OXFORD
UNIVERSITY PRESS
Bioinformatics

PlanExp: intuitive integration of complex RNA-seq datasets with planarian omics resources

Journal:	*Bioinformatics*
Manuscript ID	BIOINF-2019-0974.R1
Category:	Original Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Castillo-Lara, Sergio; Universitat de Barcelona, Genètica, Microbiologia i Estadística Pascual-Carreras, Eudald; Universitat de Barcelona, Genètica, Microbiologia i Estadística; Institut de Biomedicina de la Universitat de Barcelona (IBUB) Abril, Josep; FACULTAT de BIOLOGIA - UNIVERSITAT de BARCELONA, COMPUTATIONAL GENOMICS LAB @ GENETICS DEP.
Keywords:	Gene expression, Database, Visualization, Data integration

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Databases and Ontologies

PlanExp: intuitive integration of complex RNA-seq datasets with planarian omics resources

S. Castillo-Lara [1,2], E. Pascual-Carreras [1,2], J.F. Abril [1,2]*

[1] Computational Genomics Lab; Genetics, Microbiology & Statistics Dept.; Universitat de Barcelona, Catalonia, Spain.
[2] Institut de Biomedicina de la Universitat de Barcelona (IBUB); Universitat de Barcelona, Catalonia, Spain.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: There is an increasing amount of transcriptomic and genomic data available for planarians with the advent of both traditional and single–cell RNA sequencing technologies. Therefore, exploring, visualizing, and making sense of all these data in order to understand planarian regeneration and development can be challenging.
Results: In this work we present PlanExp, a web-application to explore and visualize gene expression data from different RNA-seq experiments (both traditional and single-cell RNA-seq) for the planaria *Schmidtea mediterranea*. PlanExp provides tools for creating different interactive plots, such as heatmaps, scatterplots, etc. and links them with the current sequence annotations both at the genome and the transcript level thanks to its integration with the PlanNET web application. PlanExp also provides a full gene/protein network editor, a prediction of genetic interactions from single-cell RNA-seq data, and a network expression mapper that will help researchers to close the gap between systems biology and planarian regeneration.
Availability: PlanExp is freely available at https://compgen.bio.ub.edu/PlanNET/planexp. The source code is available at https://compgen.bio.ub.edu/PlanNET/downloads.
Contact: jabril@ub.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Schmidtea mediterranea has emerged as a widely used model for studying whole-body regeneration in animals due to its regenerating capabilities. During the last few years, the increasing amount of currently available molecular and sequence data has boosted the knowledge of its developmental and regenerative capabilities, consolidating its position as an emerging model organism.

Many RNA-seq experiments driven by different sequencing technologies have been performed in order to understand the regenerative abilities of this animal (Sandmann *et al.*, 2011; Labbé *et al.*, 2012; Kao *et al.*, 2013; Rodríguez-Esteban *et al.*, 2015). While they brought light into the problem of planarian regeneration and stem cell biology in a whole *in-vivo* animal system; accessing, exploring, and visualizing data derived from those experiments from an integrative perspective can be quite challenging. Most of these RNA-seq datasets were bound to the respective *de-novo* transcriptome assembly, and therefore the published data refers to sequences in those assemblies, yet the planarian community has made efforts to use a unique reference transcriptome (Rozanski *et al.*, 2018). With the recent predicted gene-set computed over the newly published planarian genome (Rozanski *et al.*, 2018; Grohme *et al.*, 2018), and the expected future incremental improvements on the genomic feature annotations for this organism, providing a way to link this gene expression data to up-to-date genomic annotations is necessary.

As previously stated, a high quality genome assembly of *S. mediterranea* has been published, and is available for researchers to explore and study freely on PlanMine. In addition, PlanNET (Castillo-Lara and Abril, 2018) offers insights into the protein network dynamics of this organism, in the form of a predicted interactome that links many planarian RNA–seq datasets together with human interolog protein pathways. Both resources offer ways to explore the transcriptome dynamics of planarians, either as static data on transcript entries (in the case of PlanMine), or by mapping expression data onto protein interaction graphs (in PlanNET). Although it can be argued that both approaches are useful and necessary, they can be quite limiting when it comes to visualizing single–cell RNA sequencing data (scRNA-seq), for instance, PlanNET does not have options for filtering its networks according to tissue expression, and PlanMine relies on single transcript pages to check expression levels.

scRNA-seq technologies offer a very interesting outlook of the transcriptomic landscape of organisms, and have also been recently applied to study planarian biology (Plass *et al.*, 2018; Fincher *et al.*, 2018). These technologies have the characteristic of handling a large number of samples, increasing the size of the expression matrix dramatically; with some experiments gathering up to 50,000 cells, each one with expression data for more than 20,000 genes. Storing, visualizing, and exploring these data can be challenging, and the cited experiments performed on *S. mediterranea* provide a small web-application to that end, as this type of data necessitates specific tools to be presented in useful ways. However, these tools offer distinct features, and are often disconnected from the currently available data for this organism, such as sequence annotations, homology information, protein-protein interactions, genomic locations, and so on. This disconnect offers an additional challenge for researchers, who have to collect the data elsewhere in order to make sense of these experimental results.

Finally, genes and proteins do not act alone as the sole entities responsible for any biological processes. Genes are part of regulatory networks, or signalling cascades, while proteins form multi-protein complexes and often act together to perform their function, or to regulate the activity of other proteins. As such, a tool to investigate these relationships between genes in their biological context is necessary. To that effort, one of `PlanExp`'s main focus is the ability to visualize gene expression of multiple genes simultaneously, and to offer insights as to which genes co-express either in the same cells (in the case of scRNA-seq experiments) or in the same experimental conditions. Some tools have been developed for the inference of gene regulatory interactions from RNA-seq experiments (Huynh-Thu *et al.*, 2010), and some take advantage of the large number of samples in scRNA-seq data (Papili Gao *et al.*, 2017; Aibar *et al.*, 2017). Gene co-expression networks extracted from RNA-seq expression matrices have been used in the past to reveal biological insights in other organisms (Potier *et al.*, 2014; Taylor-Teeples *et al.*, 2015; Davie *et al.*, 2018). In this work we have also inferred genetic relationships from two scRNA-seq experiments of the planaria *S. mediterranea*, and the results have been integrated into `PlanExp`, to allow researchers to explore, study, and curate putative genetic interactions in the context of a centralized and multi-purpose web application.

A tool that brings together not only scRNA-seq expression data, but also traditional RNA–Seq data with the current state of the planarian genomic, transcriptomic, and interactomic knowledge can be of great interest for researchers, and will provide valuable insights into the biology of *S. mediterranea*.The main aim of PlanExp is to be a central application in which all this knowledge is accessible, extending PlanNET capabilities into further layers of information.

2 Methods

2.1 Data sources

Currently, six RNA–Seq experiments are publicly available on PlanExp: two of them are scRNA-seq (*"2018 Rajewsky Cell Atlas"* and *"2018 Reddien Cell Atlas"*), while the rest are traditional RNA–Seq or Digital Gene Expression experiments: *"2011 Bartscherer Time-course"* (Sandmann *et al.*, 2011), *"2012 Pearson Stem Cells"* (Labbé *et al.*, 2012), *"2013 Aboobaker Time-course"* (Kao *et al.*, 2013) and *"2015 Abril D.G.E."* (Rodríguez-Esteban *et al.*, 2015).

Raw counts of *"2018 Reddien Cell Atlas"* were downloaded from Gene Expression Omnibus (GSE111764); while the rest of expression count matrices were retrieved from the supplementary material of their respective articles.

2.2 Data processing

2.2.1 RNA–seq experiments

When possible, all differential expression comparisons were obtained from the available online sources, so as to keep the results between the publication and those available on PlanExp identical. However, this was only possible for the experiments *"2015 Abril D.G.E."* and *"2011 Bartscherer Time-course"*.

For the experiment *"2013 Aboobaker Time-course"*, the differential expression analyses were performed with `limma` (Law *et al.*, 2014). The *p*-value cut-off was selected based on the one reported in the corresponding article. In the case of *"2012 Pearson Stem Cells"* no differential gene expression analysis was performed due to the lack of replicates.

The design matrix of the experiment, as well as the contrasts performed for each one of them can be found in Supplementary Table 2. A summary of the analyses carried out can be found in Supplementary Material "*Integration of RNA-seq experiments*" section.

2.2.2 Single–cell experiments

Both scRNA-seq experiments were analyzed with the `Seurat` package (Butler *et al.*, 2018). Only those cells filtered by the respective authors were considered, and the cell cluster assignments were downloaded but not re-computed.

A differential expression analysis between each pair of clusters was performed with Seurat's function `FindMarkers`, using a log fold change cut-off of 0.2 (`logfc.threshold=0.2`), considering only genes expressed in at least 20% of both clusters (`min.pct=0.2`). Differentially expressed genes were selected by using an adjusted *p*-value cut-off of 10^{-5}.

An additional marker discovery analysis was performed using `Seurat`'s function `FindAllMarkers`. In this case, this process was performed by selecting the area under the ROC curve (AUC) as the test to use (`test.use="roc"`), with the parameters `min.pct=0.25` and `thresh.use=0.25`.

2.3 Genetic interactions prediction

A prediction of genetic interactions for both scRNA-seq experiments was calculated by `GRNBoost` from the `SCENIC` pipeline (Aibar *et al.*, 2017). The prediction was performed over all the cells in *2018 Rajewsky Cell Atlas* and *2018 Reddien Cell Atlas* separately, using raw counts as input (see Supplementary Material *Gene co-expression network* section for more information).

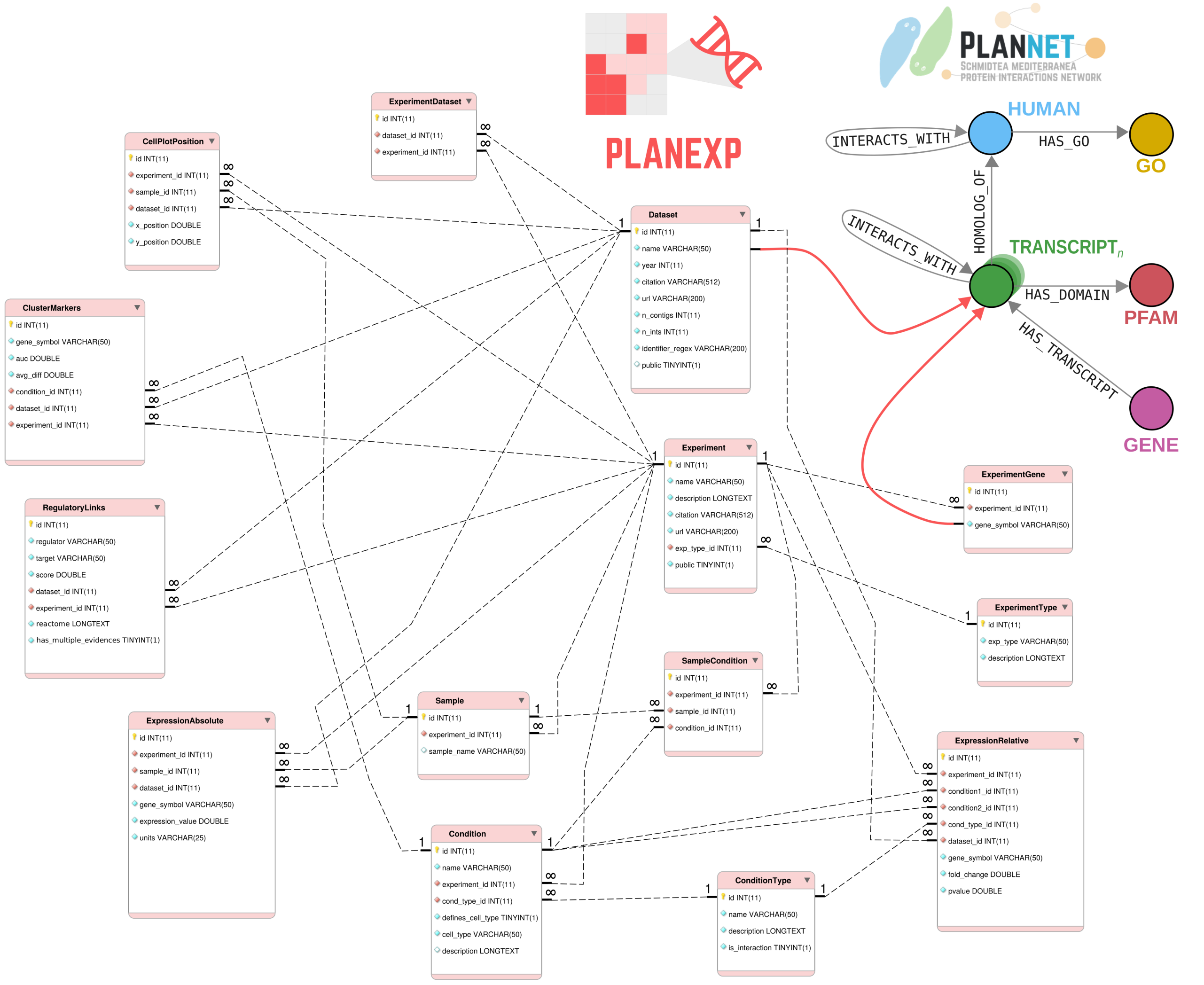


Fig. 1. PlanExp MySQL database schema to store the expression matrices, the differential expression analyses, the sample meta-data, and all the information regarding the experiments. Genes and Datasets available at PlanExp are linked to the Neo4j database of PlanNET by the MySQL columns `gene_symbol` and `dataset`, which allows PlanExp to retrieve sequence annotations, homology information, and protein-protein interactions data. While the core structure is shared for traditional RNA–Seq and scRNA-seq experiments, some tables are exclusive to the latter, namely: *CellPlotPosition*, which stores the coordinates of the cells in the t-SNE dimensions, *ClusterMarkers*, that contains the markers for each cluster computed using a ROC AUC test, and *Regulatory Links*, which stores the predicted genetic interactions by `GRNBoost`.

Regulators for the gene co-expression network inference were selected by Gene Ontology: only those planarian transcripts with a Gene Ontology annotation in PlanNET indicating a possible transcription factor activity were marked as regulators (see Supplementary Table 3).

The interactions were annotated according to the *Homo sapiens* `REACTOME` pathways (Fabregat *et al.*, 2017) that both, regulator and target, belong to. Moreover, those interactions that have been predicted in the two sets, *2018 Rajewsky Cell Atlas* and *2018 Reddien Cell Atlas*, have been tagged as having *"Multiple Evidences"*.

2.4 Database

The PlanExp database, which stores gene expression data and differential gene expression comparisons across different experiments, was implemented in `MySQL, version 5.5.49`. In order to improve the performance of the web-application, and due to the large number of dropouts in scRNA-seq experiments, the database only stores expression values different from zero, and these are inferred by the web-application when responding to the user requests. This decision allowed the larger table in the database (and thus the one with a bigger impact on performance and disk space) to shrink from more than one billion rows to only a few million.

PlanExp retrieves, for all the transcripts in the gene expression experiments, all the available sequence annotations, homology relationships, genomic location and protein-protein interactions from the PlanNET database. This database, which utilizes the graph-based database manager `Neo4j`, now also holds a recently published set of gene predictions made over the *S. mediterranea* genome (Rozanski *et al.*, 2018; Grohme *et al.*, 2018). The ten transcript datasets hosted at PlanNET were also projected over current reference genome sequences by `GMAP` (Wu *et al.*, 2016), to make them also referred through a genome browser driven by `JBrowse` (Buels *et al.*, 2016). Such a design facilitates users to access interactions, sequences, gene cards, expressions, and location from each component of the PlanNET/PlanExp applications.

A summary of the database schema used by both PlanExp and PlanNET and how they are linked together can be found in Figure 1. The whole PlanExp database can be downloaded from https://compgen.bio.ub.edu/PlanNET/downloads.

2.5 Web application

The PlanExp web-application is implemented in `python`, using the web-framework `Django`, which allowed us to take advantage of the object relational mapping (ORM) system of `Django` to easily query the database.

All the plots presented by the applications are generated with the open-source graphing library `Plotly.js` (Plotly Technologies, 2015). Thanks to this technology, all the plots are fully interactive, users can hover on the data, toggle different traces, etc. Additionally, thanks to the fact that `Plotly.js` is built upon the Web Graphics Library API, interactive visualizations with more than 50,000 points (as is the case of the cell-embedding plots for scRNA-seq), are possible without hindering the performance of the whole application.

PlanExp network visualizations are drawn with `cytoscape.js` (Franz *et al.*, 2015). On the other hand, Gene Ontology enrichment analyses are computed by the `python` module `goatools` (Klopfenstein *et al.*, 2018).

3 Results

PlanExp main page is distributed into eight sections, and each one provides a different table or visualization for users to explore any of the currently available RNA-Seq experiments.

A summary of the differential gene expression analyses can be found in Supplementary Table 1; and a full analysis example on a set of genes is detailed in Supplementary Figure 1, along with a description of all the plots available in PlanExp.

3.1 Differential gene expression

The differential expression comparisons for each one of the experiments stored in PlanExp can be accessed through a multiple dropdown selector. Each experiment's factors, and the levels of those factors, are stored in the database and retrieved when an experiment is selected in the application. Then, users can select a comparison of levels (for instance, for a given factor *Section* users may select the comparison *Head vs Tail*), and PlanExp will display a searchable table with all the differentially expressed genes in the chosen comparison, along with an interactive volcano plot.

Thanks to the PlanExp database schema, any combination of factors can be selected in the dropdown menu (for instance, a comparison of *Head 24h vs Tail 24h* is possible, which combines the factors *Section* and *Time* in a given experiment). In that way, the application is not limited by the initial design of the differential expression analyses.

In the displayed differential expression table, users can click on the differentially expressed transcript names. As a result, a gene card with all the available information for that transcript will appear on the screen, without leaving the application (Supplementary Figure 2).

Once a comparison is selected, a gene ontology enrichment analysis (GOEA) can be performed for any of the overexpressed gene sets. The results of the GOEA are displayed as a graph for each of the domains of gene ontology (biological process, molecular function and cellular component), and can be downloaded as a `CSV` file too.

Finally, in the case of scRNA-seq, a new section called "Marker genes" will be available, in which gene markers for each cell type will be displayed.

3.2 Plotting gene expression

The expression of any transcript available for the previously selected experiment can be visualized in the *Gene Expression Plot* section of PlanExp (Figure 2).

PlanExp can generate three types of plots: violin plots, heatmaps, and line plots. All are available at all times and for all experiments, and users can change the plot type for the same data on demand in order to find the most appropiate visualization.

Gene expression can be grouped by factor (or by a combination of factors if available for the experiment), via a dropdown menu. Users can select any factor in a given experiment to group the samples before plotting, and `PlanExp` will retrieve all samples matching the specified levels and compute the average expression (in the case of heatmaps, line plots, and bar plots) or plot the points individually (in the case of violin plots and scatter plots). For instance, in the case of samples collected as time-region combinations (as is the case for “*2013 Aboobaker Time-course*”), users can select to plot the expression of several genes by grouping all samples in “Head” and “Tail” levels, without taking into consideration the time at which these samples were collected. This process is done automatically at run-time by `PlanExp`, and allows users to group the samples by their factors of interest.

Alternatively, and exclusively in the case of scRNA-seq, users can plot the expression levels in each cell individually by using a heatmap where each column corresponds to a single cell and the rows to each gene/transcript.

Expression for multiple transcripts can be plotted simultaneously for all plot types, and transcripts can be searched not only by their transcript identifier, but also by their gene identifier, by their annotated PFAM domain accession, by the corresponding human homologous gene symbol, or by GO accession. PlanExp provides an autocomplete option to aid users in selecting the correct identifier or accession symbol. Thanks to the interactive nature of `Plotly` graphs, transcript traces can be toggled to hide or show them at will, plots are zoomable, they provide information on hover (median and quartiles for violin plots, for instance), they are exportable as tabular files, and can also be saved as `png` images.

3.3 Cell embedding visualization

In the case of scRNA-seq, PlanExp offers a plot of the studied cells in the experiment embedded in two dimensions using t-SNE. Cells can be colored either by condition (cell cluster, experimental condition, etc.) or by specific genes expression. In both cases, cells are always grouped in traces by condition, which allows users to toggle specific cell types (if plotted by cluster), and to hide/show only the desired cells.

