 9: R137.

12) Xu H1, Handoko L, Wei X, Ye C, Sheng J, Wei CL, Lin F, Sung WK: A signal-noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics* 2010, May 1;26(9):1199-204.

13) http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

14) Li H and Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009, Jul 15;25(14):1754-60.

15) http://broadinstitute.github.io/picard/

16) Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup: The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 2009, 25, 2078-9.

17) Li Q, Brown J, Huang H, Bickel P: Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* 2011, 5:1752–1779.

18) Salmon-Divon M, Dvinge H, Tammoja K, Bertone P: PeakAnalyzer: Genome-wide annotation of chromatin binding and modification loci. BMC *Bioinformatics* 2010, 11: 415.

19) Homer N, Merriman B, Nelson SF. BFAST: an alignment tool for large scale genome resequencing. *PLoS One* 2009, Nov 11;4(11):e7767.

20) DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A,

Cibulskis K, Gabriel S, Altshuler D, Daly M: A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 2011, 43:491-498.

21) Van der Auwera GA, Carneiro M, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K, Altshuler D, Gabriel S, DePristo M: From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics* 2013, 43:11.10.1-11.10.33.

22) McLaren W, Pritchard P, Rios D, Chen Y, Flicek P, Cunningham F: Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 2010, August 15; 26(16): 2069–2070.

23) Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, Tothill RW, Halgamuge SK, Campbell IG, Gorringe KL: CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* 2012, May 15; 28(10): 1307–1313.

24) Krueger F, Andrews SR: Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 2011. June 1; 27(11): 1571–1572.

25) Vaqué JP, Gómez-López G, Monsálvez V, Varela I, Martínez N, Pérez C, Domínguez O, Graña O, et al: PLCG1 mutations in cutaneous T-cell lymphomas. *Blood* 2014, Mar 27;123(13):2034-43.

26) Cuadrado A, Remeseiro S, Graña O, Pisano DG, Losada A: The contribution of cohesin-SA1 to gene expression and chromatin architecture in two murine tissues. *Nucleic Acids Res*. 2015, Mar 3. doi: 10.1093/nar/gkv144