Chapter 21

RAP: A Web Tool for RNA-Seq Data Analysis

Mattia D'Antonio, Pietro Libro, Ernesto Picardi, Graziano Pesole, and Tiziana Castrignanò

Abstract

Since 1950 main studies of RNA regarded its role in the protein synthesis. Later insights showed that only a small portion of RNA codes for proteins where the rest could have different functional roles. With the advent of Next Generation Sequencing (NGS) and in particular with RNA-seq technology the cost of sequencing production dropped down. Among the NGS application areas, the transcriptome analysis, that is, the analysis of transcripts in a cell, their quantification for a specific developmental stage or treatment condition, became more and more adopted in the laboratories. As a consequence in the last decade new insights were gained in the understanding of both transcriptome complexity and involvement of RNA molecules in cellular processes. For what concerns computational advances, bioinformatics research developed new methods for analyzing RNA-seq data. The comparison among transcriptome profiles from several samples is often a difficult task for nonexpert programmers. Here, in this chapter, we introduce RAP (RNA-Seq Analysis Pipeline), a completely automated web tool for transcriptome analysis. It is a user-friendly web tool implementing a detailed transcriptome workflow to detect differential expressed genes and transcript, identify spliced junctions and constitutive or alternative polyadenylation sites and predict gene fusion events. Through the web interface the researchers can get all this information without any knowledge of the underlying High Performance Computing infrastructure.

Key words HPC, Bioinformatics, Genomics, Transcriptomics, RNA-Seq, Alternative splicing sites, Fusion transcripts

1 Introduction

The analysis of transcriptome through next generation sequencing (NGS) technology is considered today a golden standard and it is completely replacing the expression profiling based on the microarray technology, that dominated the field for more than a decade. With the advent of massively sequencing, more than 10 years ago, the RNA sequencing (RNA-seq) was grown as a technological tool in molecular biology [1–3] to gain a better comprehension of the gene expression process, estimating the expressed mRNAs [4] through the sequencing of a complete transcriptome in any cell/tissue type and condition. The importance of this sequencing

Ernesto Picardi (ed.), *RNA Bioinformatics*, Methods in Molecular Biology, vol. 2284, https://doi.org/10.1007/978-1-0716-1307-8_21,

technology is due to the investigation of many aspects of molecular biology, such as the ability of studying mRNA splicing [5], the regulation of gene expression by noncoding RNAs [6] and enhancer RNAs [7]. Over 100,000 alternative isoforms were detected with RNA-seq experiments [5], but identifying functional transcripts is still challenging. In a NGS experiment a standard laboratory workflow begins with RNA extraction and ends with the preparation of the sequencing library. The library is then sequenced generating several million short reads typically hundreds of bases in length. The primary computational application of RNA-seq is to determine the quantitative changes in expression levels between experimental groups (e.g., expression at gene and/or transcript level). Other computational strategies are also investigated (alternative splicing events, alternative polyadenylation sites, fusion events, etc.) depending on the biological questions. However, as the throughput of experimental data continues to grow and bioinformatics is de facto entered in the era of big data [8], interpreting the results of the comparisons between the various RNA-seq samples becomes increasingly complex. Parallel to the technological developments of NGS, huge primary repositories of raw sequencing data (Sequence Read Archive—SRA [9], Tumor Cancer Genome Atlas—TCGA [10], Genotype-Tissue Expression—GTex [11], Cancer Cell Line Encyclopedia—CCLE [12]) have been populated with incredible speed and the data deluge. These archives contain, among different kind of omics data, a huge amount of RNA-seq samples available for new analyses [13–15] and reuse of data for in silico comparisons and validations [16].

In the last decade, several pipeline tools have been implemented for RNA-seq data analysis [17–21]. In this context, we have developed RAP, a web tool implemented in cloud on Cineca HPC infrastructure, that allows users to analyze big transcriptomic data in main model organisms [22]. A great benefit of web applications, such as RAP, is the possibility to analyze RNA-Seq data without the need of IT competences nor the knowledge about the underlined High Performance Computing infrastructure (computational resources are completely transparent to the users). In addition, a very intuitive web interface allows to customize the analysis and data results are often provided in tabular fashion allowing to further filter subsets of results. The execution of this pipeline is totally automated and optimized on HPC infrastructure. Through RAP the user is able to quickly perform a very complex transcriptome analysis identifying the differential expressed genes, transcripts, splice junctions, polyadenylation sites, and fusion events, thanks to the use of a simple and intuitive interface. The web interface is based on a cloud architecture that completely hides the underlined HPC infrastructure. Both the dedicated hardware and the software environment is described in the following Subheadings 1.1 and 1.2.

1.1 Computational Hardware

The Galileo supercomputer, whose nodes are also dedicated to transcriptome analysis, has the purpose of enabling new classes of "BigData" bioinformatic applications. It can manage and process large amount of raw data, coming from both experiments or data reuse. Galileo is composed by 1022 nodes made of 2 × 18-cores Intel Xeon E5-2697 v4 at 2.30 GHz. Therefore each node, with 128 GB of RAM, has 36 cores, for a total of 36.792 cores. Galileo was designed to optimize density and performance, allowing to analyze large data repositories in CINECA. The storage area accessed by Galileo nodes is composed by high-throughput GPFS disks for a total amount of about 2 PB. Such a storage is also connected with a large capacity tape library for a total actual amount of 12 PB.

1.2 Computational Software

RAP has been developed as an ensemble of modules interconnected through their dependencies and can be schematized as a direct acyclic graph. Its architecture allows each module to run independently, using data stored in MYSQL relational databases. The program which manages RAP modules is completely written in PHP Object Oriented, while mysql libraries are used for database interactions. RAP integrates several open source third-party analysis tools as well as in-house developed python and bash scripts into one single completely automated pipeline. All required computational tasks are managed and distributed on nodes of the clusters depending on the computational needs.

All the analysis, from the raw sequence data uploading to the achievement of final results can be handled and visualized by using an interactive, web-based graphical user interface (GUI). The RAP web-based GUI is written in PHP: Hypertext Preprocessor (PHP) language using HyperText Markup Language (HTML) and JQuery for a better user interaction. The web interface is based on Foundation 4, an advanced responsive CSS front-end framework. RAP is available at the following web address: https://bioinformatics.cineca.it/rap.

1.3 Computational Bioinformatic Workflow

The bioinformatic analysis workflow (Fig. 1) can be divided in six branches, each focused on specific biological entities:

- A main branch, mandatory, for reads mapping and expression profiling. This branch can be completed with differential analysis at transcript level (branch A) and gene level (branch B).
- Detection of fusion transcripts (branch C). The activation of this branch is optional.
- Detection and quantification of splice junctions. This branch is optional and can be completed by a differential analysis of observed junctions (branch D).

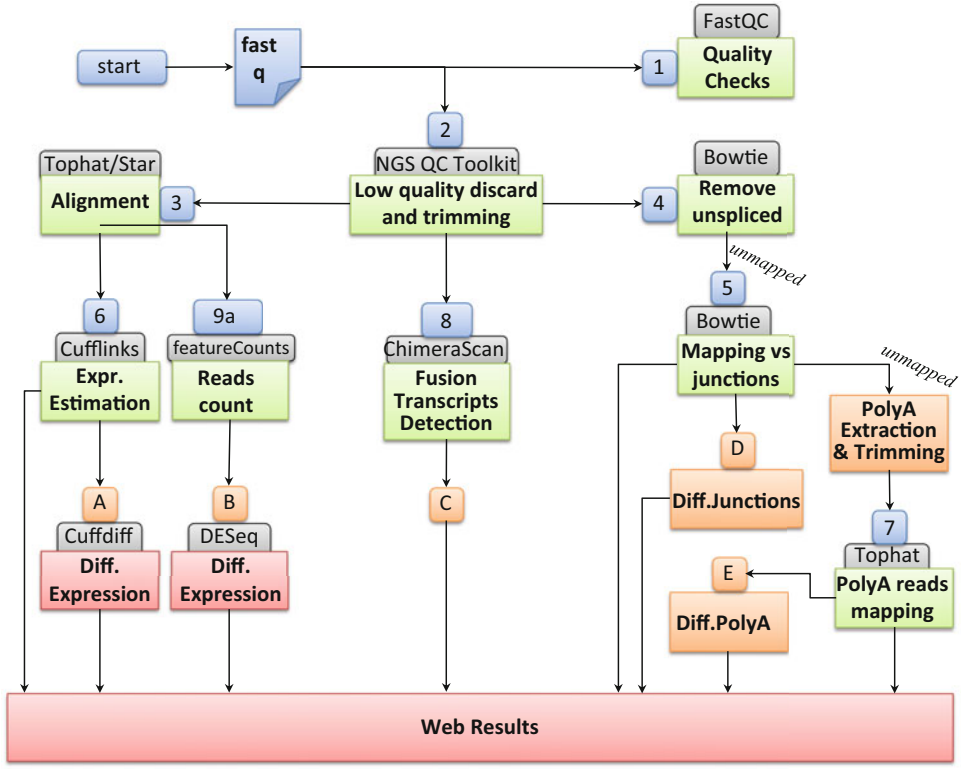


Fig. 1 Completely automated bioinformatic workflow of RAP

- Extraction of polyadenylation sites. This branch is optional and can be completed by a differential analysis of observed poly (A) sites (branch E).

