BioMedBridges

Deliverable D6.4

Project Title:	Building data bridges between biological and medical infrastructures in Europe	
Project Acronym:	BioMedBridges	
Grant agreement no.:	284209	
	Research Infrastructures, FP7 Capacities Specific Programme; [INFRA-2011-2.3.2.] “Implementation of common solutions for a cluster of ESFRI infrastructures in the field of "Life sciences"	
Deliverable title:	Mapping of standards and ontologies between the different image reference data sets	
WP No.	6	
Lead Beneficiary:	1: EMBL	
WP Title	Use case: Interoperability of large scale image data sets from different biological scales	
Contractual delivery date:	19 December 2014	
Actual delivery date:	15.12.2014	
WP leader:	Jan Ellenberg	1: EMBL
Partner(s) contributing to this deliverable:	1: EMBL 11: HMGU 16: UH	

Authors: Gabriella Rustici, Simon Jupp, Tanja Ninkovic, Philipp Gormanns, Jean-Karim Heriche, Johan Lundin, Frauke Neff, Jan Ellenberg

Contents

Figures

1 Executive Summary

The imaging use-case addresses interoperability of large-scale image data sets, which is required for reusing and comparatively analyzing data sets generated by different sources.

One of the most commonly done types of analysis of image data is annotation of cellular phenotype. However, phenotypic data associated with image datasets is often annotated using free-text which can vary widely from researcher to researcher for the same phenotype, or by using ontologies that are specific for particular species or level of annotation, thus preventing the integration of independent datasets, including those generated in different biological domains, such as cell lines, mouse and human tissues.

To harmonize the annotation of cellular phenotypes, WP6 partners have developed the Cellular Microscopy Phenotype Ontology (CMPO), a species neutral ontology for describing general phenotypic observations relating to the whole cell, cellular components, cellular processes and cell populations. CMPO is compatible with related ontology efforts, allowing for future cross-species integration of phenotypic data. CMPO is made to be used by any researcher generating phenotypic data who are annotating cellular phenotypes associated with their imaging data.

Preliminary testing of CMPO for the annotation of imaging datasets derived from different biological domains (cell lines, mouse and human tissues) is showing that CMPO is suitable to annotate cellular phenotypes observed in such images, consequently making the data interoperable.

2 Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives:

No.	Objective	Yes	No
1	Demonstrate the utility of the interoperability of large-scale image data sets from different biological scales (cell – tissue – organism)		X
2	Enable comparison of morphological image data of cellular phenotypes of individual genes with morphological image data of the diseased tissues in mouse and human	X	
3	Link imaging data with molecular data, including cancer genome sequences and cancer expression data		X

3 Detailed report on the deliverable

3.1 Background

Recent advances in imaging techniques make the study of complex biological systems feasible, particularly at the cellular level, complementing existing "omics" approaches, most notably genomics and proteomics, by resolving and quantifying spatio-temporal processes with single cell resolution [1].

Correlative analysis of cellular phenotypes, linking phenotypic data specific to individual genes to morphological imaging data from diseased tissue specimens (both human and mouse tissues) could be a powerful predictor of disease biomarkers as well as drug targets. For example, when a certain cellular phenotype, like 'mitotic delay' or 'multi-nucleated cells', observed in cells after gene knockdown experiments, is also observed in cells of a cancer tissue, we may infer that the knockdowned gene is involved in the aetiology of the disease, in that specific tissue. In this way, knowledge of the functional implications of somatic tumor mutations can thus be used to design more targeted drug therapies.

Particularly useful for this kind of analysis are data from high content screening (HCS) of biological samples. HCS is an image-based multiparametric approach that allows the study of living cells under a broad array of conditions or under exposure to multiple substances, such as small molecules or RNA interference (RNAi) reagents that can target specific genes or processes. Resulting phenotypes may include morphological changes of a whole cell, or any of its cellular components, as well as alteration of cellular processes.

Data derived from live cell imaging is typically associated with rich metadata, including genetic information, and can be more easily interpreted and linked to underlying molecular mechanisms. As we move to higher organisms, such as mouse and human, the degree of metadata available decreases (e.g. often no genetic information is available for diseased human tissues), alongside the feasibility of assays that can be carried out in such organisms (e.g. genetic engineering is only possible in cell lines and mouse models). Taking this into consideration, it becomes evident that integrating imaging datasets from different biological domains could greatly advance our understanding of the molecular mechanisms underlying specific diseases.

