OmixAnalyzer – A Web-Based System for Management and Analysis of High-Throughput Omics Data Sets

Thomas Stoltmann[1], Karin Zimmermann[1], André Koschmieder[1], and Ulf Leser

Humboldt-Universität zu Berlin, Germany
Department of Computer Science
{stoltman,zimmer,koschmie,leser}@informatik.hu-berlin.de

[1] These authors contributed equally to this paper.

Abstract. Current projects in Systems Biology often produce a multitude of different high-throughput data sets that need to be managed, processed, and analyzed in an integrated fashion. In this paper, we present the OmixAnalyzer, a web-based tool for management and analysis of heterogeneous omics data sets. It currently supports gene microarrays, miRNAs, and exon-arrays; support for mass spectrometry-based proteomics is on the way, and further types can easily be added due to its plug-and-play architecture. Distinct from competitor systems, the OmixAnalyzer supports management, analysis, and visualization of data sets; it features a mature system of access rights, handles heterogeneous data sets including metadata, supports various import and export formats, includes pipelines for performing all steps of data analysis from normalization and quality control to differential analysis, clustering and functional enrichment, and it is capable of producing high quality figures and reports. The system builds only on open source software and is available on request as sources or as a ready-to-run software image. An instance of the tool is available for testing at `omixanalyzer.informatik.hu-berlin.de`.

1 Introduction

Current projects following a Systems Biology approach to the study of biomedical phenomena typically produce a multitude of different high-throughput data sets. For instance, to study complex phenotypes such as cancer [6] and other genetic diseases [1], researchers analyze cellular samples at various levels, such as gene expression, protein expression, epigenetic status of regulatory elements in the genome, presence of differentially spliced protein isoforms, levels of metabolites etc. Managing and analyzing such diverse and heterogeneous data sets is a significant challenge; therein, analysis cannot stop at individual data sets, but needs to intelligently combine data generated by different methods [13]. In concrete projects, such technical and scientific issues are engraved by more social issues, such as highly different levels of proficiency of project members with modern methods in data analysis, problems in terms of data sharing, and unclear separations of concern between experimentalists and bioinformaticians [3].

The degree to which this harms projects varies with the size: Typically, small projects (2-4 groups) often have a clear separation of duties, a modest heterogeneity in data sets and less problems with data sharing. In large, international or national-wide projects such issues are usually resolved in a strict and manged manner by enforcing agreed policies and hierarchical organization [17]. However, the majority of typical Systems Biology projects probably are just in-between, uniting 10-20 groups with mixed expertise, having different perspectives on their subject, and gathering a heterogeneous set of experimental data. Further, these groups are often very sensitive to questions of data ownership and access authorization. Especially projects of this size can greatly benefit from central solutions to data management and analysis, as they may help to reduce cost, establish unified standard operating procedures, and foster data exchange [11].

In this paper, we present OmixAnalyzer, a system for managing, analyzing, sharing and visualizing heterogeneous -omics data sets for mid-sized projects. It uses a central database to store experimental results and sample metadata, features a full blown three-tier access rights management, and offers an intuitive web interface tailored towards biologists without specialized computer training. The OmixAnalyzer uses only open source components and builds on a plug-in architecture for adding novel data types with their individual metadata, internal data structures, and workflow-based analysis pipelines. For data analysis, the system executes configurable pipelines of R scripts, which makes changes in terms of individual analysis tools or the way algorithms are combined quite easy. Analysis results are downloadable in spreadsheet formats, can be used to generate publication-quality figures, and are stored by the system for later reuse.

We believe that the feature set of OmixAnalyzer is quite unique. For instance, systems such as Intermine [15] or Galaxy [5] are focused on genome sequences and do not target dynamic data sets typical for transcriptomics or proteomics. A variety of tools exist for transcriptome data [9], but none is suitable for our setting; for instance, using Chipster requires uploading data to a project-external server [8], while Mayday [2] clearly targets only expert bioinformaticians. Systems explicitly focusing on multiple omics data are, for instance, Vanted [12], which focuses on graph-based analysis and visualization, or Babelomics [10], which targets nation-wide projects and lacks a data access model. A number of other projects, such as SysmoSeek [16] or DIPSBC [4] only target data management but not data analysis or visualization. Overall, we are not aware of any other system that specifically targets mid-size Systems Biology projects.

