# Lawrence Berkeley National Laboratory
## Environ Genomics & Systems Bio

**Title**
Encoding Clinical Data with the Human Phenotype Ontology for Computational Differential Diagnostics

**Permalink**
https://escholarship.org/uc/item/8dt809mc

**Journal**
Current Protocols in Human Genetics, 103(1)

**ISSN**
1934-8266

**Authors**
Köhler, Sebastian
Øien, N Christine
Buske, Orion J
et al.

**Publication Date**
2019-09-01

**DOI**
10.1002/cphg.92

Peer reviewed

# Encoding Clinical Data with the Human Phenotype Ontology for Computational Differential Diagnostics

**Sebastian Köhler**[1,2,3,$], **N. Christine Øien**[4,$], **Orion J Buske**[5], **Tudor Groza**[6], **Julius OB Jacobsen**[3,7], **Craig McNamara**[6], **Nicole Vasilevsky**[3,8], **Leigh C Carmody**[3,9], **JP Gourdine**[3,8], **Michael Gargano**[3,9], **Julie McMurry**[3,10], **Daniel Danis**[3,9], **Christopher J Mungall**[3,11], **Damian Smedley**[3,7], **Melissa Haendel**[3,8,10], **Peter N Robinson**[3,9,12,*]

[1.]Charité Centrum für Therapieforschung, Charité-Universitätsmedizin Berlin Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin 10117, Germany.

[2.]Einstein Center Digital Future, Berlin 10117, Germany.

[3.]Monarch Initiative, monarchinitiative.org.

[4.]Freelance Genome Analyst, Berlin, Germany.

[5.]Gene42, Toronto, ON, Canada.

[6.]Pryzm Health, Australia.

[7.]Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK.

[8.]Oregon Health & Science University, Portland, OR 97217, USA.

[9.]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA.

[10.]Oregon State University, Corvallis, OR, USA.

[11.]Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

[12.]Institute for Systems Genomics, University of Connecticut, Farmington, CT, USA.

## Abstract

The Human Phenotype Ontology (HPO) is standardized set of phenotypic terms that are organized in a hierarchical fashion. It is a widely used resource for capturing human disease phenotypes for computational analysis to support differential diagnostics. The HPO is frequently used to create a set of terms that accurately describe the observed clinical abnormalities of an individual being evaluated for suspected rare genetic disease. This profile is compared with computational disease profiles in the HPO database with the aim of identifying genetic diseases with comparable phenotypic profiles. The computational analysis can be coupled with the analysis of whole-exome or whole-genome sequencing data through applications such as Exomiser. This protocol explains

[*]correspondence to Peter Robinson, The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA, peter.robinson@jax.org.
[$]these authors contributed equally

**Conflict of Interest.** Gene42 Inc. provides licensing, support, customization, and integration services for PhenoTips.

how to choose an optimal set of HPO terms for these use cases and to enter them with software, such as PhenoTips and PatientArchive, and demonstrates how to use Phenomizer and Exomiser to generate a the computational differential diagnosis.

## Introduction

Unambiguous, computable descriptions of disease phenotypes are critical for robust differential diagnosis and clinical care, especially for rare and genetic diseases. The fact that genetic diseases are rare, coupled with a lack of access to expert diagnosticians and/or genomic testing, means that many individuals with a rare disease go years without a diagnosis. That a substantial proportion of initial diagnoses are wrong further compounds this significant problem. Although next-generation sequencing (NGS)-based diagnostic procedures, such as whole exome and whole genome sequencing (WES, WGS), have greatly accelerated the pace of discovery of Mendelian disease-associated genes and increased our understanding of the spectrum associated with a large number of rare genetic syndromes, the molecular diagnosis rates using NGS data are still low (25–50%) (de Ligt et al., 2012; Rauch et al., 2012; Yang et al., 2013, 2014; Zhu et al., 2015; Tammimies et al., 2015; Clark et al., 2018).

We have been developing the Human Phenotype Ontology (HPO, http://www.human-phenotype-ontology.org) since 2008 to support differential diagnosis and translational research within the field of rare genetic diseases (Robinson et al., 2008). The HPO offers a computational bridge between genome biology and clinical medicine, making it a comprehensive bioinformatic resource for the analysis of human diseases and phenotypes (Robinson et al., 2008; Robinson and Webber, 2014; Köhler et al., 2017; Robinson and Mundlos, 2010; Groza et al., 2015; Köhler et al., 2014; Vasilevsky et al., 2018). It does so by offering a standardized set of phenotypic terms that are organized in a hierarchical fashion. Using standardized hierarchies enables us to put our phenotypic knowledge into an organized framework that can be analyzed by computational means. The adoption of the HPO by the 100,000 Genomes Project, SOLVE-RD, the NIH Undiagnosed Diseases Program, the NIH Undiagnosed Diseases Network, the Global Alliance for Genomics and Health (GA4GH) and many other projects (Köhler et al., 2018) underlines that it is the *de facto* standard for rare disease phenotype analysis. The HPO is the clinical flagship of the Monarch Initiative where it is embedded in a semantically unified framework of knowledge on diseases, genes, phenotypes, and model organisms with over 50,000,000 items of knowledge (Mungall et al., 2017; McMurry et al., 2016). This means that the HPO is effectively incorporated into a much larger disease biology knowledge network.

The HPO describes individual phenotypic abnormalities in a hierarchical framework of organs/body parts (terms such as *Aortic aneurysm*; HP:0004942) (Figure 1). The **HPO annotations (HPOAs)** are disease descriptions that use these standard terms. The rare genetic disease Marfan syndrome, for instance, is characterized by and therefore annotated to over 50 phenotypic abnormalities, including *Aortic aneurysm*. Each abnormality reported in a patient with that disease is represented by an HPO term. The annotations can have modifiers that describe the age of onset and the frequencies of features. For instance, the

phenotypic abnormality *Brachydactyly* (HP:0001156) is rare in hydrolethalus syndrome (3/56 according to a published study referenced in our data), but affects nearly 100% of patients in many of the other diseases characterized by this abnormality (currently 484 diseases). Algorithms can use this information to weight findings in the differential diagnosis. There are currently over 14,900 terms and 168,600 annotations to 7,370 rare Mendelian diseases. The annotations, together with additional data about age of onset, clinical modifiers, term frequency, and genes, provide computational models of the diseases.

Computable descriptions of human disease using HPO phenotypic profiles, i.e. HPOAs, have become a key element in a number of algorithms being used to support genomic diagnostics, including Exomiser (Robinson et al., 2014; Smedley et al., 2015, 2016). The HPO allows algorithms to 'compute over' clinical phenotype data in a wide variety of contexts, e.g., for the detection of shared patterns of clinical findings. The ontological structure of the HPO allows the specificity (information content) of individual terms to be quantified. This information content, when used in tandem with the structure of the HPO, enables sets of phenotypes to be fuzzy-matched (Köhler et al., 2009; Bauer et al., 2012; Schulz et al., 2009), i.e. the patterns of two phenotypic profiles can be compared and their overlap quantified. Software such as Phenomizer and Exomiser use specificity-weighted fuzzy matching of HPO terms entered by users to identify the best match to disease models, while the Matchmaker Exchange (Philippakis et al., 2015; Buske et al., 2015b) relies on fuzzy matching of HPO terms to find similar patients across databases for the purpose of novel gene discovery.

This article describes how to select HPO terms that best represent the clinical findings identified in a patient. We also describe how to use Phenomizer, PhenoTips (Girdea et al., 2013), PatientArchive, and Exomiser; four applications that utilize the HPO to provide a list of differential diagnoses.

## Basic Protocol 1. Choosing HPO Terms to Represent the Clinical Manifestations of a Proband

The purpose of this protocol is to explain how to create a list of HPO terms that describe the clinical abnormalities and other features of the proband. This protocol uses the word "proband" to denote the individual whose clinical manifestations will be analyzed with the software. The proband is any individual for whom we are seeking a diagnosis, such as a patient currently receiving medical care.

