Bioinformatics, 36(9), 2020, 2881–2887
doi: 10.1093/bioinformatics/btaa023
Advance Access Publication Date: 17 January 2020
Original Paper

Databases and ontologies

Curation and annotation of planarian gene expression patterns with segmented reference morphologies

Joy Roy, Eric Cheung, Junaid Bhatti, Abraar Muneem and Daniel Lobo *

Department of Biological Sciences, University of Maryland, Baltimore County, Baltimore, MD 21250, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on July 18, 2019; revised on December 7, 2019; editorial decision on January 7, 2020; accepted on January 14, 2020

Abstract

Motivation: Morphological and genetic spatial data from functional experiments based on genetic, surgical and pharmacological perturbations are being produced at an extraordinary pace in developmental and regenerative biology. However, our ability to extract knowledge from these large datasets are hindered due to the lack of formalization methods and tools able to unambiguously describe, centralize and interpret them. Formalizing spatial phenotypes and gene expression patterns is especially challenging in organisms with highly variable morphologies such as planarian worms, which due to their extraordinary regenerative capability can experimentally result in phenotypes with almost any combination of body regions or parts.

Results: Here, we present a computational methodology and mathematical formalism to encode and curate the morphological outcomes and gene expression patterns in planaria. Worm morphologies are encoded with mathematical graphs based on anatomical ontology terms to automatically generate reference morphologies. Gene expression patterns are registered to these standard reference morphologies, which can then be annotated automatically with anatomical ontology terms by analyzing the spatial expression patterns and their textual descriptions. This methodology enables the curation and annotation of complex experimental morphologies together with their gene expression patterns in a centralized standardized dataset, paving the way for the extraction of knowledge and reverse-engineering of the much sought-after mechanistic models in planaria and other regenerative organisms.

Availability and implementation: We implemented this methodology in a user-friendly graphical software tool, PlanGexQ, freely available together with the data in the manuscript at https://lobolab.umbc.edu/plangexq.

Contact: lobo@umbc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Understanding the mechanistic interactions of genes and their products controlling the formation of tissue patterns and shapes is a current outstanding challenge in developmental and regenerative biology (Chiou and Collins, 2018; Kicheva and Briscoe, 2015; Levin *et al.*, 2019). Toward the characterization of essential regulatory genes, surgical, genetic and pharmacological perturbations are experimentally applied to produce normal and aberrant phenotypes, together with *in situ* hybridization (Broitman-Maduro and Maduro, 2011; King and Newmark, 2018) and immunohistochemistry (Adell *et al.*, 2018; Forsthoefel *et al.*, 2018; Ramos, 2005) methods to reveal the expression patterns of specific genes. However, these rich datasets are often published and disseminated in the literature with microscopy images and textual descriptions directly embedded in the papers, which hinders the systematic analysis and knowledge extraction from the data by both human and machine-based approaches (Deans *et al.*, 2015). Furthermore, microscopy pictures contain large variations due to differences between specimens, conditions and techniques, which prevent global automated analysis and comparisons. In addition, the sites of gene expression are described with non-standardized language and their descriptions are generally incomplete, due to the lack of anatomical knowledge or from the complexities of the pattern to be summarized in short text-based descriptions (Christiansen *et al.*, 2006).

As a better alternative to gene expression pictures and natural language descriptions disseminated in the literature, centralized repositories (Deans *et al.*, 2015), anatomical ontologies (Aitken, 2005), phenotype ontologies (Dahdul *et al.*, 2018) and reference standard morphologies (Christiansen *et al.*, 2006) have been proposed for the standardization of specific datasets (Bard, 2005). Centralized repositories of gene expression patterns exist in several model organisms, such as mouse (Finger *et al.*, 2017), zebrafish (Howe *et al.*, 2017), *Caenorhabditis elegans* (Lee *et al.*, 2018), *Xenopus* (Karimi *et al.*, 2018) and fruit fly (Thurmond *et al.*, 2019), among others (Alonso-Barba *et al.*, 2016; Cicin-Sain *et al.*, 2015; Darnell *et al.*, 2007). These repositories represent a fundamental resource for the extraction of knowledge from the data, using spatial queries (Christiansen *et al.*, 2006), global spatial analysis (Frise *et al.*, 2010; Mace *et al.*, 2009; Tomancak *et al.*, 2007) and importantly, reverse-engineering methods toward the discovery of mechanistic models of development and regeneration (Crombach *et al.*, 2012; Perkins *et al.*, 2006).

However, the formalization and standardization of gene expression patterns in regenerative biology is a current challenge due to the plasticity and variability in the morphologies of model organisms in regeneration (Lobo *et al.*, 2013b). Highly regenerative organisms can drastically change their body parts both naturally and after experimental perturbations. Amphibians and insects can regenerate supernumerary limbs (Nacu and Tanaka, 2011; Turner, 1981; Yokoyama *et al.*, 1998), hydra can develop multiple heads and body axes (Duffy *et al.*, 2010; Gee *et al.*, 2010), and acoels and annelids can repeat body structures including heads and tails (Özpolat and Bely, 2016; Sikes and Bely, 2010). Furthermore, planarian worms possess the ability to restore their complete body from almost any amputation, including complete new organs such as the brain, eyes, stomach and pharynx (Cebrià *et al.*, 2018). Crucially, planarian positional information genes are essential for the regulation and signaling of the body parts to regenerate after amputations, for which they form specific expression patterns to instruct stem cells during tissue homeostasis and regeneration (Reddien, 2018). Experimentally knocking-down (RNA interference) or pharmacologically perturbing positional information genes can result in aberrant morphologies that are key for understanding the mechanisms controlling regeneration, such as worms with multiple heads (Iglesias *et al.*, 2008; Oviedo *et al.*, 2010; Petersen and Reddien, 2008) or tails (Gurley *et al.*, 2008; Petersen and Reddien, 2011). There are several transcriptomics datasets in planaria, from both RNAseq (Kao *et al.*, 2013; Labbé *et al.*, 2012; Rodríguez-Esteban *et al.*, 2015; Sandmann *et al.*, 2011) and scRNAseq (Fincher *et al.*, 2018; Plass *et al.*, 2018), as well as centralized tools to analyze them (Castillo-Lara and Abril, 2018; Castillo-Lara *et al.*, 2019), but no formalization of gene expression patterns into two-dimensional standard morphologies are available. Indeed, the plasticity and diversity of the experimental morphologies in regenerative biology in general and planarian worms in particular represent a challenge to curate and centralize a standard formalized repository of gene expression patterns.

