NIB Final Report

MILESTONE M27: User friendly data management and analysis environment for plant systems biology established and novel/improved multiomics data visualisation tool available

Genialis Llc., National Institute of Biology Version 1 FINAL

Abstract: We present the technical report regarding the implementation of the Genialis Platform at the National Institute of Biology (NIB) within the project Hyp (J4-7636).

Document administrative information:

Project acronym:	HYp
Project number:	J4-7636
Document identifier:	Hyp M27 v1.FINAL
Lead partner short name:	NIB, Genialis Llc.
Report version:	Version 1, FINAL
Report preparation date:	2018/05/31
Lead author:	Luka Ausec
Co-authors:	Barbara Jenko, Janez Kokošar, Amrita Roy Choudhury, Miha Štajdohar
Co-authors:	Kristina Gruden, Maja Križnik, Marko Petek, Živa Ramšak
Status:	Final

Application for the analysis of RNA-Seq data - Technical Report

Project ARRS J4-7636 - Spatiotemporal analysis of hypersensitive response to Potato virus Y in potato

Genialis, d.o.o. | Koprska ulica 72, 1000 Ljubljana, Slovenija
www.genialis.com

Dec 21, 2018

Povzetek

Genialis je manjši partner v projektu ARRS J4-7636 Prostorsko časovna analiza hipersenzitivnega odziva krompirja na krompirjev virus Y. Naša naloga je bila pomagati projektnim partnerjem analizirati in upravljati podatke pridobljene z metodo RNA-Seq. Genialis je zato na podlagi lastne Genialis Platform razvil aplikacijo prirejeno aplikacijo in jo ponudil v uporabo na računalniški infrastrukturi pri vodji projekta na Nacionalnem inštitutu za Biologijo.

Aplikacija je sestavljena iz treh glavnih modulov. Prvi modul je enostaven grafični vmesnik za vnos surovih podatkov. Uporabniki lahko naložene podatke analizirajo avtomatsko s prednastavljenim analitskim postopkom, ki ga je Genialis implementiral po priporočilih vodje projekta. Genialis je prispeval tudi k izboljšani anotaciji genoma krompirja. Drugi modul je namenjen upravljanju podatkov s konceptoma vzorcev in zbirk, omogoča pa tudi enostavno anotiranje podatkov, da jih uporabniki lahko kasneje hitro in zanesljivo najdejo. Tretji modul pa je sestavljen iz številnih medsebojno povezanih grafičnih komponent za vizualno raziskovanje rezultatov analiz.

V tem dokumentu natančno popišemo Genialisov prispevek k realizaciji Nalog 6 in 7, torej aplikacijo za analizo RNA-Seq podatkov. Najprej podamo splošen opis aplikacije, sledi še opis implementiranih analitskih postopkov. Poročilo lahko služi tudi kot tehnična dokumentacija za razviti produkt.

Abstract

Genialis is a minor partner in the project ARRS J4-7636 Spatiotemporal analysis of hypersensitive response to Potato virus Y in potato. Our goal was to help the project consortium manage and analyse RNA-Seq data produced within the project. To this aim, Genialis developed an application on top of our proprietary Genialis Platform, and deployed it within the computational infrastructure of the project leader at the National Institute of Biology.

The application consists of three main modules. The first provides a wizard-like interface for data input. Once raw sequence files have been uploaded the users can run the predefined analysis pipeline with a single click. Genialis implemented the tools and pipeline according to the specifications of the project leader. Moreover, Genialis contributed to the development of an improved potato genome reference. The second module allows for data management and annotations of samples and collections, and indexes the metadata to provide a rapid search functionality. Finally, the third module consists of several interlinked components for visual analysis of the data.

The purpose of this document is to thoroughly describe Genialis' deliverables within Work Packages 6 and 7, that is the RNA-Seq application. We begin by explaining the general

functionality of the software and proceed by detailed account of implemented analytical processes. This report thus also serves as technical documentation of the software.

Table of Contents

General description of the software features

This section provides a thorough description of the main software features. We begin by introducing the underlying principles such as data model and permissions management. This is followed by a detailed account of the three main modules of the application: data input (and analysis), search and results (and data management), and visualisations.