### Necessary Resources

1. Computer with internet access and a current web browser.

2. Clinical data from medical charts or comparable sources.

3. HPO Browser available through http://www.human-phenotype-ontology.org.

   a. This website lists a number of additional recommended browsers with additional search or visualization features (Köhler et al., 2018).

4. Word processor, other program, or pen and paper to record your selected HPO terms.

## Steps and Annotations

1. **How to use HPO Browser.**—Open the HPO Browser by going to http://www.human-phenotype-ontology.org. Find the box at top center. Choose 'Term' from the dropdown menu and start typing the phenotypic feature of interest into the box that says "Search for phenotypes, diseases, genes…" at top center. As illustrated below (Figure 2), the search box provides autocomplete functionality, meaning that a list of possible matches is shown in a dropdown menu while you are typing in the query. Click on the term in the dropdown menu that best describes your feature. This will take you to an HPO term overview page that provides a wealth of detail, including term synonyms (Figure 3). This term can be exported to your program of choice. If the suggested term was not what you had in mind, use the panel on the left hand side to navigate up and down the ontology. You can also start the search again in the same search box, now located at the top right.

2. **Choose the most specific HPO terms**—It is essential that you choose the most specific HPO term for the phenotypic feature you have listed. For instance, if a proband was diagnosed with an aortic aneurysm that affects the aortic arch and has saccular morphology, then the HPO term *Saccular aortic arch aneurysm* (HP:0031647) should be chosen and not the more general term *Aortic arch aneurysm* (HP:0005113), which is the parent of *Saccular aortic arch aneurysm*, or *Thoracic aortic aneurysm* (HP:0012727), which is the parent of *Aortic arch aneurysm* (Figure 1). If a specific term is not available, request it through the HPO GitHub tracker (https://github.com/obophenotype/human-phenotype-ontology/issues/new/choose). New HPO terms become available with new HPO releases, i.e. every two months. This means that you should pick the most specific HPO term available on the day you are selecting your HPO terms, even if that term is not as specific as you would have liked. If possible, the term can be replaced with the more specific term once it is available in a new HPO release.

3. **Choose important and relevant HPO terms**—Choose HPO terms that cover all the important phenotypic abnormalities observed in the proband. The precise definition of "important" will depend on clinical judgement, meaning that the definition can vary for different probands. For instance, if a hospitalized proband has multiple blood tests for electrolytes and one of these reveal a borderline high sodium level, it may be inappropriate to include the HPO term *Hypernatremia* (HP:0003228). Because some rare genetic diseases, such as nephrogenic diabetes insipidus caused by deleterious variants in the *AQP2* gene, are characterized by severe hypernatremia, it is appropriate to include the HPO term *Hypernatremia* if the proband has repeated measurements of high serum sodium. Similarly, many phenotypic abnormalities that occur rarely in the general population occur at a higher frequency in certain diseases. Low-grade myopia or scoliosis are examples of such common features. We are unable to provide a general rule for when such a term should be included. Though it may be tempting to exclude the HPO term for a mild phenotypic feature, such as *Scoliosis* (HP:0002650) in a proband with mild scoliosis, keep in mind that genetic diseases almost always have a spectrum. This means that it is appropriate to include the term

*Scoliosis* in a proband with multiple skeletal anomalies independent of its severity, however, it may be inappropriate to include it in a proband with isolated, mild scoliosis. This emphasizes the importance that the HPO needs to be considered in the context of the proband's clinical situation. An additional example is described in the note at the end of Basic Protocol 2.

**3. Add negated HPO terms to denote excluded (normal) clinical findings—** The HPO is limited to terms that describe abnormal phenotypes. This means that an organ system or physiological function examined and found to be normal can be denoted by a negated term. Many HPO term entry systems provide a "NOT" button for this purpose. In contrast to when entering a positive phenotypic feature, negated terms should be described by the most general term available. For instance, if a detailed ultrasound of the liver had normal results, the best negated HPO term is NOT *Abnormal liver morphology* (HP: 0410042) because in principle all of the morphological abnormalities denoted by more specific descendents of the term were ruled out.

**4. Choose a sufficient number of terms—**There is no general rule for the optimal number of HPO terms that should be used to describe a proband. One recent study showed five well-chosen terms to be a good threshold for phenotype-driven exome analysis (Kernohan et al., 2018). Remember that some diseases are characterized by a larger number of clinical abnormalities than others. This is illustrated by comparing a congenital polymalformation syndrome with isolated hearing loss. It is clear that a polymalformation syndrome will have to be described by a larger number of HPO terms than a disease resulting from an isolated abnormality, such as hearing loss. Probands with more than one disease may need to be described by a larger number of terms. Therefore, as a rule of thumb, enter at least five HPO terms. If fewer than five abnormalities are observed, include excluded (negated) annotations to reach a total number of at least five terms.

It is essential that you choose HPO terms that cover all relevant findings seen in the proband. Try to provide a global overview of the phenotype without limiting yourself to a single organ system. In probands with a single clinical abnormality, such as hearing loss, we recommend that you add a negated term for each organ system in which clinical abnormalities were excluded. Software can use the negated terms to narrow down the differential diagnosis. The extraction of HPO terms from electronic medical records may ease entry of HPO terms in the future (Son et al., 2018).

## Basic Protocol 2. Phenomizer

Phenomizer was the first software tool to use semantic similarity metrics to measure phenotypic similarity between queries and hereditary diseases annotated with the HPO (Köhler et al., 2009). It uses a statistical model to assign p-values to the resulting similarity scores. The resulting p-values are then used to rank the candidate diseases. This protocol describes how to enter HPO terms in Phenomizer and how to interpret and refine the results.

### Necessary Resources

1.    Computer with internet access and a current web browser.

> **2.** Clinical data from medical charts or comparable sources.
>
> **3.** Optional - List of HPO terms describing the proband's clinical abnormalities as described in Basic Protocol 1.
>
> **4.** Phenomizer is available through http://compbio.charite.de/phenomizer/.

1. **Entering HPO Terms.**

   1. Click on Phenomizer's Features tab. Find the best HPO term by typing part of the term, the term's synonym, or the HPO-identifier into the Search box. The Search box provides autocomplete functionality, meaning that a list of possible matches is shown in a dropdown menu while you are typing in the query (Figure 4).

   2. Clicking on the Search button located to the right of the search box results in a table listing the HPO terms that best match your query. You can also use the Ontology tab to use the ontology hierarchy to identify the most suitable HPO term (Figure 4). This is useful when descending the ontology hierarchy in order to identify the most specific term currently available.

   3. You can open a context menu for each feature displayed in the table on the left hand side by right-clicking on the row. The context menu shows the hierarchical position of the selected HPO term, as well as a list of annotated diseases (Figure 5).

   4. Select your chosen HPO term by double clicking on it, using drag-and-drop, or by clicking on "add to patient's features" from the context menu. Either of these actions will move the term to the Patient's Features tab on the right hand side.

   5. To delete an HPO term from the Patient's Features list, right-click on that row and select 'remove' from the context-menu. Use the 'clear'- button in the lower left hand corner of the table to delete all HPO terms from the list.

2. **Patient features**—The HPO terms that you selected in the previous step are now listed on the right-hand side of the Phenomizer window. It is important to enter as many of the patient's clinical features as possible as this usually improves the specificity of the proposed differential diagnoses. Item 4 explains how to add and remove features in order to further explore the differential diagnosis options.

3. **Interpret results**—Clicking the 'Get diagnosis' button in the lower right hand corner makes Phenomizer compute a list of differential diagnoses ranked by p-value (Figure 6), which appears under the Diagnosis tab on the right hand side. Remember that a significant p-value does not mean that the diagnosis is confirmed, simply that the diagnosis is a plausible one and that the physician should consider statistically significant differential diagnoses carefully. If Phenomizer fails to identify any conditions with a significant p-value, this is interpreted to mean that the chosen combination of HPO terms is not per se sufficient to make a diagnosis. It could also mean that the proband has a condition not yet listed in the

database used by Phenomizer. Phenomizer currently encompasses the mainly Mendelian diseases listed by Online Mendelian Inheritance in Man (OMIM) (Amberger et al., 2015).

