HPO Workbench Documentation

Release 1

Monarch Initiative

Jun 06, 2023

Contents:

CHAPTER 1

HPOannotQC

1.1 Installing and running HPO Annotation Q/C

This version of HpoAnnotQc uses phenol-1.3.2-SNAPSHOT, which needs to be installed locally with `mvn install`. Following this, to build HPO Annotation Q/C, clone the GitHub repository at https://github.com/monarch-initiative/hpoannotqc, and build HPO Workbench using maven.

```
$ git clone https://github.com/monarch-initiative/hpoannotqc.git
$ cd hpoannotqc
$ mvn clean package
```

This will create an executable jar file.

```
$ java -jar target/HpoAnnotQc.jar
    Usage: java -jar HpoAnnotQc.jar [-hV] [COMMAND]
    Variant-motif visualization tool.
    -h, --help      Show this help message and exit.
    -V, --version   Print version information and exit.
   Commands:
    download, D  download files
    big-file, B  Create phenotype.hpoa file
    gene2phen    Create genes to phenotypes file
```

We will update this as soon as phenol-1.3.2 is released in maven central.

1.2 HPO Annotation Small Files

Each annotated disease in the HPO corpus is represented in a single so-called small file.

1.2.1 Small file format

The small files have tab-separated value format, i.e., TSVs. Please note that the format is different from our main release file (the "big file", phenotype.hpoa), which is created by combining data from the small files. There are 14 fields.

Column	Item	Comment
1	diseaseID	OMIM:600269, DECIPHER:81
2	diseaseName	e.g., Neurofibromatosis type 1
3	phenotypeID	e.g., HP:0000123
4	phenotypeName	e.g., Scoliosis
5	onsetID	e.g., HP:0003581
6	onsetName	e.g., Adult onset
7	frequency	e.g., HP:0040280 or 3/7 or 24%
8	sex	Male, Female
9	negation	NOT or not
10	modifier	semicolon sep list HPO terms
11	description	free text
12	publication	e.g., PMID:123321
13	evidence	PCS, IEA, ICE, or TAS
14	biocuration	HPO:skoehler[YYYY-MM-DD]

1. **diseaseID**. This field is a string that must be one of "OMIM:id", "ORPHA:id", or "DECIPHER:id". The id portion of the name is the code given by the database, e.g., `OMIM:157000`. Additional source databases may be admitted in the future.

2. **diseaseName**. This field is a String that represents the label (name) of the disease in question, e.g., `Marfan syndrome`.

3. **phenotypeID**. This must be a valid HP id. It must be the primary id (not an alt_id) for the current version of the HPO; if not, an error must be generated by the Q/C code; the Q/C code should allow the HPO id's and the labels of affected annotations to be updated after manual inspection by the user.

4. **phenotypeName**. The label of the HPO term refered to by the phenotypeId field, e.g., `Arachnodactyly`.

5. **onsetID**. The age of onset ID, being an HPO id of a term from the `Onset` subhierarchy of the HPO. This must be the primary id (not the alt_id). This field can be left empty, in which case, the ageOfOnsetName field must also be empty.

6. **onsetName**. The label corresponding to the ageOfOnsetId. This field can be left empty, in which case, the ageOfOnsetId field must also be empty.

7. **frequency**. This column can be one of three formats: A valid HPO term from the frequency subontology, a fractional expression m/n (e.g., 4/7 meaning that 4 of 7 individuals in the cited study had the disease and the feature in question, while the feature was ruled out in the remaining 3 of 7 individuals); or a percentage value such as 47%. This column may be empty.

8. **sex**. This column may be empty or may contain the strings "MALE" or "FEMALE".

9. **negation**. This column may be empty or may contain the string "NOT"

10. **modifier**. This column may be empty of contain HPO term ids for one or more terms from the Clinical Modifier subontology. Multiple terms are to be separated by semicolons.

11. **description**. Free text. This column must not be used to store modifiers.

12. **publication**. The publication reference for the annotation assertion. Must be present and must be one of PMID:123, OMIM:123 or ?. Note: pimd:123 is not accepted. The following prefixes are allowed:

- PMID
- OMIM
- http
- ISBN
- DECIPHER

13. **evidence**. One of the three HPO evidence codes.

- IEA
- TAS
- PCS

14. **biocuration**. This field must begin with a valid reference of the form prefix:id. This can be something like `ORCID:0000-0000-0000-0123` or a database id followed by a name (usually first initial-lastname), e.g., `HPO:mmustermann`.

This field contains the date when the term was first created and must have the form yyyy-mm-dd, e.g., 2016-07-22. Multiple biocurations are separated by a semicolon, e.g., `HPO:skoehler[2013-06-25]; HPO:probinson[2015-12-06]`.

1.3 HPO Annotation File Formats

The HPO annotation files are created by editing one file per disease entries (which we will call "small" files here for brevity). These files were merged into a single file that has been called `phenotype_annotation.tab`. Starting in 2018, the HPO team is migrating to a new big file format called `phenotype.hpoa`. In this document, we will describe the format of `phenotype.hpoa`.