Here, we present a methodology combining mathematical graph representations with ontology terms to generate standard reference morphologies and register and annotate gene expression patterns. Anatomical ontology terms, including anatomical regions and organs, form the nodes of the graph, whereas the graph edges represent their spatial interrelation. The annotation of gene expression patterns can be automated for any morphology using the anatomical ontology terms encoded in the nodes of their graph representation. We show here a proof of concept with planarian worms, a highly plastic organism with an extraordinary regenerative capability, but this approach can be extended to other organisms and datasets with highly variable morphologies. This methodology was implemented in a software tool, which can be used to build centralized repositories of annotated experimental morphologies and their corresponding gene expression patterns not only for use by human scientists but also importantly as input for automated reverse engineering machine learning methods.

2 Results

To overcome the difficulty of searching, comparing and analyzing gene expression patterns in the hundreds of experimental morphologies of planarian worms, we present here a semi-automatic methodology to curate gene expression patterns in formalized standard reference morphologies and annotate them with anatomical ontology terms. The process begins with the curation of the worm morphology using a mathematical graph representation. Then, a reference model of the particular morphology is automatically generated from the graph representation, in which the gene expression pattern can be registered. The annotation with anatomical ontology terms is performed automatically by comparing the gene expression pattern with the automatically generated model of anatomical locations from the formalized morphology. In addition, pre-defined organs can be assigned and annotated manually, and further ontology terms are automatically suggested from the textual descriptions of the morphologies or gene expression patterns. A freely available software tool was developed implementing this methodology, providing a complete pipeline to curate and annotate gene expression patterns in planarian worms.

2.1 Formalizing morphologies with mathematical graphs linking anatomical terms

To create standard morphologies accounting for the diversity and plasticity of experimental morphologies of planarian worms, a mathematical graph augmented with geometric information is used to manually encode the anatomical features of reference morphologies. The graph nodes represent anatomical terms from the planarian anatomy ontology (Nowotarski *et al.*, 2019), while the graph edges represent the spatial relationships between the terms. Mathematical graphs have been used previously for defining planarian worms and limb morphologies (Lobo *et al.*, 2013a, b), and here, we augment them with the planarian anatomy ontology to specify the nodes of the graph and produce reference morphologies for the curation of gene expression patterns. A node can either be an anatomical region of the worm, such as the head, neck, pharyngeal or tail, or an organ, such as the eye, brain, pharynx, and so on, as defined in the planarian anatomy ontology. Region nodes contain information about the overall size and shape of a particular anatomical region using four to two real numeric parameters representing the length of the region from its center toward orthogonal directions, for regions being connected with none to two or more other regions, respectively. This relative definition of locations, instead of absolute positions and angles, guarantees the connectedness of the morphology and its invariance to rotations. Regions are linked between them with edges, forming the topology of the worm. Each edge linking regions contain a real numeric parameter indicating the distance between the region centers and the location of the interface between the two connected regions. Organ nodes are linked to the region nodes where they are located, and this edge contains a vector for the location of the organ with respect the center of the region. Organ nodes can be encoded as a spot (e.g. eye, brain, pharynx) including a rotation parameter indicating the orientation of the organ within the region, or they can be encoded as a line (e.g. nerve cords) including two vectors parameters for the positions of the start and end locations. In this way, a mathematical graph can encode any type of worm morphology defined with region and organ anatomy ontology terms.

Figure 1 shows graph encodings formalizing examples of worm morphologies from gene expression pattern assays using whole-mount *in situ* hybridization. The examples include wild-type morphologies of different species, experimental aberrant morphologies such as double-head and double-tail worms, and regenerating pieces. The mathematical graph is visualized with a set of labeled circles representing the graph nodes (larger for regions and smaller for organs), together with lines for the graph edges. Quantitative parameters are visualized with red circles for the sizes for the regions orthogonal to the direction of the region, green circles for the location of the intersection between regions and blue circles for the rotation of the organs. The graph encodes the main characteristics of a morphology, leaving out the small differences between individual worms; hence, the three wild-type morphologies (Fig. 1A–C) are represented with the same graph. In this way, a mathematical graph can represent unambiguously an organism anatomical features using ontology anatomical terms, their spatial interactions, and overall sizes.

2.2 Generating anatomical reference morphologies from mathematical graphs

From the mathematical graph of anatomical terms, a reference model of the morphology is automatically generated. Since the mathematical graph unambiguously links spatial mathematical