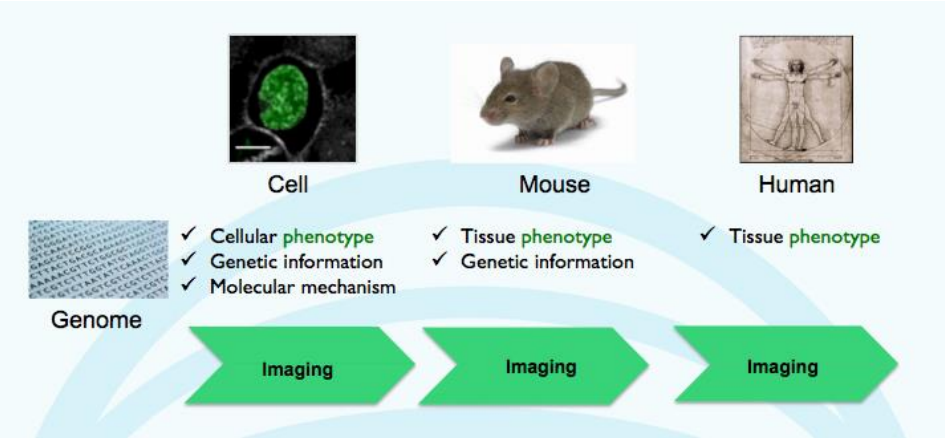

Figure 1 Different types of information and metadata are available for different types of images

Due to its late arrival on the "omics" scene, the imaging field has not yet achieved the same degree of standardization that other high-throughput approaches have already reached [1], thus hampering integration of image data with current biological knowledge. Additionally, integrating phenotypic data from imaging assays is challenging as such data is typically described using free-text. Therefore, to enable data integration, experimental imaging datasets from different sources need to be harmonized with regards to structure, formatting and annotation.

The use of ontologies to annotate data in the life sciences is now well established, particularly for non-image data, and provides a means for the semantic integration of independent datasets. Despite the availability of several species-specific ontologies for describing cellular phenotypes (e.g. the Fission Yeast Phenotype Ontology), there isn't an appropriate infrastructure in place to support the large-scale annotation and integration of phenotypes across species and different biological domains.

In order to enable integration of imaging datasets from different sources, we have developed the Cellular Microscopy Phenotype Ontology (CMPO) for the annotation of such datasets. CMPO is built using ontology design patterns that are compatible with related species-specific ontology efforts, such as Fission Yeast Phenotype Ontology [2], Ascomycete Phenotype Ontology [3] and Mammalian Phenotype Ontology [4], allowing for cross-species integration of phenotypic data.

3.2 Cellular Microscopy Phenotype Ontology

Eleven imaging datasets were sourced to collect a set of candidate phenotypic descriptions for manual ontology annotation [5-14, and van Roosmalen W et al. (manuscript sub-mitted)]. Our approach was to annotate the phenotypes with a basic Entity-Quality (EQ) pattern, describing each phenotype in terms of an Entity (E), from one of many given reference ontologies, such as Gene Ontology (GO), and an associated Quality (Q), from PATO [15]. These annotations could then be used to generate new terms with logical definitions in CMPO.

We developed a simple Web application called Phenotator for the data providers to submit and annotate their phenotypes using EQ. The Phenotator is built using services from the NCBO BioPortal [16] to generated simple drop down menus and autocomplete search functionality to guide the users in generating EQs with appropriate terms. Phenotator provides a feature to export the annotations as an ontology in OWL (Web Ontology Language) format.

127 phenotype descriptions from the original 11 datasets were entered into Phenotator, together with 41 phenotypes collected from cell migration assays (Z. Kam, personal communication) and 193 phenotypes from the GenomeRNAi database [17]. The domain experts entered EQ based descriptions for a total of 201 of these phenotypes. The EQs were transformed into an OWL file that provided the basis for the new CMPO ontology. CMPO was further refined by an ontologist who worked with the domain experts to organize the top level of the ontology into biologically meaningful categories. Additional meta-data such as full text descriptions, synonyms and literature references were also generated for each CMPO term.

A dedicated website for CMPO is hosted by the EBI[1], and the ontology is released monthly on the NCBO BioPortal[2] and the EMBL-EBI's Ontology Lookup Service (OLS)[3]. The source file is hosted on GitHub[4].

CMPO has already been integrated into the MitoSys project database[5] and the Cellular Phenotype Database[6].

