Cell, chemical and anatomical views of the Gene Ontology

David Osumi-Sutherland[1]*, Enrico Ponta[2], Melanie Courtot[1], Helen Parkinson[1] and Laura Badi[2]

[1] European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

[2] Roche Pharma Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel, Switzerland

Abstract. The Gene Ontology (GO) consists of around 40,000 terms refering to classes of biological process, cell component and gene product activity. It has been used to annotate the functions and locations of several million gene products. Much pharmacological research focuses on understanding how disease conditions differ from physiological conditions in molecular terms with the aim of finding new drug targets for therapy. Gene set enrichment analysis using the GO and its annotations provides a powerful way to assess those differences.

Roche has developed a bespoke controlled vocabulary (RCV) to support enrichment analysis. Each term is manually mapped to a list of Gene Ontology (GO) terms. The groupings are tailored to the research aims of Roche and as a result, many groupings are out-of-scope for GO classes. For example, many RCV terms group process and cell parts according to the cell type they occur in.

The manual mapping strategy is labour intensive and hard to sustain as the GO evolves. We have automated mappings between RCV and the GO via OWL-EL queries. This is made possible by extensive axiomatisation linking the GO to ontologies of cells, anatomical entites and chemicals. We can fully automate mapping for approximately one third of the terms in the RCV, with another 40% having 10 or fewer GO terms requiring manual mapping. Automated mapping uncovers many missing mappings. GSEA using the resulting, semi-automated mapping of RCV to GO detects enrichment to gene sets missed with the manual-only mapping.

The OWL query approach we describe can be used as the basis of new ways to query the GO, group annotations and carry out GSEA. Importantly, it allows the classifications used in enrichment analysis to be much more closely tailored to the needs of researchers and industry than was previously possible.

Keywords: OWL-EL, gene set enrichment analysis, gene expression, gene ontology, GO

1 Introduction

The Gene Ontology (GO) consists of almost 40,000 terms and has been used to annotate millions of gene products to record their subcellular location (e.g., lysosome), their molecular function (e.g., kinase activity) and their wider role in cellular, developmental and physiological processes (e.g., signal transduction) [4]. The classification and part hierarchies in the GO are used to group genes annotated with related terms in user-facing tools such as QuickGO [2] and AmiGO [3], and to generate gene sets for gene set enrichment analysis (GSEA) [16].

Much pharmacological research focuses on understanding the molecular differences between disease conditions and physiological conditions, with the aim of finding new drug targets for therapy. Differential expression experiments analysing disease models or pathological tissue samples are an important source of data contributing to our understanding of this. GSEA using GO derived gene sets is an efficient way to find functionally coherent gene sets that are statistically over or under-represented in gene lists derived from these experiments. GSEA results using the full GO and large numbers of genes can be difficult and slow to interpret due to high levels of overlap between gene sets. There are a number of sources of overlap: Grouping via class and part heirarchies means that gene sets derived from annotation to a class subsumes the gene sets of its subclasses and subparts; one GO class can sit in multiple branches of the heirarchy; a single gene product may be annotated to terms in multiple branches.

One way to reduce overlap is to use a flat list of high or intermediate level GO terms, commonly referred to as a slim. But for this to provide useful results, the terms in the slim need to be sufficiently descriptive to fit the experimental use cases. Rather than use a slim of GO terms, F. Hoffmann-La Roche Ltd. ("Roche"), maintains an internal controlled vocabulary (referred to hereafter as RCV) for use in GSEA. The RCV consists of around 360 terms, each of which is mapped to a set of terms from GO, just as a term in a GO slim maps to a set of subclasses and subparts. It is tailored to the research interests of Roche, and its terms were chosen with the aim of achieving gene set composition descriptive and broad enough to allow robust and statistically significant results, though not so broad and redundant in composition that it prevents easy interpretation of results. Detecting enrichment to gene products involved in anatomy, organ or cell-specific processes or components can be critical for pharmacological research, especially when working with complex tissues where there is a need to tease apart events occurring in specific tissue compartments or cell types. To support this, many RCV terms group GO terms in ways that are out of scope for classes in the GO, such as grouping processes solely by where they occur.