We introduce now with more level of detail the single steps of the pipeline.

Quality checks. The first step of the pipeline provides several quality control checks on raw sequence data by running FastQC tool [23] (Fig. 1, step 1). It produces a set of quality results formatted in HTML report pages, which give a quick overview of whether user data has any problems and therefore how to consider the rest of the downstream analysis. Some samples of the dataset could be indeed biased and contaminated.

Law quality discard and trimming. The action of filtering out low quality reads to increase the overall dataset quality, is performed by the execution of NGS QC Toolkit [24]. It allows to process in the next steps of the pipeline only high-quality reads and therefore provide more robust results in the next mapping phase. This trimming step is launched in parallel with the previous step of quality control (no dependency is applied between the steps 1 and 2). In addition both FastQC and NGSQCtoolkit are optimized to

handle several samples at a time and developed to work in a multi-threaded fashion, distributing the computational load on the available HPC resources.

Read mapping. A very crucial phase for analysis consists of aligning the reads against the reference genome and transcriptome (when an annotation is provided). A tool able to perform both kinds of alignment is TopHat2 [25] (Fig. 1, step 3) that includes, as mapping engine, Bowtie2 [26]. TopHat2 is able to use during the analysis the full-length transcripts defined by annotations in order to improve both sensitivity and accuracy. The computational strategy implemented in TopHat2 is able to align the reads with true indels (insertions and deletions), also taking advantage of Bowtie2 ability to detect short indels very accurately. The first step of the method consists in transcriptome mapping of the reads and, in a second step, those unmapped or poorly aligned are mapped against the reference genome in order to detect those reads entirely within exons. Also in this phase some reads, the multiexon spanning reads, are unmapped and, in the next step, they are split into smaller segments that mapped against the reference genome. Fragments are then aligned to the junction (splice site) flanking sequences and stitched together to shape whole read alignments. The last step consists in realigning the reads minimally overlapping the introns against the exons.

As an alternative alignment workflow RAP allows to replace Tophat2 with STAR [27] for faster but still sensitive analyses.

Gene/isoform expression quantification and differential analysis. The building of the transcriptome assembly and the evaluation of the expression level of all detected isoforms is a specific task of Cufflink, which takes as input the results of Tophat2 mapping (Fig. 1, step 6). Cufflink builds the transcript assembly starting by a gene reference annotation. In this phase the user has the opportunity to choose between two assembler algorithms: Cufflinks and RABT [28] assembler.

The first module assembles aligned RNA-Seq reads into a parsimonious set of transcripts, then estimates the abundances of the transcripts considering how many reads support each one.

The second allows to include in the calculus both reference transcripts and novel assembled genes and isoforms. This second assembler is particularly useful for those organisms where a deep annotation does not already exist. The differential analysis at transcript-level resolution of RNA-seq experiments and controls, based on expression levels calculated by Cufflinks, is performed by Cuffdiff2 (branch A). It both provides more accurate transcript-resolution estimates of changes in gene expression, performing statistical computations at isoform-level resolution, and considers the variability in measurements across biological replicates of an experiment.

At gene expression level, another parallel branch (branch B) in the workflow is executed to estimate the raw-count. The third-party software launched is HTSeq [29] that models each gene as the union of all its exons. In particular htseq-count is the script of HTSeq designed for RNA-Seq data analysis: taken a GTF file as input from the previous step, it counts for each gene how many aligned reads overlap exactly its exons whereas the reads overlapping with more than one gene are discarded. These counts will be used for gene-level differential expression analyses performed by the next step.

Fusion events detection and annotation. A further optional branch of RAP workflow, branch C, can be activated with the aim of detecting chimeric transcripts (Fig. 1, step 8), made of exons from two different genes that encode novel putative proteins. The workflow integrates ChimeraScan [30], a software built over Bowtie aligner to identify putative fusion breakpoints. The reads mapping is performed against a mixture of genome-transcriptome reference. Read pairs not aligned concordantly are split into smaller segments (default = 25 bp) and realigned. Reads that align to distinct references or distant genomic locations of the same reference are added in a list of putative 5–3 transcript pairs that could be chimera candidates. A new reference index from the list of sequences is built and candidate junction-spanning reads are realigned against this index. In order to reduce false-positive chimeras, the incorporated spanning reads are filtered, discarding those supported by few reads or those with fragment sizes greater than the range of the distribution.

Splice junction detection. The analysis of splice junctions is performed by the execution of branch E. High Quality Reads, derived from NGS QC Toolkit, are mapped with Bowtie against the reference genome and are discarded from the initial dataset (Fig. 1, step 4). The unmapped reads, that may potentially contain a splicing site, are mapped again by using always Bowtie to a custom-built splice junction library (Fig. 1, step 5). This reference is built starting from a gene annotation model in GTF format [31]. It includes two different categories of splice junctions: known junctions derived from RefSeq [32] and novel junctions, obtained through a combinatorial exon skipping procedure by considering all compatible exon skipping patterns.

Polyadenylation site detection. Another optional branch (E) is executed in cascade from the previous branch D and concerns the analysis of Polyadenylation sites. Residual reads still unmapped from alignments against genome, transcriptome and junctions may contain information about polyadenylation sites (Fig. 1, step 7). In this phase, Poly(A) tags (reads containing a stretch of A at the end of the sequence) are extracted, trimmed, and aligned to the

genome. Another spliced alignment with Tophat2 is applied to verify if the sequence also contains a splice junction related to the final exon. As final step, a parsing procedure is applied to annotate the concurrent occurrence of polyadenylation signal (PAS) sequences. In addition to the canonical polyadenylation signal (AAUAAA) a total of 10 known variants are considered [33].

Differential expression analysis. The correct identification of differentially expressed biological entities (such as, in RAP workflow, the identification of differential expressed genes, transcripts, polyadenylation sites, and splice junctions) between specific conditions, each eventually represented by more replicates, is a key in the understanding phenotypic variation. The branches deputed to this kind of analysis are respectively, branch A, B, D, and E in Fig. 1. In particular, RAP detects differentially expressed genes by using DESeq [29] taking as input raw counts calculated by HTSeq (Fig. 1, step B). According to the DESeq algorithm the variance is the sum of a term of raw variance (derived from biological variability) and end of gunshot noise (from counts uncertainty). This method allows to process data without or with very few replicates, putting genes together with similar expression levels.

Cuffdiff2 [34] is the tool used to estimate the differential expressed transcripts from transcript abundances determined in the previous step by Cufflinks (Fig. 1, step A). It calculates differential analysis at transcript-level and controls the variability across replicates and the uncertainty in abundance expression estimates caused by ambiguously mapped reads. In case of incorrect rejections of a true null hypothesis (false positives) it introduces also the Benjamini–Hochberg correction [35] for multiple testing of differential expression (false discovery rate, FDR).

The differential expression analysis of junctions and polyadenylation sites, performed in branches D, E are managed by two home-made PHP parser scripts specific to the output format results from the respectively previous steps 7 and 5 in Fig. 1.

1.4 Reference Genomes and Transcripts Included in RAP

RAP supports the analysis of RNA-seq reads from several organisms. At present, the workflow is available for *Homo sapiens* (genomes hg18, hg19, and hg38), *Mus musculus* (genomes mm9 and mm10), *Rattus norvegicus* (genome rn4), *Drosophila melanogaster* (genome dm3), *Saccharomyces cerevisiae* (genome sacCer3), *Equus ferus caballus* (genome equCab2), *Danio rerio* (genome danRer10), *Zea mays* (genomes maize3 and Mo17_v1), and *Arabidopsis thaliana* (genome tair10). Additional reference genomes and annotations can be added on the HPC cloud becoming available for further analysis, upon users' request.

2 Material

To access the RAP pipeline and results it is not required any special hardware or software, apart from a common web browser (Chrome, Firefox, Internet Explorer, and Safari are mostly all supported).

Each account is granted for a 60 days period with a limitation of 2 projects, 2 analyses per projects and 12 files per analysis. Upload of files via web is limited to 2 GB but a FTP account can be requested to be able to upload files without any size limitation. Files can be uploaded as compressed archives (zip, gz) so a compression software can be very useful to limit the bandwidth.

3 Methods

To ensure the confidentiality of data, the use of RAP requires a personal account and by default only the data owner can access to uploaded files and obtained results. All academic users who provide an institutional email address can request for an account via the registration module on the website (the logging form for the registration and authentication is highlighted in red square in Fig. 2).