When searching for multiple genes at once, users can toggle an option that, instead of showing all cells and coloring them

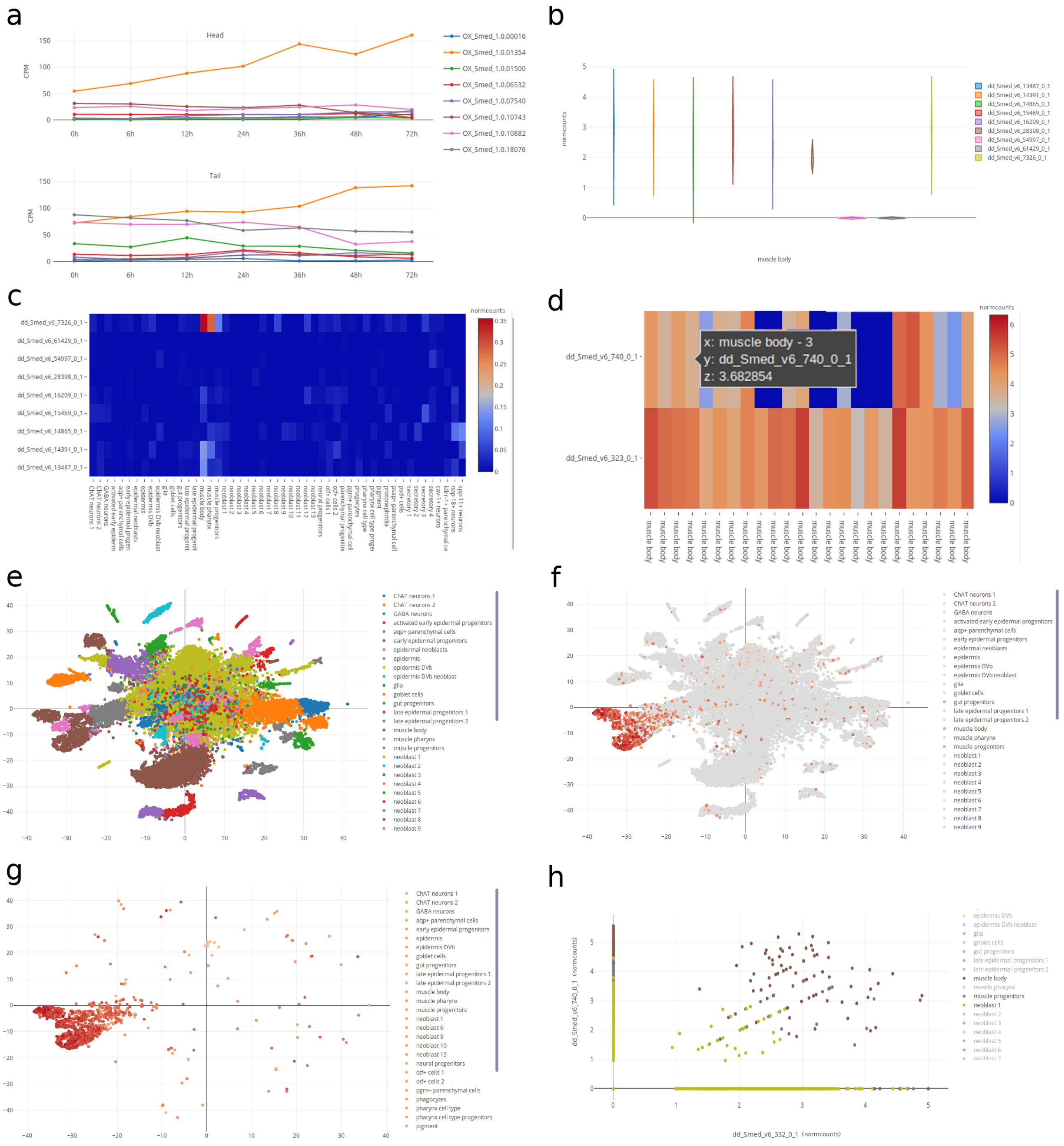


Fig. 2. Figure examples produced by PlanExp. **a.** Line-chart comparing expression of all planarian transcripts in the *Consolidated* dataset homologous to human WNT proteins in the experiment *2013 "Aboobaker Time-course"*. The interaction of the factors *Section* and *Time* was selected to see the changes in expression across time in the studied levels of section *"Head* and *Tail"*. Transcripts can be hidden or shown by clicking on their transcript identifiers on the right legend. **b.** Zoom-in on a violin-plot summarizing the expression of WNT genes in the *Dresden* dataset in a particular cell type of the experiment *"2018 Rajewsky Cell Atlas"*. **c.** Heatmap showing the expression of WNT genes across the different cell clusters in *"2018 Rajewsky Cell Atlas"*. Clicking on transcript identifiers pops up a gene card with information about the transcript (Supplementary Figure 2). **d.** PlanExp facilitates to plot a heatmap of the expression of individual cells. Users can hover on heatmaps to check the expression values directly. **e.** Cell embedding plot of *"2018 Rajewsky Cell Atlas"*, coloring cells by cluster. Groups of cells can be hidden or shown by clicking on cluster names in the legend on the right side, and users can zoom on the visualization too. **f.** Mapping expression of *dd_Smed_v6_740_0_1* on the previous t-SNE plot. **g.** Mapping the expression of multiple genes and coloring cells by the mean expression is possible when selecting a checkbox option on the main application, which will also hide all cells not expressing the searched genes. **h.** Gene co-expression can also be explored through the *Co-expression plot*, which will project each cell expression in a scatterplot where each axis represents the expression value for the selected genes. Users can choose which condition to use in order to group samples (in this case, cell cluster is the selected condition), and then groups of samples can be hidden or shown by taking advantage of the interacting visualization capabilities of `Plotly`.

by the expression of one gene, will only show cells expressing all genes and will color them by the average expression of the searched genes in each cell.

3.4 Gene co-expression plots

PlanExp is able to produce several plots with the aim to identify if two or more genes are being expressed in the same samples simultaneously. The *Gene Co-expression Plot* lets users plot the expression of two genes for scRNA-seq experiments as a scatter-plot, aiding researchers in the identification of co-expressed genes in the different cell-types (thanks to the toggling function that shows and hides cell types), and allowing them to identify specific cells by hovering over the data points.

The *Gene Co-expression Counts* section displays the number of samples that express multiple genes simultaneously. For more information, see Supplementary Figure 1.

3.5 Genetic interactions

An inference of genetic interactions over the two available scRNA-seq experiments in PlanExp was performed using `GRNBoost` of the `SCENIC` pipeline. Both predictions resulted in a similar number of predictions, and the confidence scores over both predictions are moderately correlated (Pearson's r = 0.683, see Supplementary Material “*Gene co-expression network*” section).

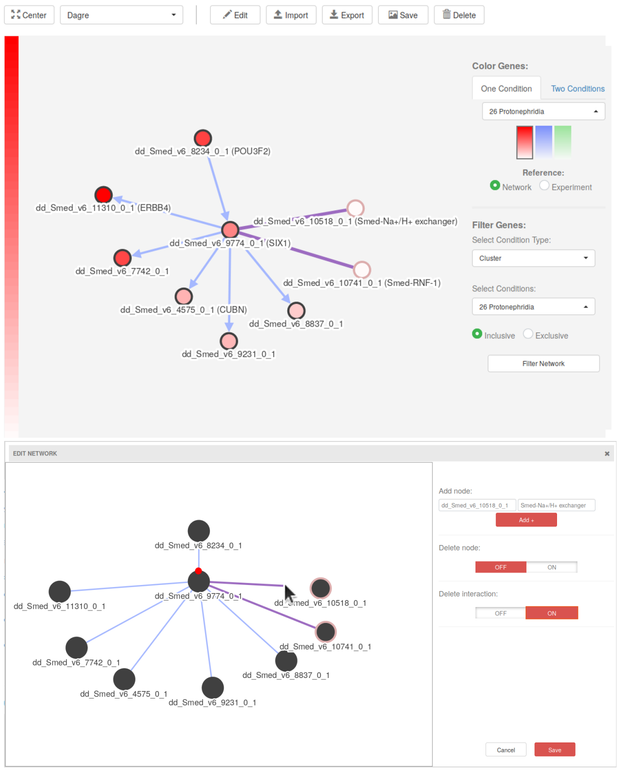
The inferred networks are available for users to explore in the corresponding section of `PlanExp`, where interactions can be retrieved by searching for gene symbols, contig identifiers, PFAM domain accessions, Gene Ontology identifiers, `REACTOME` identifiers, and `REACTOME` pathway names. The networks can also be sent to the network viewer embedded in this tool for further exploration.

3.6 Network editor and expression mapper

The last section of PlanExp provides a way for researchers to filter and visualize expression data for each of the stored experiments over a gene protein interaction network. This network viewer has an import option that allows users to use networks transferred directly from PlanNET (Figure 3). Users can also push a *Net Explorer* graph from PlanNET onto this editor, thanks to the *Export to PlanNET* button.

Gene expression for any of the available conditions in a given experiment can be mapped over the network, coloring the nodes in the graph proportionally to gene expression. Additionally, a differential expression comparison between two conditions can also be mapped, color-shading by log fold change those nodes that are significantly over or under expressed in the comparison taken.

In order to map gene expression data over any set of interacting genes and proteins (even if they are not available on PlanNET), a fully fledged network editor was implemented and is available in PlanExp. This editor allows researchers to add or remove genes to the visualization, as well as to add custom interactions to prototype pathway models. Those edges are marked to distinguish them from interactions already described in PlanNET.



Color Genes: One Condition | Two Conditions

26 Protonephridia

Reference: ● Network ○ Experiment

Filter Genes:

Select Condition Type: Cluster

Select Conditions: 26 Protonephridia

● Inclusive ○ Exclusive

Add node: dd_Smed_v6_10518_0_1 Smed-Na+/H+ exchanger

Delete node: OFF

Delete interaction: ON

Fig. 3. Network expression mapper (top panel): PlanExp facilitates mapping gene expression in a particular experiment condition, or in a comparison of two conditions, onto a network of gene-protein interactions. This network can be filtered by using the controls on the bottom right corner, highlight genes with a similar expression value (left gradient bar) or interact with the visualization in different ways by using the provided graph controls (top buttons). **Network editor** (bottom panel): Alternatively, users can create or edit their own gene-protein networks with the newly implemented network editor, which allows the creation and removal of both nodes and interactions. These networks can then be sent to the network mapper by clicking on the *Save* button.

4 Discussion

In this work, we present PlanExp, a web application that provides a universal integrated framework for exploring and visualizing both traditional RNA-Seq and scRNA-seq data of *S. mediterranea*; offering several `JavaScript`-based interactive visualizations, allowing the immediate retrieval of sequence annotations, protein-protein interactions, and genomic data.

Database scalability and performance. The ability to hold and store a diverse set of expression experiments for planarians allows researchers to plot gene expression across conditions, create tables with differentially expressed genes, and map expression data onto a gene network by using the same web-application, independently of the experiment of interest, and thus, the biological question at hand. Additionally, thanks to its database design and its scalability, PlanExp will be able to store many more gene expression experiments performed in the future with ease, independently of the experimental design of such experiments. With the pace at which new *S. mediterranea* expression data is being generated, and the increasing amount of space needed to store that information (especially with the advent of scRNA-seq technologies), the development of a new way to store all that information together, and to be able to retrieve data in a timely manner as to be useful for end users, was necessary. PlanExp stores expression data for 72,560 samples in total, mapped over five transcriptome datasets with an average transcript count of 22,456 per sample, and is able to retrieve expression values for all samples in a given condition for over a hundred genes in just a few seconds.

Dynamic visualizations. PlanExp provides new ways to explore already published RNA-seq data in the form of several plots and tables. Contrary to most other visualizations available elsewhere for this type of data, PlanExp bases all of them in `Plotly`'s graphing library, which in turn is built on top of `D3`. This technology allows the exploration of the plots by the end user in ways that static images cannot provide, enabling the ability to zoom, remove traces (such as hiding clusters in a t-SNE plot, or showing only some genes in a line plot), hovering on the visualization to check expression values, and even exporting the data as a tabular file for futher study. All this exploratory data analysis is made at the client level, freeing the server of recomputing each plot for each new request, and allowing a faster and more responsive modification of such plots by users.

Linking expression to all omics data. By including previously published traditional RNA-seq experiments on PlanExp, we have provided a new way to explore them that was not possible before, significantly reducing the barrier of entry to the retrieval of useful data from the current planarian knowledge base. Additionally, PlanExp brings the current genomic and transcriptomic annotations directly onto the tables and visualizations: with homologous human genes and transcript-gene relationships available in the differential gene expression tables; genomic locations, protein-protein interactions, and PFAM domains available from all tables and on heatmaps; and with the ability to search by gene identifiers, human genes, PFAM domains, or GO accessions in a given experiment condition to produce a variety of plots. Thanks to the ability of the PlanNET database to link previously assembled trascriptomic datasets with planarian genes, human homologs, PFAM domains and GO accessions, PlanExp provides a way to explore gene expression data based on any dataset using a variety of identifiers, ensuring that this expression data will be searchable and useful for researchers in the future to come. The addition of new sequence annotations to the genome over time will be brought to the experiments uploaded to PlanExp, as reserachers will be able to access all sequence annotations from gene information cards, and search for gene expression by gene names once new ones are annotated onto the new genome assembly.

Previous tools. While some tools have been developed in order to explore scRNA-seq data for planarians[1], PlanExp offers some advantages to researchers compared to them. First, all the visualization options in PlanExp are interactive, allowing users to explore the data more easily, by toggling gene traces, removing and adding cell types to a visualization, zooming on the plots, etc. all without waiting for a response from the server. Second, several visualizations and options, such as the network expression mapper, the co-expression plots, or the interactive heatmaps are exclusive to PlanExp. And finally, as stated previously, all the transcripts are linked to their sequence annotations in PlanNET, providing an easier way to navigate through expression data and sequence features at the same time. PlanMine[2] offers a centralized and mineable resource storing genomic, transcriptomic, and phylogenetic data, and while it provides plenty of gene expression information, the focus of `PlanExp` on the experiment instead of the gene/transcript entry allows for different visualization options and features. A comparison of PlanExp features and other applications to explore RNA-seq data in planarians can be found in Supplementary Table 4.

Network biology. Linking network biology to the gene expression data available for *S. mediterranea* is necessary in order to understand the dynamics of processes such as regeneration. PlanExp is the first online tool that allows researchers to create gene-protein network visualizations—either from currently available knowledge in the literature, from experts hand-curated networks, from predictions already available in PlanNET, or from other *in silico* predictions (Lobo and Levin, 2015)—, and then map expression data for any of the six uploaded experiments already in PlanExp. Thanks to the large number of samples available in scRNA-seq experiments, we have been able to use a gene regulatory network inference tool to predict genetic interactions for planarians.

Although `GRNBoost` was conceived as a program to retrieve genetic interactions from expression matrices, it only constitutes the first step of the `SCENIC` pipeline to retrieve genetic interactions. Subsequent steps include the filtering of the gene co-expression network by using information of cis-regulatory elements. Unfortunately, due to the novelty of the most current *S. mediterranea* genome assembly, such data is not available yet. Therefore, the predicted co-expression networks contain other types of relationships between genes apart from genetic regulatory interactions, such as protein-protein interactions, genes that belong to the same signalling pathways, or genes expressed in the same cell types (see Supplementary Material, *Gene co-expression network* section). Nevertheless, these co-expression networks are available for download so that researchers may use them for further filtering (either by following the whole SCENIC pipeline or by implementing their own filtering protocol). Given that the co-expression network prediction is the most computationally expensive step of the `SCENIC` pipeline, releasing this prediction—so that researchers may be able to implement other filtering protocols or validate the genetic interactions individually—can help the planarian community to explore and annotate many important genetic interactions.

To highlight the usefulness of our approach, we looked at a recently published experiment where an RNA-seq was undertaken comparing *soxB1-2*(RNAi) planarians to controls (Ross *et al.*, 2018). Of the 86 genetic interaction targets predicted for *soxB1-2* in `PlanExp`, 23 were shown to be differentially expressed in *soxB1-2* knockdown planarians. Three of these genes (*pkd2-1*, *pkd2-4*, and *eml-1*) presented movement and sensory defects after RNAi inhibition (Supplementary Table 6). Thus, the predicted gene co-expression relationships contain many described genetic interactions in *S. mediterranea*, and may prove a way to uncover novel regulatory links.

These co-expression interactions may also reveal unknown key elements in important planarian signalling pathways. For instance, in Supplementary Table 5, we can find relationships

[1] https://digiworm.wi.mit.edu/ and https://shiny.mdc-berlin.de/psca/

[2] http://planmine.mpi-cbg.de/planmine

between elements that belong to the *Wnt* signalling pathway as annotated in the KEGG database (hsa04310), such as *wnt11-2*, *axinA*, *wntP-2*, *sfrp1*, among others. These predictions could be used to find other unknown elements that may play a role specifically in the *Wnt* signalling pathway, or in other pathways relevant for the planarian biology.

Finally, the exploration of the gene co-expression interactions is facilitated by their annotation with the REACTOME pathways in which they appear. It is also possible for users to check which interactions were inferred independently in the prediction over the experiments *"2018 Rajewsky Cell Atlas"* and *"2018 Reddien Cell Atlas"*, restricting the search of possible edges and potentially reducing false positives. Finally, these predictions have also been integrated into the application; thus, they can be sent to the network expression mapper and editor for a deeper analysis.

5 Conclusions

Providing a tool to link expression data with the current knowledge of the different omics fields (genomics, transcriptomics and interactomics) offers researchers a unique and universal way to explore gene expression experiments performed for planarians. The large array of dynamic visualizations, searchable tables, and downloadable data that PlanExp offers will hopefully help researchers to understand the complex biology of *S. mediterranea*. Last but not least important, the application's current design will facilitate the integration of future transcriptomic datasets under such a unified interface. In that sense, researchers can send their data to be uploaded to PlanExp at:

https://compgen.bio.ub.edu/PlanNET/send_to_planexp.

Acknowledgements

We are grateful to Emili Saló for his support and discussions. We would also like to thank the reviewers for their thorough and constructive comments.

Funding

This work was supported by research grants from Spanish Ministry of Economy (BFU2017-83755P), and from Generalitat de Catalunya (2017SGR1455). Sergio Castillo-Lara is fellow of the Catalan Government "AGAUR" (FI-FDR: 2017FI_B_00191). Eudald Pascual-Carreras is the recipient of an FPI (*Formación del Profesorado Investigador*) scholarship from the Spanish *Ministerio de Educación y Ciencia* (MEC).

References

Aibar, S. *et al.* (2017). SCENIC : Single-cell regulatory network inference and clustering. *Nature Methods*, **14**(11), 1083–1086.

Buels, R. *et al.* (2016). JBrowse: A dynamic web platform for genome visualization and analysis. *Genome Biology*, **17**(1), 1–12.

Butler, A. *et al.* (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, **36**, 411–420.

Castillo-Lara, S. and Abril, J. F. (2018). PlanNET: homology-based predicted interactome for multiple planarian transcriptomes. *Bioinformatics*, **34**(6), 1016–1023.

Davie, K. *et al.* (2018). A single-cell transcriptome atlas of the aging *Drosophila* brain. *Cell*, **174**(4), 982 – 998.e20.

Fabregat, A. *et al.* (2017). The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, **46**(D1), D649–D655.

Fincher, C. T. *et al.* (2018). Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science*, **360**(6391).

Franz, M. *et al.* (2015). Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics*, **32**(2), 309–311.

Grohme, M. A. *et al.* (2018). The genome of *Schmidtea mediterranea* and the evolution of core cellular mechanisms. *Nature*, **554**(7690), 56–61.

Huynh-Thu, V. A. *et al.* (2010). Inferring regulatory networks from expression data using tree-based methods. *Plos One*, **5**(9), 1–10.

Kao, D. *et al.* (2013). The planarian regeneration transcriptome reveals a shared but temporally shifted regulatory program between opposing head and tail scenarios. *BMC Genomics*, **14**(1), 797.

Klopfenstein, D. V. *et al.* (2018). Goatools: A python library for gene ontology analyses. *Scientific Reports*, **8**(1), 10872.

Labbé, R. M. *et al.* (2012). A comparative transcriptomic analysis reveals conserved features of stem cell pluripotency in planarians and mammals. *Stem Cells*, **30**(8), 1734–1745.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, **15**(2).

Lobo, D. and Levin, M. (2015). Inferring Regulatory Networks from Experimental Morphological Phenotypes: A Computational Method Reverse-Engineers Planarian Regeneration. *PLOS Computational Biology*, **11**(6), e1004295.

Papili Gao, N., Ud-Dean, S. M. M., Gandrillon, O., and Gunawan, R. (2017). SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, **34**(2), 258–266.

Plass, M. *et al.* (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, **360**(6391).

Plotly Technologies (2015). Collaborative data science. Montreal, QC. https://plot.ly.

Potier, D. *et al.* (2014). Mapping gene regulatory networks in *Drosophila* eye development by large-scale transcriptome perturbations and motif inference. *Cell Reports*, **9**(6), 2290 – 2303.

Rodríguez-Esteban, G. *et al.* (2015). Digital gene expression approach over multiple RNA-Seq data sets to detect neoblast transcriptional changes in *Schmidtea mediterranea*. *BMC Genomics*, **16**(1), 361.

Ross, K. G. *et al.* (2018). SoxB1 Activity Regulates Sensory Neuron Regeneration, Maintenance, and Function in Planarians. *Developmental Cell*, **47**(3), 331–347.e5.

Rozanski, A. *et al.* (2018). PlanMine 3.0 - improvements to a mineable resource of flatworm biology and biodiversity. *Nucleic Acids Res.*, **47**(D1), D812–D820.

Sandmann, T. *et al.* (2011). The head-regeneration transcriptome of the planarian *Schmidtea mediterranea*. *Genome Biology*, **12**(8), R76.

Taylor-Teeples, M. *et al.* (2015). An *Arabidopsis* gene regulatory network for secondary cell wall synthesis. *Nature*, **517**(7536), 571–575.

Wu, T. D. *et al.* (2016). *GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality*. Springer New York, New York, NY.