**4. Refine query**—Phenomizer offers several strategies for refining your query.

1. Every feature can be designated as either 'observed' or 'mandatory'. If a term is denoted as 'mandatory', all conditions not having this particular feature, or a descendent thereof, will be filtered out from the generated differential diagnosis list. All features are initially listed as 'observed'. To change this to 'mandatory', click on 'observed' in the Modifier column for the HPO term of interest. This will cause a dropdown menu to appear. Click on the modifier term to select.

2. You can test different modes of inheritance. Clicking on the Mode of Inheritance box located on the lower right-hand side will bring up a dropdown list of modes of inheritance. Choose your mode of interest by clicking on it. All conditions that do not exhibit the specified mode of inheritance will be filtered out from the generated differential diagnosis list. It may be appropriate to use this filter if you have observed that the disease cosegregates according to a specific Mendelian mode of inheritance.

3. Phenomizer's "Improve differential diagnosis" functionality can be used in the event that the initial differential diagnoses ranking is inconclusive, e.g. it contains numerous entries with significant p-value or only entries without significant p-value. Compile a list of interesting diseases, e.g. the first ten in the ranking, by checking the boxes on the left-hand side of the Diagnosis box. Then click on the "Improve differential diagnosis" button on the lower right. This will bring up a pop-up box that suggests additional HPO features that are either specific to one of the selected diseases or characterize approximately half of the selected diseases. (Figure 7). To use the first option, click on the Specific Search button at the top of the box. The specific search function is designed to find features which, if present, are most specific for only one of the current differential diagnoses. To use the second option, click on the Binary Search button at the top of the box. Binary search can be an efficient way to narrow down the differential diagnosis. After entering the new feature with specific or binary search, recalculate the differential diagnosis by clicking the 'Get diagnosis' button.

**5. Download results**—The Phenomizer results can be downloaded by clicking on the Download Results button in the lower right-hand corner. The results are available as either a pdf or csv file. The downloaded file will contain the top ranked diseases and the parameters. The parameters are the HPO terms, the selected mode of inheritance, and the similarity measure. The PDF file can be used for the purposes of documentation in the patient's medical or research chart.

You can loop back to section 4 (Refine Query) to further improve the differential diagnostic process.

Note: Remember to include HPO terms that are thought to be pathological or unusual, even if they are frequent in the general population (Basic Protocol 1, Step 3). *Blue irides* (HP:0000635), for example, is a common phenotype in the general population. As a phenotype annotation, however, it refers to to the finding of markedly blue iris seen in oculocutaneous albinism and other disorders of pigmentation. Iit is a subtype of *Abnormal iris pigmentation* (HP:0008034) and should only be used when abnormal pigmentation is suspected, such as in the disease oculocutaneous albinism (Lewis, 2000).

## Basic Protocol 3. Using PhenoTips for deep phenotyping and diagnosis suggestions

PhenoTips (https://phenotips.org/) is an open-source software tool for collecting and analyzing phenotypic information from probands or families with a suspected genetic condition. The configurable patient form, family pedigree tool, and built-in gene and diagnosis suggestions are designed to help make deep phenotyping practical for clinicians and researchers. It is used by over 4,000 specialists around the world, including the NIH Undiagnosed Diseases Network (UDN) and Undiagnosed Diseases Program (UDP), the Care4Rare Canada Consortium, the RD-Connect Genome-Phenome Analysis Platform, and the PhenomeCentral matchmaking portal (Buske et al., 2015a). This protocol describes how to enter a proband's phenotypic data into PhenoTips and review the generated suggestions.

**Necessary Resources**

1. Computer with internet access and a current web browser.

2. Clinical data from medical charts or comparable sources.

3. Optional - List of HPO terms describing the proband's clinical abnormalities as described in Basic Protocol 1.

4. PhenoTips is available from https://phenotips.org. You can download and install it on your computer, or obtain a user account on an institutional or cloud-hosted instance.

**1. Create a new patient record.—**Create a new PhenoTips record from any page by clicking on the "Create…" button in the navigation bar and selecting "New patient". This creates a new, empty patient record, which can then be filled in, shared, and exported. The record has several sections which can be expanded or collapsed by clicking on the heading. The sections you see, their order, and their contents can be configured by an administrator. Save your progress by clicking on "Quick save" or save and review the patient record by clicking on "Save and view summary". Enter the patient's name, an identifier such as medical record number, sex, and date of birth in the "Patient information" section of the patient form.

**2. Record phenotypic features.—**The "Clinical symptoms and physical findings" section of the patient form provides several interfaces for following Basic Protocol 1 and recording the patient's phenotypic profile (Figure 8).

- "Quick phenotype search" provides an error-tolerant predictive search of HPO terms, their synonyms, and their definitions. Click "Y" or the phenotype name to indicate that the suggested term is present in the proband. Click "N" to denote it as absent (excluded).

- Browse phenotypic features by category below the search box. The categories and listed phenotypic features are customizable, enabling specialty clinics and research studies to provide users with a checklist of the most relevant phenotypic features for their specific context. Selecting a term from the checklist expands the more specific sub-terms, encouraging the most specific description possible.

- Click on the info button to the right of every phenotypic feature to display more information or browse the HPO structure from that term.

- All selected phenotypic features are summarized on the right, grouped by category. Click on "Add details" to add metadata to each term, e.g., age of onset, severity, or laterality, or upload supporting evidence, such as a photo or PDF report. The stars at the top of the section correspond to the specificity of a patient's phenotypic profile as computed by the Monarch specificity scorer.

**3. Optional: Using HPO terms to collect the patient's family history.**—A "Draw pedigree" placeholder is located in the "Family history and pedigree" section of the patient form. This tool can be used to record and visualize the patient's family medical history (Figure 9).

- Click on the corresponding icon to open the pedigree editor tool

- Choose whether this is a new family with this patient as the proband or this patient is an additional family member in an already existing PhenoTips family.

- Choosing to create a new pedigree will offer you the option of starting from a template family structure or importing a family tree from a file in a supported format, such as PED and GEDCOM. Add additional family members by hovering over any node and clicking on one of the relationship handles: parents, partnerships, siblings, or children.

- Click on the node to add reported medical history to an individual. A dialog windows pops up with three tabs: "Personal", "Clinical", and "Cancer". HPO terms can be added and removed in the "Clinical" tab under "Clinical symptoms". The "Cancer" tab has special questions for recording cancer diagnoses using HPO terms, along with cancer-specific metadata. HPO terms added to an individual in the pedigree will appear in the pedigree key along with the number of affected individuals in the family.

- Hover over the name to highlight affected family members.

- Drag and drop an HPO term onto any node to annotate the individual with the term.

- Once complete, click "Save" to save the pedigree and then "Close" to return to the patient record.

- Previously added pedigrees appear as a thumbnail. Existing pedigrees can be added and edited by clicking on the thumbnail and following the steps outlined above.

**4.     Optional: Adding HPO terms through abnormal measurements.**—In order to ensure that the applicable standard growth curve is used, make sure that the patient's sex and date of birth is entered into the "Patient information" section.

1.     Click on the "Measurements" section.

2.     Click on "+ New entry".

3.     Enter the date the measurements were taken.

4.     Enter all supported and available measurements.

5.     The percentile and standard deviations will be shown for each measurement. A growth curve with the patient's measurements plotted over top will be displayed.

6.     In the event of an abnormal measurement, the corresponding HPO term will be automatically added to the patient's phenotypic profile. For example, if the patient's head circumference measurement was less than 3 SD below normal for the age and sex, *Microcephaly* (HP:0000252) is automatically added.