1.3.1 phenotype.hpoa format

The first few lines present metadata (comments) preceeded by hash signs (#) at the beginning of the lines. The very next line is a header with the names of the columns.

```
#description: HPO annotations for rare diseases [7377: OMIM; 47: DECIPHER; 3300␣
↪ORPHANET]
#date: 2019-01-03
#tracker: https://github.com/obophenotype/human-phenotype-ontology
#HPO-version: http://purl.obolibrary.org/obo/hp/releases/2018-12-21/hp.owl
DatabaseID      DiseaseName     Qualifier       HPO_ID  Reference       Evidence      ␣
↪  Onset   Frequency       Sex     Modifier        Aspect  Biocuration
```

Nr	Content	Required | Example	
1	DatabaseId	Yes | MIM:154700	
2	DB_Name	Yes	Achondrogenesis, type IB
3	Qualifier	No	NOT
4	HPO_ID	Yes	HP:0002487
5	DB_Reference	Yes	OMIM:154700 or PMID:15517394
6	Evidence	Yes	IEA
7	Onset	No	HP:0003577
8	Frequency	No	HP:0003577 or 12/45 or 22%
9	Sex	No	MALE or FEMALE
10	Modifier	No	HP:0025257 (";"-separated list)
11	Aspect	Yes	"P" or "C" or "I" or "M"
12	BiocurationBy	Yes	HPO:skoehler[YYYY-MM-DD]

Explanations

1. **DatabaseId**: This field refers to the database from which the identifier in DB_Object_ID (column 2) is drawn. At present, annotations from the OMIM, ORHPANET, DECIPHER, and the HPO team are available. This field must be formated as a valid CURIE, e.g., OMIM:1547800,DECIPHER:22, ORPHANET:5431

2. **DB_Name**: This is the name of the disease associated with the DB_Object_ID in the database. Only the accepted name should be used, synonyms should not be listed here.

3. **Qualifier**: This optional field can be used to qualify the annotation shown in field 5. The field can only be used to record "NOT" or is empty. A value of NOT indicates that the disease in question is not characterized by the indicated HPO term. This is used to record phenotypic features that can be of special differential diagnostic utility.

4. **HPO_ID**: This field is for the HPO identifier for the term attributed to the DB_Object_ID. This field is mandatory, cardinality 1.

5. **DB_Reference**: This required field indicates the source of the information used for the annotation. This may be the clinical experience of the annotator or may be taken from an article as indicated by a pubmed id. Each collaborating center of the Human Phenotype Ontology consortium is assigned a HPO:Ref id. In addition, if appropriate, a pubmed id for an article describing the clinical abnormality may be used.

6. **Evidence**: This required field indicates the level of evidence supporting the annotation. The HPO project currently uses three evidence codes.

- **IEA** (inferred from electronic annotation): Annotations extracted by parsing the Clinical Features sections of the Online Mendelian Inheritance in Man resource are assigned the evidence code "IEA".
- **PCS** (published clinical study) is used for used for information extracted from articles in the medical literature. Generally, annotations of this type will include the pubmed id of the published study in the `DB_Reference` field.
- **TAS** (traceable author statement) is used for information gleaned from knowledge bases such as OMIM or Orphanet that have derived the information frm a published source..

7. **Onset**: A term-id from the HPO-sub-ontology below the term "Age of onset" (HP:0003674). Note that if an HPO onset term is used in this field, it refers to the onset of the feature specified in field 4 in the disease being annotated. On the other hand, if an HPO onset term is used in field 4, then it refers to the overall onset of the disease. In this case, no additional onset term should be used in field 8.

8. **Frequency**: There are three allowed options for this field. **(A)** A term-id from the HPO-sub-ontology below the term "Frequency" (HP:0040279). (since December 2016 ; before was a mixture of values). The terms for frequency are in alignment with Orphanet. * **(B)** A count of patients affected within a cohort. For instance, 7/13 would indicate that 7 of the 13 patients with the specified disease were found to have the phenotypic abnormality referred to by the HPO term in question in the study refered to by the DB_Reference; **(C)** A percentage value such as 17%.

9. **Sex**: This field contains the strings MALE or FEMALE if the annotation in question is limited to males or females. This field refers to the phenotypic (and not the chromosomal) sex, and does not intend to capture the further complexities of sex determination. If a phenotype is limited to one or the other sex, then the corresponding term for the "Clinical modifier" subontology should also be used in the Modifier field.

10. **Modifier**: A term-id from the HPO-sub-ontology below the term "Clinical modifier".

11. **Aspect**: one of P (Phenotypic abnormality), I (inheritance), C (onset and clinical course), M (clinical modifier). This field is mandatory; cardinality 1.

- Terms with the `P` aspect are located in the Phenotypic abnormality subontology.
- Terms with the `I` aspect are from the Inheritance subontology.
- Terms with the `C` aspect are located in the Clinical course subontology, which includes onset, mortality, and other terms related to the temporal aspects of disease.
- Terms with the `M` aspect are located in the Clinical Modifier subontology.

12. **BiocurationBy**: This refers to the biocurator who made the annotation and the date on which the annotation was made; the date format is `YYYY-MM-DD`. The first entry in this field refers to the creation date. Any additional biocuration is recorded following a semicolon. So, if Joseph curated on July 5, 2012, and Suzanna curated on December 7, 2015, one might have a field like this: `HPO:Joseph[2012-07-05];HPO:Suzanna[2015-12-07]`. It is acceptable to use ORCID ids. This field is mandatory, cardinality 1

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.