Definition of basic concepts in the platform

A **collection** is a container for samples and attached data. It is created upon data upload, and users can specify the name and a description. These become indexed fields for the collection search. Additional collection-dependent functionality is available, such as generating multi-sample QC reports and adding metadata (defining sample relations).

A **sample** is a basic operational unit in the Genialis Platform. A sample is a bunch of data objects attached to a name, and optional metadata fields (annotations). For this reason, the end user never needs to work with data objects of even files, since the system is smart enough to figure out the right inputs for a particular job when provided with a sample name.

A **data object** is created by an analytical process on the Genialis Platform and contains inputs (how the object was created), outputs (resulting files), and potentially other information (e.g. command line output of the function call, etc.). For example, an *alignment* object would contain the name of the tool, the input parameters, the .bam and .bam.bai files, the mapping stats text file, potentially tracks for genome browsers, etc. Data objects are displayed in a table in the collection details page.

A **file** is the same thing as a file in the computer science sense, e.g. a BAM file or a FASTQ file. They can be downloaded individually in a number of ways, but they are in general conveniently hidden from the end user that just wants to explore the results in the application.

Sample Basket and Genes Basket are containers that hold samples and genes, respectively. Visualisations critically depend on them, only data in baskets gets displayed. For this reason, there are a number of ways to modify the contents of the baskets, e.g. with manual input, selection from visual modules, by saving and loading gene sets and/or collections, etc.

Application currently consists of 5 **modules:** Home, Analyze, Search & View Results, Visualizations, and Bioinformatic Dashboard. Visualisation module has five **tabs**: Sample Comparison, Gene Expressions, Differential Expression, Venn View, Heat Map. Each tab may have two or more **cards.** Due to recent introduction of collection search and imbedded bioinformatic tools and pipelines, Bioinformatic Dashboard is becoming obsolete and will be removed in future versions of the software.

ReSDK is a Python-based application programming interface (API) for Genialis Platform giving users programmatic access to all of the functionality of the application. Genialis provides complete documentation and training for the more computational-savvy users.

Granular permissions management allows users to make their data public and to share their data with other users or groups. When sharing with another individual or group, the user can define permission level: is owner, can share, can edit, can download, or can view. Permissions may be managed for individual collections, samples, or data objects from the collection details page. Moreover, permissions may be quarried and changed using the Python-based ReSDK. Admin users can create user groups (for data sharing) and user types (for different UX with the software). The latter means that we can configure e.g. “normal users” and “advanced users” with different access (e.g. only advanced users can access analytical pipelines).

Data input and running analyses

Importing raw data, i.e. sequencing reads straight from the sequencer, is a three-step process. First, the users selects data for upload, and these can be files from a local machine or from BaseSpace (an online tool and repository by Illumina). In the first case, a simple drag-and-drop is all it takes, in the second, the user first needs to securely log into their BaseSpace account. The system expects single-end or paired-end compressed fastq reads generated by Illumina sequencers (typically a *.fastq.gz or *.fq.gz). If a single sample is sequenced in multiple rows, there may be multiple (pairs) of raw reads - in this case, the system will automatically concatenate these files before downstream analysis. Upload is a slow process that can literally take hours for even moderate-size batches of samples. Our upload can resume after an interruption (e.g. brake in internet connection), assuming the browser window and browser tab remain open. This is critical, since connection if irreversibly broken if users navigate away from the upload screen or if they close the tab (they are warned not to do so). Users then verify if data files are correctly assigned to samples, and edit the sample names if necessary.

There are other ways to upload your data:

- Use ReSDK, our Python-based application programming interface to automate data upload and analysis. Users can upload files directly from FTP servers, grab data from GEO, or upload other data types (e.g. BAMs or expressions).
- Use bioinformatic tools to upload other types of data, e.g. mapping files (BAM).

The second step in data import is quality control. In this step, afastqc report is generated for each uploaded file. Users can check that their data is as expected (e.g. number of reads, quality of sequencing, level of duplicates, etc.). It takes seconds to minutes for all the QC reports to be calculated. The user can also skip this step.

