**Module 2. Retrieving the disease phenotypes from HPO**

/\* First, merge the OMIM-encoded D with the HPO-encoded DP. Assemble the *2*-tuples composed of a D (as MIM id) and the D-associated DP (as HPO id). Only DP that are descendants of HP:0000118 (Phenotypic abnormality) are used, while all the descendants of the other four children of HP:0000001 (All), i.e., HP:0000005 (Mode of inheritance), HP:0012823 (Clinical modifier), HP:0040006 (Mortality/Aging) and HP:0040279 (Frequency), have been discarded.

The annotation data have been obtained from the HPO at the following web site: <http://compbio.charite.de/jenkins/job/hpo.annotations/lastStableBuild/artifact/misc/phenotype_annotation.tab> and then slightly modified (as mim2hpo.txt). \*/

.import c:/sqlite/LHPS/inputs/**mim2hpo.txt** tmp\_mim2hpo

CREATE TABLE **mim2hpo**(mim\_id INT, hpo\_id TEXT, entrez INT);

INSERT INTO mim2hpo

SELECT DISTINCT \* FROM(SELECT mim\_id, hpo\_id, entrez FROM tmp\_mim2hpo);

/\* Assemble the 2-tuples composed of a DP and its associated Informational Content (IC). Data are from file IC\_HPO.txt (from A. Gamba) and have been sligthly modified as hpo2ic.txt. The IC values are normalized in the range [0.0, 1.0]. \*/

.import c:/sqlite/LHPS/inputs/**hpo2ic.txt** tmp\_hpo2ic

CREATE TABLE **hpo2ic**(hpo\_id TEXT, ic REAL);

INSERT INTO hpo2ic SELECT DISTINCT \* FROM(SELECT hpo\_id, ic FROM tmp\_hpo2ic);

/\* Merge the two tables above (mim2hpo and hpo2ic) into the **mim2hpo2ic** table. Also, add the names of the phenotypes (from the hpo\_dictionary, originally hpo\_names.txt). The HPO dictionary was obtained by elaborating the HPO ontology (<http://www.obofoundry.org/ontology/hp.html>) as hpo.obo file. \*/

.import c:/sqlite/LHPS/inputs/**hpo\_dictionary.txt** hpo\_dictionary

CREATE TABLE **mim2hpo2ic**(mim\_id INT, hpo\_id TEXT, ic REAL, hpo\_name TEXT);

INSERT INTO mim2hpo2ic

SELECT DISTINCT \* FROM(

SELECT l.mim\_id, l.hpo\_id, ROUND(r1.ic,3), r2.hpo\_name

FROM mim2hpo l

LEFT JOIN hpo2ic r1 ON l.hpo\_id = r1.hpo\_id

LEFT JOIN hpo\_dictionary r2 ON l.hpo\_id = r2.hpo\_id

);

/\* Second, merge the DP annotations of the D with the PS annotations of the D. Before that, generate the *2*-tuples composed of a PS (as PS id) and the PS-associated D (as MIM id). \*/

CREATE TABLE **ps2mim**(ps\_id TEXT, ps\_name TEXT, mim\_id INT, mim\_name TEXT, entrez INT, symbol TEXT);

INSERT INTO ps2mim

SELECT DISTINCT \* FROM(

SELECT ps\_id, ps\_name, mim\_id, mim\_name\_morbidmap, entrez, symbol FROM ps\_dictionary

);

/\*Finally, assemble the 3-tuples composed of a PS, a D and a DP (with its IC). \*/

CREATE TABLE **ps2mim2hpo**(ps\_id TEXT, mim\_id INT, hpo\_id TEXT, ic REAL, ps\_name TEXT, hpo\_name TEXT);

INSERT INTO ps2mim2hpo

SELECT DISTINCT \* FROM(

SELECT l.ps\_id AS ps\_id, l.mim\_id AS mim\_id, r.hpo\_id, r.ic AS ic, l.ps\_name, r.hpo\_name

FROM ps2mim l

LEFT JOIN mim2hpo2ic r ON l.mim\_id = r.mim\_id

)

ORDER BY ps\_id ASC, mim\_id ASC, ic DESC;

/\* Delete the rows lacking HPO annotations of the D. \*/

DELETE FROM ps2mim2hpo WHERE hpo\_id IS NULL;

.once c:/SQLITE/LHPS/outputs/**ps2mim2hpo.txt**

SELECT \* FROM ps2mim2hpo;

DROP TABLE tmp\_mim2hpo; DROP TABLE tmp\_hpo2ic;