General comments

When testing the last update of the web-application, please make sure that you properly clear the browser's cookies and cache, as you may have an older `JavaScript` version of the libraries. This will guarantee using the last version of `PlanExp` code.

Summary of changes

A. Web application

1. Loading spinners added to all sections.
2. New *Gene Co-expression counts* section that allows users to count how many samples express multiple genes at the same time.
3. Re-ordering PlanExp sections to a more logical order.
4. Gene symbol searches are now case-insensitive (e.g.: *Wnt1* search returns same results as WNT1).
5. Gene symbol searches now work with SMESG symbols (and not only Human homolog gene symbols).
6. Implemented PFAM symbol searches (e.g.: searching for ”dead” will return the same results as searching for ”PF00270”).
7. Plots now display units when appropiate.
8. New *Send to PlanExp* page to upload gene expression datasets to PlanExp.
9. Added more clarifications to *Gene co-expression network* section (see gene network prediction changes).
10. Network viewer now displays Human homolog gene symbols on node labels.
11. Gene Co-expression network table is now fully searchable: users can search by gene symbols/PFAM domains/GO codes/REACTOME pathway names and get the predicted interactions for those genes.
12. Send to network button re-implemented and now sends only displayed gene co-expression interactions in table, allowing researchers to choose which interactions to send to network viewer.
13. *Condition Types* without differential gene expression data won't appear in the dropdown menu of the DGE section. For instance, for *2013 Aboobaker Time-course*, only the *Condition Type* "*Hour – Region*" will appear.
14. New "*Send to PlanExp*" button in PlanNET `NetExplorer`, that allows users to send PlanNET protein networks to PlanExp Network Viewer and Expression Mapper automatically, improving the integration of PlanNET and PlanExp.
15. Explanation of PlanExp plots added to Supplementary Material.
16. Small bug fixes and usability improvements.

B. Gene network prediction

1. Prediction performed with `GRNBoost` of the `SCENIC` pipeline as opposed to `GENIE3`.

2. Prediction performed over all cells in both experiments: *2018 Rajewsky Cell Atlas* and *2018 Reddien Cell Atlas*.

3. Putative transcription factors retrieved by Gene Ontology only, as opposed to Gene Ontology and PFAM domains.

4. Implemented filters described in `SCENIC` pipeline to improve the accuracy of the results and to allow future re-use of the predictions.

5. Comparison of results for both predictions: confidence score correlation and direct comparison of filtered interactions results are available in Supplementary Material. Annotation of gene links retrieved by both predictions simultaneously, to aid manual curation by researchers.

6. Renamed *Gene Regulatory Network* to *Gene Co-expression Network*, following the nomenclature used in *Aibar et al. 2017*, to avoid any confusion. Explanation of what the predicted gene links mean is available both in the corresponding section of PlanExp and in the Discussion section of the manuscript.

7. Comparison of the results predicted for a particular gene (*soxB1-2*) to candidate target genes retrieved from a recently published RNA-seq experiment (*Ross et al. 2018*). Intersection of predicted target genes available in Supplementary Material and discussed in Discussion section of the manuscript.

8. Retrieval of genes belonging to the *Wnt signalling pathway* from the network prediction, showing the usefulness of PlanExp to retrieve genes involved in the same biological processes.

9. Annotation of gene links with REACTOME pathway identifiers. Summary of results available in Supplementary Material.

10. More thorough explanation of the protocol added to Supplementary Material.

11. Complete protocol to predict gene co-expression interactions available as a markdown file, along with all the files needed to reproduce our results. Gene co-expression interactions also available for download on github site.

C. RNA-seq integration

1. Differential gene expression for *2013 Aboobaker time-course* re-computed with the `limma-trend` pipeline, due to our inability to obtain raw expression counts.

2. Differential gene expression for *2012 Pearson Stem Cells* removed due to the lack of replicates. The experiment is still accessible for exploration in the application, only the D.G.E. section is unavailable.

3. Re-named some *Condition Types* from "Section" to "Region", as per Reviewer's 1 suggestion.

4. Planarian regions re-ordered to follow the animal's anatomical order.

5. New section in Supplementary Material detailing important parameters for all computed Differential gene expression analyses.

6. All protocols and all the necessary files to reproduce the analyses are now available on our github site.

Response to reviewers

Reviewer 1

This paper presents PlanExp, a web application for centralizing and visualizing planaria RNAseq data. This is an extension of their previously published PlanNet web application for protein-protein interaction networks in planaria. The data included in the web application are two scRNAseq and four RNAseq datasets from the literature, together with further computational predictions of gene interactions derived from these datasets. The web interface is polished and easy to use. These centralized transcriptomics datasets and query tools clearly would be useful for the planarian community. However, I have some major and minor concerns as detailed below.

We would like to thank Reviewer 1 for his/her kind comments regarding our work and for his/her insightful comments. We have made an effort to adress all his/her concerns, which are listed below along with our responses.

1. The 6 datasets centralized in this resource seem to be treated in isolation. Indeed, the user needs to select which dataset to use before running any query. As such, the main advantage of having a centralized repository, that is to run meta-searches across all datasets, is not available in this web application. It doesnt seem that this capability would be very difficult to implement, and the usability of the application would increase considerably.

Indeed, users need to select a given experiment before having access to all the visualization and summarizing tools. While, the proposed functionality would be clearly useful to the planarian community, implementing it in our application would require an extensive rewrite of both the back-end and the front-end code of the application. All queries, templates, and plots expect all data to refer to a *single* experiment, and chaning them all is not feasible for us at the moment. Additionally, when designing such a tool, we would have to ensure that all expression values are comparable across experiments, and we would have to design or implement a cross-experiment normalization protocol. Finally, finding a way to map expression data from multiple experiments onto a gene-protein network, which is one of the main functionalities of `PlanExp`, can be challenging.

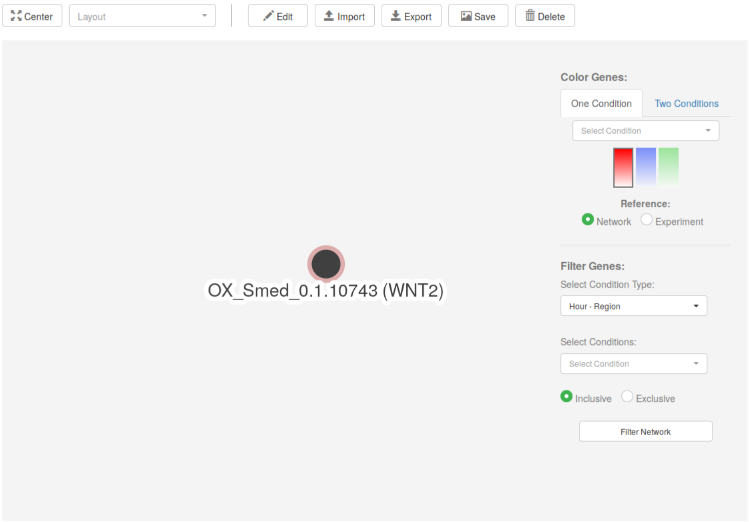
However, we will consider implementing such an extension in the future to allow for meta-searches across multiple experiments by re-using the same database for PlanExp. The requested functionality is too complex and we consider it out of the scope of PlanExp, as its focus is always a particular experiment.

2. For the most part, the genes are identified and searched in the application with their specific (and cryptic) symbol used in the dataset (such as OX_Smed_1.0.10743), instead of the common name for the gene (such as Wnt2). This makes the application difficult to use and obscures the interpretation of the graphs and their usability. See the examples in Fig. 2; if the common gene names were displayed instead (or in addition) of the cryptic gene symbols, it would be much more informative and easier to understand. The same critique can be made for the interaction networks (Fig. 3), which currently shows a network of cryptic names.

While all genes are identified with their specific symbols (such as "OX_Smed_1.0.10743"), all plots search forms and all the dynamic tables allow for searching using common gene symbols (e.g.: *WNT2*), PFAM domains (e.g.: *PF11262*) and GO identifiers (e.g.: *GO:0030182*).

While reviewing this question, we found a bug in our code that made gene names searches to be case-sensitive (i.e.: searching *Wnt2* would not return results while searching for *WNT2* would). We have now fixed this bug, and all gene name searches are now case-insensitive. The autocomplete functionality of all gene search bars now also works in a case-insensitive manner.

Additionally, we have changed the network visualization to display the homologous "*official*" gene symbol of each node in parenthesis. These symbols appear even when creating the network from scratch in the Network Editor, but also when importing networks from PlanNET. We have chosen to use the homologous gene symbol to allow full compatibility with PlanNET networks (see question 4), although we plan on adding the ability to switch to planarian gene names (as named by *Grohme et al. 2018*) by using a radio button on top of the visualization.



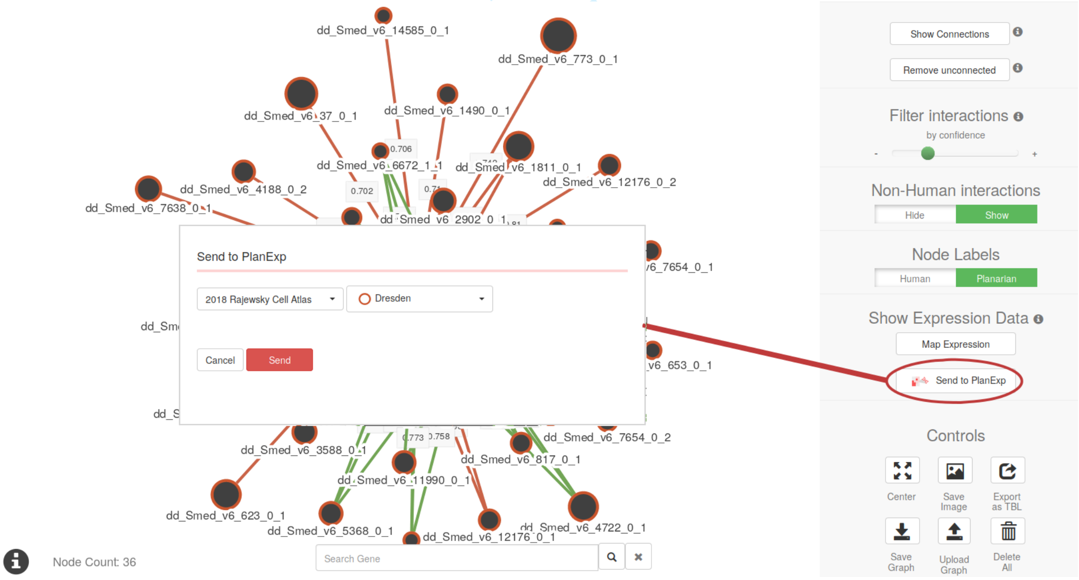
For the other visualizations, such as the Heatmaps in Figure 2, formatting issues with the plotting library `Plotly` make it difficult to display both the cryptic transcript ID and the gene name at the same time due to the length of the labels. Displaying only the gene name is unfortunately not an option, as we would not want to have repeated labels in the rows of the heatmap, for instance. However, in the case of these visualizations (contrary to the network where we now display the gene name), users can click on the transcript IDs to get a PlanNET information card with all the relevant information for a particular transcript.

3. For the prediction of genetic interactions using GENIE3, random cell samples for each cluster were used. How variable are the predicted interactions when running the algorithm again with a different random sample? A statistical analysis confirming the validity of this approach would be helpful.

We have now changed the genetic interaction inference protocol (see Summary of Changes above). Currently, we don't use any random subset of cells anymore, and instead use **<u>all cells</u>** for both predictions. This new approach makes unnecessary to subsample the datasets and bootstrap to assess any variance due to sampling bias. For more information, refer to the Supplementary Material "*Gene co-expression network*" section, where we also compare the predictions between the experiments *2018 Rajewsky Cell Atlas* and *2018 Reddien Cell Atlas*.

4. Even if this new PlanExp web application is included in the portal for planNET, it doesnt seem that there are much linking or interactions between them. For example, in order to visualize a network, a user needs to go to PlanNET and download a file to the computer, and then go to PlanExp and upload such a file there. That seems to be an unnecessary burden, and a better integration between the two applications would be helpful. Also, the button "Send to Network viewer doesnt seem to be available as far as I can tell.

We agree with Reviewer 1 that this initial approach was not integrative enough as it required too much user-dependant steps. We have now implemented a way to send PlanNET networks directly to PlanExp, and this feature is currently available on the application. Users are able to click on a button called *Send to PlanExp* which will reveal a pop-up menu to select an Experiment and a Dataset. Upon acceptance, a new tab will be opened with the selected PlanExp parameters and the PlanNET network ready for exploring in the PlanExp network viewer and expression mapper. Users will also be able to edit this network with the built-in network editor in PlanExp.



Send to PlanExp

2018 Rajewsky Cell Atlas

Dresden

Finally, a bug with the JavaScript `DataTables` library prevented the Send to Network button to appear under some circumstances. We have now changed the button implementation to ensure it is always going to appear correctly.

5. After selecting the options for generating a graph (such as the volcano plot), there is no indication that there is a computation running in the background, and instead the interface doesnt change for a few seconds, which is confusing. A processing message would be helpful here.

Some loading gifs were missing in the application. We have now added them, which should improve the user's experience, avoiding any confusion as to if a process is running or not.

6. When plotting gene expressions for worm sections, it would be more useful if the order of the sections follows the anatomical order of the worm. Also, instead of section it may be more appropriate to call them region, since the head or the tail are regions, not sections, of the worm.

We have re-ordered the levels of the ”Section” factor in all experiments to more closely follow the anatomical order of the worm. We have also re-named the ”Section” factor to ”Region”.

7. There are some grammatical mistakes and typos along the paper, which would benefit from some copy editing. Here are some examples, just from the introduction:

- ***Fincher and others., 2018, − > Fincher et al., 2018***
- ***as each cell can be considered one − > one sample?***
- ***exponentially; with − > exponentially and with (there is no verb in the second clause).***
- ***as type of data necessiates − > as this type of data necessitates***
- ***researchers; and will − > researchers and will (no semicolon)***

We have fixed all the errors listed by Reviewer 1, and we have checked for more mistakes elsewhere in the manuscript. All changes made to the text of the manuscript are colored in red.

Reviewer 2

In this manuscript, the authors have developed an open source web application, PlanExp, to explore RNAseq expression data from planarian, a tissue regeneration model organism. Currently, the different publicly available planarian RNAseq datasets are linked to their own de-novo transcriptome assemblies and annotations, and connections between the available planarian gene expression data and up-to-date reference genome and transcriptome annotations are missing. The authors propose to provide such an integrative tool with PlanExp. PlanExp has been designed to link published planarian RNAseq datasets with the currently available planarian genome annotations through another web application previously developed and published by the authors, PlanNet. PlanNet is a predictive planarian protein interactome application that is based on an interolog approach. PlanExp has been built using several packages described in the methods section. The authors used 6 available RNAseq experiments including large scale scRNA-seq files from different publications, some of which had to be re-computed by the authors, and the recent versions of the planarian reference transcriptome and planarian reference genome that have been published in 2018. The PlanExp database can be upgraded and integrate novel RNAseq experiments and up-to-date genome annotations as they become available. Amongst other interesting features provided by this application are the possibility to have access to gene cards with several search options, the expression mapper and network editor, and the dynamic visualization tool that should contribute to help researchers to explore the planarian transcriptome. With PlanExp, the authors aim to provide to the scientific community, especially working on planarians and on tissue regeneration, an upgradable tool that gathers all the available planarian expression data and sequence annotations, thus contributing in building an unified "omics" database for this model organism, a valuable application. Below is a list of specific comments on the manuscript that should be addressed by the authors prior to publication.

We want to thank Reviewer 2 for his/her comments and suggestions. We are grateful on his confidence on this tool as a valuable resource for contributing to an unified omics database for the planarian community. Below, we attempt to address his/her concerns.

A) The critical issue of dealing with different sequence annotations across several publications and the current planarian reference transcriptome and genome has be addressed more clearly by the authors.

We map all transcriptome sequence sets against a reference set, so we can compare their equivalent identifiers for a given transcript across experiments. The reference `PlanNET` network was already described. Each transcriptome defines a layer in the database linking each node/transcript to the homologous human genes, prior to the construction of the interologs network. As such, users have access to all genomic annotations stored in the `PlanNET` database through the `neo4j` connections, keeping them separate from the main `PlanExp` database (see Figure 1).

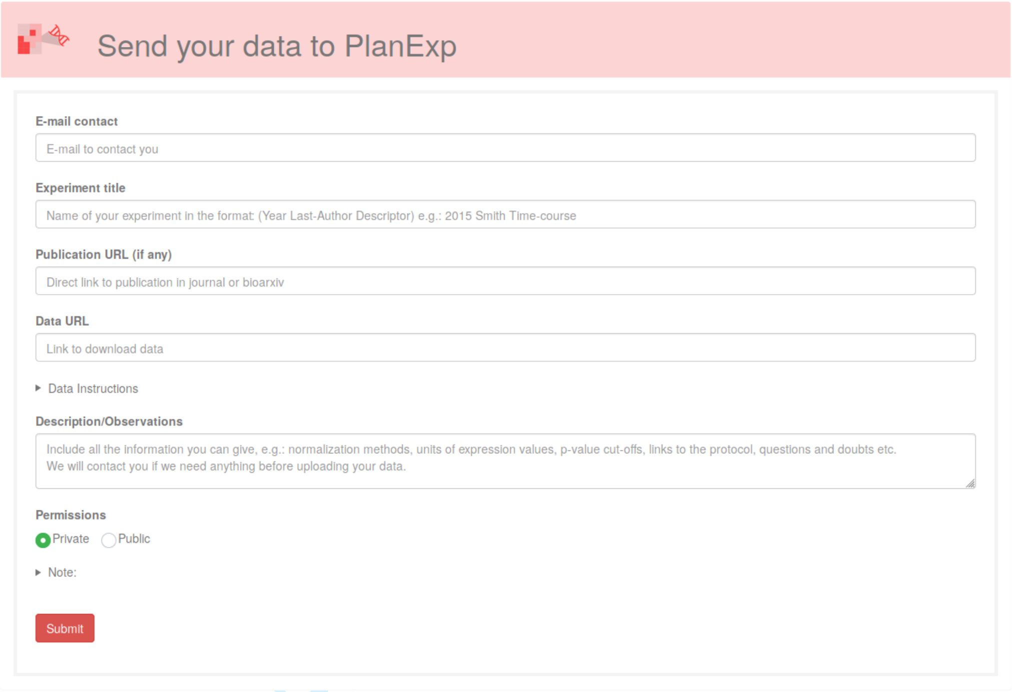
B) The link that the authors provide between expression data and evolving genomic annotation has to be clarified, and it should be explained how they will implement the application with novel annotation and novel datasets, and what end users will be able to do in this respect.

We have added some clarifications to the Discussion section of the manuscript (new text added in red):

> ***Linking expression to all omics data.*** *By including previously published traditional RNA-seq experiments on PlanExp, we have provided a new way to explore them that was not possible before, significantly reducing the barrier of entry to the retrieval of useful data from the current planarian knowledge base. Additionally, PlanExp brings the current genomic and transcriptomic annotations directly onto the tables and visualizations: with homologous human genes and transcript-gene relationships available in the differential gene expression tables; genomic locations, protein-protein interactions, and PFAM domains available from all tables and on heatmaps; and with the ability to search by gene identifiers, human genes, PFAM domains, or GO accessions in a given experiment condition to produce a variety of plots. Thanks to the capability of PlanNET database to link previously assembled trascriptomic datasets with planarian genes, human homologs, PFAM domains and GO accessions, PlanExp provides a way to explore gene expression data based on any dataset using a variety of identifiers, ensuring that this expression data will be searchable and useful for researchers in the future to come. The addition of new sequence annotations to the genome over time will be brought to the experiments uploaded to PlanExp, as reserachers will be able to access all sequence annotations from gene information cards, and search for gene expression by gene names once new ones are annotated onto the new genome assembly*

Genome annotations will be added once they are published by researchers and made available to the public. By keeping these annotations separate from the expression data in the database (see Figure 1), we can keep adding new tracks to the genome browser, new homologs, new PFAM domains, etc. and all of these will be instantly available in `PlanExp` gene cards, and instantly searchable in all `PlanExp` forms.

We intend for `PlanExp` to keep growing, both in features and in available datasets. To that end, we have implemented a new page (https://compgen.bio.ub.edu/PlanNET/send-_to_planexp) where researchers can send their data to `PlanExp`. Users have the ability to keep their data private to their own user or to make it public for all `PlanExp` users. We have added this link to the main manuscript to encourage readers to send their data, and a new link in the main `PlanExp` page has been added. We are open to collaborate with those groups that may produce novel datasets that may require further curation before uploading to `PlanExp`.



Send your data to PlanExp

E-mail contact

E-mail to contact you

Experiment title

Name of your experiment in the format: (Year Last-Author Descriptor) e.g.: 2015 Smith Time-course

Publication URL (if any)

Direct link to publication in journal or bioarxiv

Data URL

Link to download data

▸ Data Instructions

Description/Observations

Include all the information you can give, e.g.: normalization methods, units of expression values, p-value cut-offs, links to the protocol, questions and doubts etc.
We will contact you if we need anything before uploading your data.

Permissions

● Private ○ Public

▸ Note:

Submit

C) PlanExp is coupled to the PlanNet application which was designed to display an interactome for Transcription Factors. It is unclear how the TF-oriented focus of PlanNet relates to the larger scope of the PlanExp application.