**5.     Gene and diagnosis suggestions.**—PhenoTips uses the HPOAs developed and maintained by the HPO team to automatically suggests genes and OMIM disorders based on the patient's phenotypic profile. These annotations can be updated in PhenoTips through the Administrative interface. Relevant genes are shown in the "Suggested genes" section (Figure 10). The HPO terms used are shown at the top, along with the number of associated genes. Clicking on an HPO term will toggle whether or not it is included in the gene suggestions. Suggested genes are shown underneath, in decreasing order of the number of associated HPO terms. The table can be exported in a tab-delimited format compatible with Microsoft Excel by clicking on the "Download" button. A similar interface is provided for suggesting OMIM diagnoses in the "Diagnosis" section. OMIM disorders are ranked according to the similarity between the patient's phenotypic features and the documented phenotype of each OMIM disorder. The prioritization algorithm is configurable. BOQA is the default used at the time of writing (Bauer et al., 2012).

## Basic Protocol 4. Explore candidate disorders using the HPO-based phenotype profile in PatientArchive

PatientArchive is a clinical phenotyping platform underpinning some of the existing Undiagnosed Diseases Programs and rare disease sharing initiatives, such as the Initiative for Rare and Undiagnosed Diseases (IRUD) Japan, the Undiagnosed Diseases Program Western Australia, and the Australian Genomics Matchmaker Exchange node.

This protocol describes the use of a patient phenotype profile built from HPO terms to explore candidate disorders in PatientArchive.

**Necessary Resources**

1. Computer with internet access and a current web browser.

2. Clinical data from medical charts or comparable sources.

3. Optional - List of HPO terms describing the proband's clinical abnormalities as described in Basic Protocol 1.

4. PatientArchive is available through this website. http://patientarchive.org/ New users will have to register for an account by clicking on the 'Account' button on the top left-hand side.

## 1. Create a patient from the Dashboard

- Click on the 'New Patient' button located on the Dashboard. This opens the Demographics section.

- Enter standard details, e.g., name, gender, ethnicity, etc.

- Save the entered data by clicking on the 'Save' button.

- Proceed to augment the patient record with clinical data, imagery, tests, or share the patient with other platform users.

**2. Create a phenotype profile—**Use the 'Add Clinical Record' function under 'Clinical Records' in the 'Clinical Data' section to create the phenotype profile by adding a clinical note. The platform will automatically extract HPO terms from the text and assign them to the patient phenotype profile (Figure 11). The terms are structured and displayed according to the top-level abnormality in HPO - e.g., *Hearing impairment* (HP:0000365) is shown under *Abnormality of the ear* (HP:0000598). This enables a quick assessment of the complexity of an underlying disorder based on the number of anatomical structures impacted.

**3. Explore the candidate disorders space—**Exploring phenotypically related disorders as a part of the differential diagnostic process is often useful. To do so, select 'Explore disorders' from the 'Analytics' section. You can either manually enter the disorders to be ranked or explore the Top-N diseases using Orphanet as the background knowledge base. Selecting the 'Top 3 similar' function will produce the top 3 ranked disorders and the associated disease phenotype profiles (Figure 12). The ranking is computed via a series of semantic similarity measures between HPO terms. The disease phenotype profiles are extracted from the HPO annotations in the Orphanet knowledge base.

# Basic Protocol 5. Exomiser

Exomiser is a Java program designed to analyse WES and WGS VCF samples using the patient's phenotype to distinguish between causative and benign variants. It has found widespread usage throughout the rare disease interpretation and genomic diagnostic space (Bone et al., 2016; Robinson et al., 2014; Smedley et al., 2015, 2016). Once a VCF file and appropriate HPO terms (Basic Protocols 1 and 3) are available, Exomiser can highlight likely pathogenic variants within a matter of minutes. It is used in many major national and

international rare disease initiatives such as the UDN, UDP, 100,000 Genomes Project, RD-Connect and SOLVE-RD. Exomiser also underlies the phenotype matching capabilities of the Matchbox software, created as a collaboration between the Broad Institute and the Monarch Initiative as an off-the-shelf solution for groups wishing to join the Matchmaker Exchange. It is used for pheno-genomic analysis within the PhenomeCentral matchmaking portal.

Exomiser uses a set of HPO terms that describe the clinical abnormalities of the individual being investigated and a VCF file containing the individual's exome or genome data. Exomiser will also accept a multi-sample VCF file containing the proband and their genetic relatives when accompanied by a PED file containing the pedigree information of these individuals.

A demonstration version of Exomiser is available through the online version hosted at https://exomiser.monarchinitiative.org/exomiser/. This provides a demonstration VCF file and input phenotype terms. Note, that in the interests of performance, the web version does not accept VCF files in excess of 75MB or 100,000 variants and is not deployed in a suitable, secure computing environment for analyzing real patient data. The more technically capable, full-featured command-line version is available from https://data.monarchinitiative.org/exomiser/latest/. This will run locally, and requires no network connection once downloaded. This protocol will focus on the command-line version, however the suggestions for interpretation of the results apply equally well to the web version of Exomiser.

### Necessary Resources

1. Computer running a 64-bit operating system with at least 4GB of **free** RAM, Java version 1.8 or higher installed, internet access, and a current web browser.

2. List of HPO terms describing the proband's clinical abnormalities as described in Basic Protocol 1.

3. A WES or WGS VCF file containing data from the proband or a multi-sample file containing the proband and genetic relatives accompanies by a PED file.

**1. Program Setup**—Download the latest Exomiser release and data distribution from https://data.monarchinitiative.org/exomiser/latest/ At the time of writing these are https://data.monarchinitiative.org/exomiser/latest/exomiser-cli-11.0.0-distribution.zip

https://data.monarchinitiative.org/exomiser/latest/1811_hg19.zip

https://data.monarchinitiative.org/exomiser/latest/1811_phenotype.zip

This example does not require the 1811_hg38.zip file.

In the interest of brevity, we refer the reader to the README.md file found at https://data.monarchinitiative.org/exomiser/README.md for instructions on installation of the command-line client. We similarly draw your attention to the contents of the file https://data.monarchinitiative.org/exomiser/latest/

README_IMPORTANT_1811_PHENOTYPE_DATA.txt. The application will only run if you follow the instructions in this file.

**2. Analysis Setup—**You are ready to run the analysis once the program has been downloaded and set-up. The distribution includes a sub-directory called 'examples' which contains both VCF data and Exomiser analysis scripts. For the purposes of this protocol we will assume that the user is running a version of UNIX or a UNIX-like operating system and has installed the distribution in their home directory. We will now look at the analysis script for a quartet (family of four) containing a single affected individual.

```
~/exomiser-cli-11.0.0/examples/test-analysis-multisample.yml
```

The input analysis script contains the most salient parts of the analysis relating to the sample at the start:

```
analysis:
 genomeAssembly: hg19
 vcf: examples/Pfeiffer-quartet.vcf.gz
 ped: examples/Pfeiffer-quartet.ped
 proband: ISDBM322017
 hpoIds: ['HP:0001156', 'HP:0001363', 'HP:0011304', 'HP:0010055']
```

These fields tell the program the major genome assembly (genomeAssembly) which the sample VCF was called against, the full path to the sample VCF (vcf), full path to the pedigree (ped) in PED format, the sample ID of the proband in the VCF file (proband) and the proband phenotype profile encoded using the HPO (hpoIds). The user can input their own VCF and PED file (in the case of a multi-sample VCF) and tell the program the identifier of the proband in the VCF file and input the HPO IDs. In this case, the HPO identifiers are for a fictitious patient with Pfeiffer syndrome. These instructions apply equally when using HPO terms of your choice. Note: The remainder of the analysis script uses the recommended settings for performing a whole exome analysis and can be modified to analyze your input of choice.