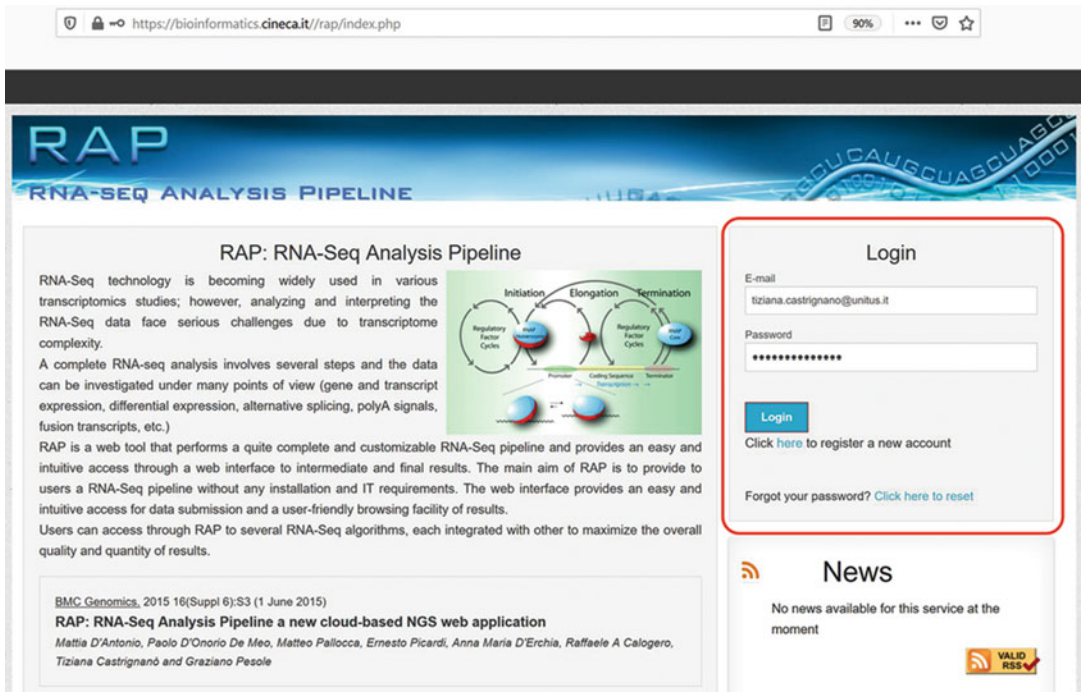


Fig. 2 RAP Login Form

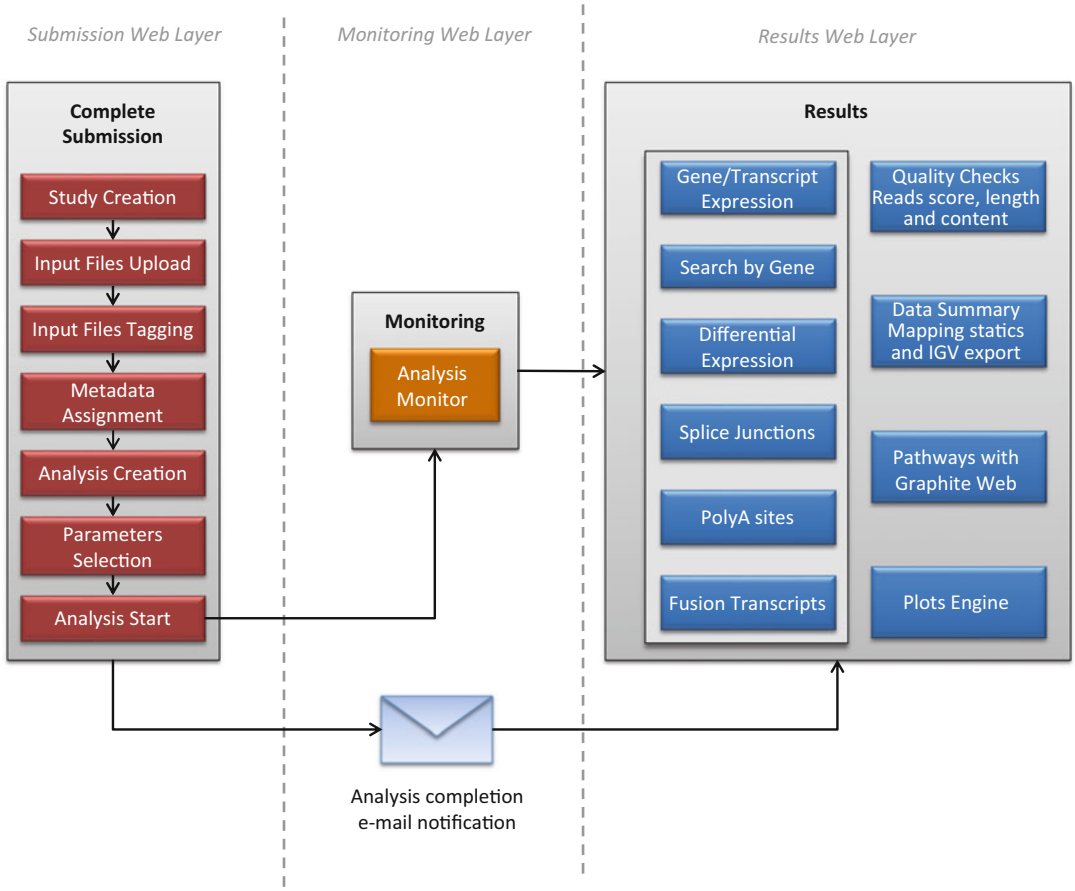


Fig. 3 RAP Web User Interface workflow

A workflow scheme of RAP web user interface is shown in Fig. 3.

The first step to submit a dataset to RAP is the creation of a new study. RAP implements a data architecture inspired by the data format standard proposed by the European Nucleotide Archive (ENA) curated by the European Bioinformatics Institute (EBI). According to ENA guidelines, a study is a homogenous collection of data about a single sequencing project. The creation of a new study in RAP only requires little information such as a title, a description and an access level (private, group, or public).

A private study can be accessed only by the owner while a public study will be accessed by everyone. After the creation (Fig. 4a), the user can open the study by clicking its name or using the view icon. The owner can also edit the given information or delete own studies (Fig. 4b).

After the Study creation, the web user interface shows a workflow composed of three steps: "Add Files," "Process Files," and "Analyze Files" (Fig. 5).

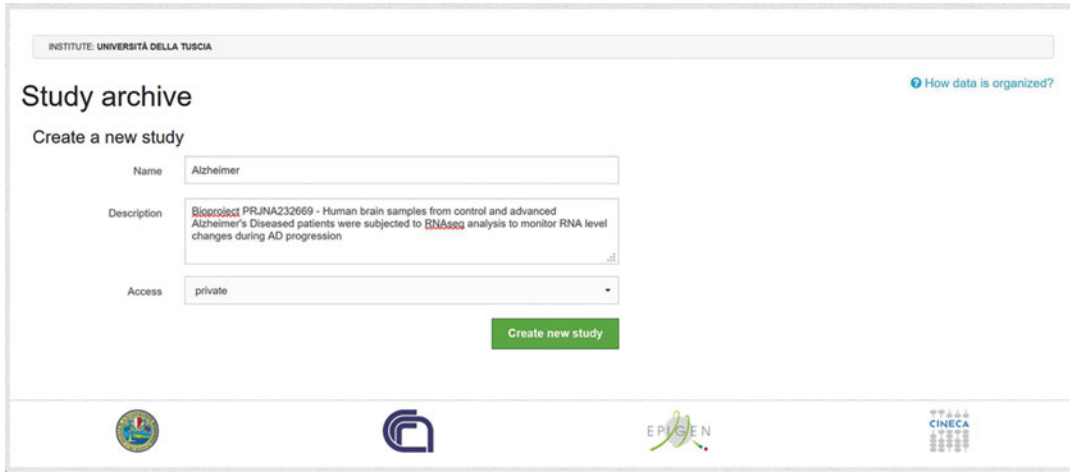


Fig. 4 (**a**) Creation of a new Study in RAP. (**b**) View of the archive on created or submitted Studies

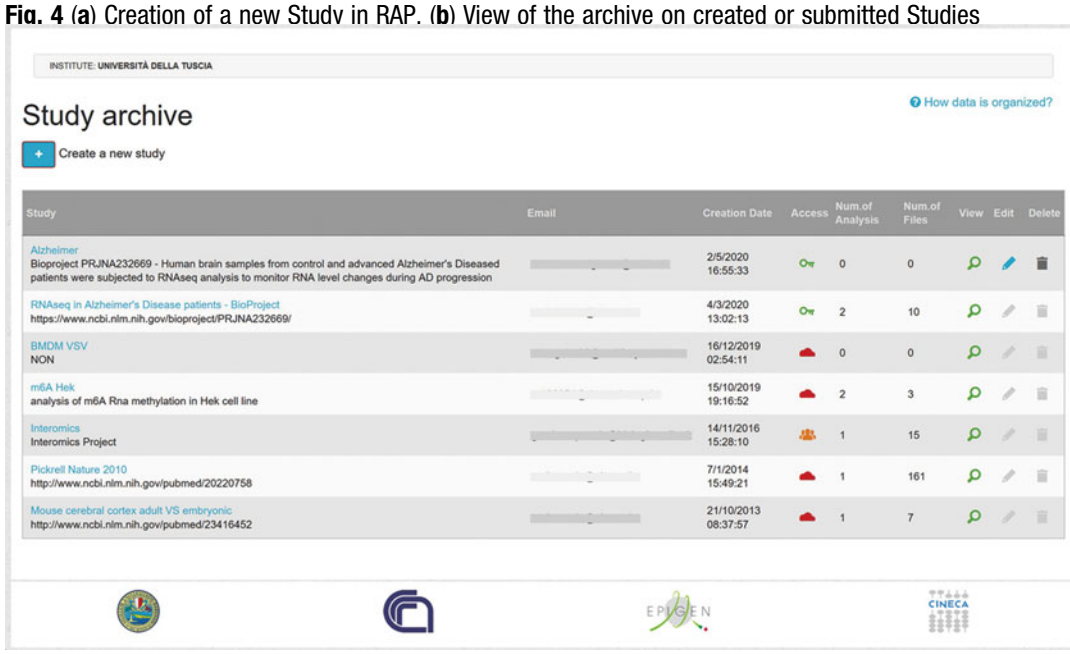


Fig. 4 (continued)

Before starting any analysis, the user has to upload one or more input files. The upload engine offers several options for the submission of the input files: Web Upload, Web Link and FTP protocol (Fig. 6).

