Development of a "Contrast and Gene Set" Toolkit-Integrated Database Platform for Downstream Computational Analysis Supporting Biological Interpretation of Gene Expression Data

C Poussin[1], L Hermida[2], A Sewer[1], S Ansari[3], S Gubian[1], J Hoeng[1]

[1]Philip Morris International R&D, Philip Morris Products SA, Neuchâtel, Switzerland; [2]Hermida Consulting, Rüschlikon, Switzerland; [3]Philip Morris International R&D, Philip Morris Research Laboratories GmbH, Cologne, Germany

Introduction

In transcriptomics, the identification of differentially expressed genes (DEG) when studying effect(s)/contrast(s) of interest constitutes the central component for further downstream computational analysis (e.g., gene over-representation / enrichment analysis or reverse engineering) leading to mechanistic insights. Therefore, it is essential to:

1) adequately store contrast data

2) automatically extract gene sets from these DEG lists in order to efficiently support downstream activities and further leverage data on a long-term basis.

We report here the development of a Contrast and Gene Set toolkit-integrated database (DB) platform implementing the requirements (1 and 2) described above. The platform also incorporates computational tools for downstream bioinformatics analysis (e.g. GSEA, Gene Set Enrichment Analysis) using internal and external gene sets as *a priori* biological knowledge to support biological interpretation of data.

Keywords

Contrast: corresponds to a quantitative estimate of the differential effect between treatment and reference conditions from microarray data. Linear models are generally used to estimate the contrast with the associated moderated statistics and significance level. The contrast and information related to significance are converted to a tabular format in order to be stored in the database.

Gene Set: collection of genes representative of a biological process/pathway, of a perturbation of interest (e.g. chemical, genetic, specific condition,…), targets of a transcription factor, chromosomal location. These gene sets are usually used as *a priori* knowledge for downstream computational analysis (e.g. GSEA) necessary to support biological interpretation.

Toolkit: suite of customized computational tools fully wrapped in Galaxy (1), an open-source workflow management and data integration system.

Methods

Requirements

The development of a "Contrast and Gene Set" **database** [DATA STORAGE]

- Design of a MySQL relational database with different tables capturing contrast and gene set data as well as meta-information about the experiment and the DEG analysis.
- Implementation of a Perl API to import data and to query the database through web services.
- Definition of a simple and standard format to structure results generated by DEG analysis performed with limma (2), SAM (3) or other statistical approach(es).
- Writing of a R (4) library to :

[Import process]

a) gather meta-information from raw data repository database.

b) convert limma or SAM R objects (contrasts) into specific format incorporating the meta-data information to generate a contrast dataset.

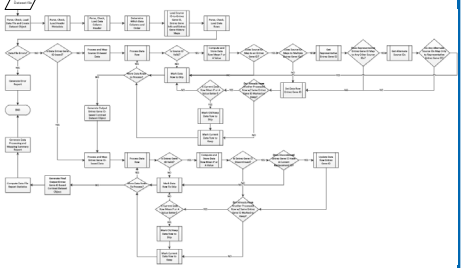
c) import contrast dataset to the database using the API.

[Export process]

a) read contrast dataset from the database using web services.

b) format results into R objects programmatically used for downstream analysis.

Figure 1. Constrast dataset import workflow with mapping and collapsing processes



- Definition of a mapping and collapsing methodology for automatic ID conversion of imported contrasts using the latest public annotations (NCBI Gene database).
- Establishment of rules for automatic extraction and storage of gene sets from imported contrasts.

The development of a **toolkit** [DATA PROCESSING]

- Integration of the GSEA Java application from the Broad Institute (5).
- Creation of a filtering tool that allows to query the database for gene sets based on specific criteria (e.g. organism, system, cell/tissue, stimulus, contrast type, gene regulation direction) and create a « gmt » file for GSEA on the fly.
- Implementation of various modules for a customized use of GSEA results (e.g. creation of a leading edge matrix for significant gene sets, merging of GSEA results from different contrasts and more).
- Design of a tool for importing and storing any type of external gene list of interest (e.g. gene signatures extracted from papers or identified by machine learning approaches) in the database, that can be further used as gene set for downstream analyses.

