ontoFAST: An R package for interactive and semi-automatic annotation of characters with biological ontologies

Sergei Tarasov[1,2] | **István Mikó**[3] | **Matthew Jon Yoder**[4]

[1]Finnish Museum of Natural History, Pohjoinen Rautatiekatu 13, FI-00014 Helsinki, Finland

[2]National Institute for Mathematical and Biological Synthesis, University of Tennessee, Knoxville, TN 37996, USA

[3]University of New Hampshire, Durham, NH, USA

[4]Illinois Natural History Survey, Champaign, IL, USA

Correspondence
Sergei Tarasov
Email: sergei.tarasov@helsinki.fi

Funding information
Academy of Finland grant to ST, Grant/Award Number: 339576; NSF Award DBI-1300426 to National Institute for Mathematical and Biological Synthesis.

1. The commonly used Entity-Quality (EQ) syntax provides rich semantics and high granularity for annotating phenotypes and characters using ontologies. However, EQ syntax might be time inefficient if this granularity is unnecessary for downstream analysis.

2. We present an R package ontoFAST that aid production of fast annotations of characters and character matrices with biological ontologies. Its interactive interface allows quick and convenient tagging of character statements with necessary ontology terms.

3. The annotations produced in ontoFAST can be exported in csv format for downstream analysis. Additinally, OntoFAST provides: (i) functions for constructing simple queries of characters against ontologies, and (ii) helper function for exporting and visualising complex ontological hierarchies and their relationships.

4. OntoFAST enhances data interoperability between various applications and support further integration of ontological and phylogenetic methods. Ontology tools are underrepresented in R environment and we hope that ontoFAST will stimulate their further development.

KEYWORDS
annotation, characters, character matrix, ontology, phenomics,

phenotypes, phylogenetics

1 | INTRODUCTION

During the past few decades a massive number of organismal phenotypes have been described by biologists for phylogenetic purposes in the form of character matrices and statements written in natural language (NL). Nowadays, ontologies are emerging as a fundamental technology for managing phenotypic data (Balhoff et al., 2010). An ontology is a computer-based representation of concepts and their logical relationships for a specific domain of knowledge (Deans et al., 2012, 2015; Balhoff et al., 2013). Ontological representation facilitates the conversion of NL into machine-parsable statements, thereby, providing new opportunities for computer-aided comparative phenomics and trait analysis (Deans et al., 2015; Dececchi et al., 2015; Burleigh et al., 2013; Tarasov, 2019).

The Entity-Quality (EQ) syntax, adopted by the Open Biological and Biomedical Ontologies (OBO) Consortium (Washington et al., 2009), and Phenoscape project (`http://phenoscape.org`), is the common convention for ontological representation of phenotypes (Balhoff et al., 2010; Gkoutos et al., 2005). EQ syntax bounds an entity, corresponding to a specific anatomical structure from an anatomy ontology with a quality term from the generic Phenotype and Trait Ontology [PATO, Mungall et al. (2010)]. The annotation of character matrices and phenotypes using EQ syntax is implemented in a comprehensive Java-based application Phenex (Balhoff et al., 2010), and an earlier software Phenote, designed for annotating mutant phenotype in model organisms (Washington et al., 2009). The flexibility of the EQ approach provides rich semantics for describing nearly any organismal phenotype at the very high level of granularity (Dahdul et al., 2018, 2010).

However, if high granularity is not needed by downstream analysis, then the use of EQ syntax might be time inefficient. For example, the ontology-informed phylogenetic method for reconstructing ancestral anatomies PARAMO (Tarasov et al., 2019) uses an input where a character statement is only tagged with one or few ontology term(s) (i.e., URI(s): uniform resource identifier). To date, there is no software that would facilitate this "light" version of phenotypic annotation, at the same time, doing this manually is laborious due to enormous amount of terms contained in any ontology.

