Intelligent workflows for automated analysis of mass spectrometry-based proteomics data

Güler, A.T.

CHAPTER 4

COMICS: Cartoon Visualization of Omics Data in Spatial Context Using Anatomical Ontologies

Dmitrii Travin[1,*], Iaroslav Popov[1,*], Arzu Tugce Guler[2], Dmitry Medvedev[1], Suzanne van der Plas-Duivesteijn[2], Monica Varela[3], Iris C. R. M. Kolder[3], Annemarie H. Meijer[3], Herman P. Spaink[3], Magnus Palmblad[2]

[1] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, 119234 Moscow, Russian Federation

[2] Center for Proteomics and Metabolomics, Leiden University Medical Center, PO Box 9600, 2300 RC, Leiden, The Netherlands

[3] Institute of Biology, Leiden University, PO Box 9502, 2300 RA, Leiden, The Netherlands

[*] shared authors

This chapter has been published in:
Journal of Proteome Research (2018) 17, 739-744
doi: 10.1021/acs.jproteome.7b00615

Abstract

COMICS is an interactive and open-access Web platform for integration and visualization of molecular expression data in anatomograms of zebrafish, carp and mouse model systems. Anatomical ontologies are used to map omics data across experiments and between an experiment and a particular visualization in a data dependent manner. COMICS is built on top of several existing resources. Zebrafish and mouse anatomical ontologies with their controlled vocabulary (CV) and defined hierarchy are used with the ontoCAT R package to aggregate data for comparison and visualization. Libraries from the QGIS geographical information system are used with the R packages "maps" and "maptools" to visualize and interact with molecular expression data in anatomical drawings of model systems. COMICS allows users to upload their own data from omics experiments, using any gene or protein nomenclature they wish, as long as CV terms are used to define anatomical regions or developmental stages. Common nomenclatures such as the ZFIN gene names and UniProt accessions are provided additional support. COMICS can be used to generate publication-quality visualization of gene and protein expression across experiments. Unlike previous tools that have used anatomical ontologies to interpret imaging data in several animal models, including zebrafish, COMICS is designed to take spatially resolved data generated by dissection or fractionation and display this data in visually clear anatomical representations rather than large data tables. COMICS is optimized for ease-of-use, with a minimalistic web interface and automatic selection of the appropriate visual representation depending on the input data.

Introduction

Ontologies

For more than a decade, ontology-based data integration has been used to merge heterogeneous data in many domains, including bioinformatics[1]. In many disciplines, ontologies have to be actively maintained to keep up with the development or new discoveries in the field. This is particularly true for the more technical ontologies used to annotate datasets in genomics or proteomics, such as the PRIDE Controlled Vocabulary[2], and ontologies used to describe bioinformatics operations as well as data types, formats and identifiers, such as EDAM[3]. However, there are also examples of mature and essentially complete ontologies. These include the anatomical ontologies of well-studied organisms, the anatomies themselves being highly conserved over time (millions of years). Simpler controlled vocabularies (CVs) may be sufficient for some purposes, such as standardizing the way datasets in public repositories are annotated with metadata. However, when comparing or integrating heterogeneous (or heterogeneously annotated) data generated in different laboratories or using different experimental protocols, such CVs lack the necessary structure. A proteomics researcher may wish to find mass spectrometry datasets from an organism of interest generated using any "electrospray ionization" (CV term ID "MS:1000073") technique to build a spectral library of comparable data. But if some such datasets are annotated as having been acquired with "microelectrospray" (MS:1000397) and others as being derived from a "nanoelectrospray" (MS:1000398) experiment, how does the software know these all qualify as "electrospray ionization" mass spectrometry datasets? This information is provided by the relationships between the terms as defined in an ontology. In this case, both the specific "microelectrospray" and "nanoelectrospray" have a direct "is a" relationship with the more general or parent "electrospray ionization". One can therefore reason that they are all "electrospray ionization" datasets, and hence compatible for this researcher's defined purpose.