The third step defines the experiment. The users creates a collection for the batch of samples. User needs to select a source of annotation (options are ITAG, PGSC, or combined

ITAG/PGSC). Finally, users can choose some advanced parameters to the pipeline, e.g. choose strandedness of their data or more relaxed/stringent trimming parameters).

With this, the user can start the analysis. They can follow the progress of each sample in Search and Results page by clicking on the progress icon. They are notified over email once the entire batch of samples is done. Several quality reports are generated for each sample, i.e. fastqc of raw reads, fastqc of processed reads, mapping statistics. The user can combine those into a single report using the multiqc tool.

Data management, searching, and retrieving results

Genialis Platform provides users with a centralized location to organize, store, protect, validate, and process their data. This enables users to access their lab's historical data at any time to validate what was done previously, reanalyze and ask new questions of the data, and compare different datasets.

For each processed sample, Genialis records the history, sequence, and runtime details of every processing step carried out on that data to ensure data provenance. Each analysis step includes runtime inputs, performance metrics, downloadable or viewable intermediate files, and standard outputs and command outputs. Sample history is key if users want to troubleshoot their analysis, publish the methods, download an intermediate results file, or how the data was processed analysis.

User-defined Annotations include information such as species, RNA source (such as tissue/cell line), treatment (such as drug A, KO, fasted, control), duration of treatment (such as 1h, 2 weeks), dosage (such as 5 mg/ml), gender, age, annotators/contributors, case vs control.

Metadata are important for several reasons. (1) you cannot publish without submitting the data AND metadata to a public repository. (2) recording metadata is essential to reproducibility and is a basic best practice (often overlooked). (3) Most metadata fields are indexed in the search engine, so they can be used later to quickly find data. (4) The color-by and group-by for the box plots and bar charts are dependent on these sample details.

- Annotator (required): who uploaded and annotated the data
- Organism (required) the organism from which the sequences were derived
- Source (required): the biological material from which the sample was derived
- Cell Type: the cell type from which the sample was obtained
- Strain: the microbial or eukaryotic strain name
- Genotype: the genetic makeup of the sample
- Molecule: the type of molecule that was extracted from the biological material
- Description: additional information about the sample that was not included in the previous fields
- Notes: additional information about the sample that was not included in the previous fields

There are two dedicated search pages, one for samples and one for collections. They both work in a similar manner. The search works on sample/collection name and metadata fields and returns the results within a second. Sample results come in a table with linked key result files, collection results come with a bit of extra information, e.g. description. In either case, clicking on a result name gives further details of that sample or collection. Currently, only searching for sub-words work (e.g. one would find “sample1-treatment-P0032” by searching for “sample1” or “P0032”, but not by searching for “ample1” or “P00”).

Interactive visualisation components

Visual exploration of results starts by populating the Sample basked from Search and View results. Sample Basket is a container for samples represented by a grey retractable side bar on the left side of the visualisations. In fact, the user can toggle between sample and genes basket, with the latter being a container for the current selection of genes that are added from anywhere within visualisations. Genes and samples follow the user through visualisations, they can always be inspected, removed, or stored (gene sets). The following sections describe the five tabs in the visualisations.

Sample Comparison view

Questions users are trying to answer here:
Are my replicates consistent? Are my experimental controls/treatments distinct from one another?

What happens here?
Sample Comparison is the first place users go to understand their data. The first pass is usually a sanity check, whereby users can immediately detect if replicates and experimental conditions group as expected. Deeper inspection yields insights into the overall similarities and differences evinced by different experimental or sampling regimen. Reproducibility between technical replicates should be generally high, but there may be larger differences between biological replicates. Thus users can easily identify if any outliers exist before proceeding with the downstream analysis.

Sample Hierarchical Clustering - Sample Hierarchical Clustering reveals the overall distance, or dissimilarity between samples. Samples or clades of samples that group more closely together, e.g. with less distance between them, may be inferred to have more similar transcriptome phenotypes. One would expect replicates to group closely within clades. Like treatments, genotypes or experimental conditions might also yield more closely ordered samples. The clustering plot includes all samples in the basket, and will update automatically when a new set of samples is added to the basket. The user may select among three distance functions—Euclidean, Pearson’s and Spearman’s—as well as three linkage types—Average, Complete, or Single. Further, the user may toggle between a dendrogram based on the entire transcriptome or a plot derived only from the genes included in the Genes module.