**3. Running an Analysis—**From your home directory (~), issue the following commands:

```
~$ cd exomiser-cli-11.0.0
~/exomiser-cli-11.0.0$ java -Xmx4g -jar exomiser-cli-11.0.0.jar --analysis ./
examples/test-analysis-multisample.yml
```

The result of running these commands will be output to the console, which contains the following information if the setup was successful:

```
Welcome to:

 |___|_ _  | ___
 | | |`-./ \_| |\/\_/|`_`-._\/,_-._
 | | | | |_/ | _> < O | | | |`\_,_/|
 |_|_|_|\_| |___/\_\__/|_| |_|_|_\_/|

A Tool to Annotate and Prioritize Exome Variants    v11.0.0

-
2019-04-04 18:16:25.476  INFO 137 --- [          main]
org.monarchinitiative.exomiser.cli.Main  : Started Main in 19.865 seconds (JVM running for
20.527)
2019-04-04 18:16:25.485  INFO 137 --- [          main] o.m.e.cli.ExomiserCommandLineRunner
          : Exomiser running...
-
o.m.e.c.analysis.AbstractAnalysisRunner  : Creating analysis results from VCF examples/Pfeiffer-
quartet.vcf.gz
2019-04-04 18:17:00.062  INFO 137 --- [          main]
o.m.e.c.analysis.AbstractAnalysisRunner  : Finished analysis in 0m 34s 365ms (34365 ms)
2019-04-04 18:17:00.071  INFO 137 --- [          main] o.m.e.c.writers.AnalysisResultsWriter
          : Writing results…
```

**4.  Interpreting results—**The output of this analysis is in the directory ~/exomiser-cli-11.0.0/results. For a simple, human-centric view open the file Pfeiffer-quartet-hiphive-exome-PASS_ONLY.html in a web browser. Navigate past all the input and summary sections to the section entitled 'Prioritised Genes'. These are the main results of the analysis and the top-ranked candidates (Figure 13).

The section is organized by gene ranked by the Exomiser score. The Exomiser score is a score calculated using a logistic regression combining the phenotype and variant scores. These two components are calculated independently from the phenotype match and the genetic information found in the VCF and pedigree and then combined.

In the 'phenotype matches' subsection the software highlights the input HPO terms (on the left) with the matched term for that model (those on the right). For example, the top phenotype match to 'Craniofacial-skeletal-dermatologic dysplasia' has a phenotype score of 0.879 using the HiPhive algorithm. These scores range from a minimum of 0 to a maximum of 1, where 1 would be a perfect self-hit. Any missing phenotypes are not displayed.

## COMMENTARY

### Background information

It can be extremely challenging to make the correct diagnosis in an individual with rare genetic disease, and yet the correct diagnosis is essential for providing the most precise clinical management. The HPO and computational tools that use the HPO are now commonly used to help manage clinical information and suggest differential diagnoses on the basis of clinical data or combined genomic and clinical data. In order to get the best results, it is essential to choose HPO terms wisely and to understand how to run computational tools and interpret their results.

**Critical parameters and troubleshooting—**A large number of academic and commercial tools use the HPO differently for the purposes of performing phenotype-driven analysis. A partial list of these tools is available in two recent review articles (Köhler et al., 2017, 2018). Computational phenotype matching is not an exact science, meaning that the applications often return different lists of matches. Indeed, even the same application can return different results depending on how the algorithm is parameterized. It is important for users to become familiar with the tools and to explore the effects of different settings.

A.  Troubleshooting Basic Protocol 1

You are very likely to come across probands with a number of clinical findings unrelated to the primary genetic diagnosis. Take the example of a recent Kabuki syndrome 2 case study that reported a patient born in the breech position (*Breech presentation*; HP:0001623). As there is no known association between Kabuki syndrome 2 and breech presentation, this was presumably a chance finding (Guo et al., 2018). Be aware that the addition of HPO terms for unrelated findings will typically cause Phenomizer and similar software to reduce the match score.

An incomplete HPOA dataset is another problem you may encounter. If a finding truly associated with the disease is absent from the HPOA database the software will rank it as false-positive. Please report the missing annotation to the HPO GitHub tracker.[1] You should consider removing the annotation from your query until the association has been added to the HPOA database.

**B.**     Troubleshooting Basic Protocol 4

The automatic extraction of HPO terms from free text depends on the underlying processing techniques and the specific terminology used by HPO. This can result in terms not being found. Rephrasing the text often leads to the desired outcome. For example, including *happy disposition* in the clinical note of a patient will lead to a false negative, while replacing the phrase with *happy demeanor* will result in the appropriate HPO term being found.

**C.**     Applying negative HPO terms

For those of you designing your own application that relies on HPO terms or using an already existing application that allows you to enter excluded HPO terms, consider making an ontology hierarchy high level "N" (NOT) annotation for phenotypic categories where the proband had normal findings. For example, if a full visual exam found no visual impairment, then go ahead and choose N-"Visual impairment".

### These tools do not replace clinical judgement

All the described tools are intended to be systems for experts rather than expert systems. That is, the tools support clinicians and researchers by providing information on the basis of data input. If the tools work well, then the correct result will be at rank one or at least among the top ranks ("on the first page"). Human judgment to assess and, in some cases, confirm the results of the computational analysis is always required.

The Human Phenotype Ontology and related tools are intended to be used by qualified and licensed physicians in order to aid in reaching the correct diagnosis in patients with hereditary diseases, in research contexts, and for use as a teaching tool. The HPO and related tools do not make diagnoses. Rather, tools that use the HPO produce a ranked list of possibilities that can be used by physicians as a part of the diagnostic workup. These tools should not be used to make medical decisions without the advice of a physician.

---

[1]https://github.com/obophenotype/human-phenotype-ontology

## Conclusion

The HPO and tools that use it are in wide use for rare-disease diagnostics and translational research. This article has explained how to choose an optimal set of HPO terms to represent the phenotypic abnormalities observed in a proband, and has explained how to use four popular HPO-based tools.

## Acknowledgements.

## References

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, and Hamosh A 2015 OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic acids research 43:D789–D798. [PubMed: 25428349]

Bauer S, Köhler S, Schulz MH, and Robinson PN 2012 Bayesian ontology querying for accurate and noise-tolerant semantic searches. Bioinformatics 28:2502–2508. [PubMed: 22843981]

Bone WP, Washington NL, Buske OJ, Adams DR, Davis J, Draper D, Flynn ED, Girdea M, Godfrey R, Golas G, et al. 2016 Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. Genetics in medicine: official journal of the American College of Medical Genetics 18:608–617. [PubMed: 26562225]

Buske OJ, Girdea M, Dumitriu S, Gallinger B, Hartley T, Trang H, Misyura A, Friedman T, Beaulieu C, Bone WP, et al. 2015a PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. Human mutation 36:931–940. [PubMed: 26251998]

Buske OJ, Schiettecatte F, Hutton B, Dumitriu S, Misyura A, Huang L, Hartley T, Girdea M, Sobreira N, Mungall C, et al. 2015b The Matchmaker Exchange API: automating patient matching through the exchange of structured phenotypic and genotypic profiles. Human mutation 36:922–927. [PubMed: 26255989]

Clark MM, Stark Z, Farnaes L, Tan TY, White SM, Dimmock D, and Kingsmore SF 2018 Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. NPJ genomic medicine 3:16. [PubMed: 30002876]

Girdea M, Dumitriu S, Fiume M, Bowdin S, Boycott KM, Chénier S, Chitayat D, Faghfoury H, Meyn MS, Ray PN, et al. 2013 PhenoTips: Patient Phenotyping Software for Clinical and Research Use. Human mutation 34:1057–1065. [PubMed: 23636887]

Groza T, Köhler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, Schriml LM, Kibbe WA, Schofield PN, Beck T, et al. 2015 The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. American journal of human genetics 97:111–124. [PubMed: 26119816]

Guo Z, Liu F, and Li HJ 2018 Novel KDM6A splice-site mutation in kabuki syndrome with congenital hydrocephalus: a case report. BMC medical genetics 19:206. [PubMed: 30509212]

Kernohan KD, Hartley T, Alirezaie N, Care4Rare Canada Consortium, Robinson PN, Dyment DA, and Boycott KM 2018 Evaluation of exome filtering techniques for the analysis of clinically relevant genes. Human mutation 39:197–201. [PubMed: 29193559]

Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine J-P, Gargano M, Harris NL, Matentzoglu N, McMurry JA, et al. 2018 Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic acids research. Available at: 10.1093/nar/gky1105.

Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GCM, Brown DL, Brudno M, Campbell J, et al. 2014 The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic acids research 42:D966–D974. [PubMed: 24217912]

Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, Mundlos C, Horn D, Mundlos S, and Robinson PN 2009 Clinical diagnostics in human genetics with semantic similarity searches in ontologies. American journal of human genetics 85:457–464. [PubMed: 19800049]

Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, Baynam G, Bello SM, Boerkoel CF, Boycott KM, et al. 2017 The Human Phenotype Ontology in 2017. Nucleic acids research 45:D865–D876. [PubMed: 27899602]

Lewis RA 2000 Oculocutaneous Albinism Type 1 In GeneReviews® (Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Stephens K, and Amemiya A, eds.) University of Washington, Seattle, Seattle (WA).

de Ligt J, Willemsen MH, van Bon BWM, Kleefstra T, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, de Vries P, Gilissen C, et al. 2012 Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. The New England journal of medicine 367:1921–1929. [PubMed: 23033978]

McMurry JA, Köhler S, Washington NL, Balhoff JP, Borromeo C, Brush M, Carbon S, Conlin T, Dunn N, Engelstad M, et al. 2016 Navigating the Phenotype Frontier: The Monarch Initiative. Genetics 203:1491–1495. [PubMed: 27516611]

Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C, Brush M, Carbon S, Conlin T, Dunn N, Engelstad M, et al. 2017 The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic acids research 45:D712–D722. [PubMed: 27899636]

Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, Brunner HG, Buske OJ, Carey K, Doll C, et al. 2015 The Matchmaker Exchange: a platform for rare disease gene discovery. Human mutation 36:915–921. [PubMed: 26295439]

Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, Albrecht B, Bartholdi D, Beygo J, Di Donato N, et al. 2012 Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. The Lancet 380:1674–1682.

Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, and Mundlos S 2008 The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. American journal of human genetics 83:610–615. [PubMed: 18950739]

Robinson PN, Köhler S, Oellrich A, Sanger Mouse Genetics Project, Wang K, Mungall CJ, Lewis SE, Washington N, Bauer S, Seelow D, et al. 2014 Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome research 24:340–348. [PubMed: 24162188]

Robinson PN, and Mundlos S 2010 The human phenotype ontology. Clinical genetics 77:525–534. [PubMed: 20412080]

Robinson PN, and Webber C 2014 Phenotype ontologies and cross-species analysis for translational research. PLoS genetics 10:e1004268. [PubMed: 24699242]

Schulz MH, Köhler S, Bauer S, Vingron M, and Robinson PN 2009 Exact Score Distribution Computation for Similarity Searches in Ontologies. Algorithms in Bioinformatics:298–309. [PubMed: 19765311]

Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, Siragusa E, Zemojtel T, Buske OJ, Washington NL, et al. 2015 Next-generation diagnostics and disease-gene discovery with the Exomiser. Nature protocols 10:2004–2015. [PubMed: 26562621]

Smedley D, Schubach M, Jacobsen JOB, Köhler S, Zemojtel T, Spielmann M, Jäger M, Hochheiser H, Washington NL, McMurry JA, et al. 2016 A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. American journal of human genetics 99:595–606. [PubMed: 27569544]

Son JH, Xie G, Yuan C, Ena L, Li Z, Goldstein A, Huang L, Wang L, Shen F, Liu H, et al. 2018 Deep Phenotyping on Electronic Health Records Facilitates Genetic Diagnosis by Clinical Exomes. American journal of human genetics 103:58–73. [PubMed: 29961570]

Tammimies K, Marshall CR, Walker S, Kaur G, Thiruvahindrapuram B, Lionel AC, Yuen RKC, Uddin M, Roberts W, Weksberg R, et al. 2015 Molecular Diagnostic Yield of Chromosomal Microarray Analysis and Whole-Exome Sequencing in Children With Autism Spectrum Disorder. JAMA: the journal of the American Medical Association 314:895–903. [PubMed: 26325558]

Vasilevsky NA, Foster ED, Engelstad ME, Carmody L, Might M, Chambers C, Dawkins HJS, Lewis J, Della Rocca MG, Snyder M, et al. 2018 Plain-language medical vocabulary for precision diagnosis. Nature genetics 50:474–476. [PubMed: 29632381]

Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu Z, et al. 2013 Clinical whole-exome sequencing for the diagnosis of mendelian disorders. The New England journal of medicine 369:1502–1511. [PubMed: 24088041]

Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y, Ward P, Braxton A, Wang M, Buhay C, et al. 2014 Molecular findings among patients referred for clinical whole-exome sequencing. JAMA: the journal of the American Medical Association 312:1870–1879. [PubMed: 25326635]

Zhu X, Petrovski S, Xie P, Ruzzo EK, Lu Y-F, McSweeney KM, Ben-Zeev B, Nissenkorn A, Anikster Y, Oz-Levi D, et al. 2015 Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. Genetics in medicine: official journal of the American College of Medical Genetics 17:774–781. [PubMed: 25590979]
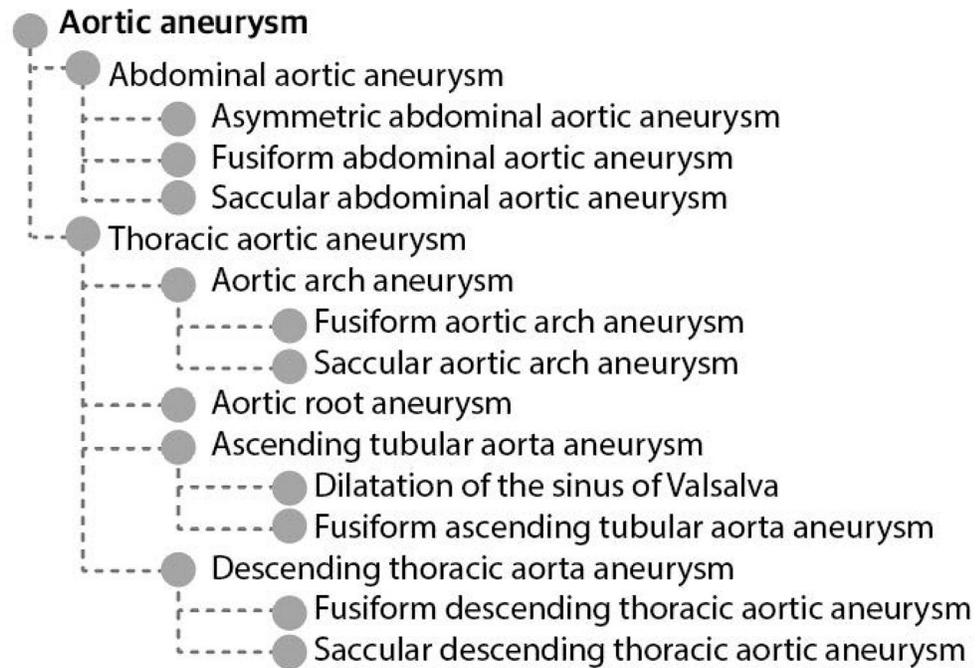
**Figure 1.**
Excerpt of the HPO showing its hierarchical structure. *Saccular aortic arch aneurysm* (HP: 0031647) is a specific form (subclass) of *Aortic arch aneurysm* (HP:0005113), which in turn is a specific form of *Thoracic aortic aneurysm* (HP:0012727), and so on.
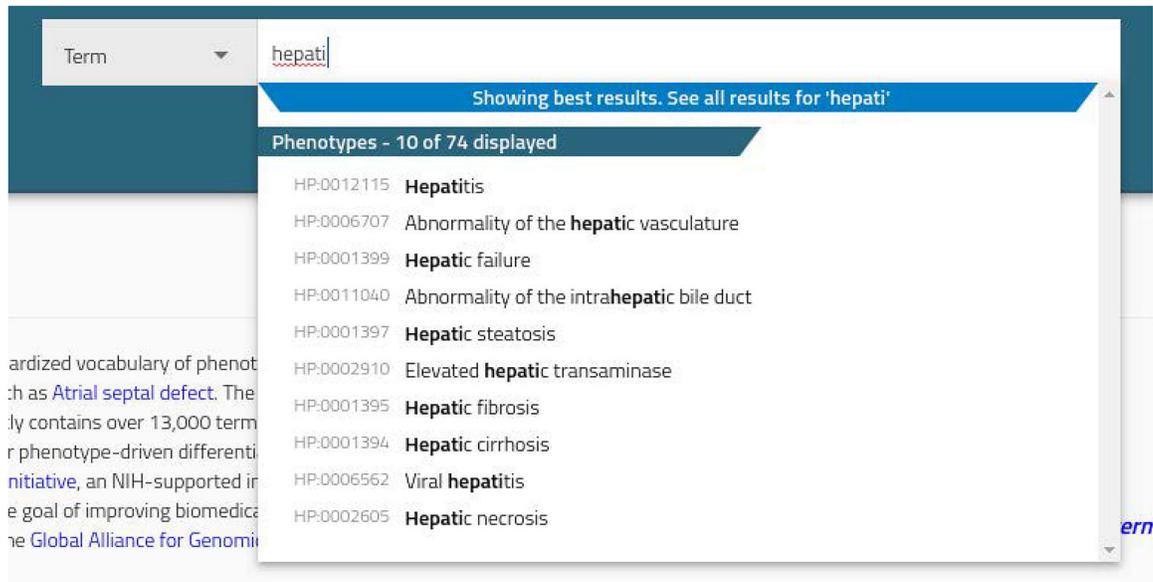
**Figure 2.**
Search and autocomplete functionality of the HPO Browser. The first ten hits that match the text entered so far are shown. To see all hits, click on the blue box with the text "Showing best results. See all results for …"

