HPO Workshop

Release 1

Monarch Initiative

Jul 20, 2022

CONTENTS:

1 HPOannotQC **3**

1.1 Introduction 3
1.2 Navigating the HPO Website 4
1.3 Phenomizer 7
1.4 Encoding Clinical Data with HPO 9
1.5 Exomiser 10
1.6 Human, Mouse & Cross-Species Comparison 10

The site contains material to help you learn about the Human Phenotype Ontology (HPO). It is intended to support workshops but can be used for self study. The workshop has a number of topics that are shown in the menu. Each topic contains one or more exercises and a link to a page with hints and answers.

CHAPTER

ONE

HPOANNOTQC

1.1 Introduction

The Human Phenotype Ontology (HPO) was launched in 2008 to provide a comprehensive logical standard to describe and computationally analyze phenotypic abnormalities found in human disease. T he HPO is now a worldwide standard for phenotype exchange. As an ontology, HPO enables computational inference and sophisticated algorithms that support combined genomic and phenotypic analyses. Broad clinical, translational and research applications using the HPO include genomic interpretation for diagnostics, gene-disease discovery, mechanism discovery and cohort analytics, all of which assist in realizing precision medicine.

We have published several articles that explain how the HPO is constructed. The 2008 paper explains how the HPO was created, and the subsequent papers describe a series of improvements and innovations. For this workshop, we recommend skimming the 2021 manuscript.

- Robinson et al (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet;83(5):610-5. [PMID:18950739]
- Köhler et al (2014) The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res;42(Database issue):D966-74. [PMID:24217912]
- Köhler et al (2017) The Human Phenotype Ontology in 2017. Nucleic Acids Res;45(D1):D865-D876. [PMID:27899602]
- Köhler et al (2021) The Human Phenotype Ontology in 2021. Nucleic Acids Res;49(D1):D1207-D1217 [PMID:33264411]

1.1.1 Scope of the HPO

The HPO was initially developed for the analysis of Mendelian disease. Subsequently, the HPO has been used in a number of other contexts. The following papers discuss an extension of the HPO to analyze common (complex) disease, an application of the HPO for leverage EHR-encoded laboratory data to search for biomarkers in asthma, and a method for using semantic clustering to characterize subtypes of long COVID.

- Groza T et al (2014) The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. Am J Hum Genet. 2015 Jul 2;97(1):111-24. [PMID:26119816]
- Zhang XA et al (2019) Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. NPJ Digit Med;2:32. [PMID:31119199]
- Reese JT et al (2022) Generalizable Long COVID Subtypes: Findings from the NIH N3C and RECOVER Programs [medRxiv]

1.1.2 Translations and Plain-language HPO

We have developed a "translation" of the HPO into plain language that is intended for use by patients and their families.

- Vasilevsky NA et al (2018). Plain-language medical vocabulary for precision diagnosis. Nat Genet;50(4):474-476. [PMID:29632381]

Collaborating groups have created translations in European, Asian, African, and Australian aboroginal languages. Details are available on the HPO Website.

1.1.3 Community collaborations

The HPO has benefited enormously from thousands of contributions from hundreds of clinicians and researchers across the world. This hasa led to a growth of the HPO from initially about 8000 terms to over 16,000 terms today. We have received suggestions for new HPO terms on our GitHub tracker and have also conducted about 40 "hackathons" with domain experts from various fields on clinical medicine. Some of this activity is summarized in the Nucleic Acids Research database articles listed above. In some cases, separate papers describing this activity were published.

- Köhler S et al (2012). Ontological phenotype standards for neurogenetics. Hum Mutat;33(9):1333-9. [PMID:22573485]
- Sergouniotis PI et al (2019). An ontological foundation for ocular phenotypes and rare eye diseases. Orphanet J Rare Dis.;14(1):8. [PMID:30626441]
- Ong E et al (2020) Modelling kidney disease using ontology: insights from the Kidney Precision Medicine Project. Nat Rev Nephrol. 2020 Nov;16(11):686-696 [PMID:32939051]
- Gasteiger LM (2020) Supplementation of the ESID registry working definitions for the clinical diagnosis of inborn errors of immunity with encoded human phenotype ontology (HPO) terms. J Allergy Clin Immunol Pract;8(5):1778 [PMID:32389282]
- Haimel M (2021) Curation and expansion of Human Phenotype Ontology for defined groups of inborn errors of immunity. J Allergy Clin Immunol:S0091-6749(21)00732-6 [PMID:33991581]
- Lewis-Smith D (2021). Modeling seizures in the Human Phenotype Ontology according to contemporary ILAE concepts makes big phenotypic data tractable. Epilepsia;62(6):1293-1305 [PMID:33949685]