Common methods for generating deep proteomics datasets often involve separation or fractionation. These can be applied on the sample level, for example, by dissection[4], cell sorting[5] or organelle fractionation[6], each defining a spatial context of subsequently generated data. Fractionation on the protein level is also commonplace,

and provide a protein-level context for peptide-level data. When comparing two such large datasets in any -omics field, we cannot assume the two datasets have been acquired in exactly the same way. Depending on the laboratory, equipment, experimental protocol, skills of the experimentalists involved, or allocated effort, the dissection or fractionation may have been done differently, altering the spatial definition of the fractions of the dataset. To integrate such datasets for the purpose of comparison of spatial expression patterns, the datasets must be annotated using something like an anatomical or cellular ontology, with defined relationships between anatomical entities. Many such ontologies already exist, including the model-system specific *C. elegans* gross anatomy (WBBT)[7], the Drosophila gross anatomy (FBbt and FBdv), also referred to as the Drosophila anatomy ontology (DAO)[8], the Mouse Adult Gross Anatomy (MA)[9], Xenopus anatomy and development (XAO)[10] and Zebrafish anatomy and development (ZFA and ZFS)[11]. There are also the more general Anatomical Entity Ontology (AEO)[12], Biological Spatial Ontology (BSPO)[13] and the general vertebrate "Uber-anatomy" ontology (UBERON)[14] currently (20170415) containing 15,036 anatomical terms. The zebrafish ZFA and ZFS ontologies contain 3175 anatomical terms (20170627 release) and the mouse MA 3257 terms (20170207 version). For comparison, the two major ontologies covering human anatomy, the Foundational Model of Anatomy (FMA)[15] and SNOMED-CT[16], contain 75,019 and 30,933 anatomical concepts respectively[17].

Anatomical visualization

In their classic 1987 paper "Why a Diagram is (Sometimes) Worth Ten Thousand Words"[18], Larkin and Simon demonstrated how well-made figures or diagrams use location to group information, reduce the need for symbolic labels and enable a large number of conceptual inferences to be made, something the human brain is extremely good at. Larkin and Simon argued that the main advantages of diagrams are *computational* - diagrams are better representations not because they contain *more* information, but because the *indexing* of this information support extremely efficient computational processes, including those carried out in the human brain upon trying to grasp the contents of a research paper. Anatomical schemata or anatomograms are now used to interact with on-line databases, such as Reactome[19], the Human Protein Atlas[20], ProteomicsDB[21] and the EMBL-EBI Expression Atlas[22].

This paper describes a new stand-alone freeware, COMICS, with an interactive web-based interface designed to fit into a niche between existing tools for combined integration and visualization of molecular expression data in some vertebrate model organisms (zebrafish, carp and mouse). The software uses the existing anatomical ontologies to map arbitrary omics data across experiments and between one experiment and a particular visualization in a data-dependent manner. The method and software can be extended to other model systems, provided the relevant ontology and visual representation (picture). COMICS is designed for simplicity-of-use, and can generate custom, publication-quality, vector graphics mapping molecular expression (such as from transcriptomics, proteomics or metabolomics) data to anatomical diagrams. In addition to molecular expression levels, the locations in the diagram immediately convey information on similarity or dissimilarity between adjacent structures or parts of an organ, such as the eye or the brain, tissue specificity (one part against the whole) and differences in expression levels between genes/proteins or between animals.

Methods

COMICS takes as input a table of numerical data (e.g., gene or protein expression values) with each row corresponding to one CV term from an anatomical ontology, such as the ZFA[11] or MA[9], and each column to one particular gene or protein, with the CV terms as row names and gene or protein identifiers as column names. If the molecular identifiers and anatomical CV terms are swapped, then COMICS will automatically detect this and transpose the matrix. COMICS requires CV terms instead of common names of anatomical features to be able to match them correctly with parts of the picture. For carp, we also apply ZFA ontology CV-terms as there is no specific ontology for this species. Both species belong to a single Cyprinidae family and are quite close in terms of tissues and organs present[23].

First, the CV terms in the data uploaded by the user are matched to CV terms with a corresponding polygon defined in the shapefile for the selected species. This is performed using the R package ontoCAT[24], which enables extracting term parents and children (generalization/specialization) as well as terms with a part of/has part (whole/part) relationship with the given term from the anatomical ontology. This is a

key step that allows any correctly annotated data to be mapped by COMICS to the anatomical representations in the shapefiles. An example of the ontology-based pre-processing and aggregation of molecular expression data is shown in Figure 4.1. For computational efficiency, a lookup table for the mapping between the ontology and visualization shapefile is computed for each ontology. This lookup table is rebuilt once for each new version of the ontology or shapefile.