The focus of PlanNET was to build a network of interologous interactomes that facilitated analyses across different transcriptomes. The sequence annotations stored in PlanExp were used to select the transcription factors for the prediction of genetic interactions, but PlanNET is not focused on transcription factors.

We have added a section in Supplementary material listing the new functionalities of PlanExp with respect to PlanNET, to contextualize the added functionalities presented in this manuscript.

D) In the Methods section it should be discussed how the different computation of sequencing data could impact the results.

It is not the scope of the manuscript to compare different sequencing technologies and the corresponding analyses that have been uploaded to PlanExp. We can just state the obvious, that single-cell experiments require more resources than D.G.E. on RNA-seq, especially when processing the new sequencing data, but also the depth and breadth of the analyses might differ. The PlanExp database contains the final processed results of those analyses, in order to navigate through the data and extract possible biological interpretations.

E) In the Discussion, the authors should emphasize further the PlanExp (and PlanNet) added value and improvements brought over existing planarian datamining tools such as PlanMine for example. A comparison of what the different web applications offer and how they complement each other could be further exposed (as a summary of supplementary table 4).

The section *Previous tools* of the Discussion section on the main manuscript offers a brief summary of the features that PlanExp offers compared to other single-cell planarian resources. We have added a sentence mentioning how the approach of PlanExp is different from that of PlanMine. Users can then refer to Supplementary Table 4 for further information about the specific features now mentioned in the Discussion.

PlanMine offers a centralized and mineable resource storing genomic, transcriptomic, and *phylogenetic data, and while it provides plenty of gene expression information, the focus of `PlanExp` on the experiment instead of the gene/transcript entry allows for different visualization options and features.*

A stated in the answer to question C, we have added a section to Supplemnentary Material highlighting the new improvements brought to `PlanNET` by `PlanExp`. Hopefully all these changes will help researchers asssess what these tools offer, and how `PlanExp` can complement them.

F) In the Results sections, 5 out of 7 sections of the application are described. A sentence concerning the two that are not mentioned would be useful.

All eight (previously seven) sections of the application are now explained in their corresponding section in the main manuscript.

The sections *Differential Expression* and *Marker genes* are detailed in the Results section *Differential gene expression*, the *Gene Expression Plot* section is explained in *Plotting gene expression*, the *t-SNE Plot* section is explained in *Cell embedding visualization*, the *Gene co-expression Plot* and *Gene co-expression counts* are detailed in the new section *Gene co-expression plot*, the *Gene Co-expression network* is explained in *Genetic Interactions*, and finally, the *Network Viewer* section is explained in *Network editor and expression mapper.*

All plots are also now explained in detail in Supplementary Figure 1.

2) The text should be checked for langage and editing issues (underlined). Some specific points are listed below.

a) In the first sentences of the Introduction, redundant words could be avoided for a better sentence flow. In some instances, synonyms could be prefered (for example "difficult, challenging" instead of "hard").

b) In p. 2 of the Introduction, the authors should explain a in short sentence why PlanMine and PlanNet are limiting for scRNA-seq visualization.

c) The sentence "These technologies have the specific characteristic..." should be rephrased. "as each cell can be considered one (sample)" adds confusion to the sentence.

d) The sentence "However these tools offer varying levels of features..." is unclear (varying or various levels of analysis features?).

e) The sentence ".. aim of PlanExp is to provide such tool and bridge this gap and to provide this missing tool.." should be rephrased within the context of the paragraph.

f) In the Discussion, in "Network Biology", the sentence "As such, linking network biology to the static nature of gene expression .." is unclear. The use of "static" in the sentence is unclear given the fact that gene expression is dynamic, by essence.

We have thoroughly checked all the text to correct editing and spelling mistakes. We would want to thank Reviewer 2 for annotating them directly on the text and detailing some of them here. All changes made to the manuscript are colored in red.

Reviewer 3

In this manuscript, Castillo-Lara and Abril extend their PlanNET database (Castillo-Lara and Abril, Bioinformatics, 2018), which is an effort to unify the emerging wealth of genomic and transcriptomic data for the planarian Schmidtea mediterranea, a prominent animal model of regeneration. In the current paper, the authors introduce PlanEXP, which adds tools for data visualization and analysis for six recent RNA-Seq experiments: two time courses of regeneration, two studies of gene expression in sorted neoblasts and their progeny, and two single cell RNA-Seq studies. The authors also (for the first time to this reviewer's knowledge) attempt to define gene regulatory networks (GRNs)/regulator-target interactions based on co-expression of transcripts with predicted transcription factors in two single cell data sets using GENIE3, and utilize the interaction map visualization tool (an implementation of Cytoscape) already available in PlanNET as a way to import and explore these GRNs.

With a few suggestions for minor improvements noted below, the tools for visualizing RNA-Seq data are implemented well, and many planarian researchers will find the ability to quickly analyze pairwise comparisons and generate volcano plots and other visualizations useful. Hopefully the authors will continue to add additional data sets, as the six chosen are only a fraction of what has now been published in this field. Non-planarian researchers may also find these tools useful, as effort has been made to attempt to identify/associate planarian transcripts with human orthologs, which can be searched in the gene expression analysis and are clickable in the "gene cards" that are hyperlinked throughout PlanEXP (and PlanNET). In addition, the versatility of plotly tools makes these sections quite functional, and the tutorial on the PlanEXP web site is good.

We would like to thank Reviewer 3 for his/her thorough and helpful suggestions and concerns, especially those referring to the regulatory network inference, which have encouraged us to improve and expand significantly. Thanks to his/her suggestions, we have changed the whole protocol for inferring genetic interactions. For a brief list of changes, refer to the *Summary of changes* section of this document. Below we have a detailed answer to Reviewer's 3 concerns, where we try to explain the changes we have made to both the protocol and the manuscript, and how we think these changes relate to dissipate his/her concerns.

However, the implementation of GENIE3 and initial attempt to identify GRNs seems preliminary (although intriguing), and of questionable value due to (a) incomplete description of methods used and poor referencing of relevant literature; (b) no attempt to evaluate and/or validate the success of the approaches; and (c) no attempt to use the GRNs to infer anything novel about the single cell data used to generate them. Furthermore, although this work seems to constitute 40%+ of the paper, it is not even mentioned clearly in the abstract or the introduction. Without better documentation and description of the methods, validation, and/or the use of more rigorous approaches, the network interactions are not only of questionable usefulness, but may even confuse less experienced researchers not familiar with these methods or the caveats in interpreting inferred interactions. This reviewer has been unable to find published examples similar to the authors' implementation of GENIE3 (and the authors do not support this approach with references) that infer GRNs from single cell data and then make no attempt to validate the results or draw new biological insights. Thus, this reviewer believes these deficiencies must be rigorously addressed.

We have attempted to address all three of the main reasons why Reviewer 3 believes `PlanExp`'s *genetic network inference* is of questionable value. We hope we provide here enough evidences that it is worth to keep these analyses as part of the `PlanExp` sections.

First, we have added a new paragraph to the introduction to contextualize our approach by referencing relevant literature. We have added a new section in supplementary material where we describe thoroughly the methods for inferring genetic interactions, as well as uploading the complete protocol as a *markdown* file. This should make our protocol completely reproducible, and we think it adds to the manuscripts' relevance in the field.

We have explored the inferred networks to highlight how they can be useful for researchers by comparing the inferred networks to a recently published paper (see answer 4) and we have curated a list of interactions related to the *Wnt signalling pathway*, hihglighting how `PlanExp` can be useful for retrieving functionally related genes. These results are discussed in the *Discussion* section of the manuscript.

We would like to point out that the main focus of `PlanExp` is not to extract or infer anything novel about the datasets included in the application. We have tried to integrate data from many different sources in a useful application, and the regulatory network inference was devised as a way to help researchers extract and infer that information from the data themselves.

Nevertheless, we agree with most of Reviewer's 3 concerns: our approach was maybe too preliminary, non-reproducible, and we didn't compare the inferred networks to any published data, which made the whole network inference not as useful, detailed and clear as it should be. However, some predicted results seem to be in agreement with experimental work, so despite being a predictive model we still think it can be useful as an exploratory complement to `PlanExp`.

M-(1) The authors have not referenced similar implementations of GENIE3 to infer GRNs from single cell data, even though several such efforts have been published, including:

- ***Ocone et al., Bioinformatics, 2015 (https://doi.org/10.1093/bioinformatics/btv257)***
- ***Aibar et al., Nature Methods, 2017 (https://doi.org/10.1038/nmeth.4463)***
- ***Gao et al., Bioinformatics, 2018 (https://doi.org/10.1093/bioinformatics/btx575)***

As a result, the reader cannot know (without doing their own searches) what progress has been made in this space recently, and how the authors' work compares. All three of these papers validate results by demonstrating the detection of known regulatory interactions and transcription factors in specific cell types, draw new biological insights from the networks discovered, and demonstrate superiority of their approaches to other methods for identifying gene networks. If the PlanEXP authors are aware of other publications that implement GENIE3 without validating the resulting networks/interactions, they should cite them. The Aibar paper also provides detailed methods for input to and output from GENIE3, including evaluation of thresholds and cutoffs for ranking predictions, which the authors should consider in improving the description of their methods so that others can reproduce their results.

We have changed the whole protocol of predicting genetic interactions; which we now refer to them as *Gene co-expression networks* in the application as this is the name the work Aibar et al. refers to the GENIE3 predictions.

We now used the program `GRNBoost` described in the `SCENIC` pipeline (*Aibar et al. 2017*) to infer these networks, which have allowed us to perform a prediction over all cells in a reasonable amount of time. We have implemented the filters detailed in Aibar et al., to improve the accuracy of our predictions, while making our apporach comparable to those detailed in the `SCENIC` pipeline.

While we believe a full validation of the results is not feasable at the moment (there is no gold-standard to compare the predictions to), we have tried to annotate the predictions further, to improve the filtering necessary for researchers to select candidates for experimental validation.

Firstly, we have annotated the gene co-expression interactions with the `REACTOME` pathway identifiers that both the *regulator* and the *target* belong to. A summary table of how many genetic links belong to the same REACTOME pathway is available in Supplementary Table 4, which we also present below.

Secondly, we have compared the predictions of `GRNBoost` over the two scRNA-seq experiments (*2018 Rajewksy Cell Atlas* and *2018 Reddien Cell Atlas*). We can see a moderate correlation of the confidence scores of both predictions, and those gene co-expression interactions that have been simultaneously predicted in both experiments have been annotated as such: they also appear first in the web application, to allow researchers to filter-out other predictions if they want to only consider them. See Supplementary Figure 3 for more information.

Thirdly, we have manually compared our inferred genetic interactions with a recently published RNA-seq experiment (1). Our comparison shows that from the 86 targets predicted for *soxB1-2*, 23 were shown to be differentially expressed in *Ross et al.*. From these 23 targets, three were shown to produce movement and sensory defects after RNAi inhibition. We believe this comparison shows how our predictions can be useful for researchers interested in exploring genetic interactions. This comparison is detailed in Supplementary Table 6 and discussed in the Discussion section of the manuscript.

Finally, we have extracted gene co-expression interactions from our predictions of genes related to the *Wnt signalling pathway*. While these genes may not be directly regulating each other's expression, we believe the ability of the protocol to retrieve many genes involved in this pathway highlights how these results may be useful not only in predicting genetic interactions, but also in relating genes that may participate in the same signalling pathways (see Supplementary Table 5 and Discussion section of the manuscript).

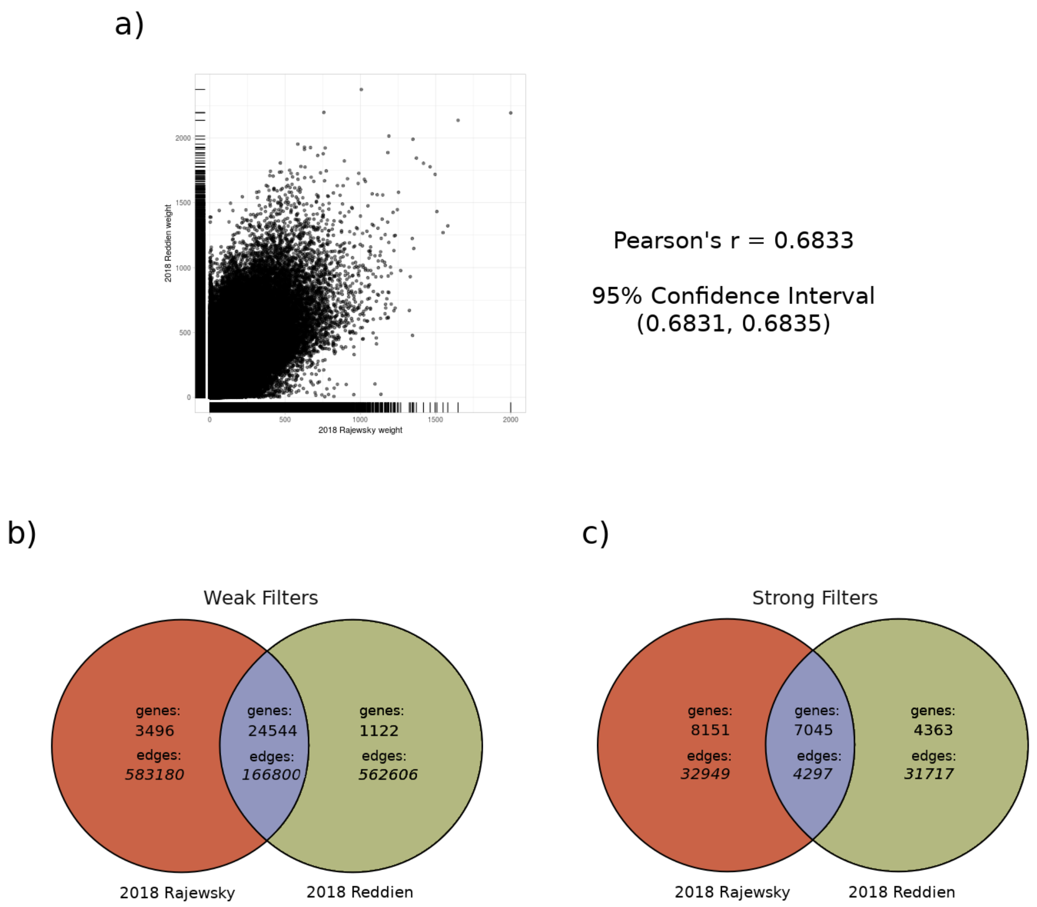
Below we have included the corresponding Supplementary tables and figures relevant for this answer:

Suppl. Table 4. Summary of the gene co-expression networks predicted for the single cell RNA-seqs experiments available on PlanExp.

	2018 Rajewsky Cell Atlas	**2018 Reddien Cell Atlas**
Total Edges	37 246	36 014
Total Genes	15 196	11 408
Edges with Reactome	3635	3490
Genes in edges with Reactome	1855	1601
Total Reactome pathways	18 671	17 215
Unique Reactome pathways	792	777

Suppl. Table 5. Retrieved gene co-expression interactions for the experiment *”2018 Rajewsky Cell Atlas”* related to the Wnt signaling pathway (`KEGG hsa04310`). Contigs were annotated using the names listed in the PlanMine database, their related genes in PlanNET and PlanMine, and an `NCBI-BLASTX` search against the non-redundant human protein sequenc es database (`e-value` < `0.01`). The whole table containing all `GRNBoost` predictions is available at https://github.com/scastlara/PlanExp-protocols.

Regulator Contig	**Target Contig**	**Regulator Gene**	**Target Gene**	**Regulator Homolog**	**Target Homolog**
dd_Smed_v6_16209_0 (wnt11-2)	dd_Smed_v6_4900_0 (axinA)	SMESG000074928.1 (ISCW-ISCW004707)	SMESG000039725.1	WNT2	AXIN1
dd_Smed_v6_16209_0 (wnt11-2)	dd_Smed_v6_7326_0 (wntP-2)	SMESG000074928.1 (ISCW-ISCW004707)	SMESG000066476.1 (MS3_03642)	WNT2	WNT2 precursor
dd_Smed_v6_16209_0 (wnt11-2)	dd_Smed_v6_4154_0 (glypican-1)	SMESG000074928.1 (ISCW-ISCW004707)	SMESG000052359.1 (MS3_08283)	WNT2	GPC4
dd_Smed_v6_4639_0 (dvl-1)	dd_Smed_v6_2413_0	SMESG000049876.1 (DVL2)	SMESG000048614.1	DVL2	PRKACA
dd_Smed_v6_63520_0	dd_Smed_v6_9669_0	SMESG000052392.1 (LEF1)	SMESG000016284.1 (PTMB.308)	LEF1	PRKACB
dd_Smed_v6_8502_0	dd_Smed_v6_3221_0	SMESG000073604.1	SMESG000049234.1 (PLCB4)	DICER1	PLCB4
dd_Smed_v6_3757_0	dd_Smed_v6_891_0	SMESG000052999.1 (BTRC)	SMESG000001466.1 (MYCBP)	BTRC	MYCBP
dd_Smed_v6_2045_0	dd_Smed_v6_9037_0	SMESG000064848.1 (PPP3CA)	SMESG000077398.1 (PPP3CA)	PPP3CA	PPP3CA
dd_Smed_v6_13487_0 (wnt2-1)	dd_Smed_v6_13985_0 (sfrp1)	SMESG000002069.1 (WNT2B)	SMESG000029446.1 (MS3_08312)	WNT2	SFRP5
dd_Smed_v6_7173_0	dd_Smed_v6_5378_0	NA	SMESG000012014.1 (PPP3CA)	PPP3CB	PPP3CA
dd_Smed_v6_5531_0 (axinB)	dd_Smed_v6_5818_0	SMESG000025925.1	SMESG000025925.1	AXIN2	AXIN1
dd_Smed_v6_10098_0	dd_Smed_v6_3221_0	SMESG000013958.1	SMESG000049234.1 (PLCB4)	PLCB1	PLCB4
dd_Smed_v6_10098_0	dd_Smed_v6_7210_0 (fz-4-4)	SMESG000013958.1	SMESG000014322.1 (FZD4)	PLCB1	FZD4
dd_Smed_v6_10098_0	dd_Smed_v6_4244_0	SMESG000013958.1	SMESG000039854.1 (PRKCA)	PLCB1	PRKCA



Suppl. Figure 3. Comparison of gene co-expression network prediction for *"2018 Rajewsky Cell Atlas"* and *"2018 Reddien Cell Atlas"* single-cell RNA-seq experiments. **a)** Distribution of relationships weights between the two predictions, showing a moderate correlation between the confidence values reported by `GRNBoost` in both experiments for the same regulator-target gene relationships. **b) and c)** Venn diagram of the filtered gene co-expression networks for both experiments using 500 as the maximum number of targets and regulators (Weak Filters) or 50 (Strong Filters). Most gene relationships are unique to each prediction, while the number of shared genes in the networks is higher, but both predictions become more sim ilar when relaxing the filters.

Suppl. Table 6. Predicted target genes for *soxB1-2* (*dd_Smed_v6_8104_0*) in the computed gene co-expression network shown to be differentially expressed in *Ross et al.* (1). When a target gene was predicted in both experiments (*”2018 Rajewsky Cell Atlas”* and *”2018 Reddien Cell Atlas”*), the score column corresponds to the mean score of both predictions as reported by `GRNBoost`.

Target Contig	**Target name**	**Predicted in**	**Score**
dd_Smed_v6_9977_0	*pkd2l-2*	Both	316.381
dd_Smed_v6_716_0	–	Both	292.228
dd_Smed_v6_10282_0	*syne1*	Both	206.841
dd_Smed_v6_3175_0	*tetraspanin homolog*	Both	203.112
dd_Smed_v6_7815_0	*ftcd*	Both	194.429
dd_Smed_v6_13327_0	*pkd2-4*	Both	176.652
dd_Smed_v6_12955_0	*pkd2-1*	2018 Rajewsky Cell Atlas	99.540
dd_Smed_v6_3331_0	–	2018 Rajewsky Cell Atlas	89.740
dd_Smed_v6_10911_0	*zfp940*	2018 Rajewsky Cell Atlas	83.010
dd_Smed_v6_9135_0	*tlcd1*	2018 Rajewsky Cell Atlas	75.098
dd_Smed_v6_7064_0	–	2018 Rajewsky Cell Atlas	68.534
dd_Smed_v6_21965_0	*semal*	2018 Reddien Cell Atlas	464.168
dd_Smed_v6_12811_0	*stum*	2018 Reddien Cell Atlas	350.143
dd_Smed_v6_3220_0	–	2018 Reddien Cell Atlas	330.581
dd_Smed_v6_3843_0	–	2018 Reddien Cell Atlas	329.732
dd_Smed_v6_3680_0	*ttpal homolog*	2018 Reddien Cell Atlas	319.375
dd_Smed_v6_3895_0	–	2018 Reddien Cell Atlas	313.629
dd_Smed_v6_7903_0	*centrin-1*	2018 Reddien Cell Atlas	309.129
dd_Smed_v6_5307_0	*cyp1a1 homolog*	2018 Reddien Cell Atlas	300.328
dd_Smed_v6_116_0	*dynll2*	2018 Reddien Cell Atlas	292.844
dd_Smed_v6_14226_0	*eml-1*	2018 Reddien Cell Atlas	291.754
dd_Smed_v6_5346_0	*rsph4A*	2018 Reddien Cell Atlas	287.853
dd_Smed_v6_10460_0	*loxhd1*	2018 Reddien Cell Atlas	287.246

M-(2) Input: What was used as the input for GENIE3? UMI counts? TPM? FPKM/RPKM? Were these values normalized (which can affect results)? If so, how? What steps were conducted for each data set, in detail? How many times was random sampling of "up to" 30 cells/cluster repeated? How reproducible were the resulting interactions over multiple replicate runs of GENIE3? And, since the authors state "up to" 30 cells, shouldn't the precise number of cells per cluster per run be reported? The authors should also consider including a graphical schematic of steps they used to implement GENIE3, as found in the three papers referenced above.

