



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

Design and development of a comprehensive
data management platform for Cytomics

~ **CytomicsDB** ~

Enrique Larios Vargas

The research developed in this thesis was performed at the Imaging and BioInformatics group (Leiden Institute of Advanced Computer Science), the Division of Toxicology (Leiden Academic Center for Drug Research) and with the support of the Database Architecture Group (Centrum Wiskunde & Informatica - Amsterdam), the Netherlands.

This research project has been supported by the BAPE program, EMECW Lot 18, Erasmus Mundus.

Design and development of a comprehensive data management platform for Cytomics: CytomicsDB

Cover design by: Terry Zabel

ISBN: 978-94-6259-950-5

Copyright © 2015 by Enrique Larios Vargas. All rights reserved. No part of this thesis may be reproduced or transmitted in any form, by any means, electronic or mechanical, without prior written permission from the author.

Printed by Ipskamp, www.ipskampdrukkers.nl

Design and development of a comprehensive data management platform for Cytomics: CytomicsDB

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 25 November 2015
klokke 10.00 uur

door

Enrique Larios Vargas

geboren te Lima, Peru
in 1977

Samenstelling promotiecommissie:

promotor:	prof. dr. J. N. Kok	LIACS, Universiteit Leiden
co-promotor:	dr. Ir. F. J. Verbeek	LIACS, Universiteit Leiden
overige leden:	prof. dr. Lucie Kubínová	Czech Academy of Sciences, Prague,
		Czech Republic
	dr. Ying Zhang	CWI, Amsterdam
	prof. dr. Bob van der Water	LACDR, Universiteit Leiden
	prof. dr. Stefan Manegold	LIACS, Universiteit Leiden

I dedicate this thesis to my mother Luveslinda, for her sacrifices and her effort for giving me the best education and also to my five angels on heaven: Padre Vicente, Madre Teodora, my cousin Ivan, my father Paco and my grandmother Rodolfa.

Table of Contents

	Page
Contents	1
List of Figures	5
List of Tables	7
1 Introduction	9
1.1 Thesis scope	9
1.2 HTS experiments workflow	10
1.2.1 Experiment Design	10
1.2.2 Wet-lab	10
1.2.3 Image acquisition	11
1.2.4 Image analysis	12
1.2.5 Data analysis	13
1.3 Data management in Cytomics	14
1.3.1 Heterogeneity of data	15
1.3.2 Data discovery	16
1.3.3 Interoperability	17
1.3.4 Understandability	18
1.3.5 Data sharing and reusability	18
1.3.6 Transport of data	19
1.4 LIM Systems	20
1.4.1 LIMS environment	20
1.4.2 How a typical LIMS works	21
1.4.3 Benefits of LIMS	22
1.5 Thesis structure	23

2	CytomicsDB architecture	25
2.1	Introduction	26
2.2	Automated Workflow of the HTS experiments	27
2.2.1	Plate Layout Design Module	28
2.2.2	Image Analysis Module	28
2.2.3	Data Management Module	31
2.2.4	Data and Pattern Analysis Module	31
2.3	System Architecture of the HTS Analysis Platform	32
2.3.1	The Presentation Layer	32
2.3.2	The Service Layer	32
2.3.3	The Persistence Layer	33
2.3.4	The Repository Layer	33
2.3.5	Flow of Control	34
2.4	Database Design	35
2.4.1	Users and Experiment Sets	36
2.4.2	Plates and Wells	37
2.4.3	Raw Images	37
2.4.4	Results of Image Analysis	38
2.4.5	Results of Data Analysis	39
2.5	Related Work	39
2.6	Conclusions and Future work	40
3	Metadata Management	41
3.1	Introduction	41
3.2	Metadata organization in an HTS experiment	43
3.2.1	Project	43
3.2.2	Experiment	44
3.2.3	Plate-Wells	44
3.2.4	Raw Images	46
3.2.5	Measurements	46
3.3	Case Study in Endocytosis of EGFR: Identification of EGFR dynamics and classification of EGFR phenotype stages	47
3.3.1	Image Acquisition	47
3.3.2	Image Analysis	47
3.3.3	Data Analysis	50
3.4	Related work	50
3.5	Conclusions and Future Work	52
4	Metadata Validation	53
4.1	Introduction	53
4.2	Critical validation in an Automated HTS workflow	54
4.2.1	Design	55

