/\* **1.The OMIM-derived genetic diseases.** \*/

/\* The first aim is to identify all the OMIM-encoded diseases with known molecular basis. To this purpose, retrieve the morbidmap.txt file from the OMIM website. \*/

/\* Import from OMIM the morbidmap.txt file and select the columns of interest (i.e., the diseases as phenotype names and the gene as MIM numbers). \*/

.open c:/sqlite/PC/pc.db

.mode tabs

.headers ON

.import c:/SQLITE/PC/inputs/morbidmap.txt tmp1\_morbidmap

CREATE TABLE tmp2\_morbidmap(phenotype TEXT, dgp\_mim INTEGER);

INSERT INTO tmp2\_morbidmap

SELECT DISTINCT \* FROM(

SELECT Phenotype AS phenotype, MIM\_number AS dgp\_mim FROM tmp1\_morbidmap

);

/\* From each phenotype, extract the starting character and the OMIM key, to retain only the real diseases (character different from ‘?’, ‘[‘ or ‘{‘) with known molecular phenotype (key = 3). \*/

ALTER TABLE tmp2\_morbidmap ADD COLUMN start TEXT;

UPDATE tmp2\_morbidmap SET start = SUBSTR(Phenotype,1,1);

ALTER TABLE tmp2\_morbidmap ADD COLUMN key TEXT;

UPDATE tmp2\_morbidmap SET key = SUBSTR(Phenotype,-3,3);

/\* Delete phenotypes related to susceptibility (start ‘{‘), non-diseases (start ‘[‘) and unconfirmed diseases (start ‘?‘). Then, from the remaining phenotypes, delete the phenotypes with unknown gene defect (key ‘(1)‘), unknown mutation (key ‘(2)‘) or caused by multiple genes (key ‘(4)‘). Retain only the phenotypes with known molecular basis (key ‘(3)’). \*/

DELETE FROM tmp2\_morbidmap WHERE(start = ‘{’ OR start = ‘[’ OR start = ‘?’);

DELETE FROM tmp2\_morbidmap WHERE key = ‘(1)’ OR key = ‘(2)’ OR key = ‘(4)’;

/\* To convert the DGP provided by OMIM (‘mim’) into the corresponding Entrez identifiers, import the mim2gene.txt file from OMIM. Also, import the (daily updated ftp version of) Entrez Gene, to get the symbols and the type of gene. \*/

.import c:/sqlite/PC/inputs/mim2gene.txt tmp\_mim2gene

.import c:/sqlite/PC/inputs/entrez\_genes.txt entrez

CREATE TABLE **mim2gene**(dgp\_mim INTEGER, type TEXT, entrez\_id INTEGER, symbol TEXT);

INSERT INTO mim2gene

SELECT DISTINCT \* FROM(

SELECT Mim\_Number, Mim\_Type, Entrez, Symbol FROM tmp\_mim2gene

);

/\* From mim2gene, delete the rows where the phenotype has been either ‘moved’ or ‘removed’, as well as the rows where no indication of the dgp as entrez\_id is provided. \*/

DELETE FROM mim2gene WHERE type = ‘moved/removed’;

DELETE FROM mim2gene WHERE entrez\_id = ‘’;

/\*.Join the diseases (from tmp2\_morbidmap) with the DGP (from mim2gene) and the Entrez symbols (from entrz), to retrieve the diseases that are mapped to a DGP. \*/

CREATE TABLE tmp3\_morbidmap(phenotype TEXT, dgp\_mim INTEGER, mim\_type TEXT, entrez\_id INTEGER, symbol TEXT, type\_gene TEXT);

INSERT INTO tmp3\_morbidmap

SELECT DISTINCT \* FROM(

SELECT a.phenotype, a.dgp\_mim, b.type, b.entrez\_id, c.Symbol, c.type\_of\_gene

FROM tmp2\_morbidmap a

JOIN mim2gene b ON a.dgp\_mim = b.dgp\_mim

JOIN entrez c ON b.entrez\_id = c.GeneID

);

/\* Drop the phenotypes, in which the gene is NOT a protein-coding gene (e.g., ncRNA, snRNA, snoRNA, pseudo, other or unknown). \*/

DELETE FROM tmp3\_morbidmap WHERE type\_gene != ‘protein-coding’;

/\* Note that entrez\_gene 6473 corresponds to two sex-related symbols, namely SHOX (dgp\_mim 312865) and SHOXY (dgp\_mim 400020). \*/

UPDATE tmp3\_morbidmap SET symbol = ‘SHOXY’ WHERE (dgp\_mim = 400020 AND entrez\_id = 6473);

/\* Retrieve the phenotype identifier, where available. To this purpose, check manually, make the necessary corrections and delete all the phenotypes, in which the rightmost characters are not 6-digit id. Proceed as follows. \*/

/\* Make the following correction. \*/

UPDATE tmp3\_morbidmap SET phenotype = ‘Paroxysmal extreme pain disorder, 167400 (3)’ WHERE phenotype = ‘Paroxysmal extreme pain disorder, 167400, (3)’;

/\* Remove the rightmost string ‘ (3)’. \*/

UPDATE tmp3\_morbidmap SET phenotype = SUBSTR(phenotype,1,LENGTH(phenotype)-4);

/\* Then, retrieve the phen\_id. Where phen\_id is available, the outcome will be a 6-digit integer. Otherwise, the outcome will be a non-integer value that will be filtered out when selecting phen\_id > 999999, i.e., a non-numeric value). \*/

ALTER TABLE tmp3\_morbidmap ADD COLUMN phen\_id INTEGER;

UPDATE tmp3\_morbidmap SET phen\_id = SUBSTR(phenotype,-6,6);

DELETE FROM tmp3\_morbidmap WHERE phen\_id > 999999;

/\* Table **omim\_def** is the final version of the morbid map complete with the dgp.\*/

CREATE TABLE **omim\_def**(phen\_mim INTEGER, phenotype TEXT, dgp\_mim INTEGER, entrez\_id INTEGER, symbol TEXT);

INSERT INTO omim\_def

SELECT DISTINCT \* FROM(

SELECT phen\_id, phenotype, dgp\_mim, entrez\_id, symbol FROM tmp3\_morbidmap

ORDER BY phen\_id ASC

);

/\* Export omim\_def as output in .txt \*/

.once c:/sqlite/PC/outputs/**omim\_def.txt**

SELECT \* FROM omim\_def;

DROP TABLE tmp1\_morbidmap; DROP TABLE tmp2\_morbidmap; DROP TABLE tmp3\_morbidmap;

DROP TABLE tmp\_mim2gene;