Additional hackathons of this nature are planned over the next five years with support from the NHGRI. Interested groups are invited to contact us for more information.

1.2 Navigating the HPO Website

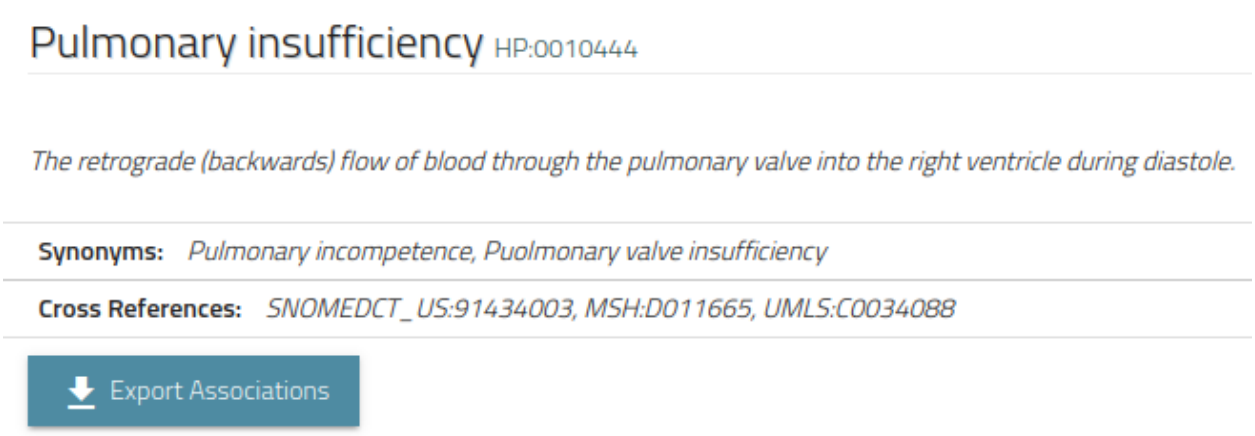
This page contains a series of exercises that illustrate how to find information on the HPO Website.

1.2.1 The basics

Go to the start page of the HPO site. You will find a search box that offers autocomplete functionality.

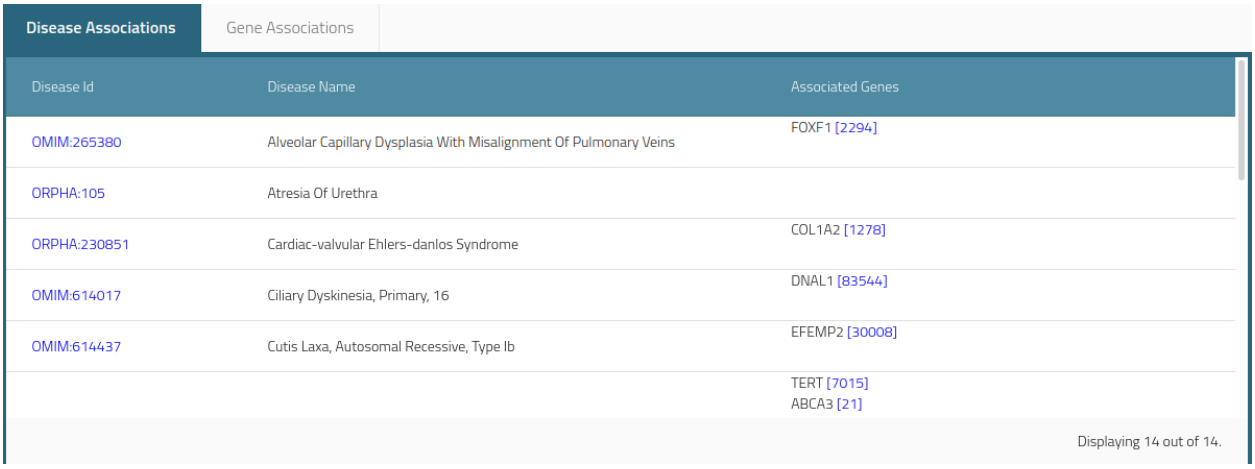
You can search for HPO terms, diseases, and genes. Let's start by searching for *Pulmonary insufficiency*. Enter it in the search window and click on the link. You should get to the page for Pulmonary insufficiency.

