Information Architecture in High Throughput Screening

High Content Analysis, Architecture and Knowledge Discovery

Abstract. This paper describes a high throughput screening architecture for functional genomics screens that use high content methods. Case studies were performed using the Yin case study approach. Additionally a detailed model is provided using the Method Engineering methodology to identify deficiencies. This study shows that current architectures lack interchangeability and functionality.

Keywords: high throughput screening, architecture, workflow, high content analysis, laboratory information management system

1 Introduction: High Throughput Screening

It is simply not practical to investigate a great quantity of reagents manually, it would be error-prone [1, 2]. High Throughput Screening (HTS) is a process in which large libraries of chemical or biological reagents can be tested for activity in assays using automated methods [1]. It is an essential experimental instrument for the analysis of many biological processes and an intensifying domain for bioinformaticians. Experiments are conducted using 96, 384 or 1536 well-plates where cells or proteins and libraries of reagents are added according a specified protocol. Reagents are stored in micro plates or 2d-barcoded mini-tubes. High content analysis is a subset of HTS in which images of cells are acquired through the use of automated microscopy. Subsequent automated image analysis can be used to generate multi-parameter numerical data. Thus, screens generate large amounts of captured data which require further analysis for the identification of germane outliers, or hits [3]. HTS is widely used within multiple disciplines including drug discovery, functional genomics and toxicology [4 - 6]. Functional genomics screens frequently make use of RNA interference technology, (RNAi) that allows for the specific reduction of the function of individual genes. Insight can be gained where domain data sets are innovatively structured and integration of various data sets is instigated [4]. Here we focus on the use of HTS for RNAi screens combined with high content analysis.

1.1 Problem description

Little literature is present on how to capture knowledge efficiently using flexible IT in an academic setting in high throughput processes for drug discovery and functional genomics. Currently, it is unknown how the information flows and the system architecture of HTS can be optimized. Are the software packages fulfilling the needs of a HTS facility? To what extent are current software packages interoperable or ready for seamless communication and is the use of the software efficient in terms of automation, consistency, redundancy, time utilization and usability? This research contributes to this functional and technical research gap by investigating the following research question: ‘What is the most efficient information architecture for HTS?’

2 Background: LIMS Tooling

Processes such as quantification, normalization and laboratory information management or statistical analysis are usually supported by (open source) software. A small literature overview of commonly used applications within HTS facilities is provided here. It is stated by [7] that a high throughput facility needs a sophisticated informatics infrastructure to support high volumes of interleaved screening projects. Additionally it requires technologies and design techniques that are anticipated to support rapid adjustments of the software [8]. The particular problem for laboratories engaged in a wide range of high throughput activities is the shortage of cohesive and easy-to-use solutions [3]. Laboratory Information Management Systems (LIMS) are required for the collection, viewing and editing of experimental information. Genetic studies for example require high throughput and large size sample sets. These samples are usually gathered from different sources and time points [7]. This requires appropriate handling otherwise it can lead to amplified errors, decreased accessibility and efficiency of data. A Laboratory Information Management System (LIMS) is a solution for these issues. SLIMS [7] is an open source LIMS which is able to cope with information considering patient samples. However, the software contains functionality that is more widely useful and has proven out to be a supportive tool in biological samples and the plates, boxes or mini-tubes that are used for physical storage of the samples. In other words, to manage what is where. Screensaver is another open source Laboratory Information Management System (LIMS) [8]. Unlike SLIMS which was designed for the management of patient sample collections only, Screensaver was specifically designed for the management of HTS. Screensaver serves different users e.g. students, post-docs and managers and as well as library management also stores experimental results, screening workflows and metadata from screens. It allows for cross-screen comparisons, reagent cherry picking and heat maps for data visualization through a web-based interface. Web CellHTS2 [3] is a web-based application for analyzing high-throughput screening data from RNAi and compound screens, implemented in R/bioconductor [9]. Quantified data can be uploaded using the graphical user interface. Both sample and control-based normalization and data analysis such as B-score normalization or Loess regression results of analyses are saved as HTML and text files, zipped and then sent by email. The application runs on a JAVA web-server using AJAX technology. R-serve is used to interact among R and the JAVA application. Omero is open-source software for managing large-scale image storage, meta-data and non-image data developed in Python, JAVA and C++. The software comes with its own application programming interface (API) and offers a web-based interface. Data can be uploaded using Bio-Formats which is able to transform >100 different file formats to a common data model. Then, Omero stores the data mapped on the Open Microscopy Environment

(OME) data model [10]. Omero supports interaction with third-party software e.g. Mat-Lab and CellProfiler using a JAVA gateway or the API. CellProfiler [11-14] is an application for extracting numerical data from microscopy images by applying automated image analysis. Cell profiler 2.0 is an open source object-oriented stand-alone application built for Windows, Linux and MacOS. A plugin is integrated to support a pipeline for ImageJ. Furthermore, CellProfiler supports a large extent of image formats via OME.