We have used raw counts for `GRNBoost` (before: `GENIE3`). From the `SCENIC` user guide:

> "*Expression units: The preferred expression values are gene-summarized counts. There is currently not a strong recommendation towards using the raw counts, or counts normalized through single-cell specific methods (e.g. Seurat).* "

We have now included a thorough description of the methods employed for inferring the networks in Supplementary Material *Gene co-expression network* section, and we have released a github site with all the protocols of `PlanExp`, including the GRN inference step. All scripts, input files and steps are detailed in a markdown file, accessible here: https://github.com/scastlara/PlanExp-protocols/co-expression_network.

As explained before in answer 1, where we discuss the change from `GENIE3` to `GRNBoost`, we now perform the predictions over all cells in each experiment, thanks to the faster implementation of `GRNBoost`, and thus detailing the precise number of cells per cluster in each random subset is no longer necessary. Running replicate runs of the protocol over different random subsets is also no longer necessary, as the new predictions use all the available data.

M-(3) Output: In the Suppl. Figure 3 legend (and in the methods), the authors state "Only [the] top 1,000 genetic interactions were considered, as opposed to an arbitrary weight cut-off, due to the fact that the weight reported by GENIE3 does not have any statistical meaning and to reduce the potential load on the application." If weights are ignored, and no attempt to validate these predictions has been made, how do we know how robust each of the top 1000 predictions are? Why not choose the top 200, or the top 500? Have other publications limited using this approach (if so, please cite them)?

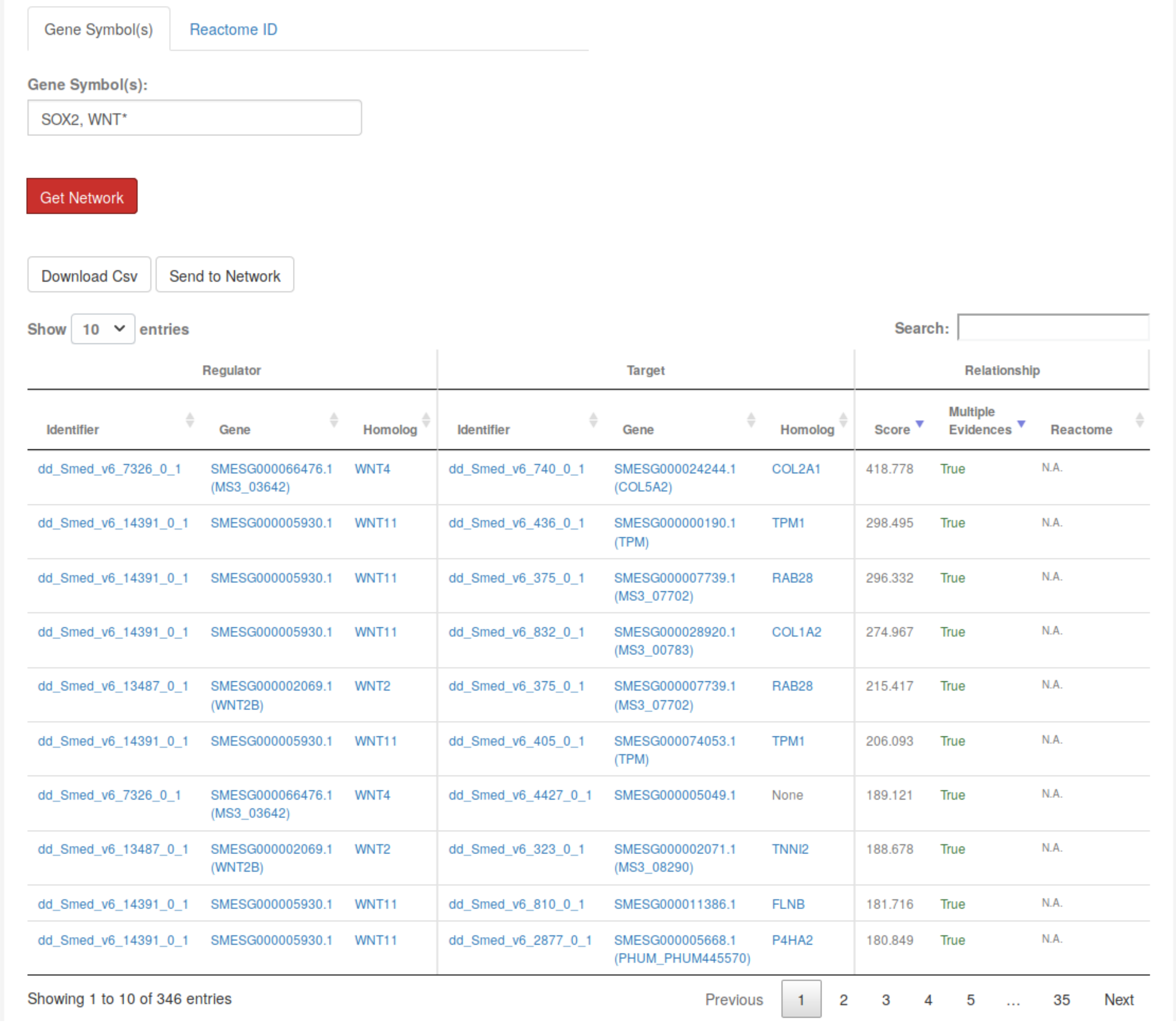
As detailed before, and explained in the new Supplementary Material section "*Gene co-expression network*", we have now applied the filters described in *Aibar et al. 2017*. However, instead of generating three possible networks, as the `SCENIC` pipeline does, we have merged the filters to produce a single, more stringent network.

Also, the authors do not report the number of networks with 1-, 5-, 10- genes, etc. Because the gene regulatory interactions section is limited to displaying only 100 interactions at a time, and does not seem to be sortable for the entire 1000, it is not possible to export only high weight interactions to the network viewer, or networks with only 5+ genes, or only interactions for a specific regulator(s).

First, due to the changes in the protocol, the connectivity of the network has changed drastically. Where before we had many small connected sub-networks (connected components from now on), we now have one big connected component (with 37,234 and 35,934 interactions for *2018 Rajewsky Cell Atlas* and *2018 Reddien Cell Atlas*, respectively) and many small connected components (11 and 41 for *2018 Rajewsky Cell Atlas* and *2018 Reddien Cell Atlas*, respectively). While these numbers are not listed in the manuscript, the protocol for inferring the gene co-expression network includes a step to compute them and they are listed in the markdown file with the protocol.

We agree that the navigation and discoverability of the network could be improved. To that effect, we have now changed the way in which `PlanExp` displays predicted interactions. Now, the network is fully searchable through two search forms: one where users can input multiple genes, and another where users can input multiple `REACTOME` pathway names and identifiers. Users are able to search for many genes simultaneously, create a dynamic DataTable with all the interactions, and then filter them by using the DataTable features. Finally, once they are pleased with the displayed interactions, they can send them to the network viewer, where expression data can be mapped over the network. We have also implemented an auto-complete feature for REACTOME pathway names, where, for instance, writing "apopto" will return all REACTOME pathway names that start with the search term prefix. As like with genes, multiple REACTOME pathways can be searched at the same time to generate a dynamic table for further filtering.

New interface for exploring the predicted gene co-expression networks. Users can search multiple genes or REACTOME pathway names at the same time with no limits, and then they can further filter the table by using the Search bar on the top-right of the table. Interactions are sorted by "Multiple Evidences" (those interactions predicted in the *2018 Rajewsky Cell Atlas* and *2018 Reddien Cell Atlas* appear first), and then by confidence score. The Search bar allows users to display only those interactions with "Multiple Evidences", those where specific genes participate, or even where both Regulator and Target belong to a particular REACTOME pathway.



Gene Symbol(s): SOX2, WNT*

Show 10 entries

Search:

Regulator			Target			Relationship		
Identifier	Gene	Homolog	Identifier	Gene	Homolog	Score	Multiple Evidences	Reactome
dd_Smed_v6_7326_0_1	SMESG000066476.1 (MS3_03642)	WNT4	dd_Smed_v6_740_0_1	SMESG000024244.1 (COL5A2)	COL2A1	418.778	True	N.A.
dd_Smed_v6_14391_0_1	SMESG000005930.1	WNT11	dd_Smed_v6_436_0_1	SMESG000000190.1 (TPM)	TPM1	298.495	True	N.A.
dd_Smed_v6_14391_0_1	SMESG000005930.1	WNT11	dd_Smed_v6_375_0_1	SMESG000007739.1 (MS3_07702)	RAB28	296.332	True	N.A.
dd_Smed_v6_14391_0_1	SMESG000005930.1	WNT11	dd_Smed_v6_832_0_1	SMESG000028920.1 (MS3_00783)	COL1A2	274.967	True	N.A.
dd_Smed_v6_13487_0_1	SMESG000002069.1 (WNT2B)	WNT2	dd_Smed_v6_375_0_1	SMESG000007739.1 (MS3_07702)	RAB28	215.417	True	N.A.
dd_Smed_v6_14391_0_1	SMESG000005930.1	WNT11	dd_Smed_v6_405_0_1	SMESG000074053.1 (TPM)	TPM1	206.093	True	N.A.
dd_Smed_v6_7326_0_1	SMESG000066476.1 (MS3_03642)	WNT4	dd_Smed_v6_4427_0_1	SMESG000005049.1	None	189.121	True	N.A.
dd_Smed_v6_13487_0_1	SMESG000002069.1 (WNT2B)	WNT2	dd_Smed_v6_323_0_1	SMESG000002071.1 (MS3_08290)	TNNI2	188.678	True	N.A.
dd_Smed_v6_14391_0_1	SMESG000005930.1	WNT11	dd_Smed_v6_810_0_1	SMESG000011386.1	FLNB	181.716	True	N.A.
dd_Smed_v6_14391_0_1	SMESG000005930.1	WNT11	dd_Smed_v6_2877_0_1	SMESG000005668.1 (PHUM_PHUM445570)	P4HA2	180.849	True	N.A.

Showing 1 to 10 of 346 entries

Finally, what percent of the interactions do the authors estimate are false positives? Some recent approaches (e.g. in the Aibar paper cited above) suggest that including cis-regulatory region analysis significantly improves GRN prediction from single cell data. While further approaches might be beyond the scope of the current paper, it seems expected, based on other publications, for the authors to provide some analysis of the quality of their predictions.

It is difficult to give an accurate estimation of how many false positives the prediction contains, as coming up with a way to validate them can be challenging (see Answer 1). However, we have added the comparison against *soxB1-2* targets studied in *Ross et al. 2018* specifically for that effect. In that case, 23/86 (∼26.74%) of the targets were detected in the RNA-seq experiment performed by *Ross et al.*. Again, we are hesitant to present this number as the "*positive predictive value*" (or *Precision*) due to the fact that the experimental conditions between the single-cell experiments and the RNA-seq are clearly different, and many relationships retrieved by `GRNBoost` that do not appear in the RNA-seq used for validation may not actually be false positives. However, we present this comparison to contextualize our results with a recently published paper, and to show the usefulness of our approach.

The `SCENIC` pipeline does in fact use cis-regulatory regions to filter the results of `GENIE3/GRNBoost`. However, performing a prediction of transcription factor binding sites for the new *Schmidtea mediterranea* genome assembly, annotating transcription factors in *S. mediterranea* with their predicted binding sites, and then filtering the results of the `GRNBoost` prediction is out of the scope of out current manuscript, as it would require extensive work for both performing the predictions but also validating these annotations. We now state the limitations of our approach and follow-up suggestions in the Discussion section of the mansucript:

> " *Although `GENIE3/GRNBoost` were conceived as programs to retrieve genetic interactions from expression matrices, `GRNBoost` constitutes only the first step of the `SCENIC` pipeline to retrieve genetic interactions. The following steps include the filtering of the gene co-expression network by including information of cis-regulatory elements. Unfortunately, due to the novelty of the most current S. mediterranea genome assembly, such data is not available yet. Therefore, the predicted co-expression networks contain other types of relationships between genes apart from genetic regulatory interactions, such as protein-protein interactions, genes that belong to the same signalling pathways, or genes expressed in the same cell types (see Supplementary Material,* Gene co-expression network *section). Nevertheless, these co-expression networks are available for download so that researchers may use them for further filtering (either by following the whole SCENIC pipeline or by implementing their own filtering protocol). Given that the co-expression network prediction is the most computationally expensive step of the `SCENIC` pipeline, releasing this prediction so that researchers may be able to implement other filtering protocols or validate the genetic links individually can help the planarian community to understand and annotate many important genetic interactions.* "

Finally, by annotating those interactions that were retrieved simultaneously in *2018 Rajewsky Cell Atlas* and *2018 Reddien Cell Atlas* we are giving researchers the ability to reduce all the predictions to only those which we expect to be more reliable. The comparison between both predictions in Supplementary Material also aids in assessing the quality of our results.

M-(4) Validation/evaluation of utility: This is possibly the most serious deficiency of the authors' effort to infer regulatory interactions. There does not seem to be any attempt to evaluate whether the GENIE3 predictions are useful or biologically relevant. Based on the number of 5-, 10-, or 20-gene networks, how do the authors' results compare to other efforts to infer GRNs from single cell data? Did the authors uncover known interactions? There are now numerous publications that report RNA-Seq data for knockdown of transcription factors known to be enriched in specific planarian cell types. Did the authors' approach detect any of these interactions? Were there any novel/unappreciated networks discovered, for example in neoblasts, muscle, or epithelial cells (currently some of the most studied planarian cell types)? While this last possibility might be a good topic for a follow-up paper, an example would also be a demonstration of the robustness of the authors' methods.

We have already discussed some of the improvements we have made to the inference of genetic interactions in the previous answers. With the changes we have implemented to both the protocol and the manuscript, we believe we have shown that our predictions are relevant and can be useful.

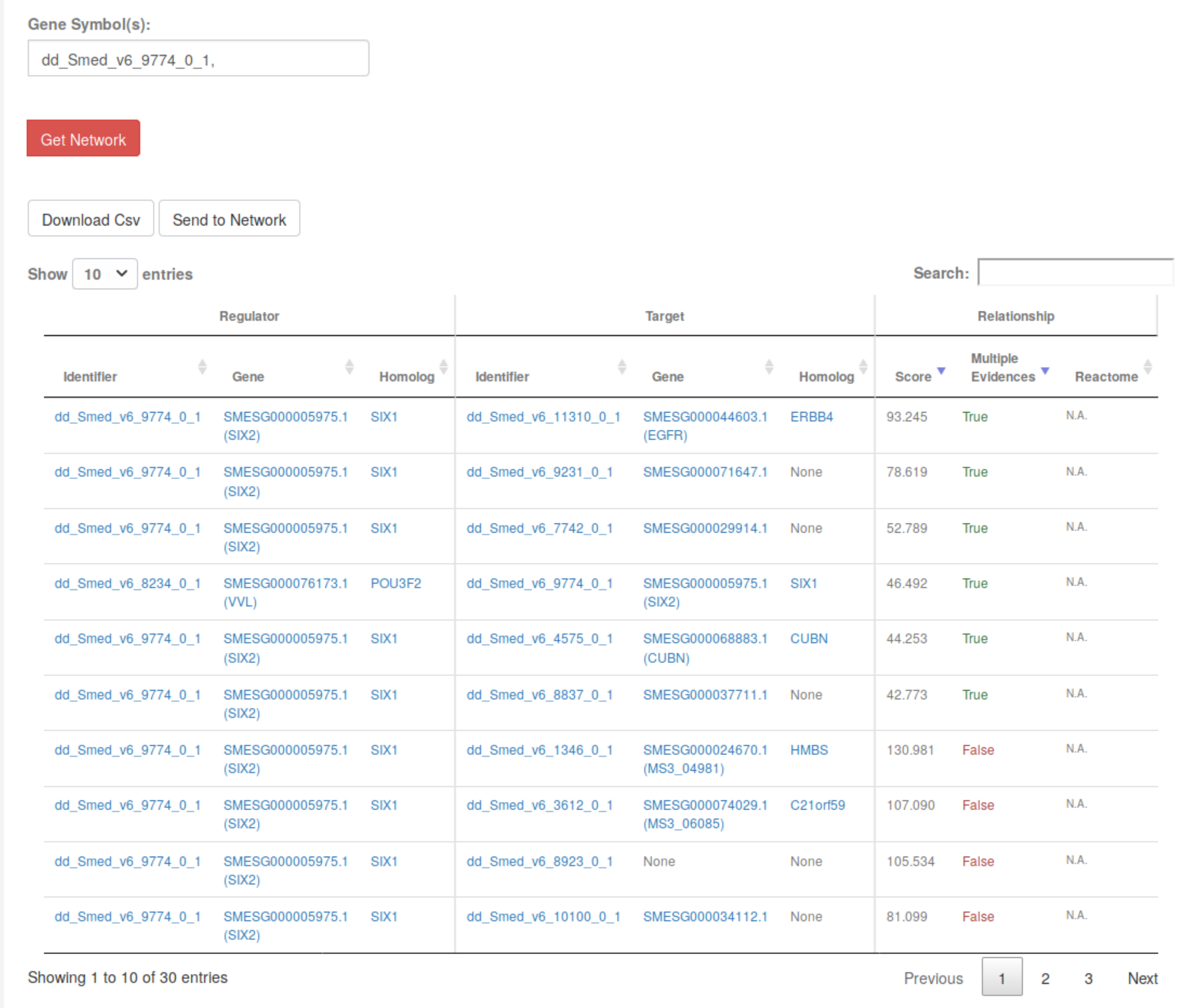
To summarize them here:

- We have uncovered known genetic interactions in *S. mediterranea*, as our protocol was able to retrieve 23 interactions out of 86 which were also detected in a recently published RNA-seq data for a knockdown transcription factor (*soxB1-2*) (1). Three of these 23 interactions were shown in Table 1 of the original article to produce sensory defects after RNAi inhibition.

- We have also included a table with elements relevant in the *Wnt signalling pathway*, showing how our predictions can aid in retrieving new candidates that may participate in relevant signalling pathways.

Although we have not included it in this manuscript, we have also looked at two other works where some genetic interactions or regulatory systems were proposed. In *Scimone et al.* (2), the gene *cubilin*, *egfr-5* and *tetraspanin* (among others) are shown to be significantly under-expressed in *six1/2* and *pou2/3* knockdowns. The interaction between *six1/2* (`dd_Smed_v6_9774_0`), *cubilin* (`dd_Smed_v6_4575_0`) and *egfr-5* (`dd_Smed_v6_11310_0`) have been inferred in both experiments (*2018 Rajewsky Cell Atlas* and *2018 Reddien Cell Atlas*). In the case of the prediction performed over *2018 Reddien Cell Atlas*, the gene *tetraspanin* (`dd_Smed_v6_9647_0`) has also been predicted to interact with *pou 2/3*. A connection between *pou2/3* (`dd_Smed_v6_8234_0`) and *cubilin* was not predicted by `GRNBoost`, but `six1/2` and `pou2/3` are listed as interacting by the prediction over both scRNA-seq experiments that we performed. In *Rink et al.* (3), the gene *egfr-5* was shown to be crucial in the regulation of the regeneration of the planarian excretory system. Note that the connection between *six1/2* and *egfr-5* is the first one listed in the results when searching for `dd_Smed_v6_9774_0_1` in *2018 Rajewsky Cell Atlas*, while the *pou2/3→six1/2* is the fourth, the *six1/2→cubilin* connection is the fifth, and *pou2/3→tetraspanin* is the ninth when searching for `dd_Smed_v6_8234_0_1` in the *2018 Reddien Cell Atlas* prediction. Given that interactions are sorted by confidence (first by those which have been predicted in both scRNA-seq experiments, and then by score), we think that our ranked predictions can be useful to select new candidates for experimental validation.

Genetic interactions predicted for *six1/2* in *2018 Rajewsky Cell Atlas.* *egfr-5* (dd_Smed_v6_11310_0), *pou2/3* (dd_Smed_v6_8234_0), and *cubilin* (dd_Smed_v6_4575_0) were shown to be significantly under-expressed in *Scimone et al. six1/2* knockdown planarians. *egfr-5* was later described to be crucial in the regeneration of the excretory system in *Rink et al.*.