The Web Upload is implemented by using JavaScript and allows the user to upload several files at once, but the size of the single file is limited to 2 Gb. The user can verify the upload progress and interact with the system by adding or removing files also during the transfer. Nevertheless, any web based upload widget has limitations, for example during the upload process the page cannot be

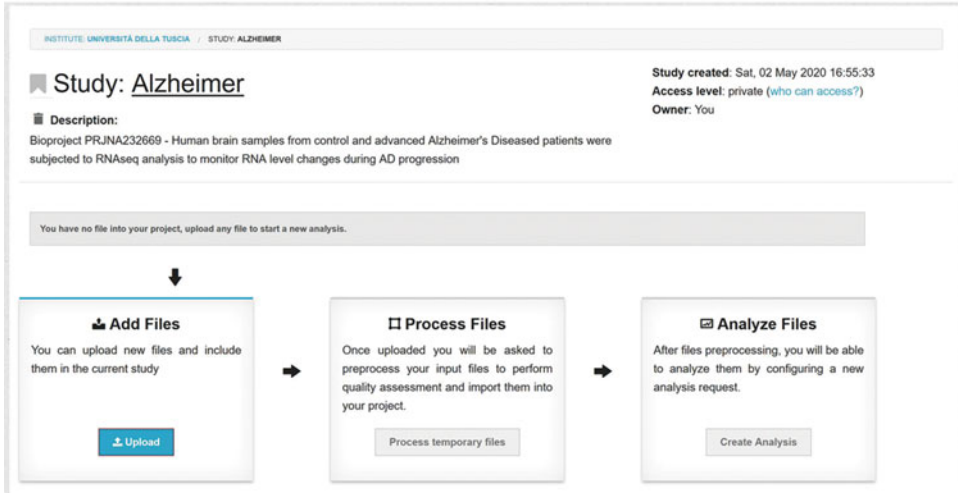


Fig. 5 Workflow of the three main steps: "Add Files," "Process Files," and "Analyze Files"

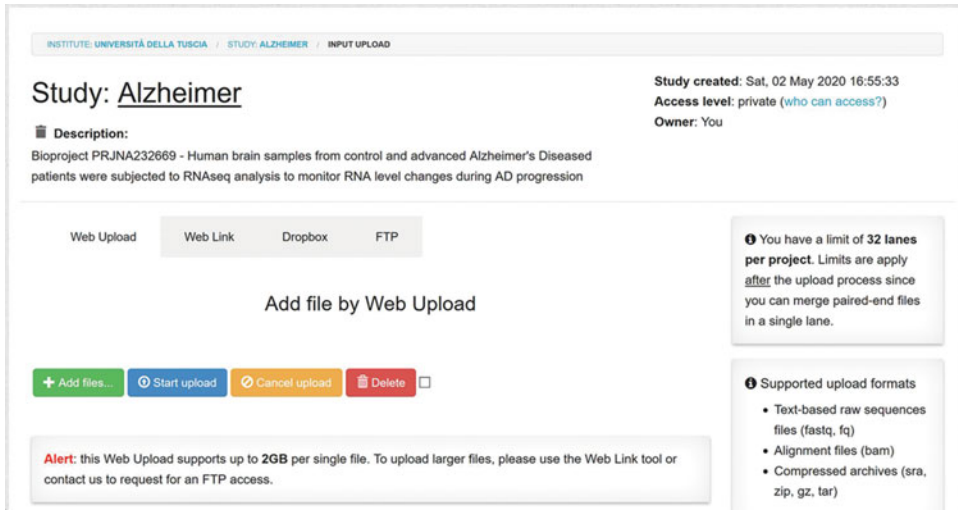


Fig. 6 Uploading options for input files

closed, otherwise the upload is interrupted. To overcome these limitations as well as the maximum allowed file size the user can provide one or more links and the system will handle the download in batch. The user will be notified by email when downloads are completed. As a third option, the user can access the RAP user space by using a FTP client.

Several input formats are supported such as text-based raw sequences produced by Illumina sequencing platforms (i.e., FASTQ), prealigned data (i.e., BAM and SAM), and compressed reads (i.e., SRA archives). The user can also upload these files in a compressed archive to speed up the uploading process (several common compressed formats are supported, such as zip, tar, gzip, and bz2).

At the end of the upload if the files are produced through a paired-end sequencing protocol, the user will be allowed to specify each pair of files that comes from the same lane, using a simple drag and drop utility provided by the GUI. The same utility can also be used to reorder lanes, if needed. After that, the user has to assign to each file (or pair of files for PE read data) a unique label (file tagging). A good label should be short and explanatory enough to recognize the input since it will be used in the results pages as replacement for the file raw name. The user can also associate one or more samples, adding information about the sequenced material. This metadata information (i.e., organism, tissue, cellular line, phenotype, and strain) can be useful to describe the input files and are especially important in the case of public experiments.

After the metadata assignment, the uploaded files are imported into the project and can be used to start a new analysis by selecting one or more inputs. The user can choose between two different workflows: one based on Tophat2 and the other based on STAR alignment algorithm. Before starting the analysis the user can accept a set of default parameters to perform a standard workflow or can customize the parameters to tailor the analysis on own data. Parameters are divided into six categories: Common parameters, Quality check and filtering, Genome spliced alignment, Transcript assembly and abundance estimation, Determination of polyA reads, Gene fusions detections in paired-end RNA-Seq datasets.

The first category contains parameters common to all or many pipelines modules, such as the reference database and the Reference-GTF. If input files have been associated with an organism during the annotation phase, only database for such organism are reported here, simplifying the analysis customization and thus preventing potential errors. A set of flags (search-junctions, search-polya, search-chimeric) enable or disable the optional branches.

With the Quality check and filtering parameters the user can modify the behavior of quality control and trimming module. With the quality and length parameters is respectively possible to modify the cutoff value for the PHRED quality score for high-quality filtering (default value is 20) and the percentage of read length that should be of given quality (default value is 70%). This module can also remove primers and adaptors by selecting one of provided libraries (Genomic DNA/Chip-Seq Library, Paired End DNA Library, DpnII gene expression Library, NlaIII gene expression Library, Small RNA Library, Multiplexing DNA Library) or uploading a file containing user defined sequences (one per line).

With Genome alignment parameters, the user can customize the mapping phase performed by TopHat2 or STAR (based on the selected workflow). An additional option (GenerateBigWig) allows the user to request the automatic conversion of alignments into

BigWig format. It is a convenient file format for display dense data to be loaded into the University of California, Santa Cruz (UCSC) Genome Browser.

In the Transcript assembly and abundance estimation section the user can enable the reconstruction of novel-transcripts or provide a GTF list (MaskFile) containing transcripts to be ignored during the assembly.

Finally, in the Determination of poly(A) reads and Gene fusions detections in paired-end RNA-Seq datasets categories, the user can customize the poly(A) extraction step and chimeric transcripts determination step, respectively.

Once configured, the analysis is automatically submitted on a queue system based on torque Resource Manager. The user can follow the analysis progress through a monitor page (that can be opened by clicking on the monitor icon, highlighted with a red square in Fig. 7a; a list of all steps are displayed (Fig. 7b) with corresponding running status (to be done, queued, running, skipped, completed, error), the used software and parameters, the intermediate output results if completed.