For the anatomical drawings, we used the mature QGIS open-source geographic information system[25] to create shapefiles. These shapefiles were constructed from simple polygons corresponding to anatomical structures such as organs or parts of organs in zebrafish and carp. These shapefiles can easily be extended to include other model systems or developmental stages for which anatomical ontologies are available. Inspiration for the anatomical illustrations was drawn from previously published work[26,27,28].

To visualize the numerical data obtained from the user on the anatomical shapefiles we used the existing *maps* and *maptools* R packages commonly used for working with maps and *gridSVG* to produce vector graphics in the SVG format. The Adobe PDF is supported by the pre-installed *grDevices* package. The range of numerical data is translated to a palette of colors forming a one-, two- or three-color gradient. COMICS has several options that enable the user to choose from several predefined color schemes or make a new one and choose the number of bins for the gradient and scaling (linear or logarithmic). In addition, the user can keep the gradient fixed across diagrams or scale it automatically for each visualization. The former option is used for comparing (absolute) expression across many diagrams. The latter automatically adapts to the minimum and maximum values in the data for each gene or protein and is optimal for looking at tissue specificity or relative expression of two genes or proteins. The expression of two entities can also be computed and compared directly in COMICS.

The cartoons can be saved individually or as a collection, as vector graphics in the PDF or SVG formats.

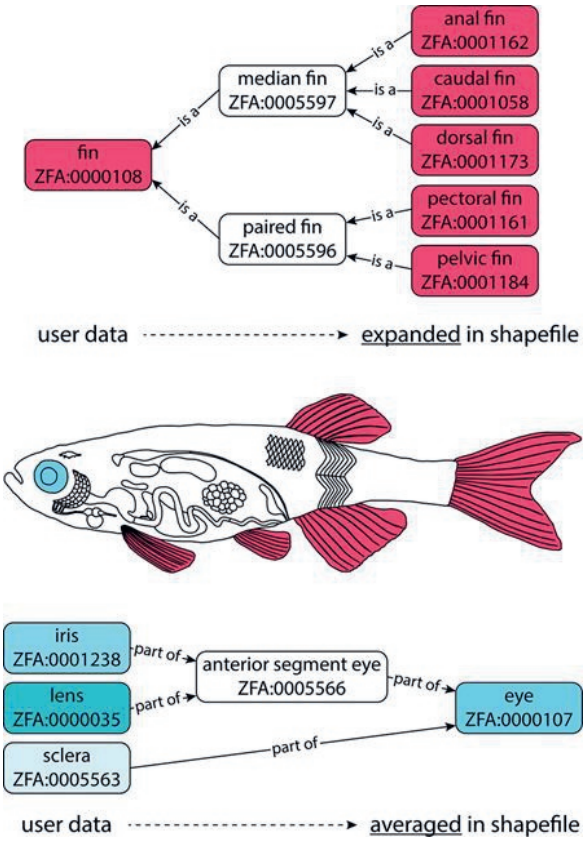

Figure 4.1. ZFA anatomical ontology is used to map scalar expression data to defined anatomical regions. This also provides a means to directly and visually compare data from different experiments and heterogeneous datasets. In this example, the user has provided data for "fin", which is then propagated to the five distinct fins visualized in the tool, through the parent-child (is a) relationships defined in ZFA. Because the fins are not distinguished in the user's dataset, the expression value provided by the user is mapped to all five visible fins. If the user provides information on a more detailed level than is visualized by COMICS, then the mean expression of all children or parts are mapped to the anatomical structure defined in the shapefile. Here, separate expression data for the iris, sclera and lens (all part of the eye) are averaged to the eye. The averaging is done once, for all parts, independent of intermediate levels in the ontological hierarchy (such as the anterior segment eye). The default shapefile corresponds to the organs and tissues that are easy to dissect for an omics experiment, although the shapefile can easily be modified to incorporate other experimental designs.

Technically, COMICS is a web application build around R scripts. For standalone usage it is containerized using Docker. The container includes all software, including source code, packages and scripts, making it very easy to install and run COMICS locally, independently of other installed software. The standalone mode enables the user to work with the application locally, without uploading datasets to any server. Links to the Docker container and locations where COMICS can be run remotely will be maintained on https://edu.nl/drrew.