Unlike its popularity in the phylogenetics community, `R` environment (R Core Team, 2020) is not broadly used for developing ontology-oriented software. This hinders creation of workflows that seek to integrate ontological and phylogenetic approaches within the same programming environment and, hence, inhibits development of new computational methods at the interference of the two fields (Tarasov, 2019). To fill up these gaps, we have created an open source `R` package `ontoFAST` that provides an interactive interface for "light" phenotypic annotation of characters with biological ontologies. Additionally, it provides functions for visualizing hierarchies of characters and ontologies using the `sunburstR` package (Bostock et al., 2020) and `Cytoscape` (`https://cytoscape.org`). Finally, `ontoFAST` provides a means to construct queries of characters against ontologies for getting a new insight into their phenotype-phenotype relationships. In turn, this enhances interoperability between ontology-oriented applications and, hopefully, will stimulate further development of ontological tools in R.

2 | MATERIALS AND METHODS

2.1 | ontoFAST Availability

The current stable version of the package ontoFAST requires R 3.5.0 and is distributed under the GPL license. The package can be downloaded from CRAN at https://cran.r-project.org/web/packages/ontoFAST/index.html, its development version is available at https://github.com/sergeitarasov/ontoFAST. The detail tutorial is given at https://github.com/sergeitarasov/ontoFAST/wiki.

2.2 | Implementation of ontoFAST

ontoFAST is developed using Shiny (RStudio, Inc, 2020) that enables building interactive web applications straight from R. The interactive interface of ontoFAST can be run either from within RStudio (RStudio Team, 2021) or any web browser. OntoFAST uses functions from ontologyIndex package (Greene et al., 2017) for parsing and manipulating ontologies. It also depends on visNetwork package (Almende B.V. et al., 2019) for interactive visualization of ontology graphs.

2.3 | Data

We tested ontoFAST by using it to annotate two character matrices. All these datasets are included in the package and tutorial. One matrix with 392 characters (dataset Sharkey_2011) from the large-scale Hymenoptera (sawflies, wasps, ants and bees) phylogeny (Sharkey et al., 2012) was annotated using the Hymenoptera Anatomy Ontology (dataset HAO) (Yoder et al., 2010); the annotations are stored in Sharkey_2011_annot dataset.

Another matrix of 232 characters from the dung beetle (Coleoptera: Scarabaeinae) phylogeny (Tarasov, 2017) was annotated using dung beetle ontology (Scarab dataset), the annotations are stored in Tarasov_2017_annot. The Scarab ontology was developed from HAO by enriching it with anatomical terms specific for dung beetle since beetles, so far, lack any comprehensive anatomy ontology. Scarab is an informal and experimental ontology and should be used with caution in other studies.