3 Method: A multiple case study approach

According to Yin (1984), there are three types of case study; explanatory, descriptive or exploratory [15]. The nature of this study is exploratory, followed by the four stages of [15]; design a case study, conduct the case study, analyze the case study evidence and develop the conclusions, recommendations and implications. Second, [15] described four different categories of case studies in a two by two matrix (see [15] for additional information). The horizontal line outlines a single or multiple-case design. The vertical line sketches a holistic or embedded approach. Therefore this study would fit best in type three, which implies a holistic or single unit of analysis and infers multiple case designs. In addition, [15] recommends four validity criteria for empirical research; Construct validity, which means that the measured concept is measured in the right manner [16]. Internal validity is defined as forming fundamental associations and avoiding false associations. External validity is the establishment of the field where findings can be generalized. Finally the empirical reliability implies that the same results can be found by repeating the study. Construct validity is tackled by using multiple sources of evidence and by sending the case report to the interviewee. Feedback was then provided, including reports and presentations on their architecture and process workflows. In this way the content was validated. Internal validity is only used in causal or explanatory case studies. In this case study, checking the internal validity is inapplicable because this study is an exploratory case study. The cases here are found to be representative for performing RNAi screens, automated microscopy and robotic liquid handling. Therefore the external validity is covered. Taking empirical reliability into account, we would expect that the results we found will be consistent when repeating this study. However, as time goes by, enhanced technology might increase the efficiency of an HTS facility. This paper focuses on the IT architecture of academic HTS facilities. A qualitative analysis is performed where semi-structured interviews at the Dutch Cancer Institute, the Leiden University Medical Center, the German Cancer Research Center, the European Molecular Biology Laboratory and the University Medical Center Utrecht were conducted to investigate the needs of a high throughput screening facility in terms of informatics architecture. The interviewees were experienced managers, IT personnel, researchers or Ph.D. students. For the semi-structured interviews a schematic overview of the HTS architecture was provided at the end (see figure 1) to see how their architecture differed from the envisioned architecture which was modeled in advance. We asked the heads of HTS facilities to provide information on the architecture and workflow of HTS through an interview. Generally they pointed out a particular researcher with the right prerequisite knowledge. Again this researcher was contacted by email with the same question. An appointment was made for a meeting using Skype or visiting them for a face-to-face interview. Notes were made during the interview and the semi-structured interview as guidance. Additional questions were asked if anything was unclear. After the interview, the notes were sent by email for validation. Finally when feedback was received, it was processed.

4 Results

Table 1 shows the development stage of the HTS facilities that were questioned. The reason of the selected aspects which are provided here is because these aspects according to the interviewees seem to be the most important aspects of a LIMS. The facilities interviewed, carry out high throughput RNAi screens using high content assays.

Table 1. Results overview of our case study.

	Image storage	Reagent Management	Screening Management	Data Analysis	Multi-parameter data
Facility 1	++	+/-	-	+/-	-
Facility 2	++	++	+/-	++	++
Facility 3	++	++	+/-	++	++
Facility 4	++	-	-	-	n.a.
Facility 5	++	-	+/-	+/-	+/-

- means working with flat files e.g. plain text or Excel sheets

+/- means semi-structured, working with an application that is supportive but still requires adjustments

++ means structured which implies working with integrated software that supports the communication of information in an interoperable and compatible fashion