Principal Component Analysis - PCA is a mathematical approach that identifies and ranks the dimensions, or principal components (PC), that account for the largest proportion of variation within a dataset. The PCA plot allows users to see how/whether their samples cluster. This can help assess the quality of their experiment by determining the reproducibility among replicates and determining possible batch effects.The PCA plot includes all samples in the basket, and will update automatically when samples are added or removed from the basket. The user may toggle between a PCA plot based on the entire transcriptome or a plot derived only from the genes included in the Genes basket. Users may also turn on or off the sample labels. Hovering over any point will also highlight that sample label.

Gene Expression view

Questions users are trying to answer here:
Do my known markers genes display expected expression patterns? How do the genes I just discovered behave across and between samples?

What happens here?
With the Gene Expressions view, users may want to check whether or not their favorite gene or set of genes is behaving as expected under certain experimental conditions. Alternatively, they may have hypothesized that their favorite gene, set of genes, and/or gene signature would behave in a certain way under given conditions, and they can quickly check that here.

Box Plot - The Tukey box plot, or box-and-whisker plot, is a familiar way to display non-parametric data without making assumptions on the underlying distribution. By default, the expression values (transcripts per million, or TPM) across all samples in the basket (y-axis) are shown for each selected gene (x-axis). The horizontal line bisecting the box marks the median value whereas the bottom and top of the box denote the 1st and 3rd quartiles, respectively. The whiskers represent the lowest datum within 1.5 interquartile range (IQR) of the 1st quartile, and the highest datum within 1.5 IQR of the 3rd quartile. When an individual gene is selected in the Genes basket, the corresponding box is shaded blue. Mouse hover over boxes reveals a tooltip with additional quantitative details. The user may toggle log2 transformation of the y-axis expression values. Further, the user may use the Group-by menu to organize the expressions by meta-data in order to visualize the distribution of various sample groupings one gene at a time.

Bar Chart - The bar plot enables a deep dive to see how selected genes are expressed (TPM) in each individual sample in the basket. The user may consider known biomarkers, novel genes of interest, or look more closely at differentially expressed genes identified in the subsequent dashboard. When an individual gene is selected in the Genes basket, the corresponding bars are shaded blue. Mouse hover reveals a tooltip with additional quantitative details. The user may toggle log2 transformation of the y-axis expression values. Further, the user may use the

Group-by or Color-by menus to organize or mark the bars by meta-data to provide more insightful visual clues.

Differential Expressions view

Questions users are trying to answer here:
Which genes are upregulated and/or downregulated in response to treatment but not the controls?

What happens here?
Here users can quickly see what genes and/or gene sets are significantly up- and downregulated. This is where they get to interact with their data, click on data points, drill-down and find out more information about a particular data point, zoom in/out, and select groups of data points to explore further. Users would first need to either create a differential expression group by clicking on the plus sign or select an existing differential expression group(s) in the DE selection card. If one group is selected, then the volcano plot will automatically update. If two groups are selected, then the differential comparison plot is automatically updated.

Differential Expression Selection - Differential Expression analysis is a powerful and now common tool for determining statistically measurable differences in gene abundances between two groups of samples. One might use this analysis to explore which genes are up- or down-regulated in response to a given experimental treatment or between two patient samples, tissues or genotypes. This table allows the user to choose one or more differential expressions (DE) datasets to visualize in the adjoining plots and tables. Selecting one DE object will populate the Volcano plot. Selecting two DE objects will populate the DE Comparison Plot. The user may inspect the existing DE details and toggle the false discovery rate (FDR) and fold change thresholds by clicking on the DE name within a row. The user may also initiate a new DE analysis by clicking the (+) icon and then define the samples for case and control groups. Currently, DESeq2 tool is supported. The user can only set the thresholds for what they define as differentially expressed gene (fold-change difference and false discovery rate FDR value), but cannot change other parameters of the analysis. Precomputed DE objects directly uploaded to the platform are displayed here automatically.