Gene Symbol(s): dd_Smed_v6_9774_0_1,

Show 10 entries Search:

Regulator			Target			Relationship		
Identifier	Gene	Homolog	Identifier	Gene	Homolog	Score	Multiple Evidences	Reactome
dd_Smed_v6_9774_0_1	SMESG000005975.1 (SIX2)	SIX1	dd_Smed_v6_11310_0_1	SMESG000044603.1 (EGFR)	ERBB4	93.245	True	N.A.
dd_Smed_v6_9774_0_1	SMESG000005975.1 (SIX2)	SIX1	dd_Smed_v6_9231_0_1	SMESG000071647.1	None	78.619	True	N.A.
dd_Smed_v6_9774_0_1	SMESG000005975.1 (SIX2)	SIX1	dd_Smed_v6_7742_0_1	SMESG000029914.1	None	52.789	True	N.A.
dd_Smed_v6_8234_0_1	SMESG000076173.1 (VVL)	POU3F2	dd_Smed_v6_9774_0_1	SMESG000005975.1 (SIX2)	SIX1	46.492	True	N.A.
dd_Smed_v6_9774_0_1	SMESG000005975.1 (SIX2)	SIX1	dd_Smed_v6_4575_0_1	SMESG000068883.1 (CUBN)	CUBN	44.253	True	N.A.
dd_Smed_v6_9774_0_1	SMESG000005975.1 (SIX2)	SIX1	dd_Smed_v6_8837_0_1	SMESG000037711.1	None	42.773	True	N.A.
dd_Smed_v6_9774_0_1	SMESG000005975.1 (SIX2)	SIX1	dd_Smed_v6_1346_0_1	SMESG000024670.1 (MS3_04981)	HMBS	130.981	False	N.A.
dd_Smed_v6_9774_0_1	SMESG000005975.1 (SIX2)	SIX1	dd_Smed_v6_3612_0_1	SMESG000074029.1 (MS3_06085)	C21orf59	107.090	False	N.A.
dd_Smed_v6_9774_0_1	SMESG000005975.1 (SIX2)	SIX1	dd_Smed_v6_8923_0_1	None	None	105.534	False	N.A.
dd_Smed_v6_9774_0_1	SMESG000005975.1 (SIX2)	SIX1	dd_Smed_v6_10100_0_1	SMESG000034112.1	None	81.099	False	N.A.

Showing 1 to 10 of 30 entries

m-(1) A more explicit, itemized list of what is new (including GENIE3-based inference of regulatory interactions) in PlanEXP should be provided in the abstract, introduction, and discussion. This will benefit the paper since PlanEXP is being added to PlanNET, and it is important to clearly distinguish the current manuscript's contribution.

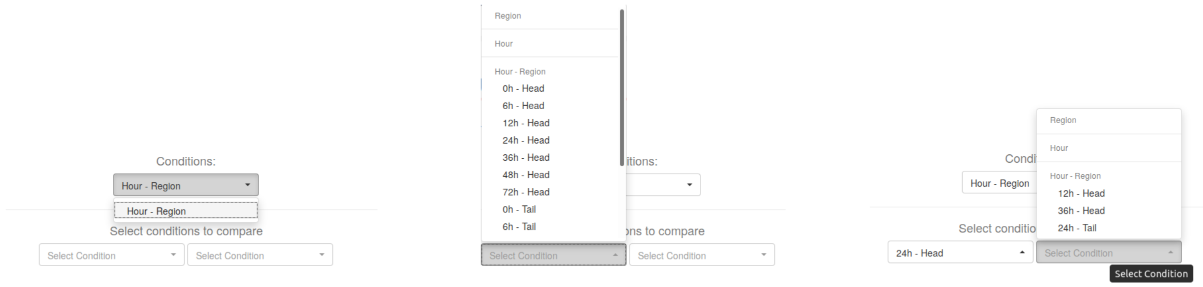
We want to remind Reviewer 1 that the whole `PlanExp` module is new to `PlanNET`, and thus all the available features in the `PlanExp` webpage are new. However, we have now listed the new features added by `PlanExp` to `PlanNET` in the Supplementary Material.

m-(2) Is it possible to limit the drop down menu options for ”Differentially expressed genes” to only the comparisons that are available? For example, for Reddien-Dresden, it is only possible to select conditions when ”Cluster” is selected in the top drop-down. If ”Section” is selected, the user can choose ”Head,” ”Pharynx,” etc. in the first drop-down, but nothing in the second drop-down. The Aboobaker data set has similar limitations – only the ”Hour-Section” Condition enables pairwise choices in the next two drop-downs. Then, even if ”0h - head” is selected in dropdown #1, only ”0h-head” or ”6h-head” are available in dropdown #2. The authors should consider testing all of the dropdown conditions for usability.

We have now changed the dropdown menu of the DGE section of `PlanExp` so that only those "*Condition Types*" with differential gene expression computed for them are shown. Once a "*Condition Type*" is selected, all the levels in that factor will appear on the two dropdowns below. Upon selection of the first dropdown, only the levels for which there are differentially expressed genes will be shown on the second.

In the case of *2013 Aboobaker Time-course*, the first dropdown (*Conditions*) only shows the condition type *Hour* − *Region*, as it is the factor that was used in the design matrix of the experiment (see Supplementary Table 2). The second pair of dropdowns will show first all the levels of the *Hour* − *Region* factor (0h − Head, 6h − Head, and so on). Upon selection of the *Condition 1* dropdown, the second dropdown's options will now be limited to the levels for which there are differentially expressed genes. So, for instance, *36h* − *Head* as *Condition 1* will allow users to select *24h* − *Head*, *24h* − *Head*, and *36h* − *Tail*. In the case of this experiment, we followed the author's design, where they compared each timepoint to the previous one in the same region (head or tail), and then they also compared each timepoint in a region to the same timepoint in the other region. That is, not all pariwise combinations are available, as these were not computed by the authors in the original publication nor by us.

We have added the following image to the `PlanExp` tutorial :



1) Only those Factors for which there are differentially expressed genes are shown

2) All levels are available in both dropdowns at first.

3) Once a level is selected in the first Dropdown, the options of the other dropdown are limited to only those with differentially expressed genes

m-(3) Also, for Aboobaker (and perhaps others), why is it possible to select only "hour" or "section" when the data were actually collected as time-section combinations? Have the authors somehow averaged head/tail data for individual times, or section data across all times? If so, detailed methods should be reported.

We have added an explanation for this in the Results section of the paper, as it is not something that we have to pre-compute, but rather, a function that `PlanExp` performs in real time when selecting a given Factor.

> *" Gene expression can be grouped by factor (or by a combination of factors if available for the experiment), via a dropdown menu. Users can select any factor in a given experiment to group the samples before plotting, and `PlanExp` will retrieve all samples matching the specified levels and compute the average expression (in the case of heatmaps, line plots, and bar plots) or plot the points individually (in the case of violin plots and scatter plots). For instance, in the case of samples collected as time-region combinations (as is the case for "*2013 Aboobaker Time-course*"), users can select to plot the expression of several genes by grouping all samples in "Head" and "Tail" levels, without taking into consideration the time at which these samples were collected. This process is done automatically at run-time by `PlanExp`, and allows users to group the samples by their factors of interest."*

m-(4) In many plots, axes and legends are displayed without any units of measurement. Can log2FC, TPM, FPKM, etc. be displayed on the plots?

Heatmaps, line plots, barplots and scatterplots now display the correct units of measurement. All plots in the manuscript and supplementary material have been updated to include the new displayed units.

m-(5) The t-SNE plot for the Fincher/Reddien data set has a very different appearance from that available at Digiworm. For example, clusters are found in different plot locations, subclusters are often separated for related cell types, and specific cell type markers map to fewer or more regions, unlike in the original publication. The obvious recommendation is to reproduce the t-SNE plot from Fincher, but if this is not possible, the authors need to detail in methods how this t-SNE plot was generated, and why the original plot is not used. Because the t-SNE plot function duplicates to some extent online resources already made available by others, the current plot will confuse non-specialists, and users will just use the Digiworm web page.

We have added a whole section in Supplementary Material (section *Integration of RNA-seq experiments*) detailing the important parameters for reproducing our analyses (including the computation of t-SNE plots). The whole protocol is available now at our github site. Unfortunately, the computation of t-SNE dimensions is difficult to reproduce, as it has an associated randomness that, even by using the same parameters, reproducing it is impossible without knowing the initial random seed. To avoid any confusion to our users, we have added the following statement in the corresponding section of the website:

> *" While cells with similar expression patterns will appear closer in the plot, the randomness associated with the dimentionality reduction method might display cells in a different position from the original publication's plot. "*

m-(6) PFAM domain IDs in gene cards should be the same as those that allow searching/display in the Gene Expression module (e.g., a search for "DEAD" does not work, but "PF00270" does). Also, there seem to be many planarian transcripts for

which human orthologs exist, but for which no PFAM domains have been predicted. The authors should check their cutoffs to be sure they are not overly stringent. Otherwise, this limits the utility. For one example, see SMESG000037282.1/dd_Smed_v6_6729_0_1, a SCRO/NKX2-4 homolog that clearly should have a homeodomain predicted (PFAM search with the longest ORF predicts a homeodomain with E-value of 1.3e-19).

We have now added the functionality to search for PFAM domain names across all `PlanExp` (and `PlanNET`), as it was only possible to search by PFAM domain accession before. Now, "DEAD" returns the same results as "PF00270".

We used an e-value cut-off of 10^{-20} for annotating PFAM domains using HMMER (detailed in the original `PlanNET` publication). Changing these cut-offs would be a difficult task that would need a database update for all `PlanNET` annotations. However, we will consider lowering the cut-offs (perhaps to 10^{-10}) in a future `PlanNET` update, which will then affect `PlanExp` automatically.

m-(7) The GO annotation tool does not always display data when complete (e.g. in my hands the X1 vs Xins for the Abril data). The authors should rigorously test their web site on multiple browsers, or specify a browser for best performance.

We have fixed a bug where plotting some G.O. visualizations would make the application crash under certain conditions (which were reproducible by using the Abril DGE experiment). We would like to thank Reviewer 3 for spotting and reporting the bug here, which allowed us to fix it.

The website should work on any modern browser with JavaScript enabled (we have tested it in Chrome, Firefox and Vivaldi).

m-(8) The authors should make sure detailed EdgeR methods are available in the supplemental table, to which I did not have access. Not only FDR/p-values are important for re-analyzed data, but also cutoffs for low expressers, normalization methods and order, etc. so that others could reproduce the authors' results if desired.

As we stated in a previous question, we have added a new section to Supplementary Material detailing the important parameters for re-analyzing the data, but we have also uploaded the complete protocol for re-analyzing the RNA-seq datasets to our github site, which is linked in the corresponding Supplementary Material section.

When creating the protocols that would allow other researchers to reproduce the results, we spotted an issue with our input tables for *2013 Aboobaker Time-course* and *2012 Pearson Stem Cells*. We did not have access to raw counts for these two experiments, which prevented us from using the program *EdgeR*. As such, we have changed the protocol to use the *limma-trend* pipeline in the case of *2013 Aboobaker Time-course*, allowing normalized values as input. The *2012 Pearson Stem Cells* differential gene expression has been removed from the website due to the lack of replicates (although all the other plots and features of `PlanExp` can still be used with this experiment). The protocol is accessible here: https://github.com/scastlara/PlanExp-protocols/blob/master/markdowns/2013_Aboobaker.md .

m-(9) More detailed legends need to be supplied for each panel/figure in the supplement. For example for gene co-expression, what does each color mean, what is being plotted (cells or genes?), and so forth.

We have added an explanation for each of the plots shown in the complete protocol figure of the supplement (Supplementary Figure 1).

m-(10) Can the authors add a "working" or "loading" indicator for all sections where there isn't one so the user has a way to know progress/status?

We have added some missing loading spinners to the web application.

m-(11) Longer term: Do the authors plan to map all the data to all the transcriptomes, or at least to the Smesgene/Smest genome and transcriptome? These seem to be

(or will be) widely used references, and would arguably benefit the most researchers.

Firstly, once we can freeze the features of `PlanExp` for some time, we want to include many more RNA-seq datasets to `PlanExp` (including some that we are currently performing in collaboration with Dr. Adell and Dr. Saló). To that end, we have created a webpage that will allow researchers to aid us in collecting such data.

Finally, when more experiments are available in `PlanExp` and we have recieved some user feedback, we will consider re-mapping all experiments (or at least some of them) to the Smesgene genome.

References

[1] K. G. Ross, A. M. Molinaro, C. Romero, B. Dockter, K. L. Cable, K. Gonzalez, S. Zhang, E. M. S. Collins, B. J. Pearson, and R. M. Zayas, "SoxB1 Activity Regulates Sensory Neuron Regeneration, Maintenance, and Function in Planarians," *Developmental Cell*, vol. 47, no. 3, pp. 331–347.e5, 2018.

[2] M. L. Scimone, M. Srivastava, G. W. Bell, and P. W. Reddien, "A regulatory program for excretory system regeneration in planarians," *Development*, vol. 138, no. 20, pp. 4387–4398, 2011.

[3] J. C. Rink, H. T.-K. Vu, and A. S. Alvarado, "The maintenance and regeneration of the planarian excretory system are regulated by egfr signaling," *Development*, vol. 138, no. 17, pp. 3769–3780, 2011.

Supplementary material

PlanExp: intuitive integration of complex RNA-seq datasets with planarian omics resources

S. Castillo-Lara, E. Pascual-Carreras, J.F. Abril

1 PlanExp web-application

Comparison with PlanNET

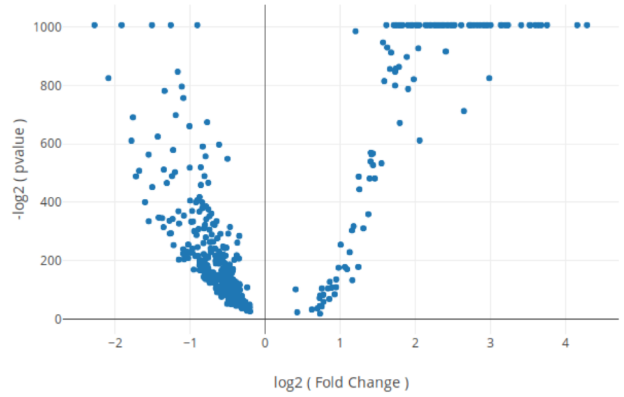
PlanNET (Castillo-Lara and Abril, 2018) is a database that holds a predicted interactome of protein-protein interactions for multiple planarian transcriptomes. Since its inception, PlanNET has had the ability to map expression data onto its networks, through the Net Explorer interface. However, the complexity of the RNA-seq datasets available for *Schmidtea mediterranea* –and specifically recently published single–cell RNA-seq datasets– made PlanNET's expression data visualization features limiting. Below, we list some of the most important features of PlanExp:

- Users can navigate through the interactome network faster on PlanNET, select a sub-network, and then transfer it to PlanExp to navigate faster through expression data.
- New plot types such as Heatmaps, bar plots, and line charts to visualize expression data for one or multiple genes.
- Ability to explore differential gene expression.
- Gene Ontology enrichment analysis.
- Database scalability due to the tabular nature of expression data: by switching from Neo4j (where the PlanNET networks are stored) to MySQL we have been able to hold many more data and query the database much faster.
- Interactive plots through the use of the JavaScript library Plotly, speeding up user's experience according to the client hardware.
- t-SNE visualizations and other features specific to single–cell RNA-seq datasets.
- Many features for exploring genes that are being co-expressed by multiple samples.
- Integration with the new SMESG gene annotations.
- Ability to store and visualize genetic interactions predicted for specific experiments.
- Built-in Network editor for creating and modifying graphs (either from PlanNET or from elsewhere).
- More powerful network expression mapper, with features such as filtering by expression, and with more coloring options.
- By focusing on the *experiment* instead of the *network* (as is the case for PlanNET) the user-interface allows for multiple visualizations and tables that refer to a specific experiment.

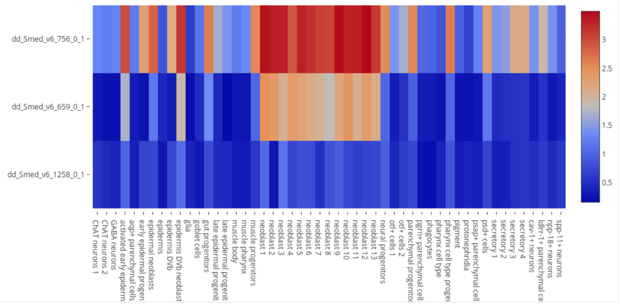
Suppl. Fig. 1: Example analysis performed on PIWIL genes expression on *"2018 Rajewsky Cell Atlas"* showing the different visualizations and features of PlanExp. These include, from top to bottom: Differential gene expression, GO enrichment analysis, gene expression plots, gene co-expression plots, t-SNE plots, identification of marker genes for cell clusters, gene co-expression network prediction, and network expression mapper.

PlanExp
Select Experiment and Database:
Summary - 2018 Rajewsky Cell Atlas
Differentially expressed genes
Volcano Plot
Perform Gene Ontology Enrichment Analysis
Marker Genes
Gene Expression
t-SNE Plot
Gene Co-expression Plot
Gene Co-expression Counts
Gene Co-Expression Network
Network Viewer

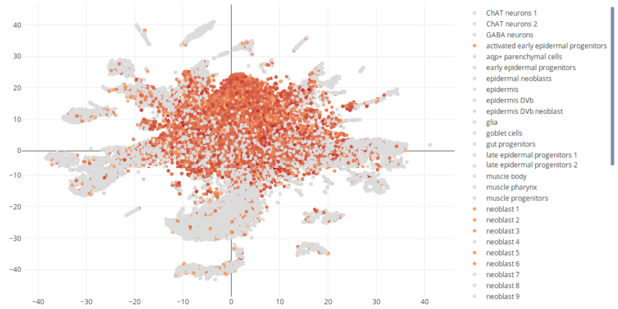
Example analysis performed on PIWIL genes expression on *"2018 Rajewsky Cell Atlas"* showing the different visualizations and features of PlanExp. These include, from top to bottom: Differential gene expression, GO enrichment analysis, gene expression plots, gene co-expression plots, t-SNE plots, identification of marker genes for cell clusters, gene co-expression network prediction, and network expression mapper.



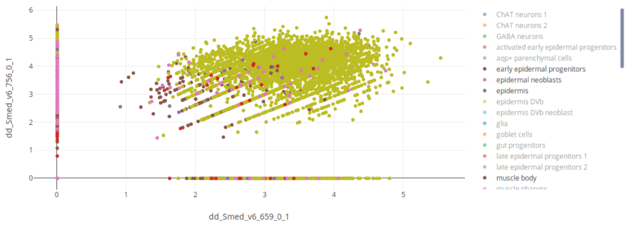
Suppl. Fig. 1A. Volcano Plot in *Differentially expressed genes* section. Each point represents a differentially expressed gene, plotted in two dimensions, with the x-axis being the logarithm of the fold change, and the y-axis the negative logarithm of the adjusted p-value. Genes with a higher y-value are more statistically significant, genes with a higher x-value are more over-expressed in the first condition of the selected comparison, while genes with a more negative x-value are more under-expressed.



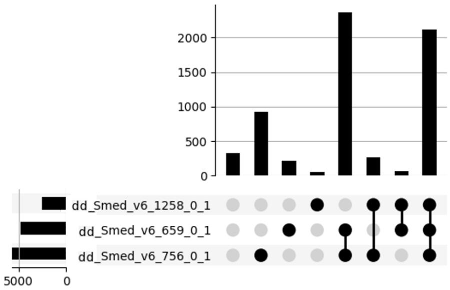
Suppl. Fig. 1B. Heatmap in *Gene Expression* section. Each row represents a gene, and each column a condition of the selected *Condition Type* (in this case, *Cluster*). The color of each cell is proportional to the average expression of each gene, in each given condition. Hovering over the cells will show the expression values, while clicking on the gene symbols will reveal a gene/contig information card.



Suppl. Fig. 1C. t-SNE Plot. Each point represents a cell plotted in a 2D space where the axis correspond to the dimensions computed by the t-SNE method. In this visualization, cells with similar expression profiles will tend to appear clustered together. However, due to the stochastic nature of the method, the coordinates and distances between cells might change with each subsequent computation of the dimensionality reduction, and thus, these coordinates might not be consistent with the same visualizations provided by the authors of the original single-cell experiments. Cells can be colored by *Condition Type* or, as is the case in this picture, by the expression of a given gene.