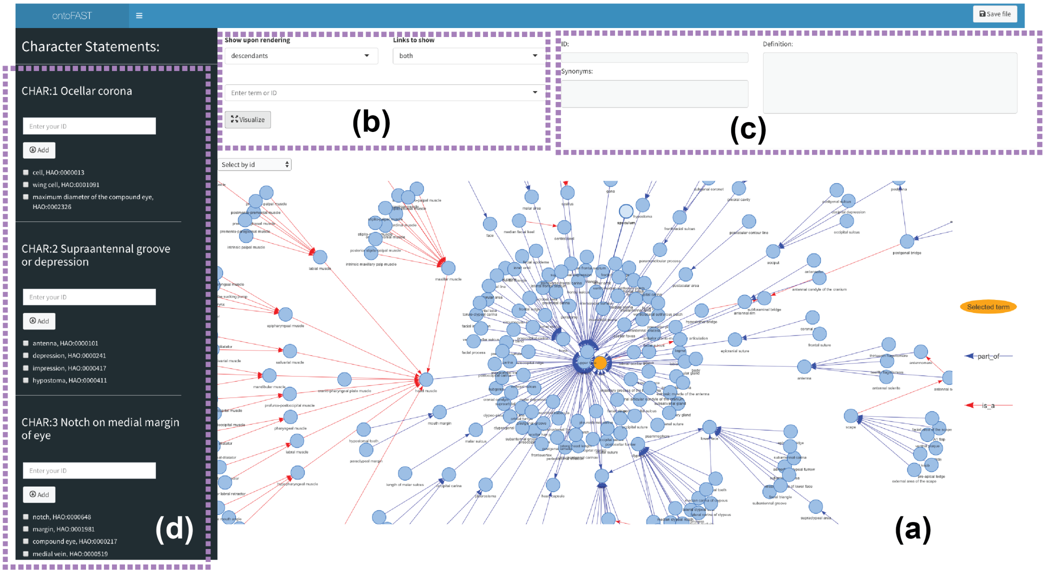
3 | RESULTS AND DISCUSSION

3.1 | Input and output data

The character annotation using ontoFAST requires two initial pieces of data: a list of character statements and a biomedical ontology. The list of character statements can be imported into R as a csv table or a vector of text strings. To import statements from a character matrix stored in the widely-used NEXUS format, one can open it in a popular software Mesquite (Maddison and Maddison, 2018) and copy character statements into any software that supports csv format (e.g. Microsoft Excel or Atom https://atom.io/).

Any organism-specific anatomy ontology or supporting ontologies [e.g. PATO, BSPO (Dahdul et al., 2014), RO (Mungall et al., 2021)] can be used for annotating characters in ontoFAST; the selected ontology should be in OBO format and can be read with get_OBO() function from ontologyIndex package.

```
install.packages("ontoFAST")
# install.packages("igraph")
```



Character Statements:

CHAR:1 Ocellar corona

Enter your ID

Add

- □ cell, HAO:0000013
- □ wing cell, HAO:0001091
- □ maximum diameter of the compound eye, HAO:0002326

CHAR:2 Supraantennal groove or depression

Enter your ID

Add

- □ antenna, HAO:0000101
- □ depression, HAO:0000241
- □ impression, HAO:0000417
- □ hypostoma, HAO:0000411

CHAR:3 Notch on medial margin of eye

Enter your ID

Add

- □ notch, HAO:0000648
- □ margin, HAO:0001981
- □ compound eye, HAO:0000217
- □ medial vein, HAO:0000519

Show upon rendering: descendants

Links to show: both

Enter term or ID

Visualize

Select by id

ID:

Definition:

Synonyms:

FIGURE 1 The interface of ontoFAST. **(a)** Ontology panel. **(b)** Customize panel. **(c)** Information panel. **(d)** Character panel.

```
library("ontoFAST")
hao_obo<-get_OBO(system.file("data_onto", "HAO.obo", package = "ontoFAST"),
extract_tags="everything", propagate_relationships = c("BFO:0000050", "is_a"))
data(Sharkey_2011)
```

The output of `ontoFAST` is a list object that contains annotations: character IDs and associated URIs (or names) of ontology terms. This list can be used for queries, exported to csv format, or, using special functions provided by `ontoFAST`, exported to third-party applications.

3.2 | Annotating characters using ontoFAST

Prior to running the interactive interface, the read-in data have to be preprocessed in R console using the following three steps. First, run `onto_process()` function to combine ontology and character statement into a single object of ontology-index class (Greene et al., 2017); this function automatically parses synonyms from the ontology, the argument `do.annot = TRUE` runs fuzzy matching of characters against the ontology and suggests candidate terms for annotation. Second, create a new environment to store a variable that will serve as an input and output for the interactive mode; the new environment should be called `ontofast` (other names will not work), it enables global usage of the input variable for the functions operating during the interactive session. Third, use the function `make_shiny_in()` to create an object in the `ontofast` environment; the name of this variable is taken as an argument by `runOntoFast()` to launch the interactive session.

```
hao_obo<-onto_process(hao_obo, Sharkey_2011[,1], do.annot = FALSE)
ontofast <- new.env(parent = emptyenv())
```

```
# creating shiny_in variable to serve as an input and output for runOntoFast()
ontofast$shiny_in <- make_shiny_in(hao_obo)
# running the interactive session
runOntoFast(is_a = c("is_a"), part_of = c("BFO:0000050"), shiny_in="shiny_in",
file2save = "OntoFAST_shiny_in.RData")
```

The interactive interface consists of four panels (Fig. 1). The **ontology panel** shows an interactive graph where nodes are the ontology classes and edges are usually *part_of* and *is_a* relationships. The **customize panel** on the top of the window allows selecting relationships to display and navigate to a required term by typing a few letters of its name. The **information panel** shows ID, synonyms, and definition of the term selected in the navigation panel.