To date, Roche has manually maintained mappings between RCV and GO. Keeping this mapping up-to-date and complete has become impractical given the evolution of the GO. Recent developments in the GO make it possible to automate mappings between the RCV and the GO. The GO has switched its underlying formalization to Web Ontology Language (OWL2) [10], and has dramatically increased the number of logical axioms [17]. The chemical participants in over 12,000 processes or functions are specified in GO via axioms referencing

chemical entities defined by Chemical Entities of Biological Interest (ChEBI) [9, 8]. Over 8000 GO classes have some direct or indirect logical link to a term from the Cell Ontology (CL) [15] or the Uber anatomy ontology (Uberon) [7]. These record, for example, the location of cellular components (e.g., the acrosome and its parts are present only in sperm), cell types that are the sole location of some process ('natural killer cell degranulation' only occurs in natural killer cells), and the products of developmental processes (bone is a product of 'bone morphogenesis'). There are also over 2500 logical axioms recording the functions of cellular components via links to molecular function and biological process terms. This axiomatisation makes it possible to construct bespoke classifications of GO classes that would be out-of-scope for named GO classes. For example, we can use OWL queries to group processes occurring in T-cells or in the pancreas, or processes involving nitric oxide or collagen fibers. Here we describe the development and testing of an automated mapping between GO and RCV, making use of OWL reasoning.

2 Methods

As the RCV is a flat list and includes classifications that are orthogonal to the classification schemes used by the GO, it is not amenable to mapping via ontology alignment techniques that use ontology structure [1]. Given the small size of RCV, it is viable to manually map each RCV term to an OWL class expression, which can then be used in conjunction with an OWL reasoner to generate lists of GO terms. The RCV does not include textual definitions definitions to clarify meaning, so for each RCV term we attempted to find a class expression (a mapping query) that reflected the intended meaning of the RCV term, as judged by the RCV term name, manual mappings and discussion with RCV developers.

2.1 Query strategy

To ensure speed and scalability, we chose to restrict mapping queries to the EL profile of OWL2, allowing us to use ELK, a fast, scaleable EL reasoner, to run queries [13]. In order to keep the mapping process simple, only a single mapping class was specified for each mapping. To compensate partially for the lack of disjunction (OR) in OWL-EL, we developed a hierarchy of high level object properties for use in queries. For example, we define **occurs_in_OR_has_participant** as a grouping relation allowing queries for processes that occur in a specified cell, or have that cell as a participant. Many RCV terms group i) processes in which a specified chemical or cell participates with ii) processes regulating those in which it participates (see table 1 for example). To support such groupings, we used an OWL property chain axiom [10] to define a relation, **regulates_o_has_participant**, which can be used to query for processes that regulate a process in which some specified entity is a participant. We then defined a super-property, **participant_OR_reg_participant**, for this new relation and **has_participant**:

These new, high-level object properties are difficult to name in a way that communicates the meanings of mapping queries clearly. In order to compensate for this, we used scripting to generate human readable descriptions for each mapping query. Compare, for example, the mapping query for the RCV term cannabinoid with its description:

> ***Mapping query*: participant_OR_reg_participant** *some* cannabinoid
> ***Description*:** "A process in which a cannabinoid participates, or that regulates a process in which a cannabinoid participates."

Table 1. Results table for RCV cannabinoid. The table shows a comparison of the manual mapping of RCV to GO terms (manual column) with the automated mappings (auto column) resulting from an OWL query for processes in which a cannabinoid participates, or that regulates a process in which a cannabinoid participates. The automated mapping found three additional GO terms compared to the manual mapping. In this case, no manually mapped terms were obsolete in GO and all automated mappings were approved.

name	ID	manual	auto	checked	black listed	is obsolete
regulation of endocannabinoid signaling pathway	GO_2000124	1	1	1	0	0
cannabinoid signaling pathway	GO_0038171	1	1	1	0	0
endocannabinoid signaling pathway	GO_0071926	1	1	0	0	0
cannabinoid receptor activity	GO_0004949	0	1	1	0	0
cannabinoid biosynthetic process	GO_1901696	0	1	1	0	0

2.2 Pipeline

Mapping queries were run using the ELK OWL reasoner [13] via calls to the OWL-API [11]. The query and results processing pipeline was written in Jython [12]. All code, mapping tables and results were maintained in a GitHub repository [5]. The mapping was specified using a single tab separated values (TSV) file in which each line maps an RCV term to an OWL-EL mapping query that includes a term from GO, ChEBI, CL, Uberon or NCBI taxonomy [18]. Query results were used

to generate a TSV file, allowing direct comparison of manual and automated mappings (see table 1 for an example). We used the GitHub API to generate tickets for each mapping, linked to the relevant TSV results file, which GitHub renders as a table. This allowed easy manual review and editing by RCV curators at Roche who used the linked tickets to discuss mapping issues and record the approval status of all mappings.

