MageComet—web application for harmonizing existing large-scale experiment descriptions

Vincent Xue1,∗, Tony Burdett2, Margus Lukk3, Julie Taylor2, Alvis Brazma2 and Helen Parkinson2

1 Department of Computer Science, Hunter College, City University of New York, NY 10065, USA, 2EMBL-EBI, WellcomeTrust Genome Campus, Hinxton, CB10 1SD and 3Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK

1 INTRODUCTION

The amount of experimental data in public repositories is accumulating at an ever increasing rate thanks to the advents of high-throughput technologies. Meta-analysis of these data are limited by the availability of harmonized, consistently annotated data in machine readable formats.

In the study of gene expression, MAGE-TAB is a commonly used file format that provides flexibility and structure for describing experimental data [1, 2]. Assays in the MAGE-TAB format can be explicitly annotated with characteristics and experimental variables, which in turn are referenced to ontologies to provide explicit meaning. Public repositories such as GEO (Fig. 1) house over 23,000 experiments in the MAGE-TAB format, which is also used by the TCGA project [3].

The ability to perform meta-analyses requires considerable investment in human re-annotation, harmonization and mapping of data from different labs to single ontologies. Due to the spreadsheet nature of MAGE-TAB, it is a convenient bridge between readability and interoperability but meta-analysis is faster if formatting, validation, ontology enrichment and common editing tasks are addressed within the application. This addresses bottlenecks in the re-annotation process and allows users to extract implicit meta-data quickly from pre-existing MAGE-TAB documents and add to this in a convenient editing environment.

We present MageComet, a web application for biologists and annotators that facilitates the re-annotation of gene expression experiments in MAGE-TAB format. It incorporates data mining, automatic annotation, use of ontologies and data validation to improve the consistency and quality of experimental meta-data from the ArrayExpress Repository.

Availability and implementation: Source and tutorials for MageComet are openly available at goo.gl/8LQPR under the GNU GPL v3 licenses. An implementation can be found at goo.gl/IdCuA

Contact: parkinson@ebi.ac.uk or xue.vin@gmail.com

Received on August 26, 2011; revised on February 29, 2012; accepted on March 24, 2012

© The Author(s) 2012. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

To whom correspondence should be addressed.
MageComet

2.4 Validation
MageComet incorporates the Limpopo (limpopo.sf.net) library into its backend for syntactic and semantic validation. Curators can edit and quickly validate changes within the application, allowing for a seamless work environment.

2.5 Autocompletion
MageComet features an auto complete widget that helps curators tag samples with consistent annotations. Leveraging the extensive synonyms available in EFO, MageComet provides a query box which returns standard labels as defined in the ontology (i.e. Homo sapiens instead of human). With MageComet’s tight integration with ontologies, it can also complete the Term Source Ref and Term Source Number columns in the SDRF for riche annotations.

3 IMPLEMENTATION
MageComet was designed as a web application to provide cross platform compatibility without installation on client machines. As a web application, it is able to synchronize with the latest ontologies and validators to ensure consistent annotations. Its integration into the web browser makes external services easily accessible. As a service, MageComet provides validation through the Limpopo parsers. It provides data mining through Whatizit and interacts with ontologies using the OntoCat library (Adamusiak et al. 2011).

4 CONCLUSION
MageComet is a web application created for harmonization of large-scale experiment annotations. It is designed to reduce the repetitive aspects of re-annotation and curation of pre-existing documents and serve as a consistent, up-to-date workspace across all platforms. Its features are designed to provide user-driven automation of repetitive tasks. It differs from standard editors through its close integration with the MAGE-TAB specification. The implementation and code can be found at goo.gl/IdCuA, and goo.gl/8LQPR, respectively.

Funding: The authors thank the Macaulay Honors College at CUNY Hunter College for funding for V.X., Duncan Odom from CRUK for funding M.L., and EMBL for J.T., H.P., and A.B.

Conflict of Interest: none declared.

REFERENCES
Adamusiak,T. et al. (2011) OntoCat—simple ontology search and integration in Java, R and REST/JavaScript. BMC Bioinformatics, 12, 218.
Hampton,T. (2006) Cancer Genome Atlas. J. Am. Med. Assoc., 296, 1958–1959.
Malone,J. et al. (2010) Modeling sample variables with an Experimental Factor Ontology. Bioinformatics, 26, 1112.
Parkinson,H. et al. (2009) ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res., 37, D868–D872.
Rayner,T.F. et al. (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data. MAGE-TAB. BMC Bioinformatics, 7, 489.