To test COMICS, we used previously published data from the public domain. Wildtype gene expression data for zebrafish was taken from ZFIN, already annotated using the ZFA[29]. Protein expression data in adult zebrafish were taken from the zebrafish spectral library[4]. Expression data from carp were taken from a recent paper on the full-body transcriptome and proteome resource for this species[30]. Mouse gene expression data was downloaded from the Mouse Atlas of Gene Expression[31], and mouse protein data was generated in-house using the same method as for the zebrafish spectral library.

Results

The main product of this work is a software tool with a simple web interface as shown in Figure 4.2. The screenshot visualizes the gene expression of the carp ortholog of zebrafish cytokeratin-8, using the ZFA ontology mapped onto the anatomy of a carp, closely resembling that of zebrafish. The interface is divided into panels containing basic information about the underlying data, image controls, the image itself and links to cross-referenced databases (here UniProt, ZFIN and NCBI). The image is interactive: as the user hovers the mouse pointer over an anatomical region, the tooltip displays the name, ontology identifier and expression level (here for the dorsal fin). Clicking on the anatomical structure will lead to the web page for this part in the online version of the corresponding ontology. The image shapefiles annotated with the ZFA and MA anatomical ontologies are available as individual files for developers who would like to integrate them in their own software.

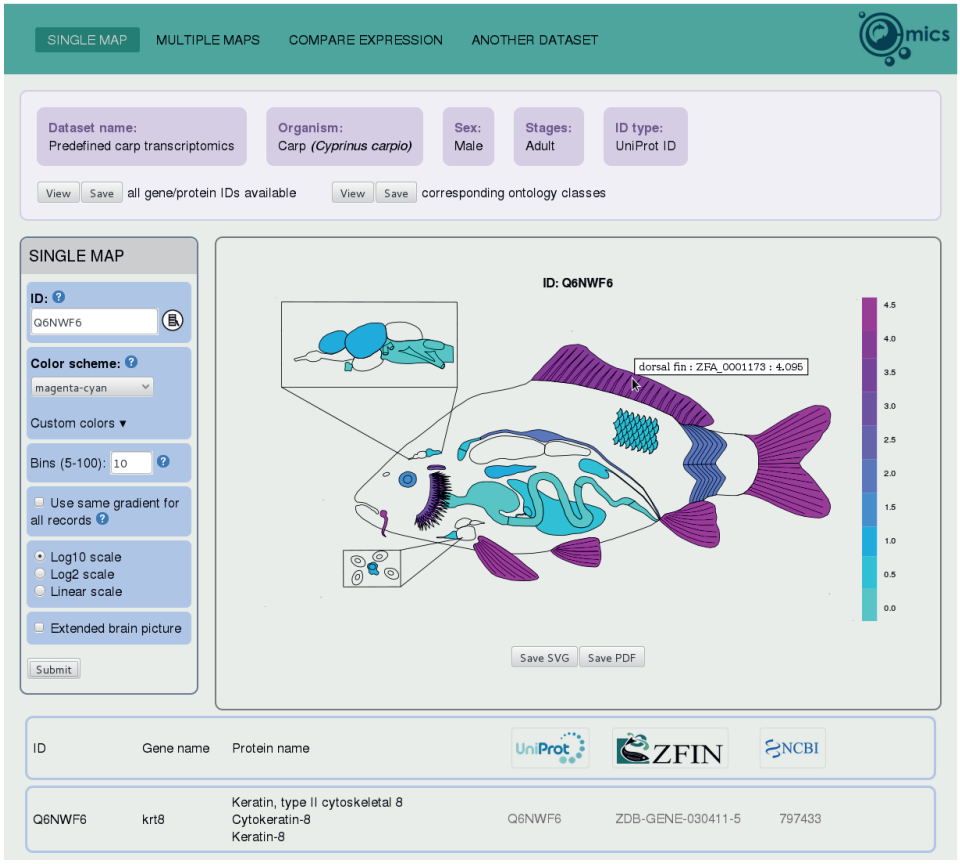

Figure 4.2. Screenshot of the COMICS interface, presenting the information about the selected dataset (top), a control panel with options and parameters for visualization (left), the generated output image (center, right) and a table containing the selected gene/protein description with links to the corresponding databases (bottom). Gene expression data[30] for the carp cytokeratin-8 (Q6NWF6) ortholog is here used as an example.