2 System Architecture

The system architecture of the OmixAnalyzer follows the Model-View-Controller pattern in most cases and is segmented into three different layers: presentation, business, and data layer. The presentation layer contains all GUI elements, i.e. the web pages the user navigates when using the system. The business layer contains the operational logic, handles the communications between the front-end and the data layer, and manages the database transactions. Also in this

layer is the analysis back-end, which contains all analysis modules available in the OmixAnalyzer as well as the data crawlers. The data layer handles the data persistance and communications with the database.

Note that the software project completely relies on open source software. Figure 1 gives an overview of the system architecture and shows the software components used.

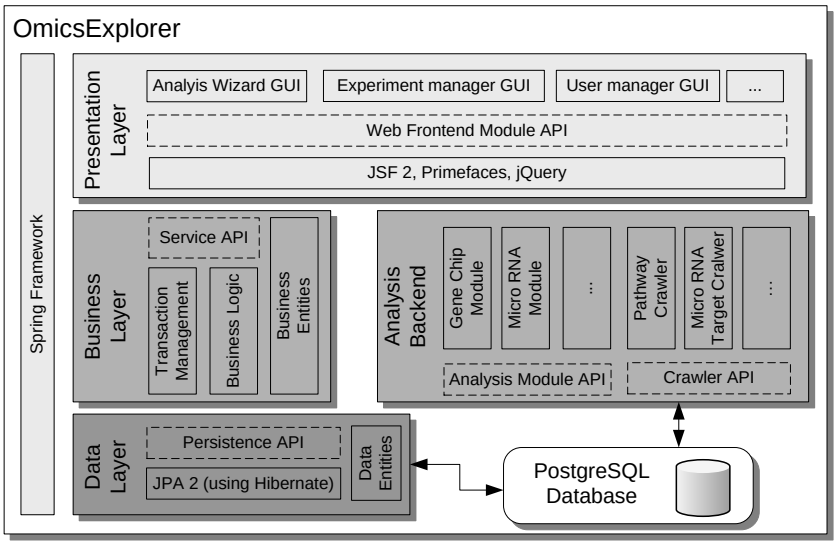


Fig. 1. System architecture of the OmixAnalyzer and software used.

The layers have been designed in a modularized manner, with each module covering a specific task. This way, new functionality can easily be added to the system by means of adding new modules. Details on how to add new modules (e.g. new data types) are covered in Section 5.

In the presentation layer, a module usually represents one or more web pages like user settings, experiment manager or analysis wizard.

The analysis backend consists of modules for data analysis and data crawlers. A data analysis module contains all available analysis workflows for a specific platform, with gene chips, micro RNA, and exon arrays currently available. A module exports description files for all supported analysis workflows providing information on required data and options for the workflow. For example, supported analysis workflows for the gene chip module are Clustering, Differential Analysis, Functional Analysis, Quality Control and Visualization. All currently implemented workflows use R as analysis backend, and methods for invoking R and handling jobs are available in the analysis API. However, new analysis modules are not required to use R as backend, but can also use Java, Perl, or other languages.

A data crawling module provides the system with data gathered from external resources. Crawlers can be run manually, or automatically at specific intervals. The crawlers provide annotation data used in the analysis workflows as well as pathway and gene identification data. Currently, the following data sources are supported: BioMART, KEGG (the last publicly available version), BioGRID, Reactome, TarBase, MirTarBase, MicroCosm, Miranda, PicTar and mapping files of different manufacturers. Additional crawling modules can easily be added.

The data layer is responsible for storing the data used in OmixAnalyzer. While some data formats like raw experimens or binary normalized expression data are stored in the file system, most data is stored in the relational PostgreSQL database. This includes all experiments and their annotations, users, groups and permissions, crawled external data, and analysis jobs and results.

3 Supported Data Analysis

A short overview of the functionality of the OmixAnalyzer demonstrates the variety of analysis possibilities. On top of supporting both the analysis of microarray and sequencing technologies, their joint analysis is a key feature of the presented software.

The analysis possibilities provided are organized in six workflows (see Figure 2(b)) according to their main topics for better overview and user guidance: Quality control, differential analysis, clustering, visualization, functional analysis and joint analysis.