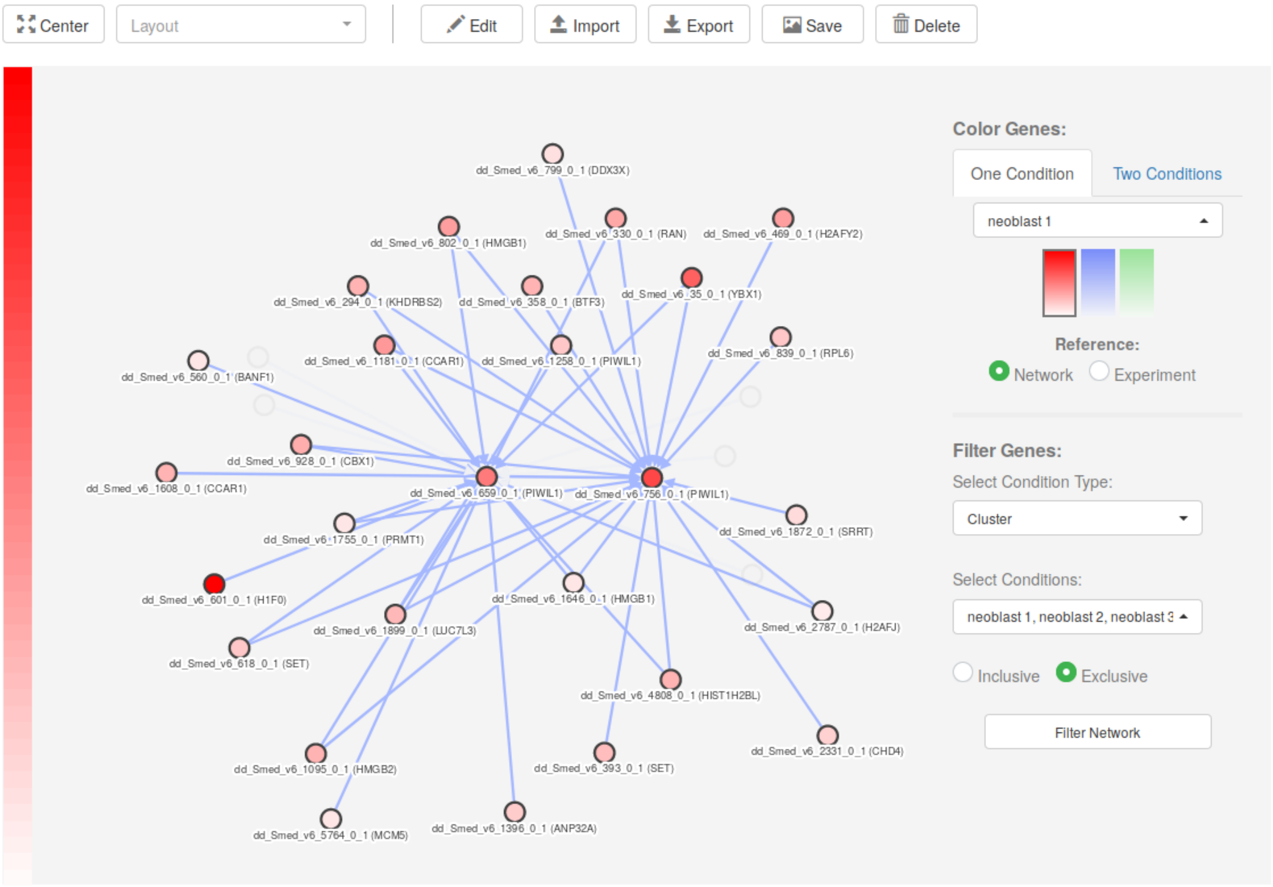


Suppl. Fig. 1D. Co-expression Plot in *Gene Co-expression Plot* section. Each point represents a cell plotted in a 2D space where the x-axis corresponds to the normalized expression of one gene (*dd_Smed_v6_659_0_1*) and the y-axis to the normalized expression of another gene (*dd_Smed_v6_756_0_1*). Cells on the upper-right express both genes simultaneously. Cells are colored according to the selected *Condition Type* (in this case, *Cluster*).



Suppl. Fig. 1E. Co-expression Counts in *Gene Co-expression Counts* section. Upset plot with the number of samples that express all the possible intersections of input genes simultaneously. Rows represent the genes, and columns the intersections. The y-axis is the number of samples of each intersection. For instance, there are 2,121 samples that express the three presented genes (last intersection).

Suppl. Fig. 1F. Network Viewer. Network visualization where nodes correspond to genes and edges to interactions. Each gene or transcript is labelled with its human homolog gene symbol in parenthesis when available. Users are able to color the genes by the expression in a particular condition or by the fold change resulting of the comparison of two conditions. Additionally, the network can be filtered by selecting one (or several) conditions in the *Filter Genes* section. Nodes will be greyed out if they are not expressed in the selected conditions. Hovering over the left color gradient legend will display the expression values ranges corresponding to each color, and will highlight the genes within that expression value range. The buttons on top of the visualization allow users to change the layout of the network; to import, export, and delete the graph; and to edit genes and interactions by using the newly implemented `Network Editor`.


Center
Layout
Edit
Import
Export
Save
Delete
dd_Smed_v6_799_0_1 (DDX3X)
dd_Smed_v6_802_0_1 (HMGB1)
dd_Smed_v6_330_0_1 (RAN)
dd_Smed_v6_469_0_1 (H2AFY2)
dd_Smed_v6_294_0_1 (KHDRBS2)
dd_Smed_v6_358_0_1 (BTF3)
dd_Smed_v6_35_0_1 (YBX1)
dd_Smed_v6_560_0_1 (BANF1)
dd_Smed_v6_1181_0_1 (CCAR1)
dd_Smed_v6_1258_0_1 (PIWIL1)
dd_Smed_v6_839_0_1 (RPL6)
dd_Smed_v6_928_0_1 (CBX1)
dd_Smed_v6_1608_0_1 (CCAR1)
dd_Smed_v6_659_0_1 (PIWIL1)
dd_Smed_v6_756_0_1 (PIWIL1)
dd_Smed_v6_1755_0_1 (PRMT1)
dd_Smed_v6_1872_0_1 (SRRT)
dd_Smed_v6_601_0_1 (H1F0)
dd_Smed_v6_1646_0_1 (HMGB1)
dd_Smed_v6_1899_0_1 (LUC7L3)
dd_Smed_v6_2787_0_1 (H2AFJ)
dd_Smed_v6_618_0_1 (SET)
dd_Smed_v6_4808_0_1 (HIST1H2BL)
dd_Smed_v6_1095_0_1 (HMGB2)
dd_Smed_v6_393_0_1 (SET)
dd_Smed_v6_2331_0_1 (CHD4)
dd_Smed_v6_5764_0_1 (MCM5)
dd_Smed_v6_1396_0_1 (ANP32A)
Color Genes:
One Condition
Two Conditions
Reference:
Filter Genes:
Filter Network

Color Genes: One Condition: neoblast 1

Reference: ● Network ○ Experiment

Filter Genes:

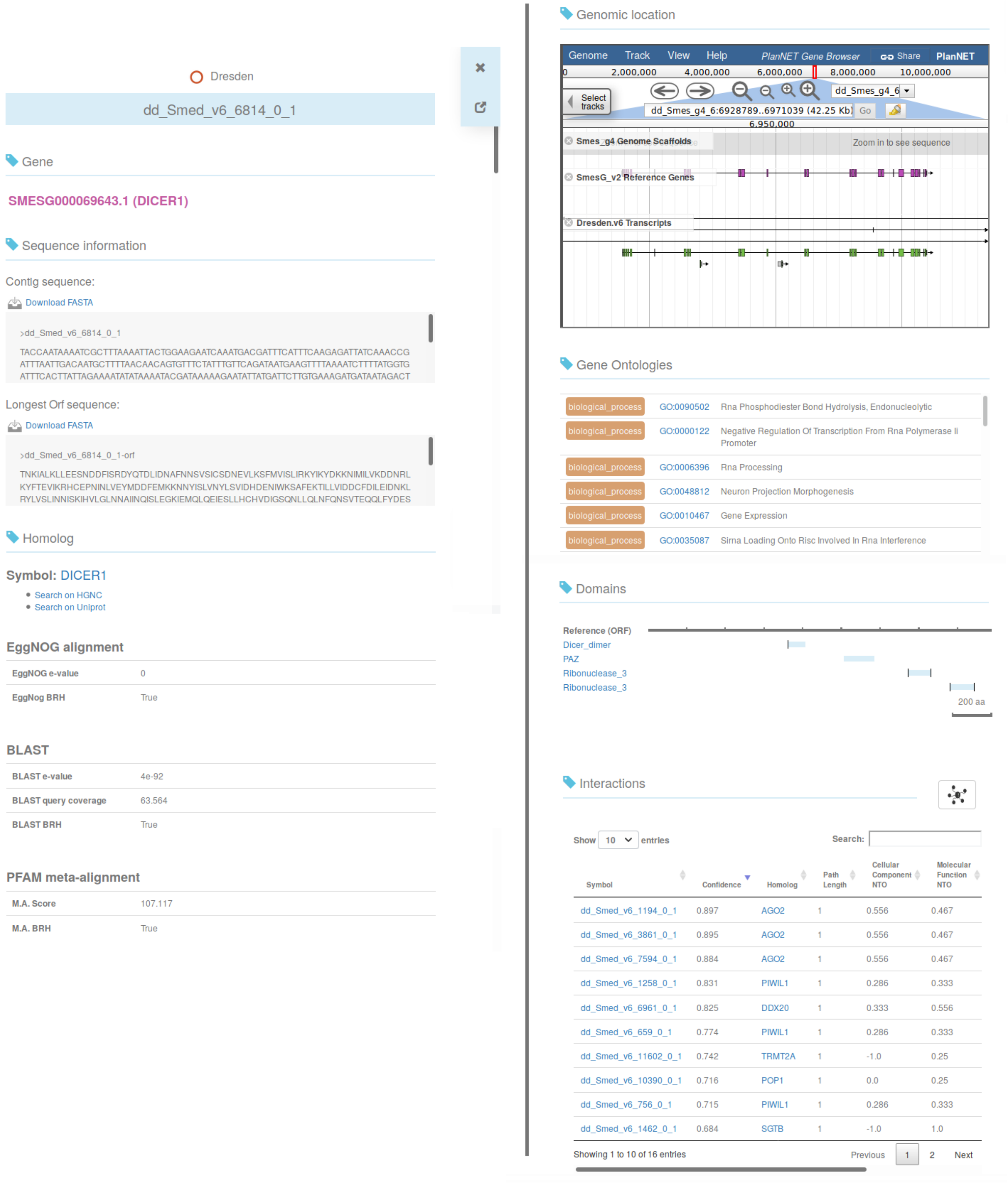
Select Condition Type: Cluster

Select Conditions: neoblast 1, neoblast 2, neoblast 3

○ Inclusive ● Exclusive

For further information refer to the PlanExp tutorial page:
https://compgen.bio.ub.edu/PlanNET/tutorial#planexp

Suppl. Fig. 2: **Gene card overlay** available on `PlanExp` when clicking on transcript or gene identifiers. All the sequence annotations, the genomic location, and protein-protein interactions for the annotated transcripts and related proteins are accessible through these gene cards, which also allow users to navigate through other `PlanNET` sections such as `NetExplorer`.



○ Dresden

dd_Smed_v6_6814_0_1

Gene

SMESG000069643.1 (DICER1)

Sequence information

Contig sequence:

Download FASTA

```
>dd_Smed_v6_6814_0_1
TACCAATAAAATCGCTTTAAAATTACTGGAAGAATCAAATGACGATTTCATTTCAAGAGATTATCAAACCG
ATTTAATTGACAATGCTTTTAACAACAGTGTTTCTATTTGTTCAGATAATGAAGTTTTAAAATCTTTTATGGTG
ATTTCACTTATTAGAAAATATATAAAATACGATAAAAAGAATATTATGATTCTTGTGAAAGATGATAATAGACT
```

Longest Orf sequence:

Download FASTA

```
>dd_Smed_v6_6814_0_1-orf
TNKIALKLLEESNDDFISRDYQTDLIDNAFNNSVSICSDNEVLKSFMVISLIRKYIKYDKKNIMILVKDDNRL
KYFTEVIKRHCEPNINLVEYMDDFEMKKNNYISLVNYLSVIDHDENIWKSAFEKTILLVIDDCFDILEIDNKL
RYLVSLINNISKIHVLGLNNAIINQISLEGKIEMQLQEIESLLHCHVDIGSQNLLQLNFQNSVTEQQLFYDES
```

Homolog

Symbol: DICER1

- Search on HGNC
- Search on Uniprot

EggNOG alignment

EggNOG e-value	0
EggNog BRH	True

BLAST

BLAST e-value	4e-92
BLAST query coverage	63.564
BLAST BRH	True

PFAM meta-alignment

M.A. Score	107.117
M.A. BRH	True

Genomic location

Gene Ontologies

biological_process	GO:0090502	Rna Phosphodiester Bond Hydrolysis, Endonucleolytic
biological_process	GO:0000122	Negative Regulation Of Transcription From Rna Polymerase Ii Promoter
biological_process	GO:0006396	Rna Processing
biological_process	GO:0048812	Neuron Projection Morphogenesis
biological_process	GO:0010467	Gene Expression
biological_process	GO:0035087	Sirna Loading Onto Risc Involved In Rna Interference

Domains

Interactions

Show 10 entries Search:

Symbol	Confidence	Homolog	Path Length	Cellular Component NTO	Molecular Function NTO
dd_Smed_v6_1194_0_1	0.897	AGO2	1	0.556	0.467
dd_Smed_v6_3861_0_1	0.895	AGO2	1	0.556	0.467
dd_Smed_v6_7594_0_1	0.884	AGO2	1	0.556	0.467
dd_Smed_v6_1258_0_1	0.831	PIWIL1	1	0.286	0.333
dd_Smed_v6_6961_0_1	0.825	DDX20	1	0.333	0.556
dd_Smed_v6_659_0_1	0.774	PIWIL1	1	0.286	0.333
dd_Smed_v6_11602_0_1	0.742	TRMT2A	1	-1.0	0.25
dd_Smed_v6_10390_0_1	0.716	POP1	1	0.0	0.25
dd_Smed_v6_756_0_1	0.715	PIWIL1	1	0.286	0.333
dd_Smed_v6_1462_0_1	0.684	SGTB	1	-1.0	1.0

Showing 1 to 10 of 16 entries

Previous 1 2 Next

Suppl. Table 1: **Comparing gene expression visualization** functionality of `PlanExp` with respect other planarian specific transcriptomic tools. Only those features relevant to the visualization and contextualization of RNA-seq and scRNA-seq experiments are listed in the table. Other applications may provide non-related functionalities, such as programatic access (API), `BLAST` search forms, etc.

Feature	**PlanExp**	**PlanMine**[1]	**Reddien**[2]	**Rajewsky**[3]
			2018	**2018**
Multiple experiments	✓	✓	✗	✗
Genomic context	✓	✓	✗	✗
Sequence annotations	✓	✓	✗	✗
Heatmaps	✓	✗	✗	✗
Violin plots	✓	✗	✗	✓
Multi-gene plots	✓	✗	✗	✗
Dynamic line/bar plots	✓	✓	✗	✗
Cell embedding plots	✓	✓	✓	✓
Dynamic cell embedding plots	✓	✗	✗	✗
Cell lineages	✗	✓	✗	✓
Mapping expression to networks	✓	✗	✗	✗
Differential gene-expression tables	✓	✗	✓	✓
Cell gene co-expression	✓	✗	✓	✗
Selection of experimental conditions to plot	✓	✗	✗	✗
Gene Ontology enrichment analysis	✓	✗	✗	✗
Body expression pattern visualization	✗	✗	✓	✗

[1] Rozanski, A. *et al.* 2018: http://planmine.mpi-cbg.de
[2] Fincher, C.T *et al.* 2018: https://digiworm.wi.mit.edu/
[3] Plass, M. *et al.* 2018: https://shiny.mdc-berlin.de/psca/

2 Integration of RNA-seq experiments

2.1 RNA-seq experiments

Differential gene expression (DGE) data for the experiments *"2011 Bartscherer Time-course"* (Sandmann *et al.*, 2011) and *"2015 Abril D.G.E"* (Rodríguez-Esteban *et al.*, 2015) were directly downloaded from the supplementary material of their respective publications. However, the complete DGE data for "2012 Pearson Stem Cells" and *"2013 Aboobaker Time-course"* (Kao *et al.*, 2013) were not available for download. As such, these had to be re-computed.

Unfortunately, due to the lack of replicates of *"2012 Pearson Stem Cells"*, we were unable to compute the DGE tables, and while the experiment is available on PlanExp for exploration, no DGE data will be shown.

For *"2013 Aboobaker Time-course"*, no raw counts data were available, and only counts-per-million (CPM) could be obtained. As such, we had to use the `limma-trend` pipeline (Law *et al.*, 2014) for computing DGE from normalized values, instead of relying on other negative binomial methods that require raw counts such as `edgeR` (McCarthy *et al.*, 2009) or `DESeq2` (Love *et al.*, 2014). Lowly expressed genes were already filtered by the authors of the original publication, removing all transcripts with less than fifty reads. For further details, refer to the original publication.

2.2 Single-Cell RNA-seq experiments

Expression data for the two single-cell experiments available on PlanNET, namely *"2018 Rakewsky Cell Atlas"* (Plass *et al.*, 2018) and *"2018 Reddien Cell Atlas"* (Fincher *et al.*, 2018), were downloaded directly from the websites provided by the authors.

In the case of *"2018 Rajewsky Cell Atlas"*, the 50 principal components resulting from the dimensionality reduction applied in their work were downloaded directly, and used for the t-SNE computation we performed. Differences between our t-SNE plot and the one available at https://shiny.mdc-berlin.de/psca/ can be attributed to the stochastic nature of the procedure. The expression counts were normalized using `Seurat`'s (Butler *et al.*, 2018) function `"NormalizeData"`, and scaled using the `"ScaleData"` function with `"nUMI"` as variable to regress. Markers were found with the `"FindAllMarkers"` function by employing the `"roc"` test, and differential expression between clusters was computed using the `"FindMarkers"` function, with a p-value threshold of 10^-5, a log fold change threshold of 0.2, and the `"min.pct"` parameter set to 0.2.

On the other hand, for the experiment *"2018 Reddien Cell Atlas"*, only the raw counts matrix and cell-cluster assignments were available for download. Data were normalized using `Seurat`'s `"NormalizeData"` function, and variable genes were found with the `"FindVariableGenes"` function, with parameters `"mean.function"` set to `"ExpMean"`, `"dispersion.function"` set to `"LogVMR"`, `"x.low.cutoff"` set to 0.2, `"x.high.cutoff"` set to 15, and `y.cutoff` set to -0.5. All parameters were set according to the ones used by the original work. Scaling was performed using the `"ScaleData"` function, regressing by the `"nUMI"`s. The dimentionality reduction was performed by PCA, using the computed variable genes, and keeping the first 150 principal components for computing the t-SNE representation.

2.3 Protocols availability

All the steps necessary to reproduce our analyses are available at https://github.com/scastlara/PlanExp-protocols, and a summary table of all experiments is shown in Supplementary Table 2. All input files are also provided on the github site, either directly or through a script that downloads them from elsewhere.

Suppl. Table 2: Summary of experiments available in PlanExp.

Experiment Name	GEO accession	Design	Samples	DGE	P-value	Description
2011 Bartscherer Time-course Sandmann *et al.* 2011	GSE30996	$\sim Time$	18	Downloaded	0.001	One-factor time course analysis of the anterior section of planarians after head amputation.
2012 Pearson Stem Cells Labbé *et al.* 2012	GSE37910	$\sim CellType$	3	N.A.	N.A.	Differential gene expression of Stem Cells, Progeny and differentiated cells.
2013 Aboobaker Time-course Kao *et al.* 2013	–	$\sim Section + Time$	28	Computed	0.01	Time course of anterior and posterior regions of planarians after head or tail amputation.
2015 Abril D.G.E. Rodríguez-Esteban *et al.* 2015	GSE51681	$\sim CellType$	3	Downloaded	0.001	Digital gene expression analysis mapped over several planarian transcriptomes to identify Neoblast specific genes.
2018 Rajewsky Cell Atlas Plass *et al.* 2018	GSE103633	$\sim Cluster + Cluster : Day$	21,612	Computed	10^{-5}	Single-Cell RNA-seq analysis that includes a non-supervised identification of cell types and a time-course differential expression analysis performed over each identified cell-type.
2018 Reddien Cell Atlas Fincher *et al.* 2018	GSE111764	$\sim Cluster$	50,562	Computed	10^{-5}	Single-Cell RNA-seq that includes a non-supervised identification of cell types.

3 Gene co-expression network

A genetics interactions co-expression network was computed for *"2018 Rajewsky Cell Atlas"* (Plass *et al.*, 2018) and *"2018 Reddien Cell Atlas"* (Fincher *et al.*, 2018) single cell experiments. The networks were predicted using the `GRNBoost` program of the `SCENIC` pipeline (Aibar *et al.*, 2017), which is a variant of the `GENIE3` program (Huynh-Thu *et al.*, 2010). These interactions provide both the most computationally expensive step of the `SCENIC` pipeline (which its results can be fed to the pipeline to predict genetic interactions from single-cell experiments), and gene co-expression relationships between genes in both experiments. The latter might reveal genes that belong to the same signaling pathways, genes co-expressed in the same cell types or that are regulated by the same transcription factors, putative protein-protein interactions, and direct or indirect genetic interactions.

The input for `GRNBoost` was a matrix of genes x cells with raw counts, and a list of putative transcription factors retrieved from `PlanNET` by looking for the contigs with an annotated Gene Ontology present in Supplementary Table 3. The list of regulators was splitted in chunks of 20 regulators to reduce the memory footprint of the `GRNBoost` program. Both analyses were run on a machine with 64 CPU cores and 512 GB of RAM.

In order to filter the results of `GRNBoost`, we chose the same set of filters applied by the `SCENIC` pipeline, but instead of generating three co-expression network sets for each analysis, we joined all the filters to produce a unique and more stringent network. These filters are:

1. Keep only edges with a weight > 0.005.
2. Keep only the best 50 targets for each regulator.
3. Keep only the best 50 regulators for each target.