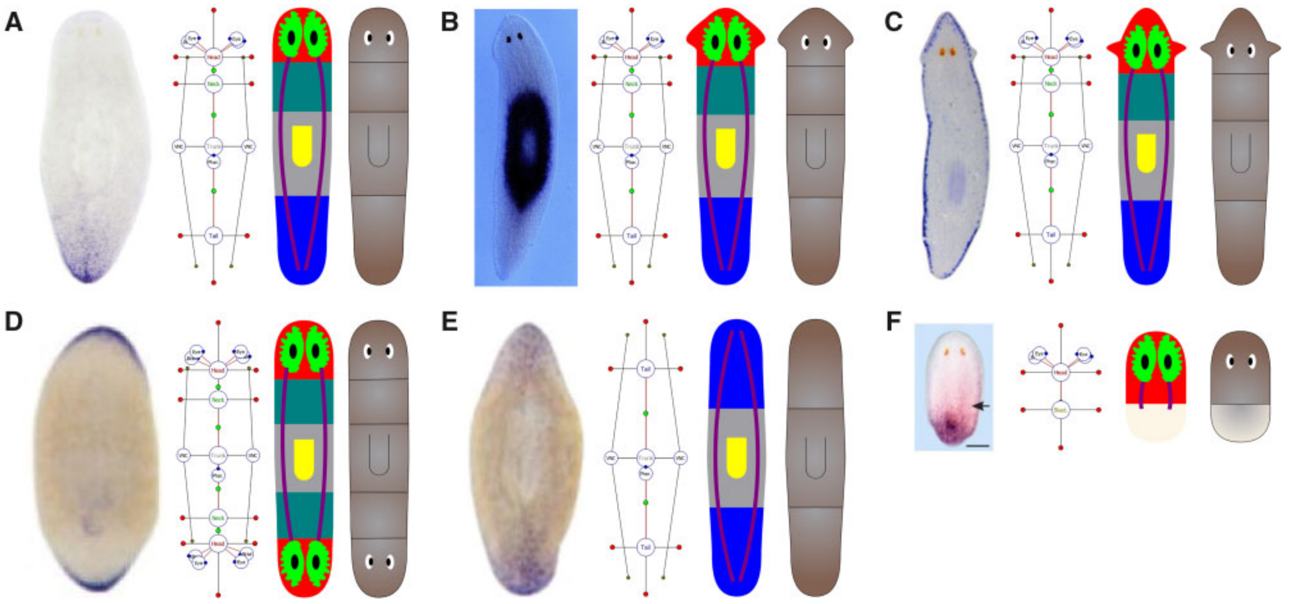

Fig. 1. Formalizing morphologies with anatomical ontology terms and automated generation of standard reference morphologies and worm diagrams. A mathematical graph describes a morphological phenotype, where nodes represent either regions or organs, as defined in the planarian anatomy ontology; links represent the topological connections between the regions or the location of the organs. The red handles define the overall size of the regions along orthogonal directions, where the green handles represent the location of the border between regions. The mathematical graphs represent the main attributes of a morphology, ignoring the unimportant differences between individual worms. An algorithm automatically converts the mathematical graph formalization into a reference morphology, containing color coded regions corresponding to an anatomical ontology term, or a worm diagram, including schematic representations of the regions and organs. Morphological differences between the species, most notably in head shape, are included in the spatial representations. Examples of formalization of morphologies from whole-mount *in situ* hybridization microscopy images for different genes and species: (**A**) wnt11-2 in wild-type *Schmidtea mediterranea*, (**B**) FoxA in wild-type *Dugesia japonica*, (**C**) ifb in wild-type *Dugesia dorocephala*, (**D**) sFRP-1 in double-head *S.mediterranea*, (**E**) fz4-1 in double-tail *S.mediterranea*, and (**F**) wntp-2 in a regenerating head fragment from *S.mediterranea*. Microscopy images adapted from Gurley *et al.* (2010) (A), Koinuma *et al.* (2000) (B), Accorsi *et al.* (2017) (C), Gurley *et al.* (2008) (D, E) and Petersen and Reddien (2009) (F). (Color version of this figure is available at *Bioinformatics* online.)

terms and define their physical properties, it precisely defines a reference model of the morphology that can be displayed with spatial regions and organ symbols (Fig. 1). Each anatomical term is then highlighted in a different color for reference mapping or in worm colors (brown) for a diagram representation of the morphology. For anatomical terms referring to regions, their graph edges and parameters define their spatial distribution and specific dimensions, respectively. From this information, an automated algorithm produces an interconnected set of spatial regions. In particular, the sizes of the regions (red circles) and the intersection between the regions (green circles) define a set of spatial points around the center of the region (region node). These points are connected by straight lines at the interface between regions and Bézier curves at the external border of the region to create a closed area for each region. A region with two or more links (such as the trunk) requires a sequence of a quadratic, a cubic, and a quadratic curve to define the border between each pair of region links. A region with a single link (such as the head or tail) requires two sequential quadratic curves to define the external border of the region. In addition, the algorithm introduces additional Bézier curves to account for the arrowhead shapes of different species (Emmons-Bell *et al.*, 2015). In particular, arrowheads (Fig. 1B, *Dugesia japonica* and Fig. 1C, *Dugesia dorocephala*) are created with a sequence of a cubic, a quadratic and a cubic curve for each side of the head, in contrast to the standard single quadratic curve used for each side of a round head (Fig. 1A, *Schmidtea mediterranea*). Different arrowheads are generated with slightly different control points predetermined for each species. For anatomical terms referring to organs, the edge linked to a region define the location with respect to that region, while its parameters define the rotation and size of the specific organ. In this way, reference morphologies and worm diagrams are automatically generated from the mathematical graphs. The reference morphologies show the different anatomical regions and organs of the worm highlighting them in different colors, while the worm diagrams show the overall morphology and the positions of the visible organs.

2.3 Annotating gene expression patterns with reference morphologies

The automatically generated reference morphologies and worm diagrams serve as standard virtual worms for describing gene expression patterns, which can then be processed for automatic annotation using anatomy ontology terms. Gene expression patterns experimentally obtained from *in situ* hybridization or immunochemistry assays recorded in a microscopy image can be mapped into the worm diagram morphologies with either manual or automated registration methods (Mace *et al.*, 2009). As a proof of concept, here we use a simple user-friendly manual registration method, in which a user directly specifies the expression levels and locations in the virtual worm. For this, the user can use a pointer device to trace the expression patterns recorded in the microscopy images directly on top of the reference morphologies. The level of gene expression and the size of the registration pen can be selected with specific sliders in the interface. Undo and clear buttons are available to help in the process of curation. Figure 2 shows examples of microscopy images of planarian gene expression patterns mapped into their worm diagram reference morphologies. The variations of a morphology shown in specific individual worms are removed when mapping the gene expression patterns to the reference morphologies. Gene expression patterns in wild-type, aberrant or regenerating morphologies—such as the double head, double tail and regenerating head fragment shown—can be encoded in the reference morphologies, since they are automatically generated from the mathematical graphs.