The COMICS tool is generic because it aggregates and displays any numerical data provided with anatomical ontology annotations linked to a shapefile. The tool can therefore be used to compare the expression of a few genes or proteins in one experiment and model system, look at the ratio of transcripts and the corresponding proteins, or compare the expression of orthologs across model systems. Figure 4.3 shows the expression of sarcosine dehydrogenase in zebrafish (*sardh* gene) and mouse (the sarcosine dehydrogenase protein), respectively, revealing the expression pattern for this pair of orthologs is conserved across the vertebrate subphylum (the

last common ancestor of the mouse and the two cyprinids lived over 400 million years ago[32]). As a final verification of the parsing of the anatomical ontology we looked at the expression of four genes with well-known spatial specificity in ZFIN (Figure 4.4). The four panels visualize gene expression, quantified as the number of experiments in which the transcript has been observed in wildtype fish and recorded by ZFIN, of four genes: rhodopsin (*rho*, ZDB-GENE-990415-271) in the eye (a), fatty acid binding protein 1a (*fabp1a*, ZDB-GENE-020318-3) in the liver (b), proopiomelanocortin a (*pomca*, ZDB-GENE-030513-2) in the brain, specifically the hypothalamus (c) and vitellogenin 2 (*vtg2*, ZDB-GENE-001201-2) in the liver and ovaries (d).

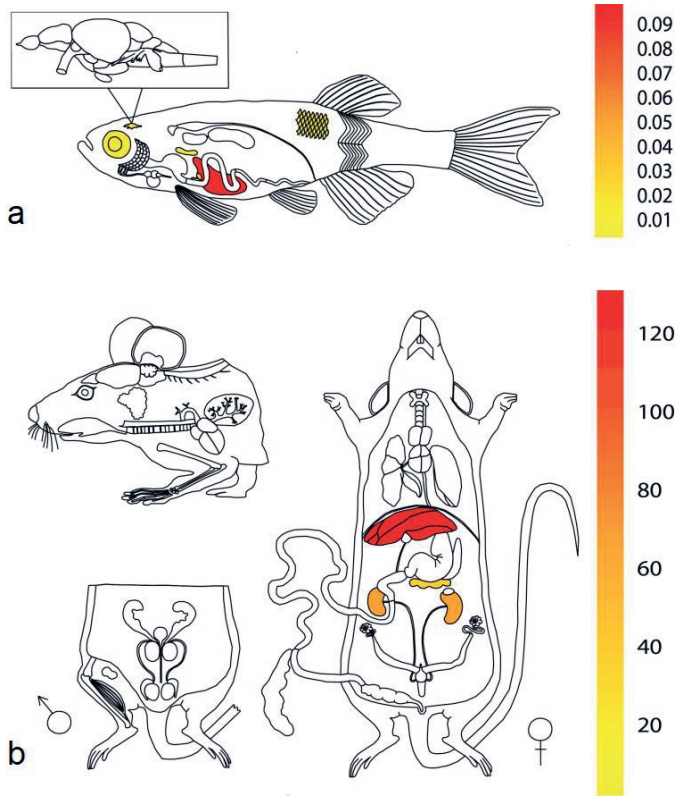

Figure 4.3. Publication-quality figures, showing the expression of Sarcosine dehydrogenase orthologs in zebrafish (*sardh* gene) **(a)**, mouse (Sarcosine dehydrogenase *protein*, UniProt accession number Q99LB7) **(b)**. The numbers on the color scales represent the fraction of experiments in ZFIN in which gene expression is observed (a) and absolute spectral counts (b)