A standard architecture for HTS was derived and is depicted in figure 1. In the course of an RNAi screen, images are generated by automated microscopy. Then a transformation from images to raw data is required which involves feature extraction from the images e.g. intracellular spot number, nuclear area and distribution of intensity. In some facilities, quantification is embedded in the microscope which is served by built-in software. Some facilities manage quantification using external software because they desire bespoke analysis algorithms. The quantified data can be managed in a screen data management system. This system provides metadata such as technical information and information concerning the screen for example when, who and what. A library information management system provides information concerning the reagent library. In the case of an RNAi screen, it will link a gene name with a particular RNAi reagent and provides the location of this reagent in various microplates. Volumes, concentrations and other information are also tracked. All facilities perform multi-step physical library management processes on their reagents before performing an assay. The system needs to deal with plate pooling, replication, dilution and compression from a 96-well format to 384-well or 1536-well format. In the case of screening with pooled reagents, it needs to assist in the deconvolution of multiplexed experiments. A more advanced system, a laboratory information management system, (LIMS) not only deals with managing reagents but has a more centered and integrated role. A true LIMS stores not just library information, (reagents), but also raw and normalized data associated project information concerning personnel and assay protocols. This advanced LIMS, is connected to the image database, and usually includes an advanced user management system. Now heat maps can show where the normalized screening results are projected on the associated well and reagents and second, the related raw data and images are linked.

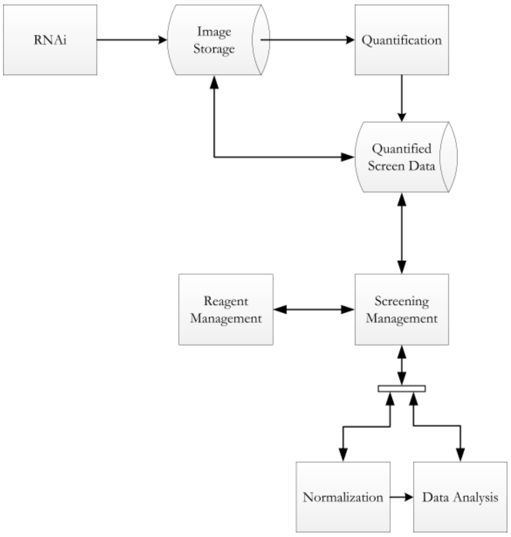
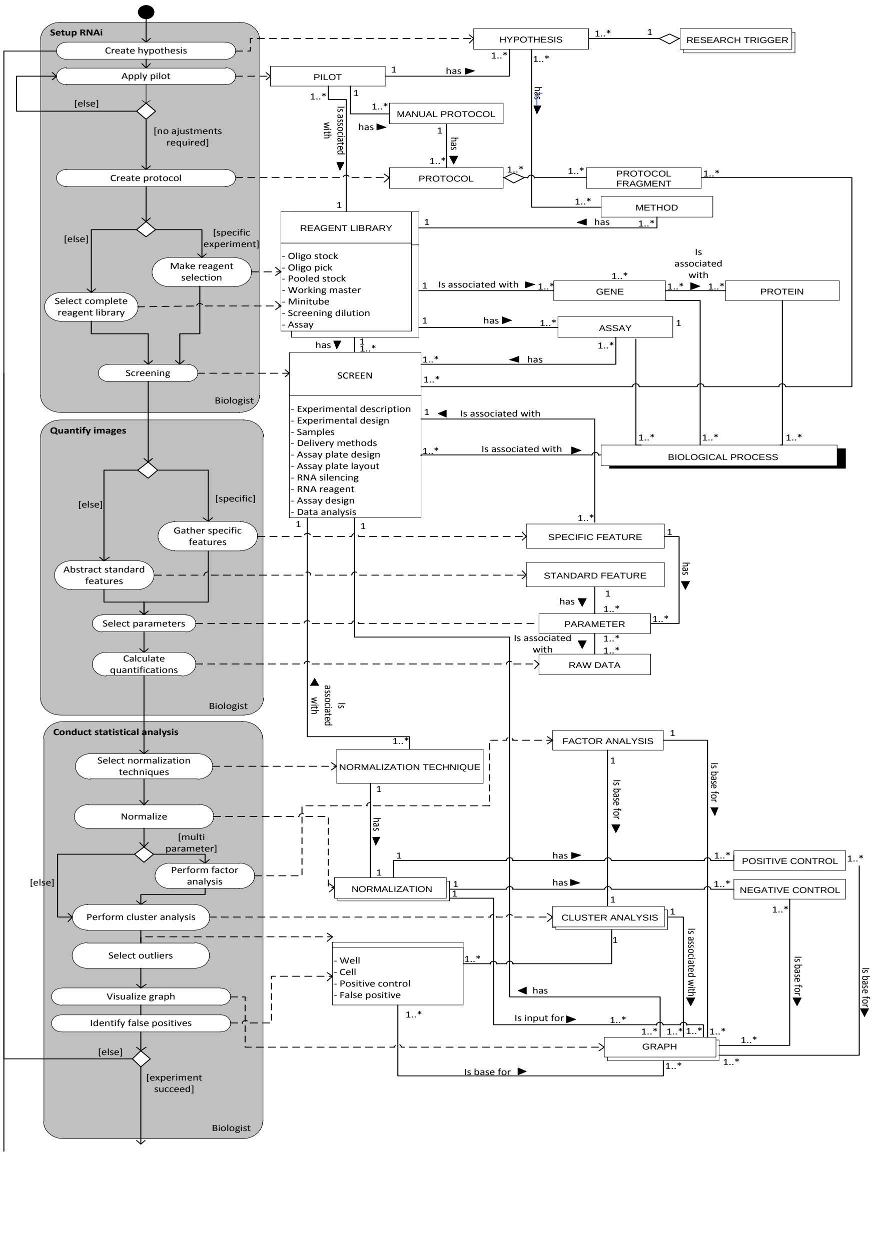


