



Universiteit
Leiden
The Netherlands

Design and development of a comprehensive data management platform for cytomics: cytomicsDB

Larios Vargas, E.

Citation

Larios Vargas, E. (2015, November 25). *Design and development of a comprehensive data management platform for cytomics: cytomicsDB*. Retrieved from <https://hdl.handle.net/1887/36426>

Version: Not Applicable (or Unknown)

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/36426>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/36426> holds various files of this Leiden University dissertation

Author: Larios Vargas, Enrique

Title: Design and development of a comprehensive data management platform for cytomics : cytomicsDB

Issue Date: 2015-11-25

Summary

In biomedical research, the study of cellular systems on large scale is called Cytomics. High-Throughput screening (HTS) is one of the most common techniques for target validation; it allows testing a large number of chemicals against disease targets for identifying hits. Moreover, an HTS environment is composed by a diversity of components that have to face challenges in terms of managing large volume of data, data heterogeneity, no integrated complex hardware architecture, diversity of software tools, managing protocols, etc. Therefore, this current scenario is pushing forward the need to develop a comprehensive data management platform for experiments in Cytomics. This new approach should be capable to enhance: (1) exploration of large volumes of data, (2) interoperability, (3) understandability, and (4) sharing and reusability of the data.

In this thesis, in four chapters our research on the development of CytomicsDB is described. This constitutes a platform which integrates the whole HTS workflow into a single system. An automated HTS workflow has been proposed to speed up the processes from plate design to image and data analysis. Chapter 2 focuses on the architecture of CytomicsDB, which relies on a modern relational database system, capable to integrate to other legacy systems used in HTS environments or other external repositories for data validation. The architecture also promotes data sharing and collaboration using several security roles, which allow scientist to publish their data or grant access to other researchers to their own experiments.

In Chapter 3, we further elaborate our approach in the metadata management performed by CytomicsDB. A semantic layer has been built in order to facilitate the understandability and exploration of the large volume of images and metadata involved in HTS experiments and at the same time allowing scientists to integrate new tools and APIs taking care of the image and data analysis. These results will become part of the experiments metadata and will be available for semantic post analysis.

In Chapter 4, we introduce the validation process as performed in CytomicsDB. Management of the HTS information is one of the key challenges for drug discovery

and in order to ensure consistency, integrity and reliability of the data stored in the platform it is compulsory to perform a strict validation process in every stage of the HTS workflow. CytomicsDB facilitates this validation process using web services, which will prove each critical entry with an internal or external repository. The metadata that we store become a key parameter for performing further image/data analysis and drill down the results of different experiments datasets.

Chapter 5 describes the integration of CytomicsDB architecture to the Leiden Life Sciences Cluster (LLSC) and the environment for data processing provided by MonetDB. In this new environment the image analysis algorithms need to be adapted, and the performance of the resulting parallelized algorithms is evaluated. The steps involved in this stage are also further elaborated in a image and data analysis data flow.

The current challenges in HTS experiments are pushing forward the need to design and develop more complex platforms for *Cytomics*, which are capable to adapt to the continuous changes presented in its environment. CytomicsDB's architecture facilitates the integration with other systems involved in HTS experiments. Moreover, the metadata has been organized in order to have a common data model, which enhances understandability, collaboration and data sharing. In CytomicsDB, the validation of the metadata stored in the repository is considered a key process. The consistency of metadata is guaranteed by accessing external biological databases that are supported and maintained by the scientific community. Finally, our platform integrates the advantages of clustering computing for the image analysis stage and also the data processing functionality provided by MonetDB.R package to speed up the data analysis avoiding the movement of huge amounts of data for processing.