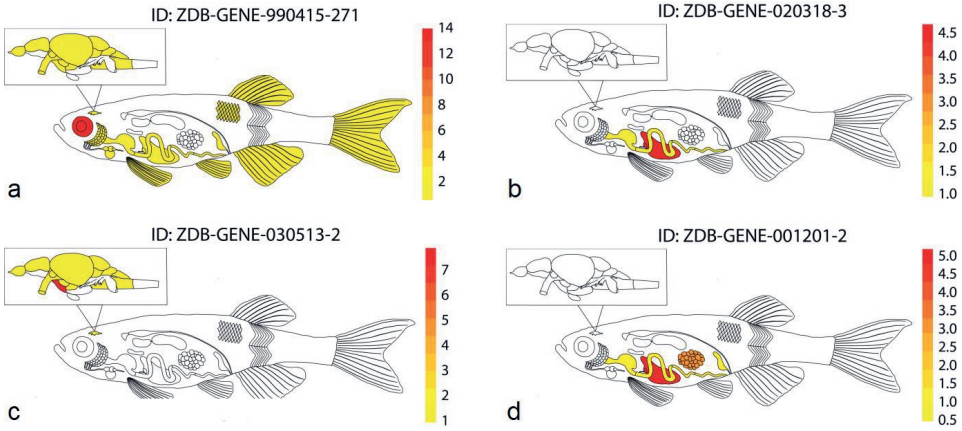

Figure 4.4. Organ-specific expression of four genes in zebrafish: rhodopsin **(a)**, fatty acid binding protein 1a **(b)**, proopiomelanocortin a **(c)**, vitellogenin 2 **(d)**, according to the number of registered detection of expression among all wildtype datasets in the ZFIN gene expression database. The color scale represent the number of experiments in ZFIN in which gene expression was observed in a particular organ or tissue.

If COMICS detects the presence of only male or female organ data, then the anatomical map will represent a single sex. If neither or both male and female organ annotations are included in the dataset, then a generic anatomical representation will be used. For mouse, a model with common superior and split inferior regions is also available.

Discussion

To summarize, COMICS is a simple, easy-to-use tool for generating visually clear, publication-quality vector graphics from arbitrary omics data using the mouse and zebrafish anatomical ontologies. COMICS should not be compared with resources pre-dating the development of these anatomical ontologies, such as the now off-line GEMS database[33], which was aimed at annotation of real images. COMICS can be used to compare the expression of a pair of genes or proteins, such as two isoforms, or the expression of a gene measured on the transcript and protein levels. In this way, one can visually inspect and quickly assess results from an ontology-based aggregation of two or more heterogeneous, spatially resolved, omics datasets. COMICS is not a tool to provide detailed and beautiful anatomical illustrations of an organism in the tradition of Vesalius[34]. Rather, we have deliberately compromised anatomical precision for

diagrammatic simplicity, ensuring the cartoons are clear also when viewed at a small scale, allowing quick side-by-side comparison of datasets. Future extensions of COMICS will include shapefiles of different embryonic and larval stages using the ZFS ontology as well as additional model systems.

Conclusions

We have here presented a simple software, COMICS, for mapping any numerical gene, protein or metabolomics data as choropleths in anatomical cartoons referred to as anatomograms. Unlike existing tools, COMICS makes full use of anatomical ontologies to integrate spatially or anatomically resolved data in several animal models, including zebrafish and mouse. COMICS is built on existing libraries and has a minimalistic web interface for selecting the appropriate visual representation and exporting publication-quality graphics. Additional model systems (as well as human anatomy or other developmental stages) are easy to add to the COMICS platform, provided an anatomical ontology in the OBO format and an organism-specific shapefile with mappings to the CV terms in the ontology are available. COMICS can be downloaded as a Docker image from https://edu.nl/drrew.

Acknowledgments

We gratefully acknowledge financial support from The Netherlands Organisation for Scientific Research (NWO) Vidi grant 917.11.398 (M.P.) and Olga Gancharova for valuable advice on the mouse anatomogram.