Quality control implements commonly used plots such as boxplot, array-array correlation plot or PCA which enable the user to estimate the quality of the data and detect potential outliers.

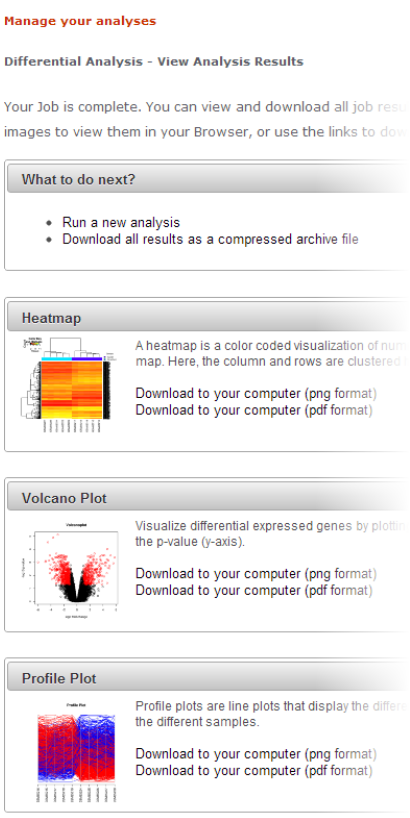
Visualization contains a potpourri of widely used plots providing a powerful tool for investigative and hypothesis generating analyses on the one hand as well as the visualization of results on the other hand.

Clustering analysis can be applied to samples or genes, or both. Hierarchical clustering using the complete linkage algorithm and euclidian distance provides a generic and powerful tool in class discovery.

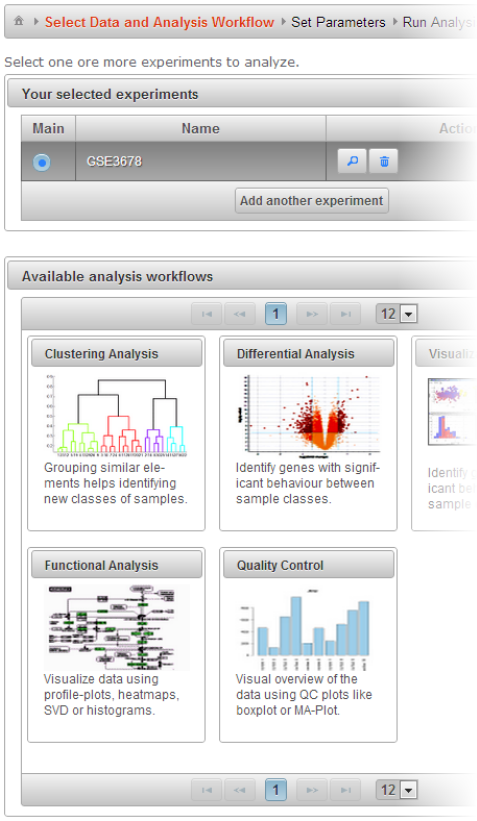
Differential analysis is a state-of-the-art implementation facilitating the detection of relevant genes based on commonly used criteria such as t-test based p-value, fold change or gene expression variance. For more than two groups, Anova is available.

A more bio-functional interpretation of the results can be obtained with Functional analysis. A pre-selected set of genes, coming from differential analysis or selected by other criteria, can be tested for significant enrichment of KEGG pathways or GO terms. This service is provided for all microarray-based data originated from chips supported by Bioconductor.

The joint analysis enables the user to compare and correlate data from different technologies as long as they contain samples which can be matched. Based on different criteria, subsets of data can be selected and combined for correlation and visualization.



2(a): Downloadable analysis results for differential analysis.



2(b): Workflows provided for the selected dataset

Each workflow leads to a result page (see Figure 2(a)) containing both a pdf file incorporating all analysis results as well as all single result files for selective download.

Additionally, every workflow provides highly specialized filtering options with respect to the data used for the concrete analysis. Clustering for example can be applied to samples, genes or both. Differential expression analysis will produce results for genes meeting certain criteria such as p-value or simply return a list and figures, and calculating the test statistic for all selected genes. A further filtering option is based on the role of genes. The user can select the later based on their membership in a certain pathway or by user-defined lists. Microarrays can even be selected based on their target genes. Due to the infinite number of combinations the filtering techniques make the OmixAnalyzer a very powerful tool by providing a maximum of creative analysis freedom.