The annotation of a gene expression pattern registered in the reference morphology is done automatically using the ontology terms in the mathematical graph encoding the reference morphology. Ontology anatomy terms define the regions and organs of the worm in the mathematical graph and hence the reference morphology contains the spatial locations of the ontology terms.

To automatically annotate genes with terms from the planarian anatomy ontology, the mapped gene expression patterns in the worm diagrams are spatially compared with the reference

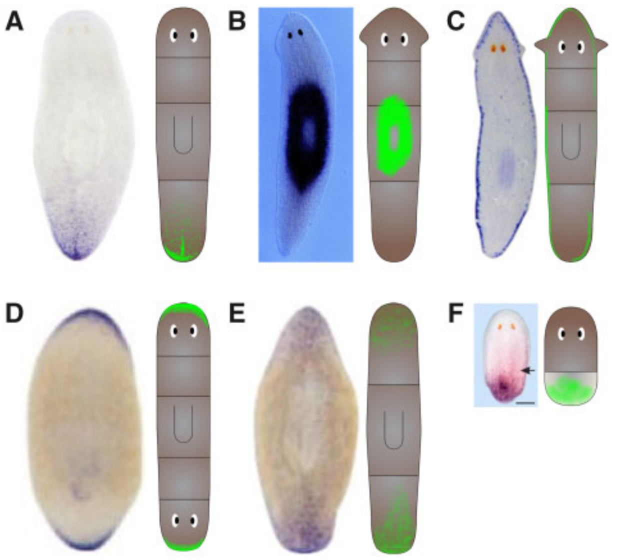

Fig. 2. Gene expression patterns registration into the worm diagrams of the reference morphologies. The worm diagrams automatically generated from the graph formalization serve as standard reference morphologies for spatially mapping the gene expression patterns from microscopy images. The registration into reference morphologies eliminates the individual variability of the experimental morphologies, as it is abstracted as a generic standard morphology. (**A–F**) Panels correspond to the morphologies in Figure 1. Microscopy images adapted from Gurley *et al.* (2010) (A), Koinuma *et al.* (2000) (B), Accorsi *et al.* (2017) (C), Gurley *et al.* (2008) (D, E) and Petersen and Reddien (2009) (F)

morphologies from the mathematical graphs: when any gene expression is found in a given morphology region, the gene is annotated with the anatomical term corresponding with that region, as defined in the mathematical graph. Figure 3 shows examples of automatically annotated expression patterns in wild-type and aberrant morphologies. The method automatically calculates quantitative measurements of the gene expression pattern as the percentage distribution of expression among the different anatomical regions. Further ontology annotations can be manually done in the tool and suggestions are provided automatically in the interface by finding keywords in the textual description of the gene expression pattern. In addition, a gene expression pattern can be manually annotated with terms from the Evidence and Conclusion Ontology (Giglio *et al.*, 2019). For example, gene expression patterns obtained with immunofluorescence assays are annotated with the term ECO_0000045, *spatial pattern of protein expression evidence*, and those obtained with *in situ* hybridization assays are annotated with the term ECO_0000047, *spatial pattern of transcript expression evidence*.

Gene expression patterns occurring in organs can be annotated with organ terms encoded in the reference morphologies. The location of an organ gene expression is pre-defined for each reference morphology and displayed in the worm diagrams. As a proof of concept in our current implementation, gene expression patterns of organs are manually selected by the curator, however, automated pattern recognition methods (Han *et al.*, 2011; Shamir *et al.*, 2010) could be used as well in the future. Figure 4 shows examples of worm diagram reference morphologies for some organs, including brain and nerve cords, eyes, pharynx and digestive system. Crucially, the use of standard reference morphologies and expression patterns for organs remove the variability between individual worms, an advantage for the application of knowledge extraction machine learning methods.

3 Implementation

We developed a database and a user-friendly tool implementing the presented methodology for the storage and curation of planarian gene expression patterns. The database stores the information for the mathematical graph encodings of the reference morphologies, the gene expression patterns extracted from the experimental assays and the ontology annotations. Information about worm species, source publications and genes are also included in the database, as well as the experimental images and their captions from the original publication, if available. Gene expression patterns are stored as 8-bit grayscale raster images, supporting 256 levels of gene expression per pixel. Genes are linked to other web databases, including SmedGD (Robb *et al.*, 2015), GenBank (Sayers *et al.*, 2019), European Nucleotide Archive (Harrison *et al.*, 2019), Uniprot (UniProt Consortium, 2019) and Quick GO (Binns *et al.*, 2009). The database is implemented with SQLite, a public domain relational database system that stores both the schema and the data into a single file. Figure 5 shows the database schema, including all the tables and their links for the storage of the reference morphologies (right, red area) and the gene expression patterns (left, blue area).

A user-friendly software tool called PlanGexQ was developed to implement the presented methodology for the curation of planarian gene expression patterns into the database. The interactive tool

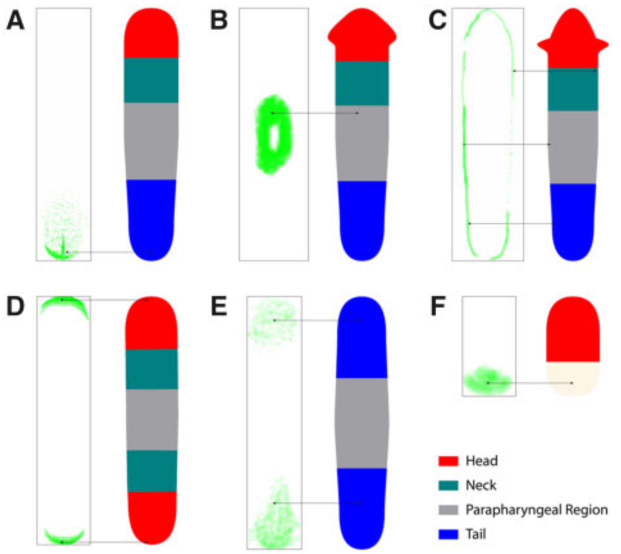

Fig. 3. Automated anatomical ontology annotation of gene expression patterns. Using the standard reference morphologies encoded in the mathematical graphs of ontology terms, genes are automatically annotated with anatomical terms from the ontology by comparing the overlapping of the registered gene expression patterns with each of the anatomical regions. The resulting quantitative distribution of the gene expression pattern among the anatomical regions are also automatically calculated. (**A–F**) Panels correspond to the morphologies in Figure 2