References

1. Wache, H. *et al.* Ontology-based Integration of Information - A Survey of Existing Approaches. in *Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing, August 4-5, 2001, Seattle, USA* 108–117 (2001).
2. Jones, P. *et al.* PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.* **34**, D659–D663 (2006).
3. Ison, J. *et al.* EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics* **29**, 1325–1332 (2013).
4. van der Plas-Duivesteijn, S. J. *et al.* Identifying proteins in zebrafish embryos using spectral libraries generated from dissected adult organs and tissues. *J. Proteome Res.* **13**, 1537–1544 (2014).
5. Bernas, T., Grégori, G., Asem, E. K. & Robinson, J. P. Integrating cytomics and proteomics. *Mol. Cell. Proteomics* **5**, 2–13 (2006).
6. Lee, Y. H., Tan, H. T. & Chung, M. C. M. Subcellular fractionation methods and strategies for proteomics. *Proteomics* **10**, 3935–3956 (2010).
7. Lee, R. Y. N. & Sternberg, P. W. Building a cell and anatomy ontology of Caenorhabditis elegans. *Comp. Funct. Genomics* **4**, 121–126 (2003).
8. Costa, M., Reeve, S., Grumbling, G. & Osumi-Sutherland, D. The Drosophila anatomy ontology. *J. Biomed. Semantics* **4**, 1–11 (2013).
9. Hayamizu, T. F., Baldock, R. A. & Ringwald, M. Mouse anatomy ontologies: enhancements and tools for exploring and integrating biomedical data. *Mamm. Genome* **26**, 422–430 (2015).
10. Segerdell, E., Bowes, J. B., Pollet, N. & Vize, P. D. An ontology for Xenopus anatomy and development. *BMC Dev. Biol.* **8**, 92 (2008).
11. van Slyke, C. E., Bradford, Y. M., Westerfield, M. & Haendel, M. A. The zebrafish anatomy and stage ontologies: Representing the anatomy and development of Danio rerio. *J. Biomed. Semantics* **5**, 12 (2014).
12. Bard, J. B. L. The AEO, an ontology of anatomical entities for classifying animal tissues and organs. *Front. Genet.* **3**, 18 (2012).
13. Dahdul, W. M. *et al.* Nose to tail, roots to shoots: Spatial descriptors for phenotypic diversity in the Biological Spatial Ontology. *J. Biomed. Semantics* **5**, 34 (2014).
14. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* **13**, R5 (2012).
15. Zhang, S. & Bodenreider, O. Aligning representations of anatomy using lexical and structural methods. *AMIA Annu. Symp. Proc.* **2003**, 753–757 (2003).
16. Côté, R. A. & Robboy, S. Progress in Medical Information Management. Systemized Nomenclature of Medicine (SNOMED). *J. Am. Med. Assoc.* **243**, 756–762 (1980).

17. Bodenreider, O. & Zhang, S. Comparing the representation of anatomy in the FMA and SNOMED CT. in *AMIA Annual Symposium Proceedings, November 11-15, 2006, Washington, DC, USA* 46–50 (2006).

18. Larkin, J. H. & Simon, H. A. Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cogn. Sci.* **11**, 65–99 (1987).

19. Fabregat, A. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **44**, D481–D487 (2016).

20. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

21. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).

22. Petryszak, R. *et al.* Expression Atlas update - An integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.* **44**, D746–D752 (2016).

23. Henkel, C. V. *et al.* Comparison of the exomes of common carp (Cyprinus carpio) and zebrafish (Danio rerio). *Zebrafish* **9**, 59–67 (2012).

24. Kurbatova, N., Adamusiak, T., Kurnosov, P., Swertz, M. A. & Kapushesky, M. ontoCAT: An R package for ontology traversal and search. *Bioinformatics* **27**, 2468–2470 (2011).

25. QGIS Association. QGIS Geographic Information System. https://www.qgis.org.

26. Davidson, A. J. & Zon, L. I. The 'definitive' (and 'primitive') guide to zebrafish hematopoiesis. *Oncogene* **23**, 7233–7246 (2004).

27. Wulliman, M. F., Rupp, B. & Reichert, H. *Neuroanatomy of the zebrafish brain: a topological atlas*. (Birkhäuser Verlag Basel, 1996).

28. Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B. & Schilling, T. F. Stages of embryonic development of the zebrafish. *Dev. Dyn.* **203**, 253–310 (1995).

29. The Zebrafish Information Network. Expression Data for Wildtype Fish. https://zfin.org/downloads/wildtype-expression_fish.txt.

30. Kolder, I. C. *et al.* A full-body transcriptome and proteome resource for the European common carp. *BMC Genomics* **17**, 701 (2016).

31. Siddiqui, A. S. *et al.* A mouse atlas of gene expression: Large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 18485–18490 (2005).

32. Broughton, R. E., Betancur-R., R., Li, C., Arratia, G. & Ortí, G. Multi-locus phylogenetic analysis reveals the pattern and tempo of bony fish evolution. *PLoS Curr. Tree Life* **5** (2013).

33. Belmamoune, M. & Verbeek, F. J. Data integration for spatio-temporal patterns of gene expression of zebrafish development: the GEMS database. *J. Integr. Bioinform.* **5** (2008).

34. Vesalius, A. *De Humani Corporis Fabrica Libri Septem.* (Padua School of Medicine, Padua, Italy, 1543).