A very straight forward but flexible and effective way to implement integrated data analysis is provided by the option to store gene lists at the end of

an analysis workflow and reuse the entities, i.e. genes, in another. If, for example epigenetic as well as expression data of the same conditions were stored in the OmixAnalyzer the set of differentially expressed genes could be tested for differential histone acetylation. Sample correspondence over experiments permits even more suffisticated options, such as miRNA to mRNA mapping by target as well as their correlation. The successful use of the OmixAnalyzer manifests in at least two publications [7, 14] accrued in the framework of the TRR54.

4 Web Interface

The web interface is entirely written in Java and XML and relies on JSF2, Primefaces, HTML5 and Ajax to offer a seamless user interface. Ajax is implemented through partial page rendering and partial page processing. Most of the client-side JavaScript code is provided through the frameworks used. The web interface is themable.

All administration of OmixAnalyzer can be done in the web interface by a user with administration privileges. The so-called administration area provides pages to manage entities within the system. The Experiment Manager Module allows to create, edit, delete and export experiments. The creation of new experiments can be done entirely using the web interface, including the upload of the experiment data. Experiments can be exported as a single zip file containing the data of the experiments in mage-tab format as well as the raw data (e. g. cel files of gene chip experiments). The user manager module contains facilities to create, edit and delete users and their roles. Additional modules are available to manage platforms, organisms, etc.

A more illustrative impression of the web interface shall be achieved by a walk-through to joint Analysis of gene chips and miRNA. The step preceding all workflows is the selection of the data set. Based on this decision, the workflows available for the selected data sets are displayed. In case of joint analysis, a miRNA and a gene chip data set need to have corresponding samples. After choosing the reviewed workflow a subset of groups can be selected. For a more specific analysis a filter can be applied to genes as well as to miRNAs. The emerging subset of samples and targets is now correlated based on an internally provided miRNA target mapping. Once the calculations are complete, the results page offers the view and download of all generated results including images in a pdf file, by single download or as an all-including compressed archive.

5 Availability and Extensibility

The OmixAnalyzer is available on request. We provide the full sources required to build and run the system, as well as a ready-to-run virtual machine image. Because the OmixAnalyzer is easy to administer but complex to install, this image can be used to set-up the system in only a few steps. The system is also available for trying out its features: `omixanalyzer.informatik.hu-berlin.de`

To run the OmixAnalyzer from the virtual machine image, VirtualBox (4.2 or higher) is required, which is freely available for most operating systems. The image contains all required software, and an installation guide for setting up the virtual machine is also provided.

Compiling and deploying OmixAnalyzer requires a few more steps. The system presupposes a Linux system with Oracle's Java JDK 7, PostgreSQL (9.0 or higher), R (2.15 or higher) and Bioconductor, ImageMagick and Maven 3. After modifying the central configuration file according to your needs, Maven will automatically download all required libraries and build a ready-to-run war file, which can be deployed on a Tomcat servlet container (version 7 or higher) or a Glassfish application server (3.1 or higher). The database can easily be installed using the supplied schema files. A detailed installation guide is also available.

With the modular architecture, new components can be added to the system with relatively small effort as plug-ins. These can add new GUI components to the system, provide more analysis methods for supported platforms, or even add a completely new platform type like for example proteomics.

To add a new GUI component, only the website layout and a Java class containing the GUI related functionality need to be written. For new platform types, a small module containing platform-specific information is required, along with some minor changes in the Data Layer. To enable users to analyze data of the new platform type, new analysis modules are required as well.

Analysis modules are the most complex type of plug-ins for OmixAnalyzer. A new analysis must contain meta information about the analysis workflow, including required input data and user options to parameterize the analysis. The actual analysis functionality can be supplied as Java, R, Perl or similar code routines.

6 Discussion

We present the OmixAnalyzer, a system for data management and analysis that specifically targets the needs of mid-size projects. For smaller groups of people, installing and maintaining a central solution like the OmixAnalyzer probably brings more burden than gain; in such settings, stand-alone systems, some of which are also available commercially, are usually the better choice. On the other hand, very large, international projects typically need a more flexible system than the OmixAnalyzer, with stronger capabilities in terms of scalable and distributed data analysis, support for automatic data and metadata ingestion, and possible direct links to LIMS systems. The OmixAnalyzer was designed to target projects just in-between, which typically can afford (limited) central and professional staff for running a data management solution.