The page offers several components. The core information about the term is shown in this box.



We see the primary identifier of the term (HP:0010444), the prefered label (Pulmonary insufficiency), a definition, and a list of synonyms (Synonyms are useful to help find terms and to support text mining).

Most HPO terms are used to annotate one or more diseases. In this case, Pulmonary insufficiency is used to annotate a number of diseases that are shown in the Disease Associations tab.



The Gene associations tab shows the genes that are associated with these diseases.

Exercise 1

Determine how many genes are associated with pulmonary insufficiency in the HPO data resource. Note that by associated we mean that if a pathogenic variant in a gene causes a Mendelian disease, one of whose manifestations is *pulmonary insufficiency*, then we regard the gene as being associated with this phenotypic feature.

There are two ways of answering this question. (1) The Gene tab has a list of associated genes. (2) You can click on the `Export associations` button shown above to download an Excel file with the genes. For this exercise, download the excel file.

The HPO hierarchy

Each term in the HPO describes a phenotypic abnormality, such as Atrial septum defect. Terms are related to parent terms by "is a" relationships. The structure of the HPO, which allows a term to have multiple parent terms, enables different aspects of phenotypic abnormalities to be explored. The true-path rule applies to the terms of the HPO. That is, if an individual has Sutural cataract, the individual can also be said to have the parent term of Sutural cataract, Zonular cataract, as well as the grand-parent term of Sutural cataract, Cataract.

Medically, this is true, because a Sutural cataract is defined as

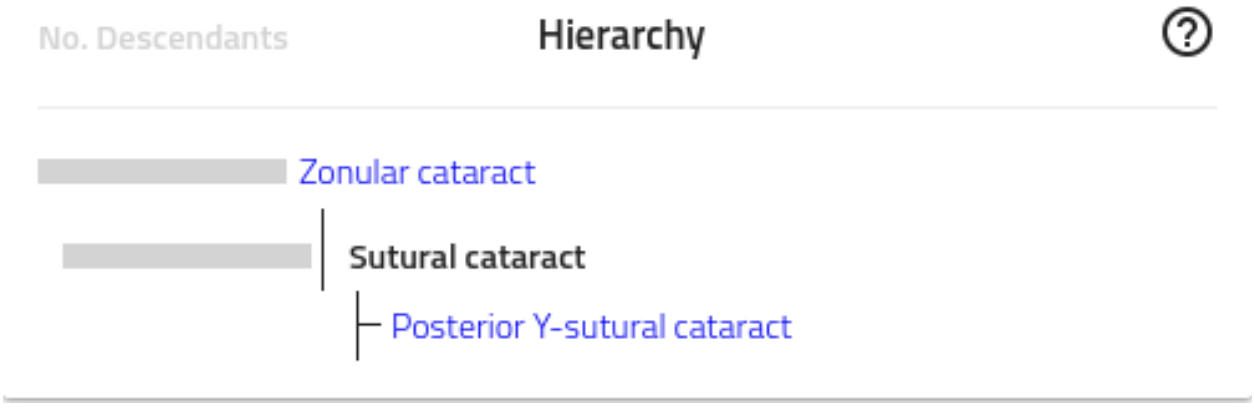
> A cataract that affects the region of the lens directly beneath the capsule of the lens.

Since a Zonular cataract is defined as a cataract that affects a specific region (zone) of the lens, this is clearly true. It is also clear that both Sutural cataract and zonular cataract are specific kinds of cataract. The HPO is constructed to capture medical knowledge in this way. It is therefore important for users to know how to navegate the hierarchy of the HPO. We will do so in the following exercise.