All the steps to reproduce the analyses, as well as all the input files and the predicted networks can be found on the following github site: https://github.com/scastlara/PlanExp-protocols.

Suppl. Table 3: Gene Ontologies used for the selection of regulators in the gene co-expression network prediction with GRNBoost.

Accession	Name	Description
GO:0003677	DNA binding	Any molecular function by which a gene product interacts selectively and non-covalently with DNA (deoxyribonucleic acid).
GO:0043565	sequence-specific DNA binding	Interacting selectively and non-covalently with DNA of a specific nucleotide composition, e.g. GC-rich DNA binding, or with a specific sequence motif or type of DNA e.g. promotor binding or rDNA binding.
GO:0006355	regulation of transcription	Any process that modulates the frequency, rate or extent of cellular DNA-templated transcription.
GO:0045893	positive regulation of transcription	Any process that activates or increases the frequency, rate or extent of cellular DNA-templated transcription.
GO:0045892	negative regulation of transcription	Any process that stops, prevents, or reduces the frequency, rate or extent of cellular DNA-templated transcription.
GO:0001217	DNA-binding transcription repressor activity	Interacting selectively and non-covalently with a specific DNA sequence in order to stop, prevent, or reduce the frequency, rate or extent of transcription by a RNA polymerase.
GO:0140110	transcription regulator activity	A molecular function that controls the rate, timing and/or magnitude of transcription of genetic information. The function of transcriptional regulators is to modulate gene expression at the transcription step so that they are expressed in the right cell at the right time and in the right amount throughout the life of the cell and the organism.
GO:0001216	DNA-binding transcription activator activity	Interacting selectively and non-covalently with a specific DNA sequence in order to activate or increase the frequency, rate or extent of transcription by a RNA polymerase.
GO:0000981	DNA-binding transcription factor activity, RNA polymerase II-specific	A protein or a member of a complex that interacts selectively and non-covalently with a specific DNA sequence (sometimes referred to as a motif) within the regulatory region of a RNA polymerase II-transcribed gene to modulate transcription. Regulatory regions include promoters (proximal and distal) and enhancers. Genes are transcriptional units.
GO:0001217	DNA-binding transcription repressor activity	Interacting selectively and non-covalently with a specific DNA sequence in order to stop, prevent, or reduce the frequency, rate or extent of transcription by a RNA polymerase.
GO:0098531	ligand-activated transcription factor activity	A DNA-binding transcription factor activity that is directly regulated by binding of a ligand to the protein with this activity. Examples include the lac and trp repressors in E.coli and many steroid hormone receptors.
GO:0005667	transcription factor complex	A protein complex that is capable of associating with DNA by direct binding, or via other DNA-binding proteins or complexes, and regulating transcription.

Continued on next page

Continued from previous page

Accession	Name	Description
GO:0051090	regulation of DNA-binding transcription factor activity	Any process that modulates the frequency, rate or extent of the activity of a transcription factor, any factor involved in the initiation or regulation of transcription.
GO:0051091	positive regulation of DNA-binding transcription factor activity	Any process that activates or increases the frequency, rate or extent of activity of a transcription factor, any factor involved in the initiation or regulation of transcription.
GO:0043433	negative regulation of DNA-binding transcription factor activity	Any process that stops, prevents, or reduces the frequency, rate or extent of the activity of a transcription factor, any factor involved in the initiation or regulation of transcription.
GO:0034246	mitochondrial sequence-specific DNA-binding transcription factor activity	Interacting selectively and non-covalently with a specific DNA sequence in order to modulate transcription by mitochondrial RNA polymerase.

3.1 Results of the prediction

The analyses produced two sets of gene co-expression relationships between a putative regulator, and a target gene which its expression is affected by the expression of the regulator. These edges were annotated with the human-planarian homologs in `PlanNET`, and their respective planarian genes. Finally, all interactions in which both the regulator and the target gene appear in the same human `REACTOME` pathway were annotated with said `REACTOME` pathway identifier. The resulting table with all the results of both analyses is available on the protocol github site [4], and a summary of the results can be found in Supplementary Table 4. A comparison of both predictions can be found in Figure S3. Gene co-expression relationships that have been retrieved by both methods appear marked on the web-application to show the two independent evidence sources, and to help researchers assess the reliability of the gene co-expression relationships. While the weights reported by `GRNBoost` are similar in both predictions (Pearson's r = 0.6833), when applying filters to restrict the networks to a smaller and more stringent set of gene relationships, these relationships vary greatly between both sets of predictions. Thus, the best 50 targets for each regulator, and best 50 regulators for each target are different in each experiment, while the overall weights of `GRNBoost`'s predictions are similar. The effects of filters 2 and 3 on the shared relationships between both predictions can be seen in the increase of the number of shared relationships when relaxing the filters (∼6.2% when using 50 as the maximum number of targets/regulators, ∼12.8% when using 500). These differences could be explained by the different experimental designs of *"2018 Rajewsky Cell Atlas"* and *"2018 Reddien Cell Atlas"*, that result in a different cell type composition and different expression profiles.

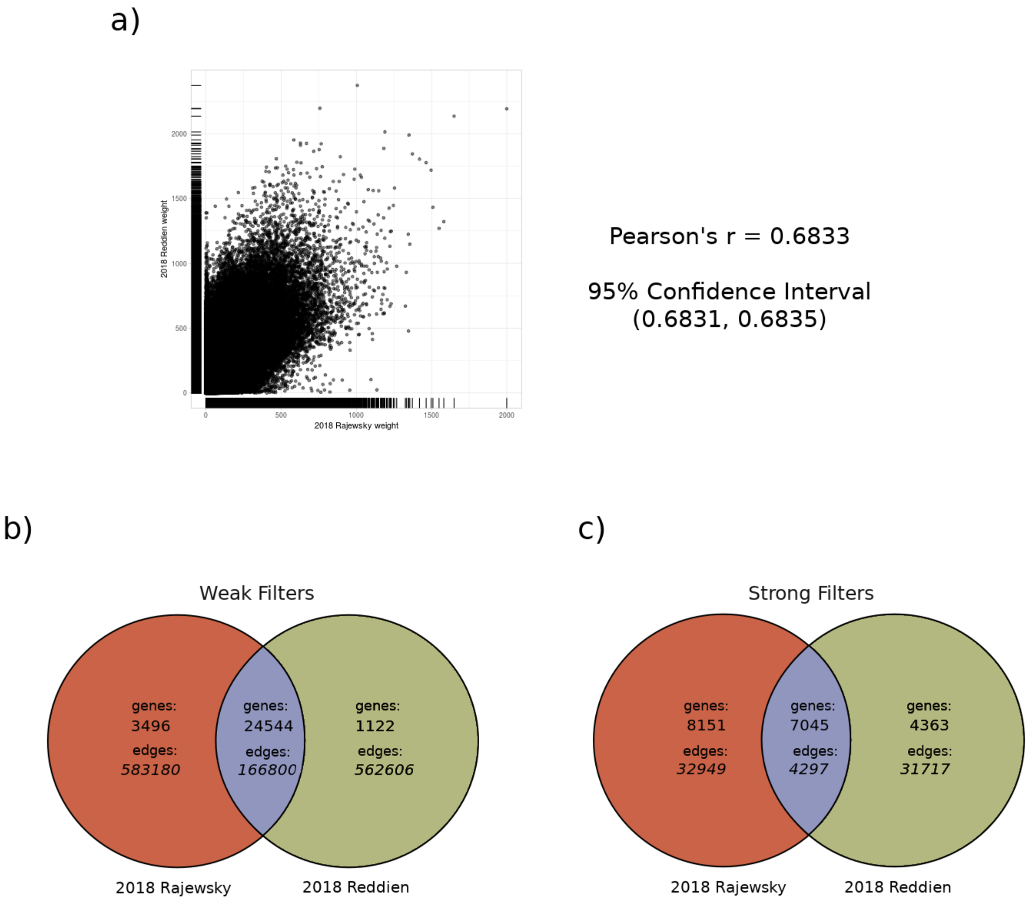
Suppl. Table 4: Summary of the gene co-expression networks predicted for the single cell RNA-seqs experiments available on PlanExp.

	2018 Rajewsky Cell Atlas	2018 Reddien Cell Atlas
Total Edges	37 246	36 014
Total Genes	15 196	11 408
Edges with Reactome	3635	3490
Genes in edges with Reactome	1855	1601
Total Reactome pathways	18 671	17 215
Unique Reactome pathways	792	777

Although `GENIE3/GRNBoost` was conceived as a program to retrieve genetic interactions from expression matrices, `GRNBoost` constitutes only the first step of the `SCENIC` pipeline to retrieve genetic interactions. As such, the predicted co-expression networks contain many types of relationships between genes. In Supplementary Table 5 we can find the genetic interactions between elements that belong to the Wnt signalling pathway as annotated in the `KEGG` database (`hsa04310`). Many key elements of the Wnt signalling pathway have been retrieved by `GRNBoost`, including: *wnt11-2* (Adell *et al.*, 2009; Gurley *et al.*, 2010; Sureda-Gómez *et al.*, 2015), which could act through the *wnt/βcat* pathway and affect *axinA* (Iglesias *et al.*, 2011); the relationship between *wnt11-5* (*wntP-2*) and *wnt11-2*, which are co-expressed in the same muscle cells (Witchley *et al.*, 2013); *wnt11-2* and *glyphican-1*, which were shown to be co-expressed in Cote *et al.* (Cote *et al.*, 2019); and a relationship between *wnt2* and a *wnt* inhibitor *sfrp1*, both of which are expressed in planarian head (Gurley *et al.*, 2010).

The protocol has also been able to identify protein-protein interactions described in humans (SMAD4, SMAD9), proteins that form complexes (MCM2, MCM5, MCM7), and genetic interactions described in humans

[4] https://github.com/scastlara/PlanExp-protocols



Suppl. Fig. 3: Comparison of gene co-expression network prediction for *"2018 Rajewsky Cell Atlas"* and *"2018 Reddien Cell Atlas"* single-cell RNA-seq experiments. **a)** Distribution of relationships weights between the two predictions, showing a moderate correlation between the confidence values reported by `GRNBoost` in both experiments for the same regulator-target gene relationships. **b) and c)** Venn diagram of the filtered gene co-expression networks for both experiments using 500 as the maximum number of targets and regulators (Weak Filters) or 50 (Strong Filters). Most gene relationships are unique to each prediction, while the number of shared genes in the networks is higher, but both predictions become more similar when relaxing the filters.

(PIM1, MAPK3). Some edges between genes that are relevant for the development of planarians that are expressed in the tail have also been retrieved. For instance, *fz4*, which has been used to characterize posterior identity (Sureda-Gómez *et al.*, 2015; Gurley *et al.*, 2008; Stückemann *et al.*, 2017); or *lox5a* which is expressed in the tail (Stückemann *et al.*, 2017), is controlled by the *wnt/βcat* pathway, and regulates posterior identity (Reuter *et al.*, 2015).

The complete list of predictions is available on `PlanExp` for them to be explored in the dynamic tables, as well as the network visualization, which allows researchers to map expression values of the experiment and to edit the interactions manually to produce relevant sub-networks.

Suppl. Table 5: Retrieved gene co-expression interactions for the experiment *"2018 Rajewsky Cell Atlas"* related to the Wnt signaling pathway (`KEGG hsa04310`). Contigs were annotated using the names listed in the PlanMine database, their related genes in PlanNET and PlanMine, and an `NCBI-BLASTX` search against the non-redundant human protein sequences database (`e-value` < `0.01`). The whole table containing all `GRNBoost` predictions is available at https://github.com/scastlara/PlanExp-protocols.

Regulator Contig	**Target Contig**	**Regulator Gene**	**Target Gene**	**Regulator Homolog**	**Target Homolog**
dd_Smed_v6_16209_0 (wnt11-2)	dd_Smed_v6_4900_0 (axinA)	SMESG000074928.1 (ISCW-ISCW004707)	SMESG000039725.1	WNT2	AXIN1
dd_Smed_v6_16209_0 (wnt11-2)	dd_Smed_v6_7326_0 (wntP-2)	SMESG000074928.1 (ISCW-ISCW004707)	SMESG000066476.1 (MS3_03642)	WNT2	WNT2 precursor
dd_Smed_v6_16209_0 (wnt11-2)	dd_Smed_v6_4154_0 (glypican-1)	SMESG000074928.1 (ISCW-ISCW004707)	SMESG000052359.1 (MS3_08283)	WNT2	GPC4
dd_Smed_v6_4639_0 (dvl-1)	dd_Smed_v6_2413_0	SMESG000049876.1 (DVL2)	SMESG000048614.1	DVL2	PRKACA
dd_Smed_v6_63520_0	dd_Smed_v6_9669_0	SMESG000052392.1 (LEF1)	SMESG000016284.1 (PTMB.308)	LEF1	PRKACB
dd_Smed_v6_8502_0	dd_Smed_v6_3221_0	SMESG000073604.1	SMESG000049234.1 (PLCB4)	DICER1	PLCB4
dd_Smed_v6_3757_0	dd_Smed_v6_891_0	SMESG000052999.1 (BTRC)	SMESG000001466.1 (MYCBP)	BTRC	MYCBP
dd_Smed_v6_2045_0	dd_Smed_v6_9037_0	SMESG000064848.1 (PPP3CA)	SMESG000077398.1 (PPP3CA)	PPP3CA	PPP3CA
dd_Smed_v6_13487_0 (wnt2-1)	dd_Smed_v6_13985_0 (sfrp1)	SMESG000002069.1 (WNT2B)	SMESG000029446.1 (MS3_08312)	WNT2	SFRP5
dd_Smed_v6_7173_0	dd_Smed_v6_5378_0	NA	SMESG000012014.1 (PPP3CA)	PPP3CB	PPP3CA
dd_Smed_v6_5531_0 (axinB)	dd_Smed_v6_5818_0	SMESG000025925.1	SMESG000025925.1	AXIN2	AXIN1
dd_Smed_v6_10098_0	dd_Smed_v6_3221_0	SMESG000013958.1	SMESG000049234.1 (PLCB4)	PLCB1	PLCB4
dd_Smed_v6_10098_0	dd_Smed_v6_7210_0 (fz-4-4)	SMESG000013958.1	SMESG000014322.1 (FZD4)	PLCB1	FZD4
dd_Smed_v6_10098_0	dd_Smed_v6_4244_0	SMESG000013958.1	SMESG000039854.1 (PRKCA)	PLCB1	PRKCA

Suppl. Table 6: Predicted target genes for *soxB1-2* (*dd_Smed_v6_8104_0*) in the computed gene co-expression network shown to be differentially expressed in Ross *et al.* 2018. When a target gene was predicted in both experiments (*”2018 Rajewsky Cell Atlas”* and *”2018 Reddien Cell Atlas”*), the score column corresponds to the mean score of both predictions as reported by `GRNBoost`.

Target Contig	Target name	Predicted in	Score
dd_Smed_v6_9977_0	*pkd2l-2*	Both	316.381
dd_Smed_v6_716_0	–	Both	292.228
dd_Smed_v6_10282_0	*syne1*	Both	206.841
dd_Smed_v6_3175_0	*tetraspanin homolog*	Both	203.112
dd_Smed_v6_7815_0	*ftcd*	Both	194.429
dd_Smed_v6_13327_0	*pkd2-4*	Both	176.652
dd_Smed_v6_12955_0	*pkd2-1*	2018 Rajewsky Cell Atlas	99.540
dd_Smed_v6_3331_0	–	2018 Rajewsky Cell Atlas	89.740
dd_Smed_v6_10911_0	*zfp940*	2018 Rajewsky Cell Atlas	83.010
dd_Smed_v6_9135_0	*tlcd1*	2018 Rajewsky Cell Atlas	75.098
dd_Smed_v6_7064_0	–	2018 Rajewsky Cell Atlas	68.534
dd_Smed_v6_21965_0	*semal*	2018 Reddien Cell Atlas	464.168
dd_Smed_v6_12811_0	*stum*	2018 Reddien Cell Atlas	350.143
dd_Smed_v6_3220_0	–	2018 Reddien Cell Atlas	330.581
dd_Smed_v6_3843_0	–	2018 Reddien Cell Atlas	329.732
dd_Smed_v6_3680_0	*ttpal homolog*	2018 Reddien Cell Atlas	319.375
dd_Smed_v6_3895_0	–	2018 Reddien Cell Atlas	313.629
dd_Smed_v6_7903_0	*centrin-1*	2018 Reddien Cell Atlas	309.129
dd_Smed_v6_5307_0	*cyp1a1 homolog*	2018 Reddien Cell Atlas	300.328
dd_Smed_v6_116_0	*dynll2*	2018 Reddien Cell Atlas	292.844
dd_Smed_v6_14226_0	*eml-1*	2018 Reddien Cell Atlas	291.754
dd_Smed_v6_5346_0	*rsph4A*	2018 Reddien Cell Atlas	287.853
dd_Smed_v6_10460_0	*loxhd1*	2018 Reddien Cell Atlas	287.246

References

Adell, T., Salò, E., Boutros, M., and Bartscherer, K. (2009). Smed-Evi/Wntless is required for Îš-catenin-dependent and -independent processes during planarian regeneration. *Development*, **136**(6), 905–910.

Aibar, S., González-blas, C. B., Moerman, T., Huynh-thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-c., Geurts, P., Aerts, J., Oord, J. V. D., and Atak, Z. K. (2017). SCENIC : Single-cell regulatory network inference and clustering. *Nature Methods*, **14**(11), 1083–1086.

Butler, A. *et al.* (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, **36**, 411–420.

Castillo-Lara, S. and Abril, J. F. (2018). "plannet: homology-based predicted interactome for multiple planarian transcriptomes". *Bioinformatics*, **34**(6), 1016–1023.

Cote, L. E., Simental, E., and Reddien, P. W. (2019). Muscle functions as a connective tissue and source of extracellular matrix in planarians. *Nature Communications*, **10**(1), 1–13.

Fincher, C. T. *et al.* (2018). Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science*, **360**(6391).

Gurley, K. A., Rink, J. C., and Alvarado, A. S. (2008). Îš-catenin defines head versus tail identity during planarian regeneration and homeostasis. *Science*, **319**(5861), 323–327.

Gurley, K. A., Elliott, S. A., Simakov, O., Schmidt, H. A., Holstein, T. W., and Alvarado, A. S. (2010). Expression of secreted wnt pathway components reveals unexpected complexity of the planarian amputation response. *Developmental Biology*, **347**(1), 24 – 39.

Huynh-Thu, V. A. *et al.* (2010). Inferring regulatory networks from expression data using tree-based methods. *Plos One*, **5**(9), 1–10.

Iglesias, M., Almuedo-Castillo, M., Aboobaker, A., and Saló, E. (2011). Early planarian brain regeneration is independent of blastema polarity mediated by the wnt/Îš-catenin pathway. *Developmental Biology*, **358**(1), 68 – 78.

Kao, D. *et al.* (2013). The planarian regeneration transcriptome reveals a shared but temporally shifted regulatory program between opposing head and tail scenarios. *BMC Genomics*, **14**(1), 797.

Labbé, R. M. *et al.* (2012). A comparative transcriptomic analysis reveals conserved features of stem cell pluripotency in planarians and mammals. *Stem Cells*, **30**(8), 1734–1745.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, **15**(2).

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**(12), 1–21.

McCarthy, D. J. *et al.* (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

Plass, M. *et al.* (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, **360**(6391).

Reuter, H., März, M., Vogg, M. C., Eccles, D., Grífol-Boldú, L., Wehner, D., Owlarn, S., Adell, T., Weidinger, G., and Bartscherer, K. (2015). β-Catenin-Dependent Control Of Positional Information Along The AP body axis in planarians involves a teashirt family member. *Cell Reports*, **10**(2), 253–265.

Rodríguez-Esteban, G. *et al.* (2015). Digital gene expression approach over multiple RNA-Seq data sets to detect neoblast transcriptional changes in *Schmidtea mediterranea*. *BMC Genomics*, **16**(1), 361.

Ross, K. G., Molinaro, A. M., Romero, C., Dockter, B., Cable, K. L., Gonzalez, K., Zhang, S., Collins, E. M. S., Pearson, B. J., and Zayas, R. M. (2018). SoxB1 Activity Regulates Sensory Neuron Regeneration, Maintenance, and Function in Planarians. *Developmental Cell*, **47**(3), 331–347.e5.

Sandmann, T. *et al.* (2011). The head-regeneration transcriptome of the planarian *Schmidtea mediterranea*. *Genome Biology*, **12**(8), R76.

Stückemann, T., Cleland, J. P., Werner, S., Thi-Kim Vu, H., Bayersdorf, R., Liu, S. Y., Friedrich, B., Jülicher, F., and Rink, J. C. (2017). Antagonistic Self-Organizing Patterning Systems Control Maintenance and Regeneration of the Anteroposterior Axis in Planarians. *Developmental Cell*, **40**(3), 248–263.e4.

Sureda-Gómez, M. *et al.* (2015). Posterior wnts have distinct roles in specification and patterning of the planarian posterior region. *International Journal of Molecular Sciences*, **16**(11), 26543–26554.

Witchley, J. N., Mayer, M., Wagner, D. E., Owen, J. H., and Reddien, P. W. (2013). Muscle cells provide instructions for planarian regeneration. *Cell Reports*, **4**(4), 633–641.