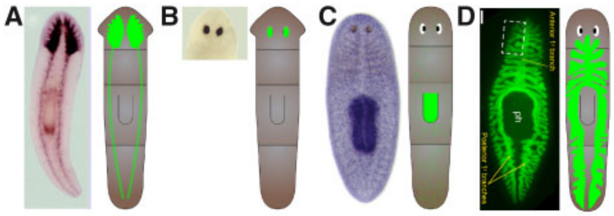

Fig. 4. Gene expression patterns in organs can be registered with pre-defined patterns. The standard reference morphologies include the location of the organs in the planarian anatomy ontology, which can be selected for mapping specific gene expression patterns. Each species has a set of pre-defined gene expression patterns corresponding to each organ in the anatomy ontology. (**A**) Central nervous system in wild-type *Dugesia japonica* with whole-mount *in situ* hybridization of *PC2*. (**B**) Eyes in wild-type *D.japonica* with whole-mount *in situ* hybridization of *ops*. (**C**) Pharynx in *Schmidtea mediterranea* with whole-mount *in situ* hybridization of *sFRP-3*. (**D**) Digestive tract in *S.mediterranea* with fluorescent whole-mount *in situ* hybridization of *pk1*. Planarian images adapted from Koinuma *et al.* (2003) (A), Mannini *et al.* (2004) (B), Gurley *et al.* (2010) (C) and Barberán *et al.* (2016) (D)

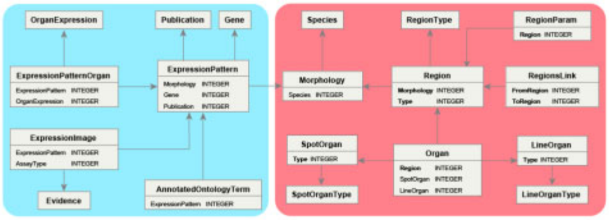

Fig. 5. Schema of the database storing the gene expression pattern information and the reference morphologies encoded with mathematical graphs. The blue area (left) shows the tables for gene expression patterns and their ontological annotations and the red area (right) shows the tables for the formalization of reference morphologies. For simplicity, only the foreign keys are displayed on each table. (Color version of this figure is available at *Bioinformatics* online.)

allows any user to define reference morphologies with a drag-and-drop interface for specifying the mathematical graphs and their parameters, import expression patterns images, map them into a standard morphology, and annotate them with ontological terms, both with the presented automated methods and manually. Furthermore, keywords from textual descriptions of the microscopy images, such as figure captions, are automatically scanned and parsed by the program to find the ids of the gene or protein corresponding to the expression pattern and add gene annotations with additional ontology terms from the anatomy ontology. Importantly, the tool includes a search module for finding specific genes, patterns, morphologies, publications, and so on stored in the database by specifying textual keyworks such as gene names or numerical parameters such as the number of head regions in a morphology. Figure 6 shows screenshots of the software tool, including the interfaces for defining reference morphologies, map gene expression patterns, and link gene information into other databases. The software is a standalone Windows desktop application developed in C++ using the Qt libraries and it is freely available at https://lobolab.umbc.edu/plangexq.

4 Discussion

Highly plastic organisms such as planaria—able to regenerate ectopic regions including multiple heads and tails—represent a challenge for the unambiguous description, encoding of their gene expression patterns, and annotation with ontology terms. This lack of formalization prevents the systematic and computational analysis of these rich datasets as they are dispersed in the literature and described with pictures and natural language. Toward the centralization, formalization and annotation of planarian gene expression patterns, here we present a curation methodology to unambiguously encode worm reference morphologies as mathematical graphs linking anatomical ontology terms. The mathematical graphs describe the most important aspects of the worm morphology, including their regions and organs and can produce standard reference morphologies for any wild-type or aberrant experimental morphology, such as double head or regenerating fragments. These standard reference morphologies, based on the planaria anatomy ontology, can be used to register gene expression patterns that can then be automatically annotated with anatomical ontology terms.

We implemented the proposed methodology in a freely available database and user-friendly software tool for the curation and annotation of planarian gene expression patterns. Using an interactive graphical interface, any user can create reference morphologies, register gene expression patterns, and annotate them with ontological terms. The program suggests ontological terms by directly analyzing the gene expression pattern and comparing it with the reference morphology, as well as by finding keywords in any description entered in the interface, such as those from figure captions in published papers. In this way, the software tool paves the way toward the comprehensive curation and centralization of gene expression patterns from the diverse morphologies of planaria by any user with minimal training.