Mapping queries were selected, tested and the results reviewed against manual mappings to decide which patterns were most appropriate. Once a mapping query was chosen, corrections and/or additions to the GO were made where results were wrong or incomplete. At this point, any clear errors in the manual mapping were blacklisted. Review of automated mappings was then passed to Roche who approved or blacklisted individual classes (see table 1 for an example). When satisfied with the results, the corresponding GitHub ticket was closed, thereby indicating the mapping as approved. Results approved by Roche were combined to produce a new RCV mapping table[1].

2.3 Gene set enrichment analyses

GSEA was performed using an open dataset comparing gene expression in adult liver and embryonic cells of mice [14]. Genes were ranked according to how much more highly they were expressed in liver vs embryonic cells and vice versa. GSEA enrichment scores were computed using GSEA software from the Broad Institute [19] with an up-to-date set of GO annotations to mouse genes [2]. The results were analysed using the Enrichment Map plugin for Cytoscape [16], which provides a graphical representation of enrichment results.

3 Results

3.1 Mapping results

We developed mapping queries for 308/364 RCV terms. Over a third (104) of the mapping queries were sufficient - meaning that no manual maintenance is required. A further 40% of the mappings (148) had 10 or fewer additional manual mappings (figure 1A) and most of these (114) had fewer than 5.

Mapping queries identified many GO terms that were not in the manual mapping (figure 1B). In some cases (e.g., leukocyte activation), over 1000 new mappings were found. Only 8 automated mappings had blacklisted terms, reflecting minor differences between the meaning of the mapping query and the intended meaning of the RCV term. 56 terms were not mapped. Some were were judged to be semantically equivalent to other RCV terms. The rest were rejected as currently not mappable due to the lack of suitable terms or axiomatisation within the GO. For example, RCV has terms for aerobic and anaerobic metabolic

[1] Available from: `https://github.com/GO-ROCHE-COLLAB/Roche_CV_mapping/blob/master/mapping_tables/results/combined_results.tsv`

[2] Available from `http://geneontology.org/page/download-annotations`

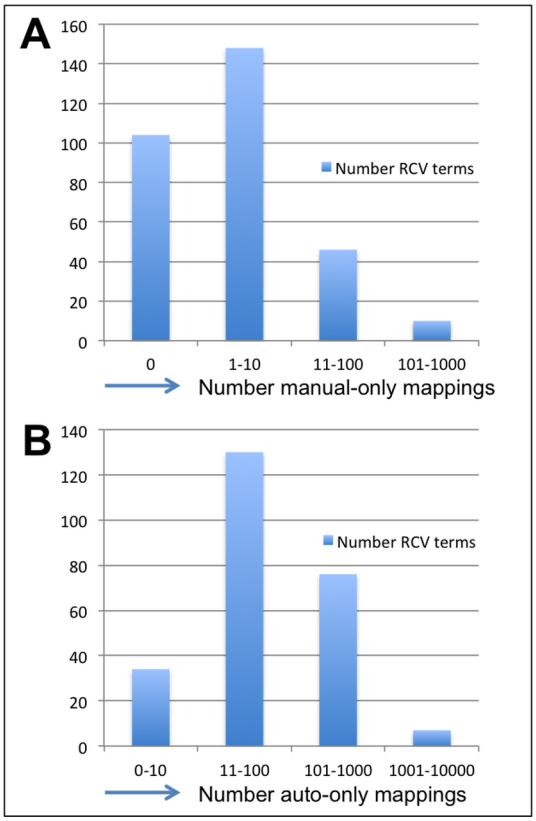


Fig. 1. Summary or mapping results *A. Distribution of manual only mappings.* X axis = number of manual-only mappings. Y axis = Number of RCV terms. Over 80% of mappings are completely automated or require less than 10 manual mappings. *B. Distribution of auto only mappings.* X axis = number of auto-only mappings. Y axis = Number of RCV terms. Many new mappings were uncovered by automation.

processes, but GO currently has no formal way to group these and no sustainable mechanisms for grouping them manually. Further formalisation of the GO is likely to improve the number of RCV terms that can be mapped.