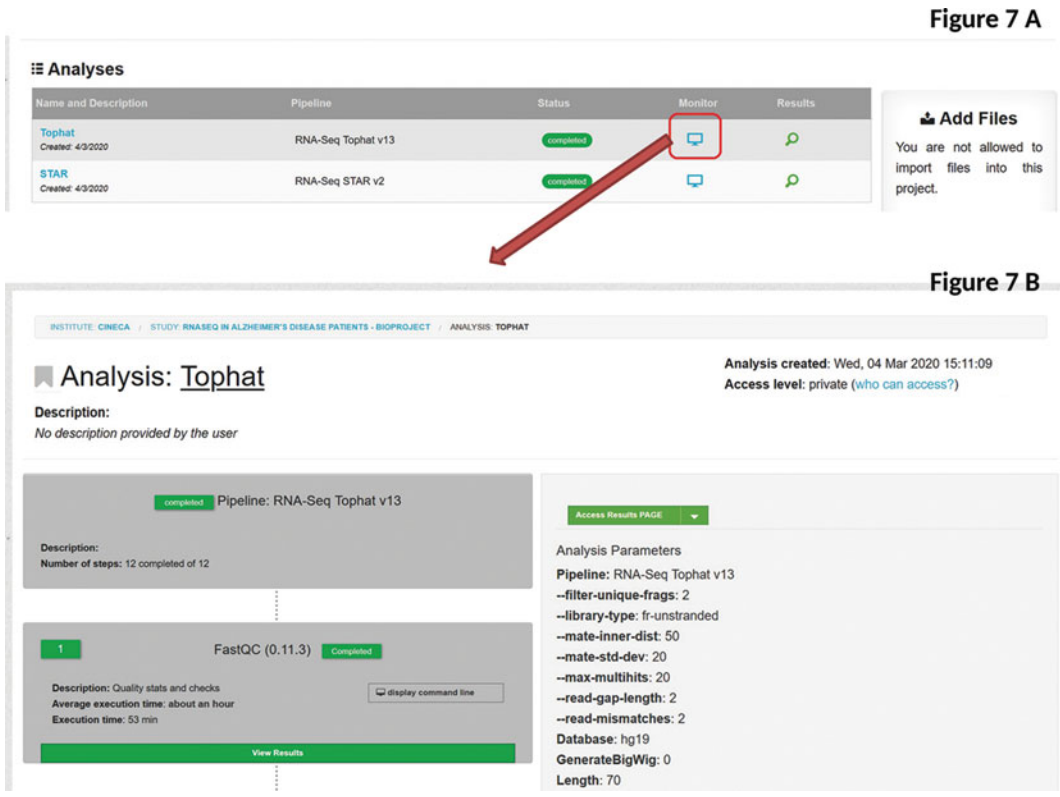


Fig. 7 (**a**) list of the launched analyses (based on Tophat and Star aligners respectively with monitor icon squared in red); (**b**) a screenshot of the monitoring page associated to Tophat analysis

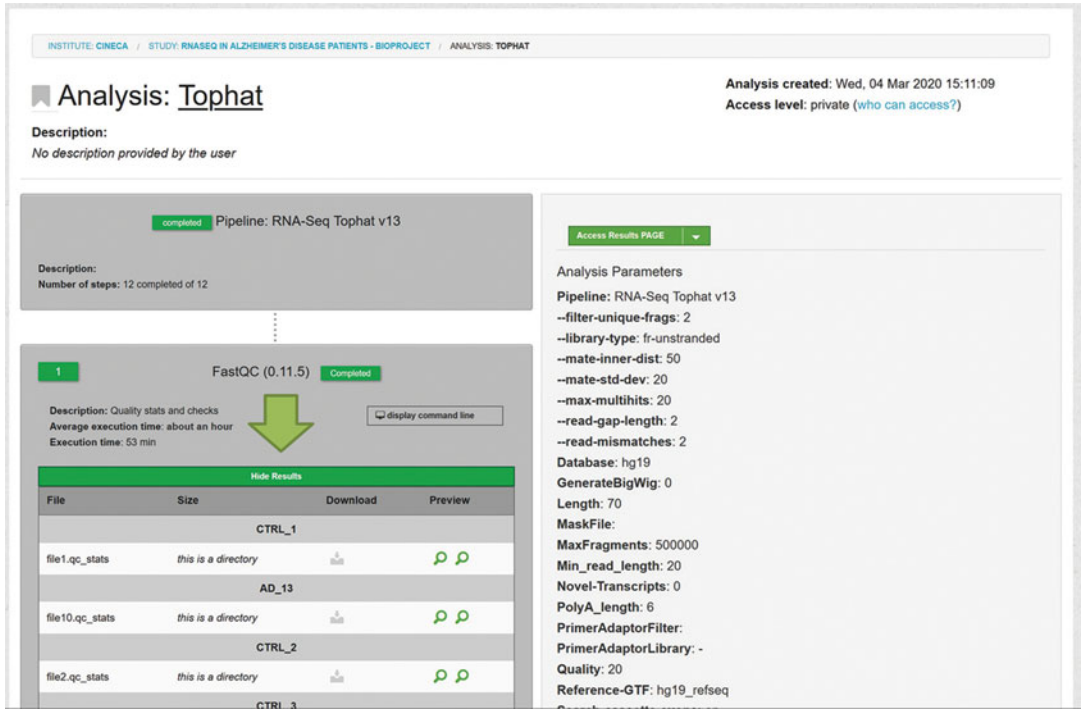


Fig. 8 Example of list of result files produced by the FastQC step; the results are available after the completion of the step and can be visualized by clicking on the green bar (pointed by the big green arrow into the Monitor Page)

After the analysis completion of each step the result files can be accessed from the monitoring page (*see* Fig. 8) and the analysis summary table in the project page. Analysis results can be accessed from the analysis summary table (*see* Fig. 9a) and the output from Quality checks is shown by default. Further results can be visualized by enabling the side menu (*see* Fig. 9b).

For a better comprehension about the output organization and visualization of RAP, real data set have been used from project "RNA seq in Alzheimer's Disease patients" https://www.ncbi.nlm.nih.gov/bioproject/PRJNA232669. The output summary results of the analysis of the bioprojet PRJNA232669 are available for users at the following URL:

- https://bioinformatics.cineca.it/rap/results.php?a=7282#data_summary for a summary of RNA-seq metrics.
- https://bioinformatics.cineca.it//rap/results.php?a=7282&gene=%09RPPH1&submit=Search#gene_expression for "Gene and transcript expression summary".
- https://bioinformatics.cineca.it//rap/results.php?a=7282&gene=%09RPPH1&submit=Search#differential_expression for "Differential gene and transcript expression".

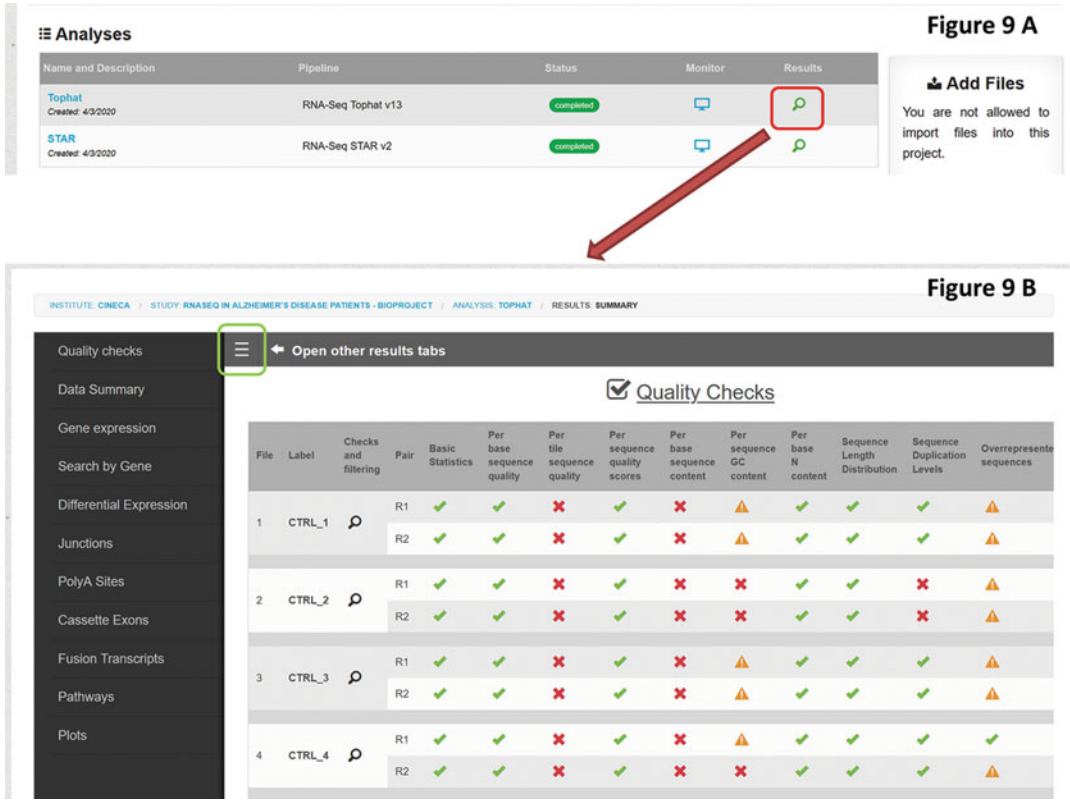


Fig. 9 (**a**) List of results in case of Tophat and Star aligners. (**b**) Result summary page for Quality Checks opened by clicking on the lent icon in the list

- https://bioinformatics.cineca.it//rap/results.php?a=7282#junctions for a "Summary of splice junction".
- https://bioinformatics.cineca.it//rap/results.php?a=7282#fusion_transcripts for "Fusion transcripts".

In this analysis samples named CTRL_* are "control" samples whereas samples named AD_* are Alzheimer's Disease samples.

Each of the URLs, loading a summary page report, allow users to go in details of results of interest. By clicking on each summary count number in the summary table a page opens visualizing the result of interest (genes, transcripts, fusion transcripts, junctions, etc).