Exercise 2

For this exercise, you are asked to determine the number of links (hops) from Sutural cataract to the root of the phenotypic abnormality section of the HPO, Phenotypic abnormality. The latter term is the parent to all pghenotypic abnormalities. Please take a few minutes to explore the children of this term in order to get a sense of the breadth of the HPO.

To calculate the number of links, note that there is one link between Sutural cataract and Zonular cataract. This can be seen in the correspond part of the page for sutural cataract.



Each term page of the HPO shows an excerpt of the entire hierarchy. The gray bar to the left of the terms in this view will show the number of descendent terms if you hover the mouse over the bar. For instance, Sutural cataract contains a single descendent term, Posterior Y-sutural cataract, while Zonular cataract has a total of 21 descendents (including Sutural cataract).

You can navigate from a term to its parent term using the links in this box, and can continue to do so until you have reached Phenotypic abnormality. The number of clicks will equal the number of links between Sutural cataract and Phenotypic abnormality, i.e., the `depth` of the term Sutural cataract.

The HPO website chooses to show only an excerpt of the hierarchy for simplicity's sake. Other ontology browsers show the entire path from the selected term to the root. We recommend the Ontology Lookup Service (OLS) of the European Bioinformatics Institute (EBI). Here is a link to the OLS page for Sutural cataract, which allows you to count the number of links directly.

Wrap-up

In this module, you should have learned to recognize the basic elements of an HPO term (id, label, definition, and in many cases comment, synonyms, and cross references). You should now be able to search for HPO terms using the HPO webpage, how to determine how many diseases and genes are associated with a term, and how to navigate the hierarchy of the HPO to determine a term's depth in the ontology.

If you had trouble with any of the exercises, see rstwebsiteanswers.

1.3 Phenomizer

The Phenomizer is a web-based application that provides clues to the differential diangosis of an individual with suspected rare disease based on the observed phenotypic abnormalities.

- Köhler S et al (2009) Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet;85(4):457-64 [PMID:19800049]

The differential diagnostic process attempts to identify candidate diseases that best explain a set of clinical features. This process can be complicated by the fact that the features can have varying degrees of specificity, as well as by the presence of features unrelated to the disease itself. Depending on the experience of the physician and the availability of laboratory tests, clinical abnormalities may be described in greater or lesser detail. We have adapted semantic similarity metrics to measure phenotypic similarity between queries and hereditary diseases annotated with the use of the Human Phenotype Ontology (HPO) and have developed a statistical model to assign p values to the resulting similarity scores, which can be used to rank the candidate diseases.

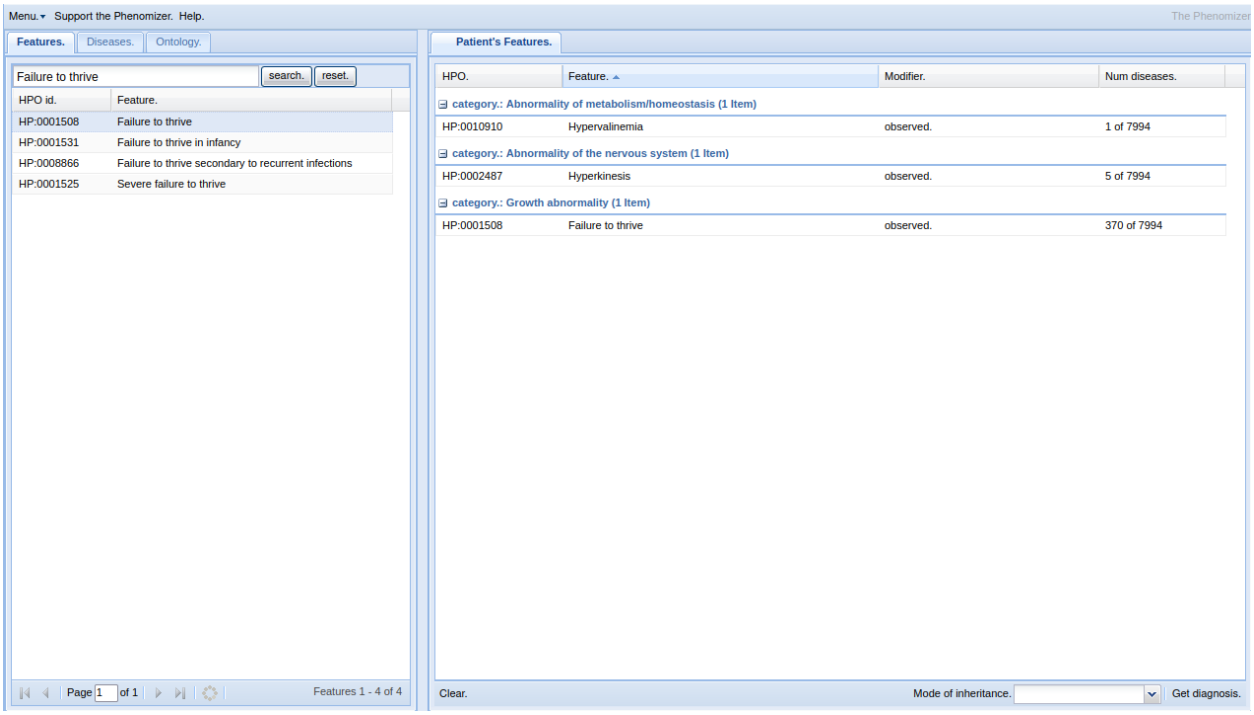
The Phenomizer has a short manual that can be downloaded from the `help` menu of the Phenomizer web application. Before doing the exercise, read the manual to familiarize yourself with the application.