The leftmost **character panel** shows the character statements. There are three ways to annotate them: (1) if you ran fuzzy matching with `onto_process()` the candidate terms are shown below the "Add" button and can be selected by checking the respective box(es); (2) click on a node in the **ontology panel**, move the cursor to the **character panel** and click the respective "Add" button; (3) paste term URI right in the **character panel**. Every character can be annotated with more than one term.

Upon the annotation is complete you can close the window and return to the console mode. The characters and their annotation are stored in the lists `ontofast$shiny_in$terms_selected` and `ontofast$shiny_in$terms_selected_id`. Consider saving your data, by clicking the "Save file" button in the top right corner, while in the interactive mode. This will help to avoid risk of loosing annotations, if R session crashes; ontoFAST saves data to the file specified via `file2save` argument in `runOntoFast()`. The created annotations can be further used in downstream analyses or saved as csv files.

```
out <- list2edges(ontofast$shiny_in$terms_selected_id)
write.csv(out, "annotations.csv")
```

3.3 | Visualizing and queering annotations

Visualizing with sunburstR and Cytoscape. Having characters linked with ontology may provide new insight into their relationships. We use the annotations produced with `ontoFAST` for Hymenoptera and dung beetles (see the Data section) to demonstrate it. The hierarchical structure can be visualized using sunburst plot from the `sunburstR` package (Bostock et al., 2020). This plot shows relational hierarchy using a series of rings; each ring corresponds to a level in the ontological hierarchy – the inner circles represent ontology classes and outermost circle represents the annotated characters (Fig. 2b-d). The function `paths_sunburst()` automatically convert `ontoFAST` data to `sunburstR` format.

The annotations and ontologies can be also exported to `Cytoscape` using `export_cytoscape()` function for further manipulation and visualization (Fig. 2a-b). `Cytoscape` is an open source software for visualizing complex networks and integrating them with any type of attribute data.

Querying. Our package has a set of functions for running simple queries with the annotated characters. The function `chars_per_term()` calculates the number of characters for each ontology class; `get_ancestors_chars()` return all shared ancestral ontology classes for a set of characters; and, `get_descendants_chars()` returns all characters descending from a given ontology class.