Full integration in **Galaxy** [TOOL INTEGRATION]

The elements of the toolkit allowing the access to the Contrast and Gene set DB as well as the collection of tools developed for downstream analyses (FDR threshold- and microarray platform-free analysis) are integrated in Galaxy and streamlined within defined workflow(s) (e.g. to run several GSEA in parallel).

Results

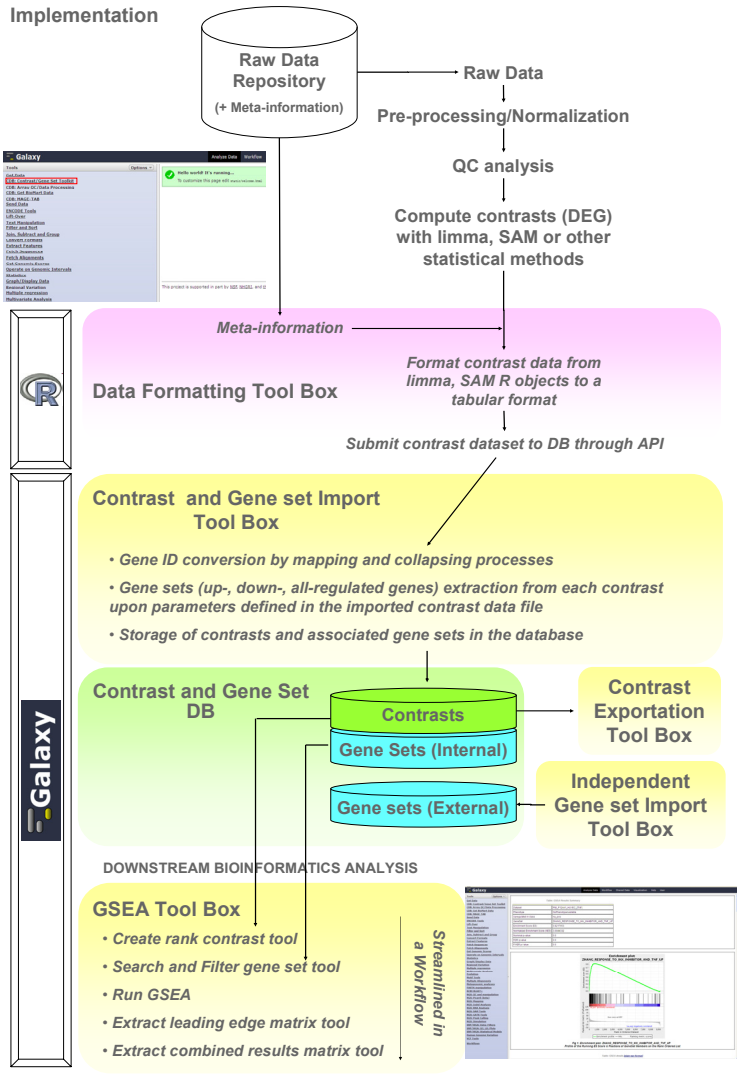


Figure 2. Schematic view and snapshots of the Contrast and Gene Set toolkit-integrated database platform

Conclusion

The Contrast and Gene Set toolkit-integrated database platform provides a unique and flexible environment to support downstream computational analyses enabling biological interpretation of data. The system has been designed in order to provide researchers with a simple, efficient, and extensible (integration of other "Omics" data types and development of new toolboxes) open source solution to store and exploit analyzed data in a sustainable manner. The software is planned to be released as a Galaxy module.

References

1. Goecks J et al, Genome Biology 2010; 11(8):R86
2. Smyth GK, Bioinformatics and Computational Biology Solutions using R and Bioconductor 2005; Eds Springer, New York, p.397-420
3. Tusher V et al, PNAS 2001, 98 (9): 5116-21
4. R Development Core Team; http://www.R-project.org
5. Subramanian et al, PNAS 2005; 102 (43): 15545-50

PMI RESEARCH & DEVELOPMENT

Philip Morris International Research & Development, Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland
T: +41 58 242 21 11, **F:** +41 58 242 28 11, **W:** www.philipmorrisinternational.com