1.3.1 Exercise 1

We will use the Phenomizer to search for the correct diagnosis of an individual observed to have the following phenotypic features

- Valinuria
- Hyperkinesis
- Failure to thrive

Enter these terms in the Phenomizer (the easiest way is to copy the terms from this webpage and paste them into the autocomplete field of Phenomizer. The app will ask you if you would like to use symmetric mode (click yes; details about this are in the paper cited above). You should now see something like the following.



Now click on the `Get diagnosis` button and examine the differential diagnosis window.

What is the top candidate proposed by Phenomizer? Why?

1.3.2 Exercise 2

In some cases, the initial workup of a patient may not provide sufficient detail to guide the differential diagnosis. Let us simulate this sitation by entering only the following two terms into a new Phenomizer session.

- Multiple cafe-au-lait spots
- Scoliosis

If we click on the `Get diagnosis` button and examine the differential diagnosis window, we will see that none of the proposed differential diagnoses is significant. There are many ways to use Phenomizer to narrow down the differential diagnosis. Let us imagine we have examined a child with neurofibromatosis type 1, but are unaware of the diagnosis. In principle, we might use tools such as Phenomizer to find phenotypic abnormalities, which, if present, would most improve the differential diagnosis. This works because in the current list, there are many diagnoses with the same relatively unspecific match. If we can identify one more HPO terms in our patient that is specific for one or other disease, then the diagnosis should move to the top of the list. The manual of the Phenomizer describes the two search modes - binary and specific.

We suggest that you add the specific additional term Axillary freckling. If you would like to work more on this example, you might want to consult the diagnostic criteria for NF1 and Legius syndrome to find additional appropriate terms.

1.3.3 Wrap-up

In this module, you have gotten familiar with the Phenomizer and the basics of HPO-based semantic similarity analysis for differential diagnostic support.

If you had trouble with any of the exercises, see rstphenomizeranswers.

1.4 Encoding Clinical Data with HPO

We recommend that clinicians, genetic counselors, and other healthcare professionals who will be entering HPO terms as a part of clinical care consult this detailed protocol about how to choose optimal HPO terms in various clinical situations.

- Köhler S, Øien NC, et al (2019). Encoding Clinical Data with the Human Phenotype Ontology for Computational Differential Diagnostics. Curr Protoc Hum Genet. 2019 Sep;103(1):e92 [PMID:31479590]

1.4.1 Exercise 1

This exercise may be difficult for those without medical training. We will extract a list of HPO terms from a published case report about an individual with X-linked Megalocornea.