Volcano Plot - The Volcano Plot has become an indispensable means of visualizing differentially expressed genes. Every dot is a gene, with relative fold change (log2 FC) on the x-axis and the statistic -log10 FDR on the y-axis. Thus the further from zero a gene is displayed, the greater the difference in expression level and/or statistical confidence we may infer. Dots representing genes in the Genes basket are shown larger and darker than background, with the actively selected gene in Blue. Mouse hover over any dot reveals additional information. FC and FDR thresholds may be modified, thus moving the corresponding lines on the plot. Outlier genes with log2(FC) > 10 are stacked by default, but may be toggled to display at their actual value. Genes of interest may be selected by drawing a box around those dots with the mouse. The resulting

pop-up enables the user to inspect those genes and append or overwrite them to the Genes basket.

Differential Expression Comparison Table - Data is displayed here when a user adds genes to the Genes basket. The table displays the gene symbol, gene name, and fold change value. The log2(FC) values are shaded on a color scale (e.g. like a heatmap) to facilitate detection of markedly different abundance patterns. The data are sortable by column, such that the gene list can be ranked for any of the selected DE analyses. This an alternative way to view which genes are up- and downregulated.

Differential Comparison Plot - By selecting two DE objects from the DE selection table above, the user may simultaneously compare the relative fold change of all the genes in the transcriptome. The axes of this plot represent the log2(FC) from the two selected DE analyses. Genes that fall along the x=y diagonal display similar abundance patterns in the chosen comparisons, while off-diagonal genes vary between experiments. Points representing genes in the Gene basket are shown as larger, darker circles, with any Gene basket selection colored blue.

Venn view

Questions users are trying to answer here:
How many of the up or downregulated genes are unique to my treatment group? (Answered by the Venn diagram) Do these genes represent any biological function? (Answered by the GO Enrichment Analysis)

What happens here?
Users select gene sets to compare in order to identify relationships between them. From there, they can select unique gene sets and discover if any of those genes represent a biological function. In short, the Venn diagram enables users to organize information visually so they are able to quickly see relationships between two or three different gene sets and identify similarities and differences between them.

Saved Venn Overlaps - Users can select previously saved Venn diagrams. Selecting one updates the Venn diagram.

Venn Diagram - Users can create Venn diagrams from saved gene sets or saved Venn overlaps. Up to four gene sets can be compared. Selecting segments of the Venn diagram allows the user to save new gene sets, view gene expression patterns in the heatmap and assign functional annotation from Gene Ontology Enrichment. By default the area of each Venn region is displayed log-proportional to the number of genes represented, but the user can also chose areas of equal size. New Venn diagrams may be created by clicking the (+) and selecting gene sets from the pop-up. Highlight one or more segments of the Venn diagram by clicking

them in sequence. De-select a segment by clicking it again. Toggle the Venn from proportional to equal-sized areas.

MapMan Ontology Enrichment Analysis is a common method to determine the functional profile of a gene set of interest. The significantly enriched MapMan Ontology (GoMapMan) terms, identified by statistical tests, are indicative of the underlying biological processes and pathways affected by the gene set. The GoMapMan enrichment analysis table shows the following columns: 1) p-value which determines whether or not the over- or under-representation of a particular GoMapMan term is significant, 2) the enrichment score which is the ratio between the number of genes from input gene set that are assigned to specific GoMapMan term divided by the number of genes in input gene set and the number of genes in that GoMapMan term divided by the number genes in genome, thus a score greater than 1 indicates over-representation of input gene set, whereas a score between 0 and 1 indicates under-representation, 3) N which is the number of genes in the set associated with the particular GoMapMan term / total number of genes in the genome associated with the same GoMapMan term. The GoMapMan enrichment analysis is performed automatically for the genes in the highlighted segments of the Venn diagram. The user may also add/remove genes in the Genes basket or select a previously saved gene set to trigger a new analysis. Any of the three GoMapMan term categories can be used. The enriched terms are listed along with their associated p-values and scores. The user may also choose a desired p-value threshold.