CMPO is currently being used to annotate cellular phenotypes from diseased tissue images from both mouse models and human individuals. Five mouse datasets could be annotated with the extended version to demonstrate the interoperability of CMPO. The annotation of mouse and human images has shown that CMPO is suitable to annotate images derived from different biological domain, making this data interoperable.

When applying CMPO to tissue images, we realized that in tissue the number of phenotypes that can occur in just a single sample is extensive. Normally researchers will focus on certain parts or processes of the cells and annotate

[1] http://www.ebi.ac.uk/cmpo
[2] http://bioportal.bioontology.org/ontologies/CMPO
[3] https://www.ebi.ac.uk/ontology-lookup
[4] https://github.com/EBISPOT/CMPO
[5] http://www.mitosys.org
[6]http://www.ebi.ac.uk/fg/sym

only those that are relevant for the process they are observing. For the purposes of the WP6 tasks and deliverables, and to mitigate the problem of a vast number of phenotypes, we have decided to restrict the annotations to nuclear morphologies related to mitosis during the first stage of implementation of CMPO on human tissue samples. As a starting point we will use phenotype terms from the Mitocheck project and add to CMPO terms related to nuclear mitotic processes that occur in fixed human tissue samples. Annotations are done using the BioMedbridges WebMicroscope platform[7], on mouse and human tissue samples uploaded to the image server. Currently a number of human tissue samples (n=31) are available for annotation and additional images will be added to represent cancer cases with mutation in mitosis related genes. Most of the human tissue samples will be obtained through The Cancer Genome Atlas (TCGA) whole-slide image archive[8]. Also, additional mouse tissue image data sets will be annotated in the coming months, all leading to WP6 deliverable 6.5.

While applying CMPO to tissue images, we also noticed that certain CMPO terms were more suitable for tissue annotations than others, as certain phenotypes are observed more often in cells that are still incorporated in their native environment and involved in the cell-cell and cell-interstitial interactions. Other CMPO terms and corresponding phenotypes (e.g. related to dynamic processes that can be observed in cell culture and freely standing/migrating cells) can rarely be observed in tissues because fixation of the material brings dynamic process to a halt.

To further optimize CMPO to be more suitable for the use of tissue samples, we will consider adding ontology terms from other upcoming ontologies such as for example the ontology of biological attributes[9].

[7] http://biomedbridges.webmicroscope.net
[8] http://cancer.digitalslidearchive.net/
[9] http://www.obofoundry.org/cgi-bin/detail.cgi?id=oba

4 References

[1] Lock, J.G. and S. Stromblad, *Systems microscopy: an emerging strategy for the life sciences.* Exp Cell Res, 2010. **316**(8): p. 1438-44.

[2] Harris, M.A., et al., *FYPO: the fission yeast phenotype ontology.* Bioinformatics, 2013. **29**(13): p. 1671-8.

[3] Engel, S.R., et al., *Saccharomyces Genome Database provides mutant phenotype data.* Nucleic Acids Res, 2010. **38**(Database issue): p. D433-6.

[4] Smith, C.L., C.A. Goldsmith, and J.T. Eppig, *The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information.* Genome Biol, 2005. **6**(1): p. R7.

[5] Di, Z., et al., *Automated analysis of NF-kappaB nuclear translocation kinetics in high-throughput screening.* PLoS One, 2012. **7**(12): p. e52337.

[6] Fuchs, F., et al., *Clustering phenotype populations by genome-wide RNAi and multiparametric imaging.* Mol Syst Biol, 2010. **6**: p. 370.

[7] Gudjonsson, T., et al., *TRIP12 and UBR5 suppress spreading of chromatin ubiquitylation at damaged chromosomes.* Cell, 2012. **150**(4): p. 697-709.

[8] Moudry, P., et al., *Nucleoporin NUP153 guards genome integrity by promoting nuclear import of 53BP1.* Cell Death Differ, 2012. **19**(5): p. 798-807.

[9] Neumann, B., et al., *Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes.* Nature, 2010. **464**(7289): p. 721-7.

[10] Ritzerfeld, J., et al., *Phenotypic profiling of the human genome reveals gene products involved in plasma membrane targeting of SRC kinases.* Genome Res, 2011. **21**(11): p. 1955-68.

[11] Rohn, J.L., et al., *Comparative RNAi screening identifies a conserved core metazoan actinome by phenotype.* J Cell Biol, 2011. **194**(5): p. 789-805.