The systems supports both computer-illiterate and savvy users. Biologists can take advantage of the built-in pre-processing, quality-control and data analysis options to work with their data and to generate results and figures for publications. Bioinformaticians may exploit the plug-in architecture to easily adapt

the system to specific needs or to add specific extensions, like novel analysis pipelines or support for other data types.

We are currently extending the system to also support proteomics data. Furthermore, we plan to provide a more comprehensive set of joint analysis methods.

7 Acknowledgements

We acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG) through Transregio TRR-54. We thank Y. Mayer, L. Sousa and M. Pichotta for contributions, and M. Hummel and the other members of the Transregio for feedback.

References

1. S. E. Baranzini, J. Mudge, J. C. van Velkinburgh, P. Khankhanian, I. Khrebtukova, et al. Genome, epigenome and rna sequences of monozygotic twins discordant for multiple sclerosis. *Nature*, 464(7293):1351–6, 2010.
2. F. Battke, S. Symons, and K. Nieselt. Mayday - integrative analysis for expression data. *BMC Bioinformatics*, 11:121, 2011.
3. E. Birney, T. J. Hudson, E. D. Green, C. Gunter, S. Eddy, et al. Prepublication data sharing. *Nature*, 461(7261):168–70, 2009.
4. F. Dreher, T. Kreitler, C. Hardt, A. Kamburov, R. Yildirimman, et al. Dipsbc-data integration platform for systems biology collaborations. *BMC bioinformatics*, 13(1):85, 2012.
5. J. Goecks, A. Nekrutenko, and J. Taylor. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010.
6. M. Joosten, V. Seitz, K. Zimmermann, A. Sommerfeld, E. Berg, et al. Histone acetylation and dna demethylation of t-cells result in an anaplastic large cell lymphoma-like phenotype. *Haematologica (accepted)*, 2012.
7. M. Joosten, V. Seitz, K. Zimmermann, A. Sommerfeld, E. Berg, et al. Histone acetylation and dna demethylation of t cells result in an anaplastic large cell lymphoma-like phenotype. *haematologica*, 98(2):247–254, 2013.
8. M. A. Kallio, J. T. Tuimala, T. Hupponen, P. Klemela, M. Gentile, et al. Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics*, 12:507, 2011.
9. A. Koschmieder, K. Zimmermann, S. Trissl, T. Stoltmann, and U. Leser. Tools for managing and analyzing microarray data. *Brief in Bioinf*, 13:46–60, 2012.
10. I. Medina, J. Carbonell, L. Pulido, S. Madeira, S. Goetz, et al. Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucl Acids Res*, 38 Suppl:W210–3, 2010.
11. N. W. Paton. Managing and sharing experimental data: standards, tools and pitfalls. *Biochem Soc Trans*, 36(Pt 1):33–6, 2008.
12. H. Rohn, A. Junker, A. Hartmann, E. Grafahrend-Belau, H. Treutler, et al. Vanted v2: a framework for systems biology applications. *BMC Syst Biol*, 6(1):139, 2012.
13. E. E. Schadt, M. D. Linderman, J. Sorenson, L. Lee, and G. P. Nolan. Computational solutions to large-scale data management and analysis. *Nat Rev Genet*, 11(9):647–57, 2010.

14. V. Seitz, P. Thomas, K. Zimmermann, U. Paul, A. Ehlers, et al. Classical hodgkin's lymphoma shows epigenetic features of abortive plasma cell differentiation. *haematologica*, 96(6):863–870, 2011.
15. R. N. Smith, J. Aleksic, D. Butano, A. Carr, S. Contrino, et al. Intermine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. 28(23):3163–5, 2012.
16. K. Wolstencroft, S. Owen, F. du Preez, O. Krebs, W. Mueller, et al. The seek: a platform for sharing data and models in systems biology. *Methods in enzymology*, 500:629, 2011.
17. W. Wruck, M. Peuker, and C. R. Regenbrecht. Data management strategies for multinational large-scale systems biology projects. *Brief in Bioinf*, 2012.