3.2 Testing the implication of automated mapping for gene set enrichment analyses

We tested the the revised, semi-automated RCV mapping by performing GSEA comparing the transcriptome of adult mouse liver against embryonic mouse cells using a standard GO slim (Figure 2A), the original, manually mapped RCV (RCV_man; Figure 2B) and the new partially automated RCV GO mapping (RCV_auto; Figure 2C). Profiles of gene enrichment in liver compared to undifferentiated cells are potentially useful benchmarks for the development and

testing of stem cell derived liver *in vitro* systems increasingly employed in toxicology testing. Results were loosely grouped by experts at Roche to provide a preliminary, biologically plausible interpretation. All approaches detected enrichment to gene sets involved in cell division and gene expression in the embryonic sample, but there were dramatic differences in detection of enrichment in the liver sample.

GSEA with the standard GO slim detects enrichment in the liver to a few sets of genes involved in metabolic processes that are known to be up-regulated in the liver. GSEA with RCV_man also detects enrichment to many more metabolic processes that are specific to or upregulated in the liver, and to a much finer level of detail. It also detects enrichment of genes involved in immune cell related process, consistent with detection of the resident immune system in the liver. The results of enrichment with RCV_auto are similar, but provide much more detail. For example, GSEA with RCV_auto detects enrichment to gene sets involved in detoxification (important for toxicology use cases) and a wider range of immune cell processes. There is also increased overlap between enriched gene sets compared to RCV_man, but at a level that is potentially informative (see edges between nodes in 2C). For example, there is overlap between sets of genes involved in both chemotaxis and processes involving types of immune cell that are known to be capable of chemotaxis.

3.3 Improvements to the GO

While GO has extensive axiomatisation linking processes to cells, anatomical structures and chemicals, this is not always complete. In mapping from the RCV to GO we found and corrected over 200 omissions in the axiomatisation. This included missing links from processes to participant cell types, anatomical structures, chemicals, cell components and transcript types. We also found and corrected a number of errors, including errors in axiomatisation of developmental processes that led to incorrect inferences for RCV anatomy terms.