**Figure 3.**

Overview page for the HPO term *Atrial septal defect* (HP:0001631).
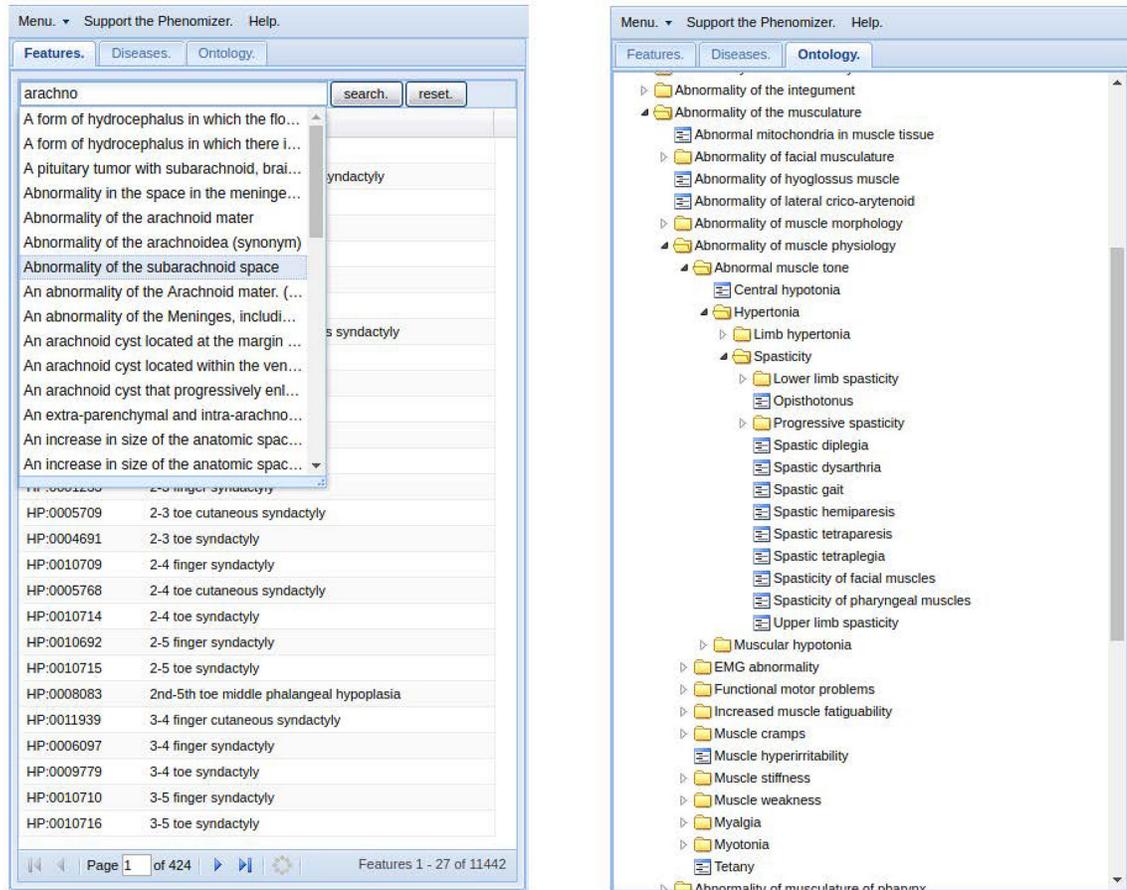
**Figure 4. Phenomizer input window.**

Users can use the autocomplete field (left) or the HPO ontology hierarchy (right) to select the best HPO term. Double clicked terms are selected and appear in the Patient's Features window.
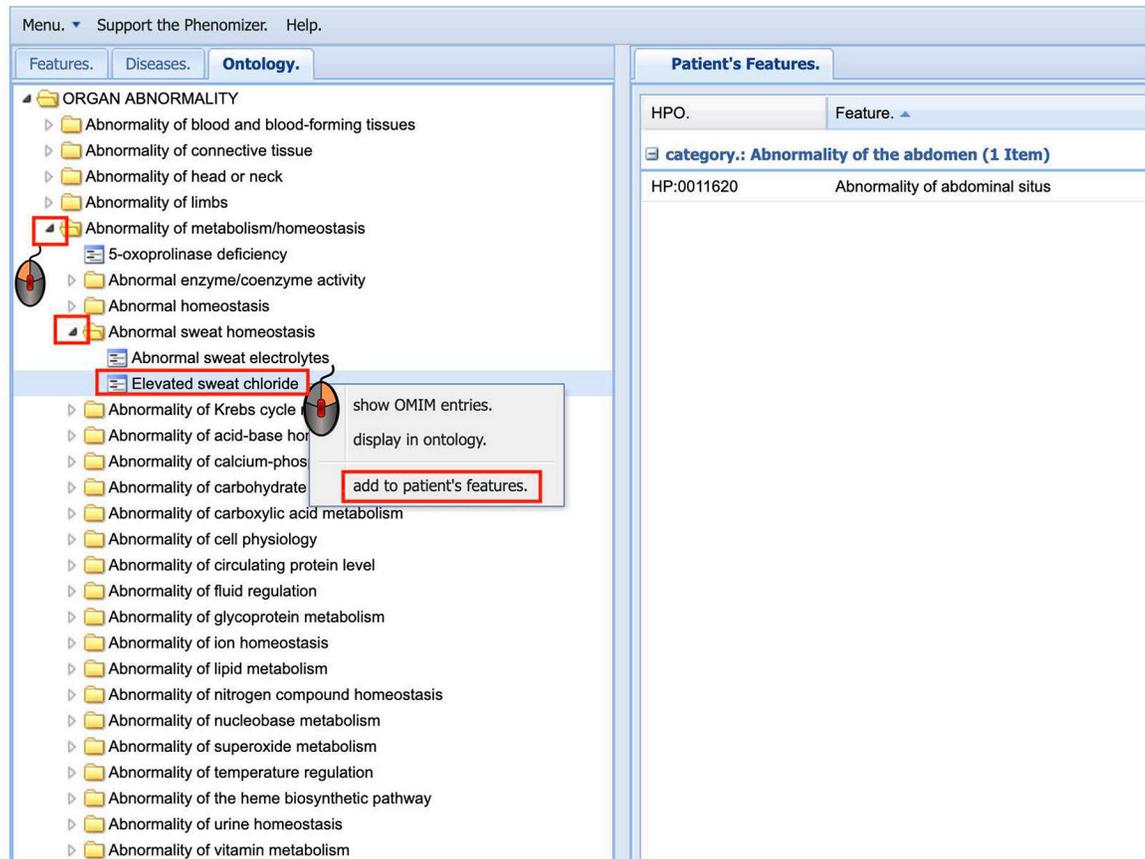
**Figure 5. Phenomizer context menu.**

Right clicking on an HPO term causes a context menu to appear. The subcommand "show OMIM entries" causes a list of diseases annotated to the HPO term to appear. The OMIM disease identifiers are also shown. The subcommand "display in ontology" displays the subset of the HPO from term up to the root of the ontology. Finally, the subcommand "add to patient's features" causes the HPO term to be appended to the list of HPO terms the user is building.