- Han J, et al (2015) X-linked Megalocornea Associated with the Novel CHRDL1 Gene Mutation p.(Pro56Leu*8). Ophthalmic Genet.;36(2):145-8 [PMID:24073597]

Go to the clinical vignette in this article, and identify the phenotypic abnormalities. Use the HPO website to search for the corresponding HPO terms. Write down the list of terms.

1.4.2 Exercise 2

In practice, many people will use text mining approaches to help identify HPO terms in clinical texts. For this exercise, we will try another published case report:

- Brizola E, et al Variable clinical expression of Stickler Syndrome: A case report of a novel COL11A1 mutation. Mol Genet Genomic Med. 2020 Sep;8(9):e1353. [PMID:32558342]

Try this tool to do the text mining.

- doc2hpo

doc2hpo has a nice online tutoral with more information about how to use the tool.

Wrap-up

In this module, you have practiced how to extract HPO terms from clinical texts. We have used published case reports to demonstrate the process. Analogous steps would be performed for real clinical data.

If you had trouble with any of the exercises, see rstchoosetermsanswers.

1.5 Exomiser

Exomiser is an application that prioritizes genes and variants in next-generation sequencing (NGS) projects for novel disease-gene discovery or differential diagnostics of Mendelian disease.

- Robinson PN, et al (2014) Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res;24(2):340-8. [PMID:24162188]
- Smedley D (2015) Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc;10(12):2004-15. [PMID:26562621]

Exomiser was among the first bioinformatics tools of its kind, published in 2014 as a freely available Java program. It requires as input (i) a variant call format (VCF) file with the called variants of a rare disease patient (or optionally a multi-sample VCF and pedigree (PED) file if family members have also been sequenced) and (ii) a set of HPO terms to describe the corresponding patient's phenotype.

A demo version of Exomiser is available on the Monarch Initiative website.

To try out Exomiser, download the example file with a causative FGFR2 variant for the autosomal dominant Pfeiffer syndrome added to exome of a healthy individual. Add HPO terms representing the Phenotype of Pfeiffer syndrome as you have done with the Phenomizer. You may want to consult the HPO page for Pfeiffer syndrome to find appropriate HPO terms.

You can run Exomiser with the default settings or adjust them (in the latter case, you may want to consult the consult the above cited papers to learn about the meaning of the parameters).

1.5.1 Exercise 1

Examine the output of Exomiser. Examine the top 5 candidates and explore the phenotypic and genetic evidence for or against their candidacy.

1.6 Human, Mouse & Cross-Species Comparison

The use of model organisms as tools for the investigation of human genetic variation has significantly and rapidly advanced our understanding of the aetiologies underlying hereditary traits. However, while equivalences in the DNA sequence of two species may be readily inferred through evolutionary models, the identification of equivalence in the phenotypic consequences resulting from comparable genetic variation is far from straightforward,

There are three major methodologies to identify phenotypes in the mouse that are relevant to a human disease (Robinson PN, Webber C, PMID:24699242).

- **(A)** Classical approach. A mouse model is made or identified that possesses a genotype equivalent to a penetrant mutation that in human underlies the disease of interest (termed construct validity). The mouse model is examined for phenotypes that resemble those that define the human disorder (face validity).
- **(B)** Phenolog mapping. A group is formed containing candidate genes for a disease of interest. The respective mouse models for the orthologues of these genes are then examined for any unusually overrepresented phenotypes among them and these phenotypes (termed phenologs) are deemed relevant to the disease.
- **(C)** Direct phenotype mapping. Given the phenotype(s) that describe a human disease, the corresponding phenotypes in mouse are inferred by means of computational reasoning using interspecies phenotype ontology analysis. In the example shown, the HPO term Aortic stenosis is defined on the basis of the PATO term constricted and aortic valve from the cross-species anatomy ontology UBERON. Similarly, the MPO term aortic valve stenosis is defined using the same PATO term constricted and aortic valve. Automatic reasoning therefore places the HPO term Aortic stenosis and the MPO term aortic valve stenosis in the direct vicinity of one another in a cross-species phenotype ontology.

1.6.1 Monarch Initiative

The Monarch Initiative integrates information on genes, variants, genotypes, phenotypes and diseases in a variety of species, and allows powerful ontology-based search. [Shefchek KA, et al, Nucleic Acids Res. 2020 PMID:31701156]. In this exercise, we will explore how to use the Monarch Web app to explore the cross-species inference algorithms that are used in the Exomiser tool. The following Figure summarizes the cross species-matching approach; for more details, please consult the Shefchek et al. paper.