Heatmap view

Questions users are trying to answer here:
Do any of the selected genes show a striking visual pattern?

What happens here?
The heatmap plots quantitative differences in expression levels of selected genes in each individual sample in the basket to allow qualitative overview of the transcriptomic landscape. The user may consider known biomarkers, novel genes of interest or look more closely at differentially expressed genes identified in the DE dashboard. In other words, the heatmap enables users to visually compare gene expression across multiple samples.

Gene Sets - Lists previously saved gene sets. If a user selects one of these, then the genes basket and heatmap update accordingly.

Expression Heatmap - The heatmap updates automatically when genes are added/removed from the genes basket. The rows and columns in the heatmap represent the selected genes and the samples in the basket, respectively. These can be clustered to reveal which genes display the most similar expression patterns, and which samples are most alike for this selection of genes. Hierarchical clustering in this plot is based on Euclidean distance as that measure is applicable regardless of the data transformation approach and robust to non-normal data distributions. The user may transform the data by row-wise Z-score (default), log2, or Z-score of

log2. Log2 data transformation is typically performed in order to stabilize the variance, compress the range of values, and distribute the data more normally to satisfy statistical assumptions. Z-score transformation, also known as the standard score, calculates the probability of a score occurring within the normal distribution which enables comparisons between two scores that are from different normal distributions. Selecting a data transformation will recompute clustering (if selected) and affect the color scale accordingly. The user may also choose to display the gene names. Mouse hover over any cell reveals additional information.

Downloading results

Downloading results from the Search and Results page.
Apart from the QC report, Results table gives direct access to the mapping and expressions results. Expressions is a plain tab-separated text file (essentially a table that can be readily imported to Excel) that lists gene ID, gene symbols, raw counts and normalised expressions (TPM and FPKM) for all the genes. The user can access other files by selecting (multiple) samples using check boxes and clicking the download button - they get to select what files to download, e.g. BAMs or QC reports. If they select expression, they'll also get a multisample TMP expressions data matrix with genes in rows and samples in columns.

Downloading results from Sample Details page.
Sample History on Sample details page lists all the inputs and outputs of each step in the pipeline. All outputs are either directly viewable or downloadable. Examples of downloadable result files include:

- Trimmed FASTQ (fastq.gz file format) - Downloadable after QC and trimming, processed FASTQ files are smaller than the raw FASTQ files, and contain only quality bases and no sequencing artefacts. They may be used for storage and reanalysis with external tools.
- BAM (mapping, binary format) - BAM files are huge (GB) and in some sense even more valuable than fastq - it's easy to extract fastq reads from them (so in theory you don't need to keep both) and they have all the alignments, so they can be used with external tools for quantification of expressions, all sorts of diagnostics, and more. Often, users will want to visualize them in a genome browser. For that, they'll need the .bam.bai index file, too. That one is just a tiny (a couple of kb) file but essential. While fastq files are human-readable text files (but compressed), BAM files are binary files (only machine-readable), but this makes them really size-efficient. One could turn them into compressed human-readable files, too (=SAM), but they take more than twice the storage space.
- .bam.bai - index file for BAM - Needed to visualize BAM in genome browsers
- Mapping statistics (.txt file format) - Mapping stats is a tiny text file with just a couple of lines of summary statistics of mapping, e.g. number of reads mapped and not mapped, numbers of pairs identified and mapped (if paired-end sequencing), etc. You wouldn't put this in a paper, but you would normally look at this to make sure things are as expected

(e.g. 10 percent not mapped reads is OK, 20 percent would raise eyebrows, 30% or more would make you want your money back from the sequencing core).

- Expressions and differential expressions (tab-separated text files) - Expressions and DE are small text files (MB), tab-formatted, usually with many rows (one row per gene or transcript, N>20k) and very few columns (gene name, count, p-value or FDR, perhaps one or two more tool-specific columns). People can view these files in Excel, plot figures from them with Excel or R, or run additional tools on them such as pathway enrichment.

