# Implementation of a Systems Biology **Data Integration Platform**

Sam Ansari<sup>1†</sup>, Leandro Hermida<sup>2</sup>, Svenja Diehl<sup>1</sup>, Carine Poussin<sup>3</sup>, Alain Sewer<sup>3</sup>, Yang Xiang<sup>3</sup>, Filipe Bonjour<sup>3</sup>, Komanda Phanzu<sup>3</sup>, Sylvain Gubian<sup>3</sup>, Bruce O'Neel<sup>3</sup>, Julia Hoeng<sup>3</sup> and Manuel Peitsch<sup>3</sup>

<sup>1</sup>Philip Morris International R&D, Philip Morris Research Laboratories GmbH, Cologne, Germany <sup>2</sup>Leandro Hermida Consulting, Rüschlikon, Switzerland <sup>3</sup>Philip Morris International R&D, Philip Morris Products S.A., Neuchâtel, Switzerland

<sup>†</sup>E-mail: sam.ansari@pmintl.com

### **Abstract**

To enable and support a systems biology approach to research, one requires an underlying infrastructure to manage, integrate, and share high-throughput functional genomics data and workflows from data production through annotation, analysis, and knowledge acquisition. At its core, there should be a comprehensive data management and annotation system and data repository that fully support publicly established standards for storing and reporting high-throughput functional genomics investigations. Such a system will serve as the platform's central hub and it will integrate with data analysis, visualization and mining tools. It will also enable collaboration between internal teams, publishing of internal investigations and data to public repositories, and incorporation of

public investigations and data into the platform for internal comparison and analysis.

Here, the implementation of a systems biology data integration and knowledge management platform to support experimental and computational workflows, examining in vivo and in vitro generated systems response profiles (gene expression, microRNA, comparative genomic hybridization, and reverse-phase protein array proteomics data) is reported. The platform utilizes open-source, freely available components where suitable, featuring caArray and caGrid [1] from the National Cancer Institute Biomedical Informatics Grid (NCI caBIG®) [2] software family as its core data management and annotation infrastructure. For data exchange, the community standard MAGE-TAB [3] format was used. caArray is integrated with GenePattern [4], an open-source bioinformatics workflow management system for integrative genomics, which is used for quality control, data analysis, and visualization purposes. caArray is also integrated with several commercial data analysis and biological pathway inferencing systems. Gene-centric, cross-investigation data mining capabilities are provided by the BioMart and InterMine open-source data warehouse systems. Under development are several other modules to integrate the platform with existing laboratory information management systems, as well as additional features to contribute the open-source caArray project.

## **Implementation**

- Data production platforms: Affymetrix®, Zeptosens, Exiqon
- Functional genomics data types: Gene expression, microRNA expression, aCGH, SNPs, protein expression
- Data exchange: MAGE-TAB
- Data management and annotation: caArray
- LIMS with planned full integration with caArray
- Analysis tools: R/Bioconductor [5], GenePattern
   Data mining and integration of public databases: BioMart, InterMine
- Programmatic integration: caGrid API, caArray API
   Sharing and publishing: caGrid, MAGE-TAB export ArrayExpress [6], SOFT export GEO [7]
- Various integrated proprietary and commercial tools for analysis and pathway inferencing

## **Objective and Requirements**

The objective is to establish and support a systems biology data integration and analysis environment. The following requirements are defined for the core infrastructure:

- Open-source
- Standards compliant (MIAME)
- Easy-to-use and user friendly (e.g., wizards for experiment creation)
- Customizable user interfaces
- Scalable, enterprise-ready architecture
- Support for standard data exchange format import/export (e.g., MAGE-TAB, MAGE-ML) for publishing to public repositories and incorporating public data
- Comprehensive programmatic API for data access and integration with other systems Data analysis and visualization tool integration
- Under active development and support
- Comprehensive documentation

## Selection of Major Components

After evaluation of various open source systems biology data management systems, caArray, along with the caBIG® infrastructure and integrated tools, was determined to best meet these requirements and will serve as the core of the systems biology data integration platform. caArray is an open-source, web and programmatically accessible array data management system. It is developed and currently supported by the U.S. National Cancer Institute as part of caBIG®, an initiative to build an information and bioinformatics framework to enable collaborative and integrative approaches to biomedical research. caArray provides full support for MAGE-TAB, curation capabilities as well support for ontologies and controlled vocabularies to ensure that the functional genomics data stored in its repository are of high quality and abide by community standards for data sharing and exchange

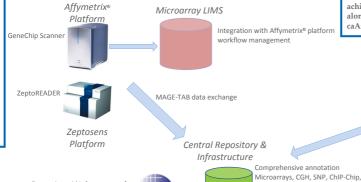
#### References

- [1] Saltz et al., Bioinformatics 22:1960-6, 2006 [2] Osters et al., J Am Med Inform Assoc. 15:138-49, 2008 [3] Rayner et al., BMC Bioinformatics 7:499, 2006 [4] Reich et al., Nat Genet. 38:500-1, 2006 [5] Gentleman et al., Genome Biol. 5:R80, 2004 [6] Farzma et al., Nucleic Acids Res. 31:88-71, 2003 [7] Edgar et al., Nucleic Acids Res. 30:207-10, 2002

#### **Abbreviations**

API: Application Programming Interface caArray: Array data management system caGrid: Network architecture providing the basis for connectivity of caBIG® tools ChIP-Chip: chromatin immunoprecipitation with microarray technology LIMSS: Laboratory information Management System MAGE-TAB: MicroArray Gene Expression Tabulator MAME: Minimum Information About a Microarray Experiment QC: Quality Control RPA: Reverse Phase Protein Array by Zeptosens SNP: Single-nucleotide polymorphism

Data coming from Microarray and RPA platforms will be nanaged by LIMS, LIMS will be integrated with caArray directly via MAGE-TAB data exchange and import using the caArray Batch Importer.



RPA workflow capture and annotation achieved using standard MAGE-TAB along with minor customizations to caArray.

> Data exchange: MAGE-TAB import for publicly available studies and data



Data Minina/Knowledge

Integration

NCBI GEO

caArray natively stores and exports MAGE-TAB which is used to publish experiments to ArrayExpress. Public data from ArrayExpress in MAGE-TAB format is imported into caArray for internal comparison.

Relevant public

integrated with a data mining

systems biology

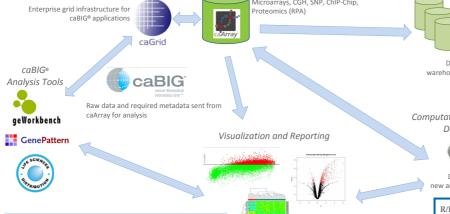
approach for

research.

system to facilitate a

GenePattern provides a bioinformatics workflow management platform and simplifies execution and sharing of

analysis code



Data from caArray to build data marts and warehouse: Web user interfaces for scientists to

launch cross-study queries

Computational Biology/Bioinformatics Development Platform





Development and implementation of new analysis algorithms and visualizations

R/Bioconductor scripts are packaged as GenePattern modules and pipelines which are then used for QC as well as analysis of high-throughput source data and metadata coming from caArray. Processed raw data is then stored back into caArray.



Data visualization is provided by R/Bioconductor packages and various commercial tools. An internal

cross-site interactions and collaborations

collaboration system and wiki enables cross-functional,