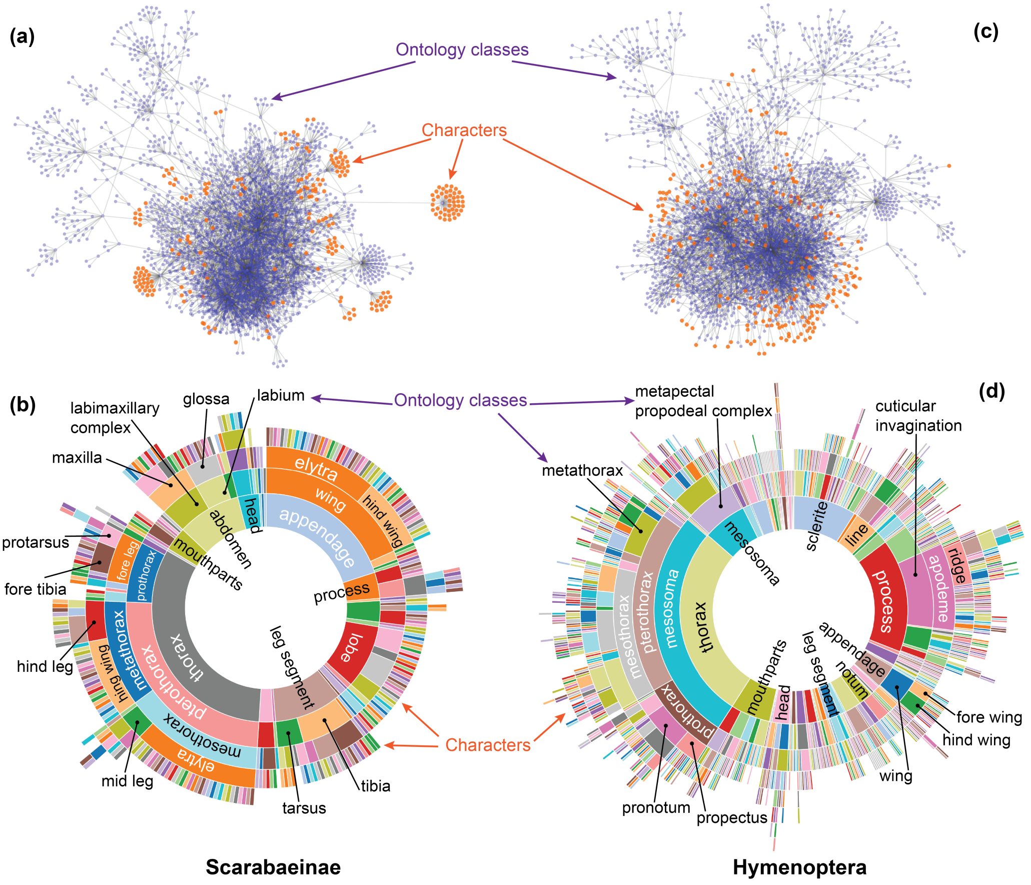


FIGURE 2 Visualization of characters annotated using ontoFAST for Hymenoptera and Scarabaeinae (dung beetles). **(a-b)** The network of ontology classes and the linked characters produced using Cytoscape. **(b-d)** The plots show hierarchy of characters and ontological classes, produced using sunburstR package.

4 | AUTHORS' CONTRIBUTIONS

ST conceived and designed the package; IM and ST annotated Hymenoptera and Scarabaeinae datasets; all authors contributed to designing presented annotation procedure and wrote the paper.

5 | DATA ACCESSIBILITY

The code of ontoFAST version xxx including all data will be archived on Zenodo upon acceptance.

references

Almende B.V., Thieurmel, B. and Robert, T. (2019) *visNetwork: Network Visualization using 'vis.js' Library*. URL: `https://CRAN.R-`

project.org/package=visNetwork. R package version 2.0.9.

Balhoff, J. P., Dahdul, W. M., Kothari, C. R., Lapp, H., Lundberg, J. G., Mabee, P., Midford, P. E., Westerfield, M. and Vision, T. J. (2010) Phenex: ontological annotation of phenotypic diversity. *PLoS One*, **5**, e10500.

Balhoff, J. P., Mikó, I., Yoder, M. J., Mullins, P. L. and Deans, A. R. (2013) A semantic model for species description applied to the ensign wasps (hymenoptera: Evaniidae) of new caledonia. *Systematic Biology*, **62**, 639–659.

Bostock, M., Rodden, K., Warne, K. and Russell, K. (2020) *sunburstR: Sunburst 'Htmlwidget'*. URL: https://CRAN.R-project.org/package=sunburstR. R package version 2.1.5.

Burleigh, J. G., Alphonse, K., Alverson, A. J., Bik, H. M., Blank, C., Cirranello, A. L., Cui, H., Daly, M., Dietterich, T. G., Gasparich, G. et al. (2013) Next-generation phenomics for the tree of life. *PLoS Currents*, **5**.

Dahdul, W., Manda, P., Cui, H., Balhoff, J. P., Dececchi, T. A., Ibrahim, N., Lapp, H., Vision, T. and Mabee, P. M. (2018) Annotation of phenotypes using ontologies: a gold standard for the training and evaluation of natural language processing systems. *Database*, **2018**.

Dahdul, W. M., Balhoff, J. P., Engeman, J., Grande, T., Hilton, E. J., Kothari, C., Lapp, H., Lundberg, J. G., Midford, P. E., Vision, T. J. et al. (2010) Evolutionary characters, phenotypes and ontologies: curating data from the systematic biology literature. *PLoS One*, **5**, e10708.