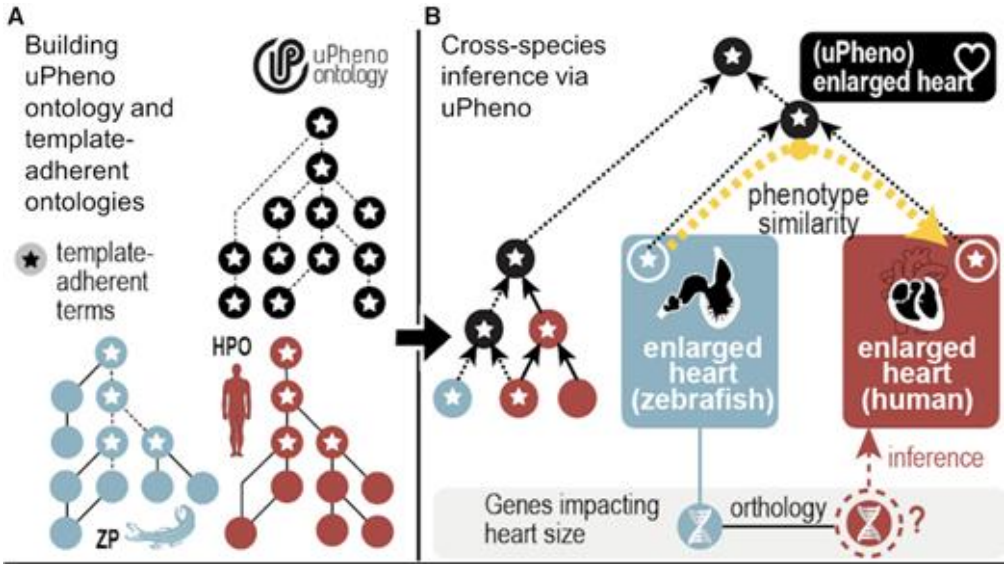


Fig. 1: uPheno template-driven ontology development and harmonization. uPheno templates are used to define phenotypes according to agreed upon design patterns. (A). Computable definitions specified using uPheno templates are used to automate classification of uPheno and parts of the Zebrafish Phenotype Ontology (ZP; dashed lines). (B). Computable definitions also drive automated classification of HPO and ZP classes under uPheno classes. For example, enlarged heart in ZP (defined using the zebrafish anatomy heart term) and enlarged heart in HPO are both classified under uPheno enlarged heart (defined using Uberon heart). Algorithms can use this classification under uPheno to predict that human orthologs of zebrafish genes annotated to enlarged heart may cause enlarged heart in humans.

Exercise 1

The Alliance of Genome Resources (AGR) is a consortium of the major model organism databases [PMID:31552413]. The AGR and Monarch Initiative websites offer good portals for exploring cross-species phenotype data. For these exercises, we will explore the AGR pages related to Noonan syndrome 1, which is caused by deleterious variants in the *PTPN11* gene.

Go to the corresponding PTPN11 page. Answer the following question:

- How many human diseases are associated with mutation in *PTPN11*? (Feel free to work with another gene of your choice).

Exercise 2

- What are some phenotypic categories that are abnormal in human and mouse?

For this, let's use the Phenogrid tool of the Monarch Initiative.

1. Open the Phenogrid entry page: https://monarchinitiative.org/analyze/phenotypes
2. Click, "No, I'll need some help"
3. Click "Generate a list from a gene"
4. Enter *PTPN11* (human)
5. Click "Compare profile" (at bottom of page)
6. Click "Show me everything"
7. Click taxon mouse

1.6.2 Wrap-up

In this module, you have learned to search for cross-species phenotype data.

If you had trouble with any of the exercises, see rstxspeciesanswers.

This application is designed to transform our internal HPO Annotation files (the small files) together with the Orphanet XML file into the `phenotype.hpoa` file. It performs extensive Q/C on the annotation files. By default it updates TermIds in the Orphanet files that have been updated.