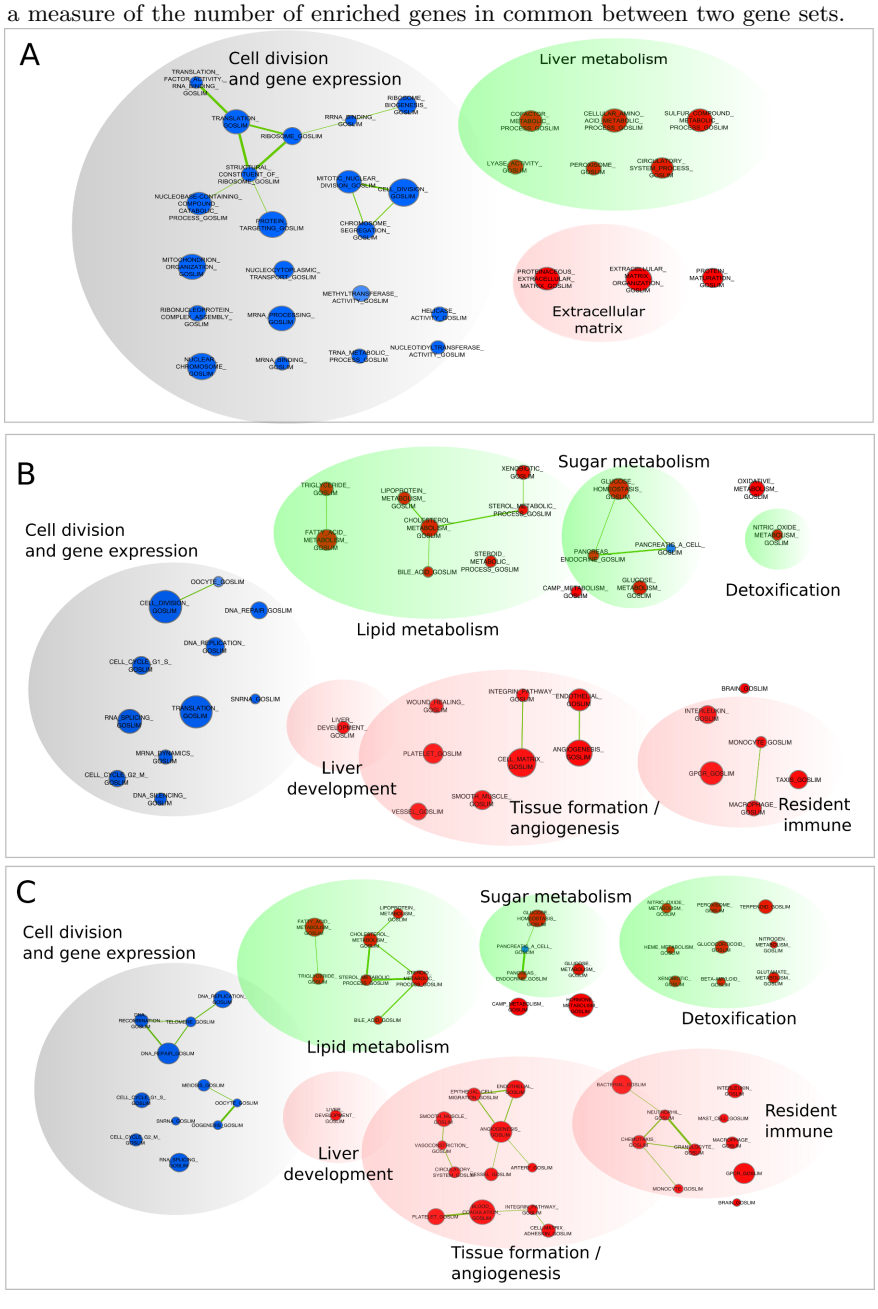
4 Discussion and future directions

This work demonstrates how the logical structure of the GO can be used to achieve biologically meaningful mappings between GO terms and terms from external controlled vocabularies defined with reference to ryps of cells, chemicals or anatomical structures. Mappings are straightforward to specify and the reasoning system used is fast and scalable [13, 17]. All mappings that are fully automated can be automatically updated as the GO changes, simply by running the mapping pipeline. Errors found during manual review were sufficiently rare that this step will not be used in future updates.

4.1 Improving the RCV mapping to GO

48% of mapped RCV terms have 10 or fewer manual mappings. We are reviewing all of these cases to decide whether to drop manual mappings or whether

Fig. 2. GSEA comparing expression between embryo and adult liver using gene sets derived from the generic GO slim (Panel A), manually mapped RCV (panel B), and semi-automated RCV (panel C). Red nodes indicate gene sets enriched in liver compared to embryos. Blue nodes gene sets enriched in embryos compared to liver. The size of the node is proportional to the size of the gene set. Connecting edge thickness is a measure of the number of enriched genes in common between two gene sets.



complete automation might be achieved by a different query strategy. In some cases, a more complete mapping could be achieved by combining the results of multiple mapping queries. For example, RCV terms for chemical metabolism are all manually mapped to GO terms for both metabolism and transport. A more complete mapping could be achieved by combining the results of separate OWL mapping queries for GO transport and GO metabolic process [3].

4.2 Alternative views of the GO and its annotations

The OWL axioms used to automate RCV mapping to GO can also be used to provide alternative views of the GO and its annotations. This is already reflected in some of the newer functionalities of the GO browsing tool AMIGO, which now displays inferred annotations to cell-types based on axioms in GO recording where processes occur[4].

4.3 Improving mechanisms for extending RCV

The system described here was designed to be lightweight and flexible, allowing maximum interaction between the designers of RCV at Roche and GO editors with minimal development overhead. Where new terms following mapping query patterns already used, they can be added via the same mechanism.

The system described bears some relationship to TermGenie [6] which is already used to generate 80% of new GO terms. One possible approach to fulfilling the needs of external groups for types of classification not included in the GO would be to offer a TermGenie-like system to create bespoke terms.

Funding This work was supported by direct funding from F. Hoffmann-La Roche Ltd and by European Molecular Biology Laboratory (EMBL) core funding. The Gene Ontology Consortium is supported by a P41 grant from the National Human Genome Research Institute (NHGRI) [grant 5U41HG002273-14].