[12] Schmitz, M.H., et al., *Live-cell imaging RNAi screen identifies PP2A-B55alpha and importin-beta1 as key mitotic exit regulators in human cells.* Nat Cell Biol, 2010. **12**(9): p. 886-93.

[13] Simpson, J.C., et al., *Genome-wide RNAi screening identifies human proteins with a regulatory function in the early secretory pathway.* Nat Cell Biol, 2012. **14**(7): p. 764-74.

[14] Winograd-Katz, S.E., et al., *Multiparametric analysis of focal adhesion formation by RNAi-mediated gene knockdown.* J Cell Biol, 2009. **186**(3): p. 423-36.

[15] Gkoutos, G.V., et al., *Using ontologies to describe mouse phenotypes.* Genome Biol, 2005. **6**(1): p. R8.

[16] Whetzel, P.L., et al., *BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W541-5.

[17] Schmidt, E.E., et al., *GenomeRNAi: a database for cell-based and in vivo RNAi phenotypes, 2013 update.* Nucleic Acids Res, 2013. **41**(Database issue): p. D1021-6.

5 Delivery and schedule

The delivery is delayed: □ Yes ☑ No

6 Adjustments made

No adjustments were made to the deliverable.

7 Background information

This deliverable relates to WP 6; background information on this WP as originally indicated in the description of work (DoW) is included below.

WP 6 Title: Interoperability of large scale image data sets from different biological scales
Lead: Jan Ellenberg (EMBL)
Participants: EMBL, HMGU, CIRMMP

Work package number	WP6	**Start date or starting event:**	month 13
Work package title	PhenoBridge-crossing the species bridge between mouse and human		
Activity Type	RTD		
Participant number	1:EMBL	11:HMGU	20: CRIMMP
Person-months per participant	37	27	16

Objectives

This work package will demonstrate the utility of the interoperability of large scale image data sets from different biological scales (cell – tissue – organism) to enable drug target and biomarker discovery for human disease with cancer as an example. Based on the standards and services developed in WP2 and 3, we will use integrated access to systematic imaging data of disease gene function in cultured human cells (EMBL-Heidelberg) and systematic imaging data available from tissue microarrays of diseased tissue from both human patients (FIMM, U Helsinki) and mouse models (HMGU).

This use case will thus link the four BMS ESFRI infrastructures Euro-BioImaging (EMBL-Heidelberg), BBMRI (U Helsinki), EATRIS (U Helsinki-FIMM) and Infrafrontier (HMGU) with the standards and services provided by ELIXIR (EMBL-EBI) and require strong links to ELIXIR’s molecular data resources. The comparison of morphological image data on cellular phenotypes of individual genes, with morphological image data of the diseased tissues in mouse models and human patients could create a powerful predictor of optimized biomarkers as well as drug targets in cancer. Linking these imaging data with molecular data including the cancer genome sequence and cancer expression data, will allow in silico validation of the predictions and prioritization of biomarkers for validation in clinical research.

Description of work and role of participants

Task 1.1. Implementation of interoperability standards and ontologies for reference image data sets

(Leader: U. Helsinki-FIMM, Participants: EMBL-Heidelberg, U Helsinki, EMBL-EBI, HMGU)

The first task to make integrated image data access usable is to map the (meta)data standards and ontologies present within each image data domain (cell, human and mouse tumor tissue) onto each other to enable correlative analysis. In line with the standards and services developed in WP2, 3 and where applicable respecting the secure access to medical data developed in WP4, we will implement unambiguous maps between the respective metadata. For this we will select high throughput imaging reference data sets with cancer related assays (e.g. www.mitocheck.org, www.cellmigration.org/resource/discovery/#genes) as well as tumor tissue and clinical data (we will make these available at: fimm.webmicroscope.net/)

Task 1.2. Prediction of novel cancer biomarkers (e.g. breast and prostate cancer) (Leader: EMBL-Heidelberg,Participants: FIMM, U Helsinki, HMGU, EMBL-EBI).

Correlative analysis of interoperable cell and tissue image datasets with their associated annotation and metadata will be mined with state of the art bioinformatic tools to predict novel biomarker candidates. In particular we will focus on genes with function in cell cycle and cell division control as well as invasive behavior, for which comprehensive molecular and cellular datasets are available. An initial set of predicted biomarkers will then be further cross-validated against the biomolecular databases hosted by ELIXIR, drawing on cancer genome and expression data, as well as general sequence and structural properties of the identified genes to also explore their potential as drug targets.