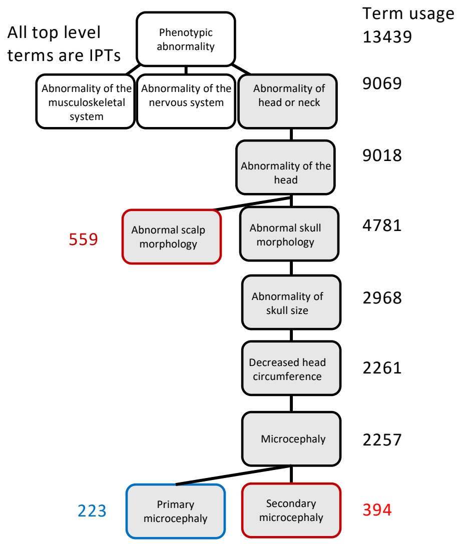
Tutorial on creating sets of Informative Phenotypic Terms

Finding IPTs in the DDD database

Informative Phenotypic Terms (IPTs) were derived from a binary matrix of terms annotated to all DDD individuals (individuals in rows, HPO terms in columns) and an OBO format version of the HPO ontology (that of 02/08/2021). An extended matrix of propagated annotations was then created (direct annotations were propagated to parent terms using the ontology structure). The usage of HPO terms was then found for all HPO terms making no distinction between direct and inferred use.

Beginning with top level phenotypic terms (those under “Phenotypic abnormality” HP:0000118), all such terms were included as IPTs, and each was descended in turn to select subterms meeting the criterion of use in the extended annotation matrix (above 250 and below 1500 uses in the entire DDD dataset of 13439 individuals). Subterms with usage above the upper threshold were further expanded, and the criterion applied at that level. As there can be multiple paths to a term from the top level, an IPT can be found under multiple top level terms – these were detected and the IPT retained under a single top level term (in a list of lists data structure). The final step was to detect IPTs that, through the ontology structure, had parent terms that were also IPTs. To remove the resulting correlation in annotation of the parent to the child, the extended annotation matrix was modified to remove the annotation to the parent term for individuals with an annotation to the child IPT in question. This modified matrix was used to compute term frequencies in gene models.


All top level terms are IPTs
Phenotypic abnormality
Term usage
13439
Abnormality of the musculoskeletal system
Abnormality of the nervous system
Abnormality of head or neck
9069
Abnormality of the head
9018
559
Abnormal scalp morphology
Abnormal skull morphology
4781
Abnormality of skull size
2968
Decreased head circumference
2261
Microcephaly
2257
223
Primary microcephaly
Secondary microcephaly
394

HPO terms were expanded (black) or retained as IPTs (red) according to usage.

The procedure does not depend on any specific version of HPO. The version of HPO specified can be found here: https://bioportal.bioontology.org/ontologies/HP

Running the resource creation scripts

The following steps are run once for a database of HPO annotations made to individuals, generating a matrix that is used by the HPO classifier. This procedure does not require individuals to have diagnoses (diagnoses are considered in model learning and classification).

Implementing the steps described above, the IMPROVE_resource.r code begins by loading an OBO format of the HPO ontology using the ontoCAT R library (line 27).

In place of the DDD phenotype data (which is available on request), a matrix of individual annotations *hpo* is created from the data in *pheno* (line 67) after finding all terms used in annotation *allHPO*. The following steps can be run on data from another database providing the individual to HPO annotation is expressed as a binary matrix (individuals in rows, HPO terms in columns).

The expanded matrix *hpoe* is created from *hpo* using the *expandToParents()* method of the ontoCAT library (line 101). The entire set of expanded terms (*allHPOExpanded*) is found in a preliminary step in order to define *hpoe*.

The data structure *informativeSpecificTerms* is initialised from terms under HP:0000118 (“Phenotypic abnormality”) and each top level term descended to identify terms meeting usage < *upper_annotation_threshold* and >= *retain_threshold* (lines 126 onwards). In the DDD data, we found 10% and 2% of total annotations were suitable thresholds.

As there can be multiple paths to an IPT, multiple instances may occur in *informativeSpecificTerms*. These are detected by tabluating terms and removing second and third occurrences to create *informativePhenotypicTerms*. This was performed intially after manual inspection (line 196) but is now automated (line 154).

The method *getDuplicatesAndParents()* returns information on any child-parent edges within *informativePhenotypicTerms* (line 255). A modified annotation matrix *hpou* is created from *hpoe* by removing parent annotation in individuals where there is an annotation to the child (lines 270). The *hpou* matrix is the resource used in the classifier. The HPO itself is not needed in classification