Fig. 1. The envisioned high throughput screening architecture

The last step in the interviewed facilities was normalization or statistical analysis. This was generally done by using R or CellHTS2. Depending on the normalization techniques e.g. median normalization, B-score normalization, loess regression or robust local fit regression [17-19], applications can provide predefined normalization techniques. Otherwise R or Mat-Lab can be useful tools for non-standard normalization techniques or data analysis. CellHTS2 lacks an option for normalizing multi-parameter data at this time. R is competent to manage multi-parameter data but is a command-line application which requires strong prerequisite knowledge. In addition, the results of CellHTS2 in HTML format are not structured in a way that makes further analysis easy. Therefore a database in which the results are stored in a regulated manner is required for enrichment of data or additional data analysis. Figure 2 outlines a more detailed process analysis using the Method Engineering modeling method [20]. After statistical analysis, the results can be enriched using public chemical and biological repositories (see figure 2). New knowledge can also be extracted using other methods e.g. pathway analysis. Currently this only is done manually. Automated enrichment requires very structured data, stored in a semantic manner. This is a substantial challenge in informatics. The last step is a reporting



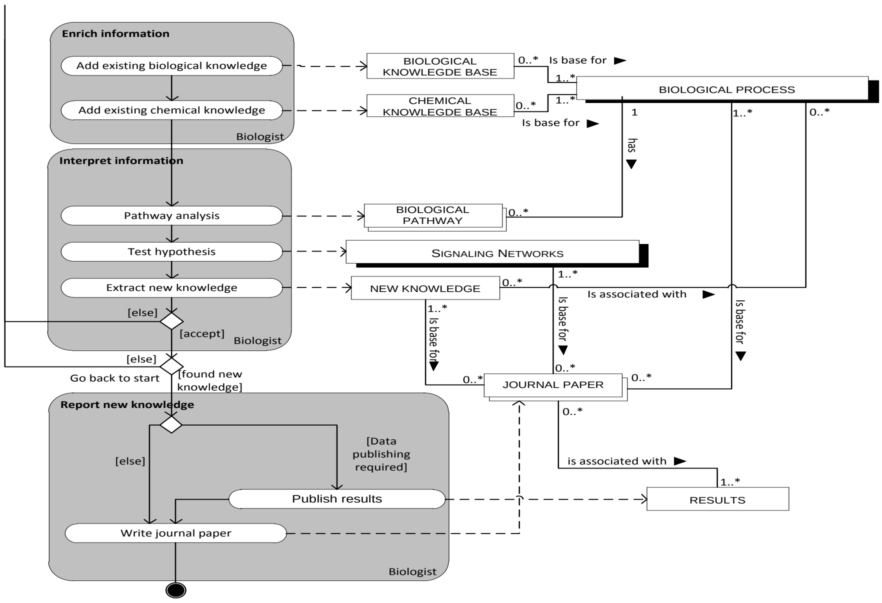


Fig. 2. Diagram illustrating how to conduct an RNAi screen using high content analysis. The right area shows all the concepts and their relations to the activities which are depicted on the left.

phase where tools should be used to clearly visualize findings. This simplifies the sharing of knowledge and provides input for publications. It was observed during interviews that business processes are not clear to key users. In the PDD it can be seen that work processes bifurcate in a number of places. Architecture and software need to be adapted to take these into account.