Other kind of results ("Gene Expression," "Search by Gene," "Differential Expression," and so on) can be consulted by the user, clicking on the menu item "Open other results tabs," in the top-left corner of the gray frame. For instance, if the user is interested in "Gene Expression" results tab, clicking on the specific tab, the system will show the below navigable results page (Fig. 10):

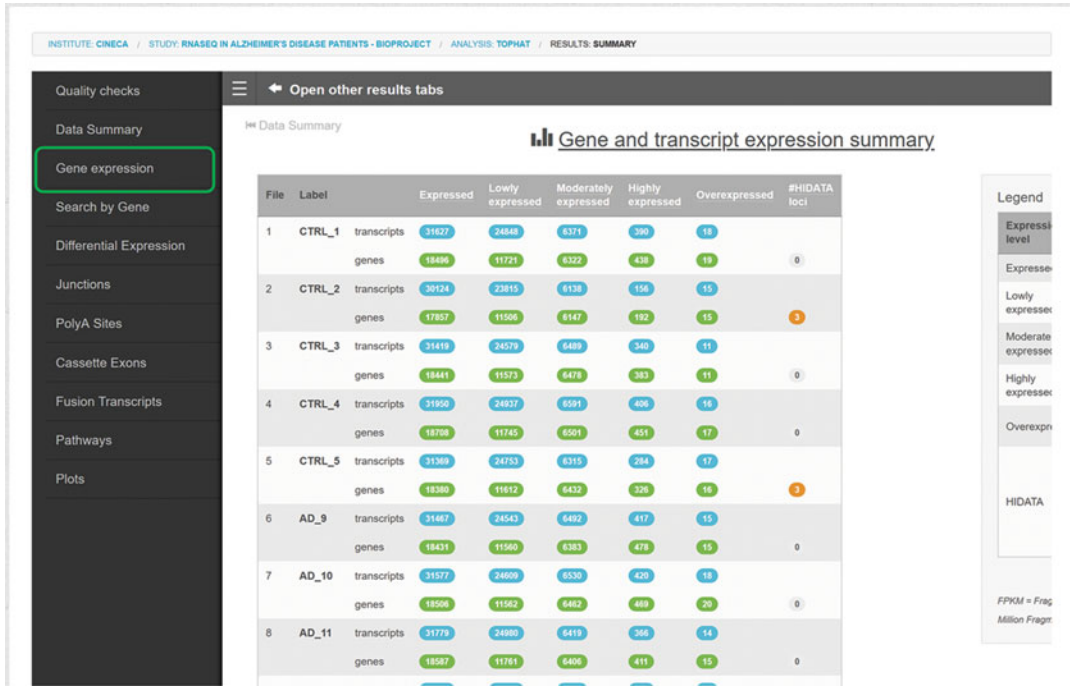


Fig. 10 Gene and Transcript Expression summary page for the analyzed samples

In case the user is interested about "Highly expressed transcripts" in sample named "CTRL_1", by clicking on the blue icon with the number "390" in the first row of Fig. 10, a page opens with the list of the transcripts of interest (genes, transcripts, genomic position, strand, transcript length, number of exons, FPKM, associated Coverage, class compared to reference annotation, ID of reference transcript, and biotype based on gencode annotation) (Fig. 11).

3.1 Analysis Results and Differential Expression

After the completion of the analysis, all output files are parsed into a MySQL database for ease of visualization from the web interface. Results are divided into several sections: Quality checks, Data Summary, Gene Expression, Search by Gene, Differential Expression, Junctions, PolyA Sites, Fusion Transcripts, Pathways, and Plots.

The first section (Quality checks) reports the output provided by FastQC and NGS QC Toolkit, arranged in a comprehensive summary table with color-coded labels to give to the user a prompt quality overview (Fig. 12). Green labels indicate passed filters, red labels point to failed filters and the orange is the color used to report filters warnings. Each label can be explored visualizing the corresponding FastQC output.

The Data Summary section reports a quantitative analysis of obtained results with metrics such as the total amount of short reads (both raw and high-quality reads as filtered by NGS QC

Click on a column title to order this table

UID	gene	transcript	Genomic position	strand	TLen	#Exons	FPKM↓	Coverage	Class	Ref.Transcript	Biotype
270	SCARNA7	NR_003001	chr3:160232695-160233024	-	330	1	968.61	2616.27	=	NR_003001	-
286	SPARCL1	NM_004684	chr4:88394488-88450655	-	2904	11	895.87	2420.65	=	NM_004684	-
232	SNAP25	NM_130811	chr20:10199477-10288066	+	2054	8	880.06	2331.79	=	NM_130811	-
367	SNORA43	NR_002975	chr9:139620556-139620689	-	134	1	829.71	2242	=	NR_002975	-
371	TMSB4X	NM_021109	chrX:12993226-12995346	+	629	3	828.7	2239.27	=	NM_021109	-
138	MT3	NM_005954	chr16:56623267-56625000	+	582	3	810.95	2190.89	=	NM_005954	-
225	SCARNA6	NR_003006	chr2:234197322-234197587	+	266	1	786.85	2126.19	=	NR_003006	-
86	SNORA53	NR_003015	chr12:98993413-98993662	+	250	1	783.8	2117.95	=	NR_003015	-
53	FTH1	NM_002032	chr11:61731757-61735132	-	1229	4	763.07	2061.69	=	NM_002032	-
79	RPL41	NM_021104	chr12:56510374-56511616	+	469	4	710.4	1915.45	=	NM_021104	-
221	RPL37A	NM_000998	chr2:217363520-217366188	+	418	4	702.3	1897.09	=	NM_000998	-
19	RPS27	NM_001030	chr1:153963239-153964631	+	351	4	687.56	1855.77	=	NM_001030	-
56	SLC1A2	NM_004171	chr11:35272752-35441105	-	12006	11	685.88	1848.93	=	NM_004171	-
329	ACTB	NM_001101	chr7:5566779-5570232	-	1812	6	656.06	1772.69	=	NM_001101	-
104	CALM1	NM_006888	chr14:90863327-90874619	+	4256	6	636.82	1720.73	=	NM_006888	-
108	SCARNA13	NR_003002	chr14:95999692-95999966	-	275	1	619.01	1672.66	c	NR_003138	-
323	EEF1A1	NM_001402	chr6:74225473-74230755	-	3512	8	612.32	1654.31	=	NM_001402	-
212	TMSB10	NM_021103	chr2:85132763-85133799	+	482	3	606.02	1637.18	=	NM_021103	-
166	SCARNA16	NR_003013	chr17:75085389-75085575	+	187	1	580.16	1567.69	=	NR_003013	-
101	RPS29	NM_001032	chr14:50050290-50053094	-	296	3	567	1532.11	=	NM_001032	-
346	CLU	NM_001831	chr8:27454434-27472328	-	2860	9	550.14	1486.56	=	NM_001831	-
196	FTL	NM_000146	chr19:49468566-49470136	+	871	4	548.72	1482.54	=	NM_000146	-
66	HSPA8	NM_006597	chr11:122928200-122932901	-	2318	9	509.33	1376.22	=	NM_006597	-

Fig. 11 Details on highly expressed transcripts found for '*CTRL_1*'

Quality Checks

Data Summary

File	Label	Checks and filtering	Pair	Basic Statistics	Per base sequence quality	Per tile sequence quality	Per sequence quality scores	Per base sequence content	Per sequence GC content	Per base N content	Sequence Length Distribution	Sequence Duplication Levels	Overrepresented sequences	Adapter Content	Kmer Content	Quality Summary
1	CTRL_1		R1	✔	✔	✖	✔	✖	⚠	✔	✔	✔	⚠	✔	✖	
			R2	✔	✔	✖	✔	✖	⚠	✔	✔	✔	⚠	✔	✖	
2	CTRL_2		R1	✔	✔	✖	✔	✖	✖	✔	✔	✖	⚠	⚠	✖	
			R2	✔	✔	✖	✔	✖	✖	✔	✔	✖	⚠	⚠	✖	
3	CTRL_3		R1	✔	✔	✖	✔	✖	⚠	✔	✔	✔	⚠	✔	✖	
			R2	✔	✔	✖	✔	✖	⚠	✔	✔	✔	⚠	✔	✖	
4	CTRL_4		R1	✔	✔	✖	✔	✖	⚠	✔	✔	✔	✔	✔	✖	
			R2	✔	✔	✖	✔	✖	✖	✔	✔	✔	⚠	✔	✖	
5	CTRL_5		R1	✔	✔	✖	✔	✖	⚠	✔	✔	✔	⚠	✔	✖	
			R2	✔	✔	✖	✔	✖	⚠	✔	✔	⚠	⚠	✔	✖	

Fig. 12 The comprehensive summary table of Quality Check results

Toolkit), mapping statistics, junction alignment metrics, information from poly(A) extraction phase, the distributions of mapped reads across functional gene regions and transcripts coverage evaluation. Furthermore this section contains a utility to load data into the Integrative Genomics Viewer (IGV), a free Java desktop application supporting interactive exploration of large-scale genomic data sets on standard desktop computers.

The Gene Expression section reports expression values as estimated by Cufflinks. The summary table reports, for each lane, colored-boxed numbers of both expressed genes (in green) and transcripts (in blue). Four different default FPKM cutoffs are proposed, to give a comprehensive idea about the number of low- or high-expressed genes and transcripts (Fig. 10). Each colored-boxed can be clicked to open the expression overview, a detailed list of all expressed genes (or transcripts) in a given sample with detailed information associated (Fig. 11). This set of results (as any other overview table describe below) can be filtered using customizable thresholds to facilitate the identification of functionally significant variants. Every column can be used to filter results and filters can be combined to produce complex queries. The output tables, as reported after the application of a set of filters, can be exported as textual/Excel files for offline downstream analyses.