4.2.2	Image Acquisition	55
4.3	Validation strategies	55
4.3.1	Strategies and algorithms	56
4.4	Implementation	58
4.4.1	Validation subjects	58
4.4.2	The validation workflow	61
4.4.3	The architecture	68
4.5	Analysis of conflict solving strategies	72
4.5.1	Criteria for selecting a strategy	72
4.5.2	Selection of strategies	76
4.6	Conclusions and Future work	78
5	Cluster Integration for Image processing and Pattern recognition	79
5.1	Introduction	79
5.2	The Leiden Life Sciences Cluster	80
5.3	Image Analysis	81
5.4	Data Analysis and Pattern recognition	82
5.5	Implementation	83
5.5.1	The architecture	83
5.5.2	The image and data analysis data flow	86
5.5.3	Image Analysis setup and Results	88
5.5.4	Data Analysis setup	90
5.6	Conclusions and Future work	92
6	Conclusions and Future work	93
6.1	Research questions and Problem statement	93
6.1.1	Designing a software architecture for cytomics	93
6.1.2	Addressing the metadata organization	94
6.1.3	Metadata validation	94
6.1.4	Speeding up data processing	95
6.1.5	Addressing the Problem Statement	95
6.2	Contributions	96
6.3	Future work	96
	References	99
	Summary	107
	Samenvatting	109
	List of Publications	113
	About the Author	115

List of Figures

1.1	Design of a 96 wells plate for an experiment using MCF7 wt cells	10
1.2	96 wells plate for HTS.	11
1.3	Microscopes for HTS.	12
1.4	Segmentation and subsequent image analysis	13
1.5	Data analysis	14
1.6	Hardware and Software components in an HTS workflow	15
1.7	Data Lifecycle in a workflow	16
1.8	Diagram of a typical LIMS environment	21
1.9	Sample work flow diagram	22
2.1	Automated workflow of an HTS Experiment	27
2.2	Web plate layout design GUI	28
2.3	Image and coarse regions	29
2.4	Using our image analysis solution, the phenotypic measurements of (a) live cells and (b) adhesion can be extracted	29
2.5	Phenotypic characterization of the Epidermal Growth Factor (EGF) treatment using a highly aggressive cancer cell line, the illustrated features are picked by branch-and-bound feature selection. The EGF-treated cell group shows a significant increased migration velocity	30
2.6	The HTS analysis platform architecture	32
2.7	Flow of control of the HTS platform	34
2.8	Database schema for Project Metadata	35
2.9	Database schema for Experiment Metadata	36
2.10	Database schema for Plate-Well Metadata	36
2.11	Database schema for Raw Images and Measurement Metadata	37
3.1	Structure of the metadata in an HTS experiment	44
3.2	Structure of the Experiment and Plate metadata	45

3.3	Structure of the Raw Images metadata	45
3.4	Structure of the Measurements metadata	46
3.5	Web plate layout	48
3.6	Example of set of images generated during the Image Analysis stage . .	49
4.1	Relationship between the entity, its attributes and fact values from data sources	56
4.2	Get candidates	62
4.3	Screening and Marking	63
4.4	Clean up	64
4.5	Update duplex marks	65
4.6	Components diagram of the validation workflow	69
4.7	Interfaces for validating a new siRNA entry	70
4.8	Validation result interface	70
5.1	CytomicsDB Architecture - Cluster Integration	84
5.2	Web Interface for Analysis	84
5.3	Image and Data analysis data flow	87
5.4	Segmentation results using up to 24 nodes	89
5.5	Partitioned concurrent tracking and join times	90
5.6	Comparison of results with three phenotypic groups	92

List of Tables

4.1	Query result after checking for <i>ETOPOSIDE</i> in the NCBI database. . . .	59
4.2	siRNA example provided by the researchers.	59
4.3	Query results from all external data sources.	60
4.4	Supported types of attributes in external Data Sources	63
4.5	siRNA example uploaded by the user	65
4.6	List obtained after first stage of siRNA validation	66
4.7	List obtained after second stage of siRNA validation	66
4.8	List of siRNAs candidates retrieved from the NCBI db.	67
4.9	Fourth step result - Matrix of similarities	67
4.10	Functions used to implement each strategy	72
4.11	Conflict handling functions and their applicable properties	73
4.12	Strategies and Functions available for siRNAs and Compounds	73
4.13	Strategies and their information availability	74
4.14	Strategies and criterias	78
5.1	Algorithms tested	89