5 Conclusions

From an information science point of view, potential for improvements are present when focusing on the architecture that is studied in this research. Advances in HTS architecture can be identified by process modeling of workflows. A great variation in the quality of architecture was found in different case instances. There are facilities that are served by bioinformaticians who have built tools for their own needs. These have implemented LIMS in a central and integrated approach. There are, however, other facilities which do not possess documented business processes. Additionally they do not have a solid architecture and are struggling with implementing open source software. There is a need for software that can be used to manage micro-plate and mini-tube based reagent libraries in academic and small- to medium-sized commercial organizations. Currently available solutions have been designed for the management of clinical samples and lack much of the functionality that is needed in screening facilities that deal with libraries in larger scale operations. Activities such as reagent pooling, and 96 to 384-well compression are not available in many commercial solutions. One key differentiating feature of library information management should be the ability to create library processing pipelines that allow the library management processes to be easily reflected in the relational database. Another key differentiator should be the ability of library information management to easily and efficiently handle mini-tube collections for hit-picking and custom library generation.

6 Discussion: Towards Ultimate Throughput Screening

The Ultimate goal within High Throughput Screening (HTS) is that every piece of information from the output of a particular screen should be rapidly put in the appropriate biological and/or chemical context so as to enable decision-making and the generation of new hypotheses. Traditionally this has been done by humans, but if the data is integrated with the appropriate ontology, it can be done in an automated fashion. Additionally, external data is easily added to the existing knowledge which further enriches the processed data. All applications should be connected in a sense that they can easily transfer processed data from one application to another automatically. Applications and steps in this process include automated microscopy, assay quantification, statistical analysis, library information management (LibIM), data-mining, merging of external knowledge, automated reporting and biological and chemical understanding. Furthermore, it is essential for scientists that the associated images or animations are available in an on-demand fashion whether it concerns kinetic, confocal, multi-parameter, individual cell-level or movie experiments. Moreover informatics workflows should be simple to build and accessible for re-use and modification. Historical insight is also desirable, so that scientists can have a vision on what they have been doing, in which order and how it was performed. In a nutshell, aspects such as interoperability, flexibility, consistency but also semantics are enormously important for enhancing high throughput screening in a more developed maturity stage as it is right now. There are (open source) applications out there which can meet certain needs but none of these are consecutive. Each application has its own purpose but the applications cannot be reciprocally interlinked. The consequence is working with flat files and using import and export functionality. This even might require adjustments before uploading e.g. Perl scripting or recalculations by using R or Mat-Lab. Screening facilities are conducting statistical analysis using R, Excel or cellHTS2 on their data. However, none of the screening facilities in this study were performing data-mining. Most interviews revealed that reporting was done by publishing a journal paper. Some facilities used tools for reporting results such as Excel, Image J, Adobe Photoshop and R. Facilities at different stages of development were analyzed during this case study which may have influence on the architecture or PDD that is offered. Also the gathering of data has been conducted using face-to-face conversations and Skype meetings which might have effect on how answers were interpreted. The number of HTS facilities that were questioned in this research was limited to five. This analysis was projected primarily towards RNAi screening and therefore cannot be generalized to compound screening. However many of the concepts are very similar. Modeling of the architecture and processes at a high conceptual level was found to be very useful as it makes it easier to identify gaps.

6.1 Recommendations

First, as mentioned in the results, most HTS facilities do not possess a well-developed workflow of the HTS processes. Also, key-users do not fully understand how information flows. Therefore it is recommended for every HTS facility that there is an up-to-date modeled business process present. Second, gene names, features and plate identifiers need to be consistent and compatible so that other software packages are handling this data. Third, implement a workflow management system which can be changed flexibly. The cognitive load on certain processes e.g. mini-tube handling is so extensive that human errors can be made very easily. Therefore a system which incorporates and helps users to remind what to do next can be very helpful. Finally, a screening management system or LIMS (Laboratory Information Management System) should be the central mediator in managing all the information concerning a screen. A Library Information Management System (LibIMS) should be a subset of a screening management system or LIMS. We suggest a project ID to associate every piece of information e.g. researchers, used reagents, plates, gene names, anti-bodies or compounds and images. Furthermore we suggest a phase ID to identify which step is currently applicable within a project as is shown in the gray brackets at the left in fig 2.