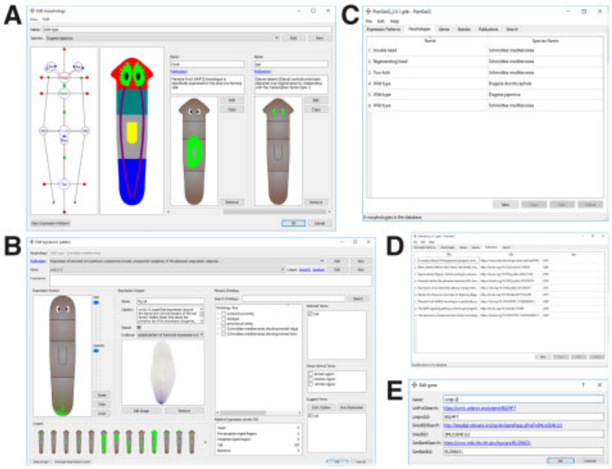

Fig. 6. Screenshots of the user-friendly software tool to curate, register and annotate planarian gene expression patterns. A user can interactively define the morphological graphs with a drag and drop interface and register the gene expression patterns in the standard morphologies. The software automatically assigns the anatomical ontology terms and calculates the coverage of the genes in each of the anatomical regions. (**A**) Visualization of a reference morphology and two gene expression patterns registered. (**B**) Interface to curate a gene expression pattern, register it to the standard morphology, and assign ontology annotations. (**C**) List of reference morphologies. (**D**) List of curated publications. (**E**) Editing the information of a gene, including its ids linking it to other databases

We limited the description of reference morphologies to the two dimensions along the anterior–posterior and the medial–lateral axes, whereas ontology terms can be used to annotate an expression pattern with dorso-ventral anatomical terms. The planaria being a flatworm, most available data in the literature are dorsal or ventral two-dimensional views for both morphologies and gene expression images; however, expression patterns along the dorsal–ventral axis are also possible (Wurtzel *et al.*, 2017). Extending the mathematical graphs describing the reference morphologies and the gene expression patterns to three dimensions would be straightforward, although the user interfaces to define the mathematical graphs and to register and visualize the gene expression patterns would be more challenging. To our knowledge, there are no datasets of planarian gene expression patterns in three dimensions. As future work, three-dimensional imaging of expression patterns in planaria, such as confocal microscopy (Trusk, 2018) or optical projection tomography (Sharpe *et al.*, 2002), together with automated registration methods (Wang *et al.*, 2015) will be employed toward the goal of creating a three-dimensional atlas of planarian gene expression patterns.

Creating a formalized and centralized database of planarian gene expression patterns with the proposed methodology would not only benefit human scientists but also will pave the way for the application of automated machine learning methods for the extraction of knowledge directly from this rich dataset currently dispersed in the literature and described with hard to analyze images and natural language descriptions. Machine learning methods can take as input formalized datasets of functional morphologies and ontological annotations as those proposed here to predict regulatory interactions and mechanistic models explaining the phenotypes (Kulmanov *et al.*, 2018; Lobo and Levin, 2015, 2017; Lobo *et al.*, 2016; Sharpe, 2017). This approach will streamline the discovery of mechanistic comprehensive models of growth and shape (Ko and Lobo, 2019) and characterize the genes of theoretical models in planaria (Herath and Lobo, 2019) toward novel biological knowledge and the much sought-after applications in regenerative biomedicine.

We applied here the proposed methodology to planarian worms, but other organisms with plastic morphologies could also be formalized with this approach. Whole-body regenerative organisms such as hydra, starfish and annelids, or specific regenerative structures such as salamander limbs, insect appendages, fish fins and roots would benefit from the presented methodology to unambiguously describe reference morphologies and map their gene expression patterns. For example, the datasets of limb regeneration, including genetic and graft manipulations resulting in complex limbs with multiple appendages, could be formalized with a similar approach using mathematical graphs (Lobo *et al.*, 2014a, b) and extended with the methodology presented here for the formalization of gene expression patterns to better inform optimization machine learning methods to discover mechanistic models of limb regeneration (Uzkudun *et al.*, 2015).

5 Conclusion

Here, we showed how combining mathematical graph representations of morphological data with terms from anatomical ontologies can readily define reference morphologies in highly plastic organisms such as planaria. Gene expression patterns can then be registered in these reference morphologies and be automatically annotated with anatomical ontology terms. This approach will pave the way for the curation of centralized and formalized datasets of gene expression patterns useful for both human and machine scientists, toward the discovery of comprehensive models in developmental and regenerative biology.

Acknowledgements

We thank the members of the Lobo Lab and the planarian regeneration community for helpful discussions.

Funding

This work was supported by the National Science Foundation [IIS-1566077] and the University of Maryland, Baltimore County with an Undergraduate Research Award.

Conflict of Interest: none declared.

References

Accorsi,A. *et al.* (2017) Hands-on classroom activities for exploring regeneration and stem cell biology with planarians. *Am. Biol. Teach.*, **79**, 208–223.

Adell,T. *et al.* (2018) Immunohistochemistry on paraffin-embedded planarian tissue sections. In: Rink, J. (ed.) *Planarian Regeneration: Methods and Protocols*. Humana Press, New York, NY, pp. 367–378.

Aitken,S. (2005) Formalizing concepts of species, sex and developmental stage in anatomical ontologies. *Bioinformatics*, **21**, 2773–2779.

Alonso-Barba,J.I. *et al.* (2016) MEPD: medaka expression pattern database, genes and more. *Nucleic Acids Res.*, **44**, D819–D821.

Barberán,S. *et al.* (2016) The EGFR signaling pathway controls gut progenitor differentiation during planarian regeneration and homeostasis. *Development*, **143**, 2089–2102.

Bard,J.B.L. (2005) Anatomics: the intersection of anatomy and bioinformatics. *J. Anat.*, **206**, 1–16.

Binns,D. *et al.* (2009) QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*, **25**, 3045–3046.

Broitman-Maduro,G. and Maduro, M.F. (2011) In Situ Hybridization of Embryos with Antisense RNA Probes. In: Rothman, J.H., *et al.* (eds.) *Caenorhabditis elegans: Molecular Genetics and Development*. Academic Press, Waltham, MA, pp. 253–270.

Castillo-Lara,S. and Abril,J.F. (2018) PlanNET: homology-based predicted interactome for multiple planarian transcriptomes. *Bioinformatics*, **34**, 1016–1023.

Castillo-Lara,S. *et al.* (2019) PlanExp: intuitive integration of complex RNA-seq datasets with planarian omics resources. *Bioinformatics*, btz802.

Cebrià,F. *et al.* (2018) Rebuilding a planarian: from early signaling to final shape. *Int. J. Dev. Biol.*, **62**, 537–550.