The Search by Gene section allows to query simultaneously all expression results. With this form the user can retrieve the expression values of a given gene or transcript obtaining as result a table reporting, for each lane and each isoform the genomic position, the gene FPKM, the transcript FPKM and the transcript coverage. In case of resulting transcripts belong to the same a graphical gene structure is also displayed (Fig. 13).

The Junctions section reports all results obtained by the mapping on the splice junctions library (Fig. 14).

The detail page lists the whole set of observed junctions, each annotated by many information. Both gene and transcript common names are dynamically linked to NCBI, respectively to Gene and Nucleotide databases. Then, several coordinates identify the precise junction by reporting chromosome, start and end position of junction, upstream exon and downstream exon (Fig. 15).

The Poly(A) Sites Section reports a summary table of the number of observed event and coupled with a graphical view of the chromosomal distribution. The user can access to the sortable, filterable, downloadable details page. The frequency of polyadenylation sites (PAS) hexamers is also shown. The canonical PAS (AAUAAA) is clearly expected to overcome the other known variants but, with a sufficient number of sites, the expected relative frequencies should be observed. In the case of bioproject PRJNA232669 the list polyadenylation sites is not available since the library was built without enriching fragments that carry poles (A) tail.

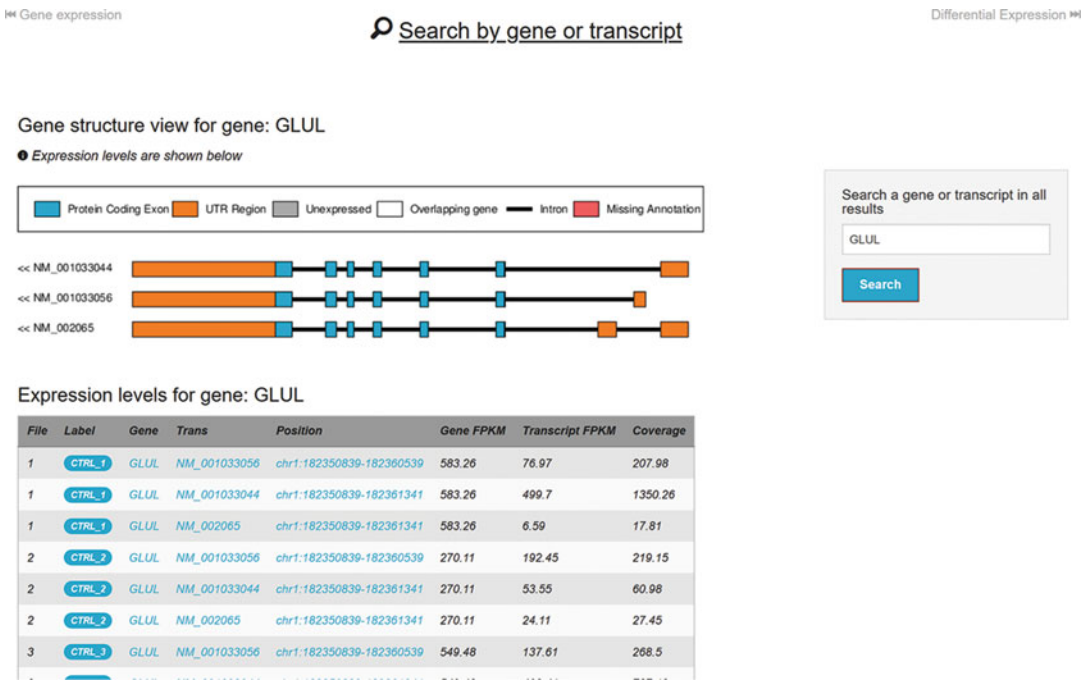


Fig. 13 Result of Search by Gene by querying GLUL. The gene structure and expression levels of the three isoforms are shown

The Fusion Transcripts section reports chimeric isoforms as detected by ChimeraScan. The information schema is the same already proposed in the previous sections: a summary table reports the number of events for each input (Fig. 16) and a details page gives information and annotations for each chimera (Fig. 17). Several information is associated to each chimeric transcript such as the coordinates and gene and transcript ids of both 5′ partner and 3′ partner; the number of supporting reads; and the fusion type (read-through, intrachromosomal, or interchromosomal).

The Pathways section is a utility page used to link results obtained by RAP to an external service, Graphite Web, allowing for pathway analyses and network visualization by using expressed genes as input.

The Plot section imports data obtained from both expression and differential expression sections exploiting the filtering engine and allows the user to created several plots by leveraging the CummeRbund package. Typical plots included in the suite are PCA (principal component analysis), MDS-plot (MultiDimensional Scaling plot), boxplots, scatterplots, density plots, heatmaps, and sashimi plots.

RAP implements several differential expression (DE) analyses, such as transcripts expression, gene expression, Poly(A) site usage, and alternative exon skipping events. RAP allows for the execution

⏮ Differential Expression

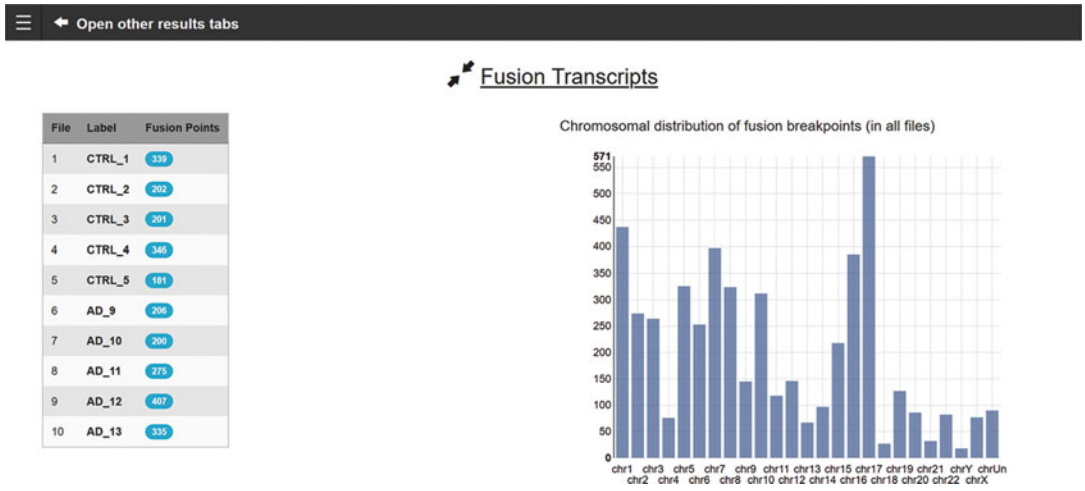
File	Label	RefSeq junctions	Novel junctions
1	CTRL_1	139347	7777
2	CTRL_2	121706	5707
3	CTRL_3	135688	6293
4	CTRL_4	144284	10336
5	CTRL_5	136351	7635
6	AD_9	137924	7618
7	AD_10	137764	7275
8	AD_11	142037	8703
9	AD_12	144264	9274
10	AD_13	139056	8866

Fig. 14 Summary results of detected junctions

Click on a column title to order this table

UID	Gene↓	trans	chr	Junction	From Exon	To Exon	Junction Type	Sample1	Sample2	baseMean	baseMeanA	baseMeanB	foldChange	log2FoldChange	abs log2
188068	1/2-SBSRNA4	NR_039978	chr4	110351787 110352902	110351119 110351936	110352848 110352902	refseq	CTRL	AD	0.10192	0	0.20385	0	0	0
184253	1/2-SBSRNA4	NR_039978	chr4	110352848 110354973	110352848 110352902	110354890 110354973	refseq	CTRL	AD	0.09213	0	0.18426	0	0	0
7778	A1BG	NM_130786	chr19	58858246 58858868	58858172 58858395	58858719 58859006	refseq	CTRL	AD	5.04	3.87	6.2	0.62467	-0.67883	0.678
7779	A1BG	NM_130786	chr19	58862904 58863798	58862757 58863053	58863649 58863921	refseq	CTRL	AD	0.08757	0.17514	0	0	0	0
161314	A1BG	NM_130786	chr19	58858857 58861885	58858719 58859006	58861736 58862017	refseq	CTRL	AD	0.23441	0.46881	0	0	0	0
171553	A1BG	NM_130786	chr19	58861868 58862906	58861736 58862017	58862757 58863053	refseq	CTRL	AD	0.06319	0.12637	0	0	0	0
171554	A1BG	NM_130786	chr19	58864414 58864693	58864294 58864563	58864658 58864693	refseq	CTRL	AD	0.12637	0.25274	0	0	0	0
184254	A1BG-AS1	NR_015380	chr19	58864261 58864840	58863336 58864410	58864745 58864840	refseq	CTRL	AD	0.16808	0	0.33615	0	0	0
171555	A1BG-AS1	NR_015380	chr19	58865080 58865884	58865080 58865223	58865735 58866549	refseq	CTRL	AD	0.37187	0.25274	0.491	0.51475	-0.95805	0.958
7780	A1BG-AS1	NR_015380	chr19	58864745 58865223	58864745 58864840	58865080 58865223	refseq	CTRL	AD	1.05	1.02	1.08	0.9445	-0.08238	0.082
7781	A2LD1	NM_001195087	chr13	101236102 101241046	101236079 101236251	101240995 101241046	refseq	CTRL	AD	0.33906	0.35029	0.32784	1.07	0.09556	0.095
171556	A2LD1	NM_001195087	chr13	101184706 101236228	101182418 101184855	101236079 101236251	refseq	CTRL	AD	0.71162	0.49344	0.92979	0.5307	-0.91404	0.914
184255	A2LD1	NM_033110	chr13	101184706 101185966	101182418 101184855	101185817 101186056	refseq	CTRL	AD	0.09213	0	0.18426	0	0	0
				9242952	9242952	9243797									