References

1. Siham Amrouch and Sihem Mostefai. Survey on the literature of ontology mapping, alignment and merging. In *Information Technology and e-Services (ICITeS), 2012 International Conference on*, pages 1–5, March 2012.
2. David Binns, Emily Dimmer, Rachael Huntley, Daniel Barrell, Claire O'Donovan, and Rolf Apweiler. Quickgo: a web-based tool for gene ontology searching. *Bioinformatics*, 25(22):3045–3046, 2009.
3. Seth Carbon, Amelia Ireland, Christopher J. Mungall, ShengQiang Shu, Brad Marshall, Suzanna Lewis, the AmiGO Hub, and the Web Presence Working Group. Amigo: online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289, 2009.

[3] This is not formally equivalent to running OWL queries with disjunction (OR), but we expect few, if any, differences in results given the current GO axiomatisation.

[4] `http://amigo.geneontology.org/amigo/term/CL:0000084`

4. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, 43(Database issue):D1049–1056, 2015.
5. David Osumi-Sutherland. The GO-Roche project - available at `https://github.com/GO-ROCHE-COLLAB/Roche_CV_mapping`., September 2015.
6. H. Dietze, T. Z. Berardini, R. E. Foulger, D. P. Hill, J. Lomax, D. Osumi-Sutherland, P. Roncaglia, and C. J. Mungall. TermGenie - a web-application for pattern-based ontology class generation. *J Biomed Semantics*, 5:48, 2014.
7. M. A. Haendel, J. P. Balhoff, F. B. Bastian, D. C. Blackburn, J. A. Blake, Y. Bradford, A. Comte, W. M. Dahdul, T. A. Dececchi, R. E. Druzinsky, T. F. Hayamizu, N. Ibrahim, S. E. Lewis, P. M. Mabee, A. Niknejad, M. Robinson-Rechavi, P. C. Sereno, and C. J. Mungall. Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *J Biomed Semantics*, 5:21, 2014.
8. J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, and C. Steinbeck. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, 41(Database issue):D456–463, Jan 2013.
9. David P Hill, Nico Adams, Mike Bada, Colin Batchelor, Tanya Z Berardini, Heiko Dietze, Harold J Drabkin, Marcus Ennis, Rebecca E Foulger, Midori A Harris, Janna Hastings, Namrata S Kale, Paula de Matos, Christopher Mungall, Gareth Owen, Paola Roncaglia, Christoph Steinbeck, Steve Turner, and Jane Lomax. Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology. *BMC genomics*, 14(1):513, January 2013.
10. Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F. Patel-Schneider, and Sebastian Rudolph, editors. *OWL 2 Web Ontology Language: Primer*. W3C Recommendation, 27 October 2009. Available at `http://www.w3.org/TR/owl2-primer/`.
11. Matthew Horridge and Sean Bechhofer. The owl api: A java api for owl ontologies. *Semant. web*, 2(1):11–21, January 2011.
12. Jython maintainers. The Jython project - available at `http://www.jython.org`., September 2015.
13. Yevgeny Kazakov, Markus Krötzsch, and František Simančík. Elk reasoner: Architecture and evaluation. *CEUR Workshop Proceedings*, 858, 2012.
14. R. Lowe. Sexually dimorphic gene expression emerges with embryonic genome activation and is dynamic throughout development (rna-seq). - available at `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58733`, 2015.
15. T. F. Meehan, A. M. Masci, A. Abdulla, L. G. Cowell, J. A. Blake, C. J. Mungall, and A. D. Diehl. Logical development of the cell ontology. *BMC Bioinformatics*, 12:6, 2011.
16. D. Merico, R. Isserlin, O. Stueker, A. Emili, and G. D. Bader. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE*, 5(11):e13984, 2010.
17. C. Mungall, H. Deitze, and D. Osumi-Sutherland. Use of OWL within the Gene Ontology. In C. Maria Keet and V. Tamma, editors, *Proceedings of the 11th International Workshop on OWL: Experiences and Directions (OWLED 2014)*, volume 1265 of *CEUR workshop proceedings*, pages 25–36, 2014.
18. National Institutes of Health National Center for Biotechnology Information (NCBI), National Library of Medicine. The ncbi entrez taxonomy homepage.
19. Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.