6.2 Future research

HTS is used in several other fields besides functional genomics. Applications with similar workflows include drug discovery, where large libraries of small molecule compounds or natural product mixtures are screened. The types of biological systems in the assays can range from the use of purified proteins to microorganisms such as pathogenic bacteria. A recent innovation is the development of assays for the screening of organoid cultures. HTS can also be applied in genetic screens using model organisms such as yeast and zebrafish. All of these applications share certain aspects of the workflow described here, as shown in fig 2. Therefore improvements in software supporting HTS processes are broadly supported by other disciplines because of commonalities in workflow and data management.

References

1. Persidis, A.: High-throughput screening. Nature Biotechnol 16:488–489 (1998)
2. Chris Allan et. al.: OMERO: flexible, model-driven data management for experimental biology Nature Methods 9, 245–253 (2012)
3. Pelz, O., Gilsdorf, M., Boutros, M.: web cellHTS2: a web-application for the analysis of high-throughput screening data. BMC Bioinformatics. 11:185 (2012)
4. Johann, D., McGuigan, M., Tomov, S., Petricoin, E., Liotta, L.: Towards a systems biology toolkit. In: Proceedings of the 17th IEEE symposium on computer-based medical systems, pp. 500–5 (2004)

5. Bajorath, J.: Integration of virtual and high-throughput screening. Nature Rev. Drug Discov. 1, 882–894 (2002)
6. Maurer, HH.: Screening procedures for simultaneous detection of several drug classes used in the high throughput toxicological analysis and doping control. Comb Chem High Throughput Screen, 461-74 (2003)
7. T. et al. SLIMS–a user-friendly sample operations and inventory management system for genotyping labs. Bioinformatics, 26: 1808–1810 (2010)
8. Tolopko, A., Sullivan, J., Erickson, S., Wrobel, D., Chiang, S., Rudnicki, K., Rudnicki, S., Nale, J., Selfors, L., Greenhouse, D., Muhlich, J., Shamu, C.: Screensaver: an open source lab information management system (LIMS) for high throughput screening facilities. BMC Bioinformatics, 11**:**260 (2010)
9. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., Zhang, J.: Bioconductor: opensoftware development for computational biology and bioinformatics**.** Genome Biology, 5:R80 (2004)
10. Goldberg, I.G. et al.:.The open microscopy environment (OME) data model and XML file: open tools for informatics and quantitative analysis in biological imaging. Genome Biol. 6, R47 (2005)
11. Lamprecht, M.R., Sabatini, D.M., Carpenter, A.E.: CellProfiler: free, versatile software for automated biological image analysis. Biotechniques.;42:71–75 (2007)
12. Jones, T.R., Kang, I.H., Wheeler, D.B., Lindquist, R.A., Papallo, A., Sabatini, D.M., Golland, P., Carpenter, A.E. CellProfiler Analyst: data exploration and analysis software for complex image-based screens. BMC Bioinformatics 9(1):482 (2008)
13. Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., Golland, P., Sabatini, D.M.: CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biology 7 (2006)
14. Kamentsky, L., Jones. T.R., Fraser, A., Bray, M., Logan, D., Madden, K., Ljosa, V., Rueden, C., Harris, G.B., Eliceiri, K., Carpenter, A.E.: Improved structure, function, and compatibility for CellProfiler: modular high-throughput image analysis software. Bioinformatics (2011)
15. Yin, R.K.: Case Study Research, Design and Methods, Sage Publications, Beverly Hills, California, (1984)
16. Jansen, S.: Applied multi-case research in a mixed-method research project: Customer configuration updating improvement. In Information Systems Research Methods, Epistemology and Applications, A.C. Steel and L.A. Hakim (eds.), (2008)
17. Boutros, M., Brás, L.P., Huber, W. Analysis of cell-based RNAi screens. Genome Biol. 7:R66 (2006)
18. Malo, N., Hanley, J.A., Cerquozzi, S., Pelletier, J., Nadon, R.: Statistical practice in high-throughput screening data analysis. Nat Biotechnol. 24:167-75 (2006)
19. Birmingham, A., Selfors, L.M., Forster, T., Wrobel, D., Kennedy, C.J., Shanks, E., Santoyo-Lopez, J., Dunican, D.J., Long, A., Kelleher, D., Smith, Q., Beijersbergen, R.L., Ghazal, P., Shamu, C.E.: Statistical methods for analysis of high-throughput RNA interference screens. Nat Methods 6:569-75 (2009)
20. Weerd, van de, I., Brinkkemper, S. Metammodeling for Situational Analysis and Design Methods, pages 38–58. Hersey: Idea Group Publishing (2008)