Downloading results from the Visualisations.
Results of very tab in the visualisations can be downloaded as a self-contained archive bundle. It not only contains images in raster (png) and vector (svg) formats, but comes with supplementary information, data tables and figure captions.

Bioinformatic tools and pipelines

This section gives technical details about the analysis pipeline, such as versions of tools, and parameters passed to these tools. Specific on the combined reference genome are given.

Description of the RNA-Seq pipeline

Adapter removal and quality trimming in the raw fastq Illumina reads (single-end or paired-end) is done with Cutadapt (Martin, 2011). A selection of Illumina adapters is already available on the platform. Quality report of raw and trimmed reads is done using fastqc (accessible at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The resulting processed file is then passed on to the STAR aligner tool (Dobin et al., 2013) which maps reads to a reference genome. These references are already available on the platform: PGSC v 4.04, iTAG, the former two combined (see description below). Quantification of gene expression is done with featureCounts from the Subread package (Liao et al., 2014) using annotations of respective genome versions. A custom script is used to calculate normalised expression values (FPKM and TPM). Several analysis statistics (read quality, mapping stats) can be aggregated using multiqc (Ewels et al., 2016). DESeq2 (Love et al., 2014) can be used to run differential expression analysis between groups of samples.

Here are the details on the parameters used for the various processes (listed are those that are different from defaults or those that are key to understanding the workflow):

1. Cutadapt (v. 1.16)
    - QC parameters:
        - -q <int> (min read QV allowed; default: 20) (leading parameter in our process)
        - --minimum-length <int> (default: 50)
        - --max-n <int> (max number of N-s in the sequence allowed; default: 2)
    - Parameters for paired end data:

    - --paired-output
    - --pair-filter=both

2. STAR (v. 2.5.4b)

User can chose reference files (PGSC v 4.04, iTAG, combined)

- --outSAMtype BAM SortedByCoordinate
- --outReadsUnmapped
- --twopassMode

3. featureCounts (v. 1.6.0)

- –fracOverlap <float> (minimum fraction of overlapping bases required for assigning a read to a feature)
- -M (to fully count every alignment reported for a multi-mapping read; default FALSE)
- --fraction (count each alignment fractionally; default FALSE)
- For PE data::
    - -p (If specified, fragments (or templates) will be counted instead of reads.)
- For strand specificity:
    - -s <int> (0 (unstranded), 1 (stranded) and 2 (reversely stranded). 0 by default. For paired-end reads, strand of the first read is taken as the strand of the whole fragment.)

4. DESeq2 (v.1.18.1)

- cooksCutoff=FALSE
- Expression matrix is filtered prior to processing with DESeq2. Features with sum of counts across all samples < 10 are excluded from the analysis.

Genome reference

Solanum tuberosum genome sequence in *.fasta* format was uploaded to Genialis platform. Genome indices for PGSC, ITAG and combined PGSC+ITAG annotation source were prepared using the STAR aligner indexing tool with the default parameters. The structure of the combined PGSC+ITAG annotation source file was initially prepared by the project leader, and was modified by Genialis to ensure the compatibility with the downstream featureCounts quantification tool. The feature IDs, feature symbols and matching descriptions for all three annotation sources were inserted in the application's Gene Knowledge Base so that these features can be queried within visualisations.

Bioinformatics tools

Other bioinformatic tools are available within the software. They can be accessed from the Collection Details page. The card is named Actions and it has several tabs listing tools for upload and processing of RNA-Seq and other types of data. Additional tools are available in the Tool Catalogue. Each tool has a graphical user interface with inbuilt help that guides the user through selection of input data files and process parameters. Tools create new data objects that can be used in subsequent analysis steps by other tools, results can be extracted for download or displayed in visualisations (depending on the object type).

References

Dobin A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Ewels, P., Magnusson, M., Lundin, S., Käller, M. (2016) MultiQC: Summarize analysis results for multiple tools and samples in a single report, Bioinformatics 32(19), doi: 10.1093/bioinformatics/btw354

Liao Y, Smyth GK and Shi W. (2014) featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7):923-30. DOI: 10.1093/bioinformatics/btt656

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology, 15(12), 550.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 17(1), pp-10.