Chiou,K. and Collins,E.-M.S. (2018) Why we need mechanics to understand animal regeneration. *Dev. Biol.*, **433**, 155–165.

Christiansen,J.H. *et al.* (2006) EMAGE: a spatial database of gene expression patterns during mouse embryo development. *Nucleic Acids Res.*, **34**, D637–D641.

Cicin-Sain,D. *et al.* (2015) SuperFly: a comparative database for quantified spatio-temporal gene expression patterns in early dipteran embryos. *Nucleic Acids Res.*, **43**, D751–D755.

UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.

Crombach,A. *et al.* (2012) Efficient reverse-engineering of a developmental gene regulatory network. *PLoS Comput. Biol.*, **8**, e1002589.

Dahdul,W. *et al.* (2018) Annotation of phenotypes using ontologies: a gold standard for the training and evaluation of natural language processing systems. *Database (Oxford)*, **2018**, bay110.

Darnell,D.K. *et al.* (2007) GEISHA: an in situ hybridization gene expression resource for the chicken embryo. *Cytogenet. Genome Res.*, **117**, 30–35.

Deans,A.R. *et al.* (2015) Finding our way through phenotypes. *PLoS Biol*, **13**, e1002033.

Duffy,D.J. *et al.* (2010) Wnt signaling promotes oral but suppresses aboral structures in Hydractinia metamorphosis and regeneration. *Development*, **137**, 3057–3066.

Emmons-Bell,M. *et al.* (2015) Gap junctional blockade stochastically induces different species-specific head anatomies in genetically wild-type *Girardia dorotocephala* flatworms. *Int. J. Mol. Sci.*, **16**, 27865–27896.

Fincher,C.T. *et al.* (2018) Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science*, **1736**, eaaq1736.

Finger,J.H. *et al.* (2017) The mouse Gene Expression Database (GXD): 2017 update. *Nucleic Acids Res.*, **45**, D730–D736.

Forsthoefel,D.J. *et al.* (2018) Fixation, processing, and immunofluorescent labeling of whole mount planarians. In: Rink, J. (ed.) *Planarian Regeneration: Methods and Protocols*. Humana Press, New York, NY, pp. 353–366.

Frise,E. *et al.* (2010) Systematic image-driven analysis of the spatial *Drosophila* embryonic expression landscape. *Mol. Syst. Biol.*, **6**, 1–15.

Gee,L. *et al.* (2010) Beta-catenin plays a central role in setting up the head organizer in hydra. *Dev. Biol.*, **340**, 116–124.

Giglio,M. *et al.* (2019) Eco, the evidence & conclusion ontology: community standard for evidence information. *Nucleic Acids Res.*, **47**, D1186–D1194.

Gurley,K.A. *et al.* (2010) Expression of secreted Wnt pathway components reveals unexpected complexity of the planarian amputation response. *Dev. Biol.*, **347**, 24–39.

Gurley,K.A. *et al.* (2008) β-Catenin defines head versus tail identity during planarian regeneration and homeostasis. *Science*, **319**, 323–327.

Han,L. *et al.* (2011) Automatically identifying and annotating mouse embryo gene expression patterns. *Bioinformatics*, **27**, 1101–1107.

Harrison,P.W. *et al.* (2019) The European nucleotide archive in 2018. *Nucleic Acids Res.*, **47**, D84–D88.

Herath,S. and Lobo,D. (2019) Cross-inhibition of Turing patterns explains the self-organized regulatory mechanism of planarian fission. *J. Theor. Biol.*, **485**, 110042.

Howe,D.G. *et al.* (2017) The Zebrafish Model Organism Database: new support for human disease models, mutation details, gene expression phenotypes and searching. *Nucleic Acids Res.*, **45**, D758–D768.

Iglesias,M. *et al.* (2008) Silencing of Smed-βcatenin1 generates radial-like hypercephalized planarians. *Development*, **135**, 1215–1221.

Kao,D. *et al.* (2013) The planarian regeneration transcriptome reveals a shared but temporally shifted regulatory program between opposing head and tail scenarios. *BMC Genomics*, **14**, 797.

Karimi,K. *et al.* (2018) Xenbase: a genomic, epigenomic and transcriptomic model organism database. *Nucleic Acids Res.*, **46**, D861–D868.

Kicheva,A. and Briscoe,J. (2015) Developmental pattern formation in phases. *Trends Cell Biol.*, **25**, 579–591.

King,R.S. and Newmark,P.A. (2018) Whole-mount *in situ* hybridization of planarians. In: Rink, J.C. (ed.) *Planarian Regeneration: Methods and Protocols*. Humana Press, New York, NY, pp. 379–392.

Ko,J.M. and Lobo,D. (2019) Continuous dynamic modeling of regulated cell adhesion: sorting, intercalation, and involution. *Biophys. J.*, **117**, 2166–2179.

Koinuma,S. *et al.* (2000) Planaria FoxA (HNF3) homologue is specifically expressed in the pharynx-forming cells. *Gene*, **259**, 171–176.

Koinuma,S. *et al.* (2003) The expression of planarian brain factor homologs, DjFoxG and DjFoxD. *Gene Expr. Patterns*, **3**, 21–27.

Kulmanov,M. *et al.* (2018) Ontology-based validation and identification of regulatory phenotypes. *Bioinformatics*, **34**, i857–i865.

Labbé,R.M. *et al.* (2012) A comparative transcriptomic analysis reveals conserved features of stem cell pluripotency in planarians and mammals. *Stem Cells*, **30**, 1734–1745.

Lee,R.Y.N. *et al.* (2018) WormBase 2017: molting into a new stage. *Nucleic Acids Res.*, **46**, D869–D874.

Levin,M. *et al.* (2019) Planarian regeneration as a model of anatomical homeostasis: recent progress in biophysical and computational approaches. *Semin. Cell Dev. Biol.*, **87**, 125–144.

Lobo,D. and Levin,M. (2015) Inferring regulatory networks from experimental morphological phenotypes: a computational method reverse-engineers planarian regeneration. *PLoS Comput. Biol.*, **11**, e1004295.