Dahdul, W. M., Cui, H., Mabee, P. M., Mungall, C. J., Osumi-Sutherland, D., Walls, R. L. and Haendel, M. A. (2014) Nose to tail, roots to shoots: spatial descriptors for phenotypic diversity in the biological spatial ontology. *Journal of Biomedical Semantics*, **5**, 1–13.

Deans, A. R., Lewis, S. E., Huala, E., Anzaldo, S. S., Ashburner, M., Balhoff, J. P., Blackburn, D. C., Blake, J. A., Burleigh, J. G., Chanet, B. et al. (2015) Finding our way through phenotypes. *PLoS Biol*, **13**, e1002033.

Deans, A. R., Mikó, I., Wipfler, B. and Friedrich, F. (2012) Evolutionary phenomics and the emerging enlightenment of arthropod systematics. *Invertebrate Systematics*, **26**, 323–330.

Dececchi, T. A., Balhoff, J. P., Lapp, H. and Mabee, P. M. (2015) Toward synthesizing our knowledge of morphology: Using ontologies and machine reasoning to extract presence/absence evolutionary phenotypes across studies. *Systematic biology*, **64**, 936–952.

Gkoutos, G. V., Green, E. C., Mallon, A.-M., Hancock, J. M. and Davidson, D. (2005) Using ontologies to describe mouse phenotypes. *Genome biology*, **6**, 1–10.

Greene, D., Richardson, S. and Turro, E. (2017) ontologyx: a suite of r packages for working with ontological data. *Bioinformatics*, **33**, 1104–1106.

Maddison, W. and Maddison, D. (2018) Mesquite: a modular system for evolutionary analysis. version 3.40. available from: h ttp. *mesquiteproject. org (accessed 15 January 2018)*.

Mungall, C., Osumi-Sutherland, D., pgaudet, Overton, J. A., Matentzoglu, N., Clare72, Balhoff, J., Harris, N., Brush, M., Touré, V., Sinclair, M., Poelen, J., Bretaudeau, A., Cain, S., Haendel, M., Vasilevsky, N., diatomsRcool, Hammock, J., Laporte, M.-A., Jensen, M. and Larralde, M. (2021) oborel/obo-relations: 2021-03-08 release. URL: https://doi.org/10.5281/zenodo.4589530.

Mungall, C. J., Gkoutos, G. V., Smith, C. L., Haendel, M. A., Lewis, S. E. and Ashburner, M. (2010) Integrating phenotype ontologies across multiple species. *Genome biology*, **11**, 1–16.

R Core Team (2020) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org.

RStudio, Inc (2020) *shiny: Easy web applications in R*. URL: http://shiny.rstudio.com.

RStudio Team (2021) *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA. URL: `http://www.rstudio.com/`.

Sharkey, M. J., Carpenter, J. M., Vilhelmsen, L., Heraty, J., Liljeblad, J., Dowling, A. P., Schulmeister, S., Murray, D., Deans, A. R., Ronquist, F. et al. (2012) Phylogenetic relationships among superfamilies of hymenoptera. *Cladistics*, **28**, 80–112.

Tarasov, S. (2017) A cybertaxonomic revision of the new dung beetle tribe parachoriini (coleoptera: Scarabaeidae: Scarabaeinae) and its phylogenetic assessment using molecular and morphological data. *Zootaxa*, **4329**, 101–149.

— (2019) Integration of Anatomy Ontologies and Evo-Devo Using Structured Markov Models Suggests a New Framework for Modeling Discrete Phenotypic Traits. *Systematic Biology*. URL: `https://doi.org/10.1093/sysbio/syz005`.

Tarasov, S., Mikó, I., Yoder, M. J. and Uyeda, J. C. (2019) Paramo: A pipeline for reconstructing ancestral anatomies using ontologies and stochastic mapping. *Insect Systematics and Diversity*, **3**, 1.

Washington, N. L., Haendel, M. A., Mungall, C. J., Ashburner, M., Westerfield, M. and Lewis, S. E. (2009) Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol*, **7**, e1000247.

Yoder, M. J., Miko, I., Seltmann, K. C., Bertone, M. A. and Deans, A. R. (2010) A gross anatomy ontology for hymenoptera. *PloS one*, **5**, e15991.