**Figure 6.**
Results of Phenomizer analysis are sorted by p-value.
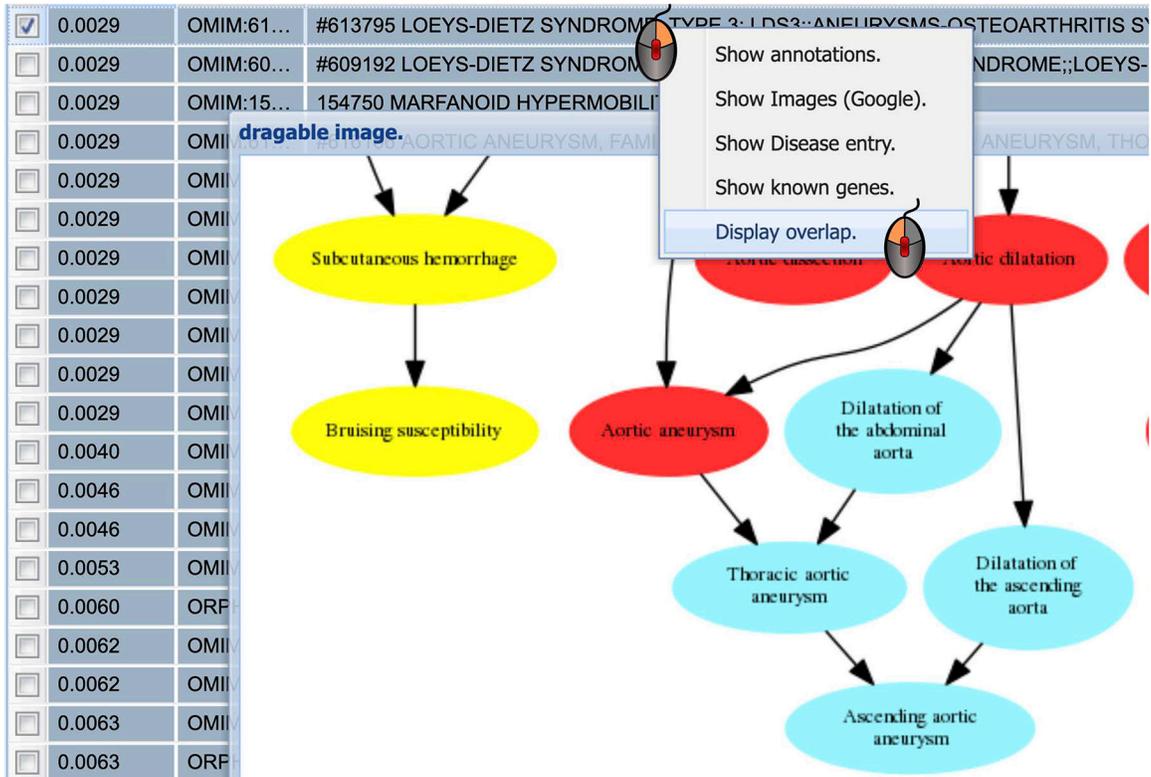
**Figure 7.**
Each syndrome in the result list has a context menu that allows you to display the overlap between the proband's HPO terms and the syndrome's features. The query terms are displayed in blue, the terms belonging to the disease in yellow, and terms shared by query and disease in red. In this example, *Aortic aneurysm* and the other red terms are shared by the disease and the query. *Bruising susceptibility* (HP:0000978) and the other yellow terms are used to annotate the disease but were not among the query terms. The blue term *Ascending aortic aneurysm* was used in the query but is not used to annotate the disease. Note that the overlap image is draggable.

**Figure 8.**
Using PhenoTips for deep phenotyping using the HPO.

**Figure 9.**
Using PhenoTips to draw a pedigree and record family history.

**Figure 10.**
Using PhenoTips to generate real-time gene suggestions based on the patient's phenotypic profile.

**Figure 11.**
Using PatientArchive to create a HPO phenotype profile from a free text clinical note.
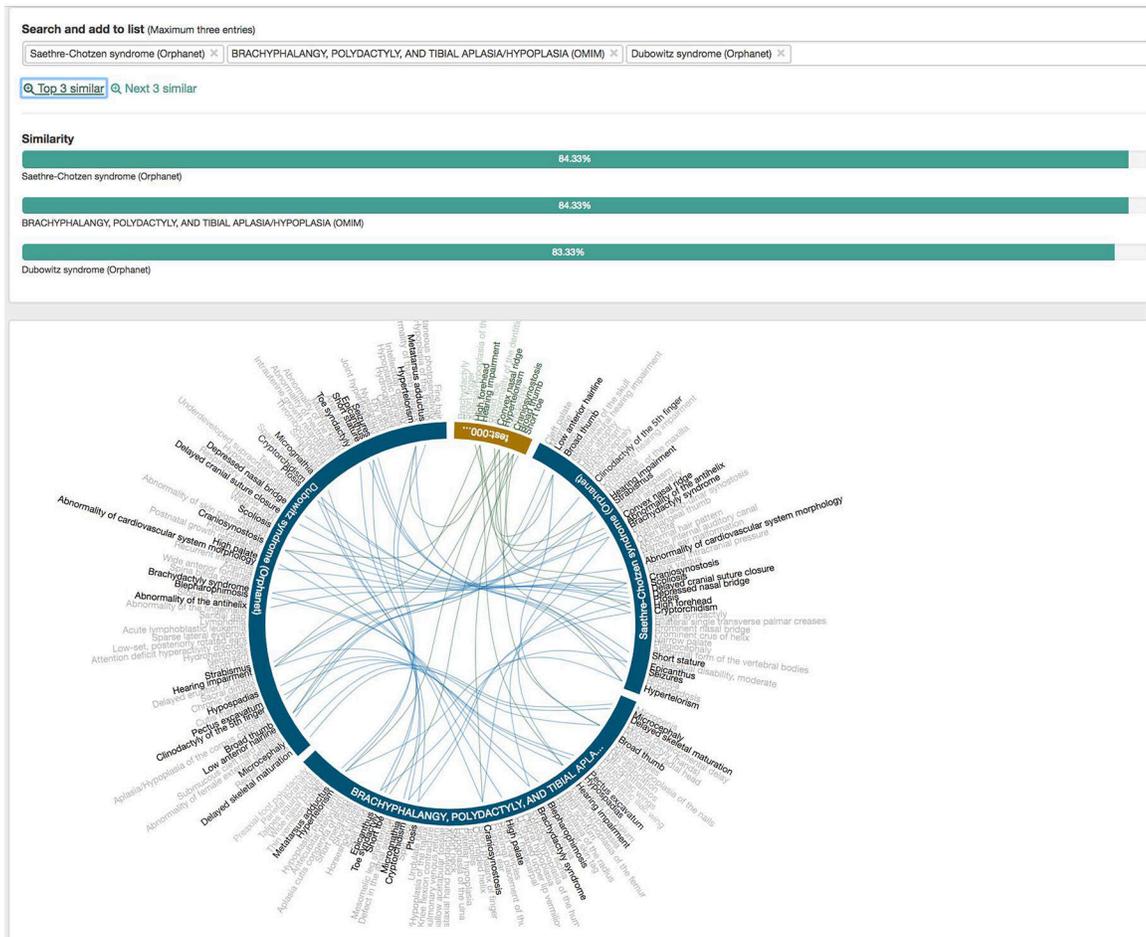
**Figure 12.**
Using PatientArchive to explore the top 3 candidate disorders associated with the patient phenotype profile.

**Figure 13.**
Exomiser output page.