Fig. 15 The report page of observed junctions, each annotated by many information



File	Label	Fusion Points
1	CTRL_1	339
2	CTRL_2	202
3	CTRL_3	201
4	CTRL_4	346
5	CTRL_5	181
6	AD_9	206
7	AD_10	200
8	AD_11	275
9	AD_12	407
10	AD_13	335

Fig. 16 Summary table of the number of gene fusion events for each input and dynamical chromosomal distribution of fusion breakpoints

Click on a column title to order this table

UID	5' partner	3' partner	5' Fraction	3' Fraction	Cluster	Type	Distance	Score	spanning frag.s	unique align.s	5' / 3' Localization
141	chrX:9754495-9912952 + SHROOM2 NM_001649 :0-4673	chrX:10031484-10112517 + WWC3 NM_015691 :104-6453	0.33333	0.5	CLUSTER218	Intrachromosomal	66313	2	1	2	unknown unknown
209	chrX:148611049-148621311 - LOC100131434 NR_027455 :0-1517	chrX:148560294-148584974 - IDS NM_001166550 :457-5807	1	0.04651	CLUSTER98	Read_Through	-22247	2	0	2	unknown unknown
101	chrX:101975641-101976018 + BHLHB9 NM_030639 :0-242 NM_001142528 :0-178 NM_001142526 :0-178 NM_001142525 :0-242	chrX:102081896-102140337 + LOC100287765 NR_038988 :335-2116	0.6	0.21429	CLUSTER236	Read_Through	16725	3	0	3	unknown unknown
165	chrX:47842684-47863376 - ZNF182 NM_001007088 :0-450	chr19:21216891-21242851 + ZNF430 NM_025189 :404-3918	0.33333	0.05263	CLUSTER24	Interchromosomal	0	2	0	2	unknown unknown
233	chrUn:105423-118779 + RN45S NR_046235 :0-13356	chr3:133543079-133614690 - RAB6B NM_016577 :0-5559	1	1	CLUSTER137	Interchromosomal	0	2	0	2	unknown unknown
185	chrUn:105423-118779 + RN45S NR_046235 :0-13356	chr11:10529433-10530722 - MTRNR2L8 NM_001190702 :0-1289	1	1	CLUSTER61	Interchromosomal	0	2	0	2	unknown unknown

Fig. 17 Detailed information and annotations for each gene fusion event

of differential analyses only starting from a completed main analysis, through the result sections, because DE is based on a statistical comparison of results obtained from two or more independents inputs.

A specific section of results (named Differential Expression) is devoted to configure and visualize differential expression operations. At first, this section only contains the list of input files, the user can request for a differential expression by selecting inputs to be included in the operation and assigning them to custom groups.

The user can select the type of differential expression operation by choosing between transcript level (based on Cuffdiff2), gene level (based on DESeq2) and both. Once configured the requested operation, the section is enriched by the differential expression request by reporting the list of selected files with corresponding assigned groups as well as the job execution status. After the operation completion, the *expand results* button allows to visualize results. This button opens an operation matrix reporting all analyzed pairs and a summary of configured parameters.

The differential usage of Poly(A) sites can be requested starting from the PolyA Sites section. The input selection and grouping strategy is the same already described above. Differential usage of polyadenylation sites is computed by a combination of custom scripts and DESeq2.

Acknowledgments

This work was supported by ELIXIR-IIB.

References

1. Emrich SJ, Barbazuk WB, Li L, Schnable PS (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res 17:69–73
2. Nagalakshmi U et al (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344–1350
3. Mortazavi A, Williams BA, Mccue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA- Seq. Nat Methods 5:621–628
4. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63
5. Wang ET et al (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456:470–476
6. Morris KV, Mattick JS (2014) The rise of regulatory RNA. Nat Rev Genet 15:423–437
7. Li W, Notani D, Rosenfeld MG (2016) Enhancers as non-coding RNA transcription units: recent insights and future perspectives. Nat Rev Genet 17:207–223
8. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ et al (2015) Big data: astronomical or genomical? PLoS Biol 13(7): e1002195. https://doi.org/10.1371/journal.pbio.1002195
9. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration (2011) The sequence read archive. Nucleic Acids Res 39:D19–D21
10. TCGA, Tumor Fusion Gene Data Portal @ONLINE. http://54.8 4.12.177/PanCanFusV2//. Aug 2017
11. GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45(6):580–585. https://doi.org/10.1038/ng.2653
12. CCLE, Broad Institute portal—CCLE Repository. https://portals. broadinstitute.org/ccle/home. Oct 2016
13. Picardi E, Manzari C, Mastropasqua F, Aiello I, D'Erchia AM, Pesole G (2015) Profiling RNA editing in human tissues: towards the inosinome Atlas. Sci Rep 5:14941
14. Licht K, Kapoor U, Amman F et al (2019) A high resolution A-to-I editing map in the mouse identifies editing events controlled by pre-mRNA splicing. Genome Res 29 (9):1453–1463
15. Cirilli M, Flati T, Gioiosa S, Tagliaferri I, Ciacciulli A, Gao Z, Gattolin S, Geuna F, Maggi F, Bottoni P, Rossini L, Bassi D, Castrignanò T, Chillemi G (2018) PeachVar-DB: a curated collection of genetic variations for the interactive analysis of Peach Genome Data. Plant Cell Physiol 59:1–9. ISSN: 0032-0781

16. Gioiosa S, Bolis M, Flati T, Massini A, Garattini E, Chillemi G, Fratelli M, Castrignanò T (2018) Massive NGS data analysis reveals hundreds of potential novel gene fusions in human cell lines. GIGASCIENCE 7:1–8. ISSN: 2047-217X
17. Gatto A, Torroja-Fungairino C, Mazzarotto F, Cook SA, Barton PJ, Sanchez-Cabo F, Lara-Pezzi E (2014) FineSplice, enhanced splice junction detection and quantification: a novel pipeline based on the assessment of diverse RNASeq 17. alignment solutions. Nucleic Acids Res 42(8):e71
18. Kalari KR, Nair AA, Bhavsar JD, O'Brien DR, Davila JI, Bockol MA, Nie J, Tang X, Baheti S, Doughty JB et al (2014) MAP-RSeq: Mayo Analysis Pipeline for 18. RNA sequencing. BMC Bioinformatics 15(1):224
19. Boria I, Boatti L, Pesole G, Mignone F, Hong D, Rhie A, Park SS, Lee J, Ju YS, Kim S, Yu SB, Bleazard T, Park HS, Rhee H et al (2012) FX: an RNA-Seq analysis 19. Tool on the cloud. Bioinformatics 28(5):721–723
20. RNA Bioinformatics (ed) (2015) Editor Ernesto Picardi. "Exploring the RNA editing potential of RNA-seq data by ExpEdit". Mattia D'Antonio, Ernesto Picardi, Tiziana Castrignanò, Anna Maria D'Erchia, and Graziano Pesole. Methods Mol Biol 1269:365–378
21. Picardi E, D'Antonio M, Carrabino D, Castrignanò T, Pesole G (2011) ExpEdit: a web server to explore human RNA editing in RNA-Seq experiments. Bioinformatics 27 (9):1311–1312
22. D'Antonio M, D'Onorio De Meo P, Pallocca M, Picardi E, D'Erchia AM, Calogero R, Castrignanò T, Pesole G (2015) RAP: RNA-Seq Analysis Pipeline, a new cloud-based NGS web application. BMC Genomics 16:S3
23. FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
24. Patel RK, Jain M (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7(2):e30619
25. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14(4):R36. https://doi.org/10.1186/gb-2013-14-4-r36
26. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9 (4):357–359. https://doi.org/10.1038/nmeth.1923
27. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635. Epub 2012 Oct 25
28. Roberts A, Pimentel H, Trapnell C, Pachter L (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics 27(17):2325–2329
29. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11(10):R106. https://doi.org/10.1186/gb-2010-11-10-r106
30. Iyer MK, Chinnaiyan AM, Maher CA (2011) ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 27(20):2903–2904. https://doi.org/10.1093/bioinformatics/btr467
31. GFF/GTF file format—Definition and supported options. http://www.ensembl.org/info/website/upload/gff.html
32. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35 (Database):D61–D65
33. Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D (2000) Patterns of variant polyadenylation signal usage in human genes. Genome Res 10(7):1001–1010
34. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31 (1):46–53. https://doi.org/10.1038/nbt.2450. Epub 2012 Dec 9
35. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. J R Stat Soc B Met 57(1):289–300