Lobo,D. and Levin,M. (2017) Computing a worm: reverse-engineering planarian regeneration. In: Adamatzky, A. (ed.) *Advances in Unconventional Computing. Volume 2: Prototypes, Models and Algorithms*. Springer International Publishing, Switzerland, pp. 637–654.

Lobo,D. *et al.* (2013a) Planform: an application and database of graph-encoded planarian regenerative experiments. *Bioinformatics*, **29**, 1098–1100.

Lobo,D. *et al.* (2013b) Towards a bioinformatics of patterning: a computational approach to understanding regulative morphogenesis. *Biol. Open*, **2**, 156–169.

Lobo,D. *et al.* (2014a) A bioinformatics expert system linking functional data to anatomical outcomes in limb regeneration. *Regeneration*, **1**, 37–56.

Lobo,D. *et al.* (2014b) Limbform: a functional ontology-based database of limb regeneration experiments. *Bioinformatics*, **30**, 3598–3600.

Lobo,D. *et al.* (2016) Computational discovery and in vivo validation of hnf4 as a regulatory gene in planarian regeneration. *Bioinformatics*, **32**, 2681–2685.

Mace,D.L. *et al.* (2010) Extraction and comparison of gene expression patterns from 2D RNA in situ hybridization images. *Bioinformatics*, **26**, 761–769.

Mannini,L. *et al.* (2004) Djeyes absent (Djeya) controls prototypic planarian eye regeneration by cooperating with the transcription factor Djsix-1. *Dev. Biol.*, **269**, 346–359.

Nacu,E. and Tanaka,E.M. (2011) Limb regeneration: a new development? *Annu. Rev. Cell Dev. Biol.*, **27**, 409–440.

Nowotarski,S. *et al.* (2019) Planarian anatomy ontology. *Zenodo*. doi: 10.5281/zenodo.2575043.

Oviedo,N.J. *et al.* (2010) Long-range neural and gap junction protein-mediated cues control polarity during planarian regeneration. *Dev. Biol.*, **339**, 188–199.

Özpolat,B.D. and Bely,A.E. (2016) Developmental and molecular biology of annelid regeneration: a comparative review of recent studies. *Curr. Opin. Genet. Dev.*, **40**, 144–153.

Perkins,T.J. *et al.* (2006) Reverse engineering the gap gene network of *Drosophila melanogaster*. *PLoS Comput. Biol.*, **2**, 417–428.

Petersen,C.P. and Reddien,P.W. (2008) Smed-catenin-1 is required for anteroposterior blastema polarity in planarian regeneration. *Science*, **319**, 327–330.

Petersen,C.P. and Reddien,P.W. (2009) A wound-induced Wnt expression program controls planarian regeneration polarity. *Proc. Natl. Acad. Sci. USA*, **106**, 17061–17066.

Petersen,C.P. and Reddien,P.W. (2011) Polarized notum activation at wounds inhibits Wnt function to promote planarian head regeneration. *Science*, **332**, 852–855.

Plass,M. *et al.* (2018) Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, **360**, eaaq1723.

Ramos,J.A. (2005) Technical aspects of immunohistochemistry. *Vet. Pathol.*, **426**, 405–426.

Reddien,P.W. (2018) The cellular and molecular basis for planarian regeneration. *Cell*, **175**, 327–345.

Robb,S.M.C. *et al.* (2015) SmedGD 2.0: the *Schmidtea mediterranea* genome database. *Genesis*, **53**, 535–546.

Rodríguez-Esteban,G. *et al.* (2015) Digital gene expression approach over multiple RNA-Seq data sets to detect neoblast transcriptional changes in *Schmidtea mediterranea*. *BMC Genomics*, **16**, 361.

Sandmann,T. *et al.* (2011) The head-regeneration transcriptome of the planarian *Schmidtea mediterranea*. *Genome Biol.*, **12**, R76.

Sayers,E.W. *et al.* (2019) GenBank. *Nucleic Acids Res.*, **47**, D94–D99.

Shamir,L. *et al.* (2010) Pattern recognition software and techniques for biological image analysis. *PLoS Comput. Biol.*, **6**, e1000974.

Sharpe,J. (2017) Computer modeling in developmental biology: growing today, essential tomorrow. *Development*, **144**, 4214–4225.

Sharpe,J. *et al.* (2002) Optical projection tomography as a tool for 3D microscopy and gene expression studies. *Science*, **296**, 541–545.

Sikes,J.M. and Bely,A.E. (2010) Making heads from tails: development of a reversed anterior-posterior axis during budding in an acoel. *Dev. Biol.*, **338**, 86–97.

Thurmond,J. *et al.* (2019) FlyBase 2.0: the next generation. *Nucleic Acids Res.*, **47**, D759–D765.

Tomancak,P. *et al.* (2007) Global analysis of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol*, **8**, R145–R124.

Trusk,T.C. (2018) 3D reconstruction of confocal image data. In: Jerome, W. *et al.* (eds.) *Basic Confocal Microscopy*. Springer International Publishing, Cham, Switzerland, pp. 279–307.

Turner,R.N. (1981) Probability aspects of supernumerary production in the regenerating limbs of the axolotl, *Ambystoma mexicanum*. *J. Embryol. Exp. Morphol.*, **65**, 119–126.

Uzkudun,M. *et al.* (2015) Data-driven modelling of a gene regulatory network for cell fate decisions in the growing limb bud. *Mol. Syst. Biol.*, **11**, 815–815.

Wang,C.W. *et al.* (2015) Fully automatic and robust 3D registration of serial-section microscopic images. *Sci. Rep.*, **5**, 1–14.

Wurtzel,O. *et al.* (2017) Planarian epidermal stem cells respond to positional cues to promote cell-type diversity. *Dev. Cell*, **40**, 491–504.e5.

Yokoyama,H. *et al.* (1998) Multiple digit formation in Xenopus limb bud recombinants. *Dev. Biol.*, **196**, 1–10.