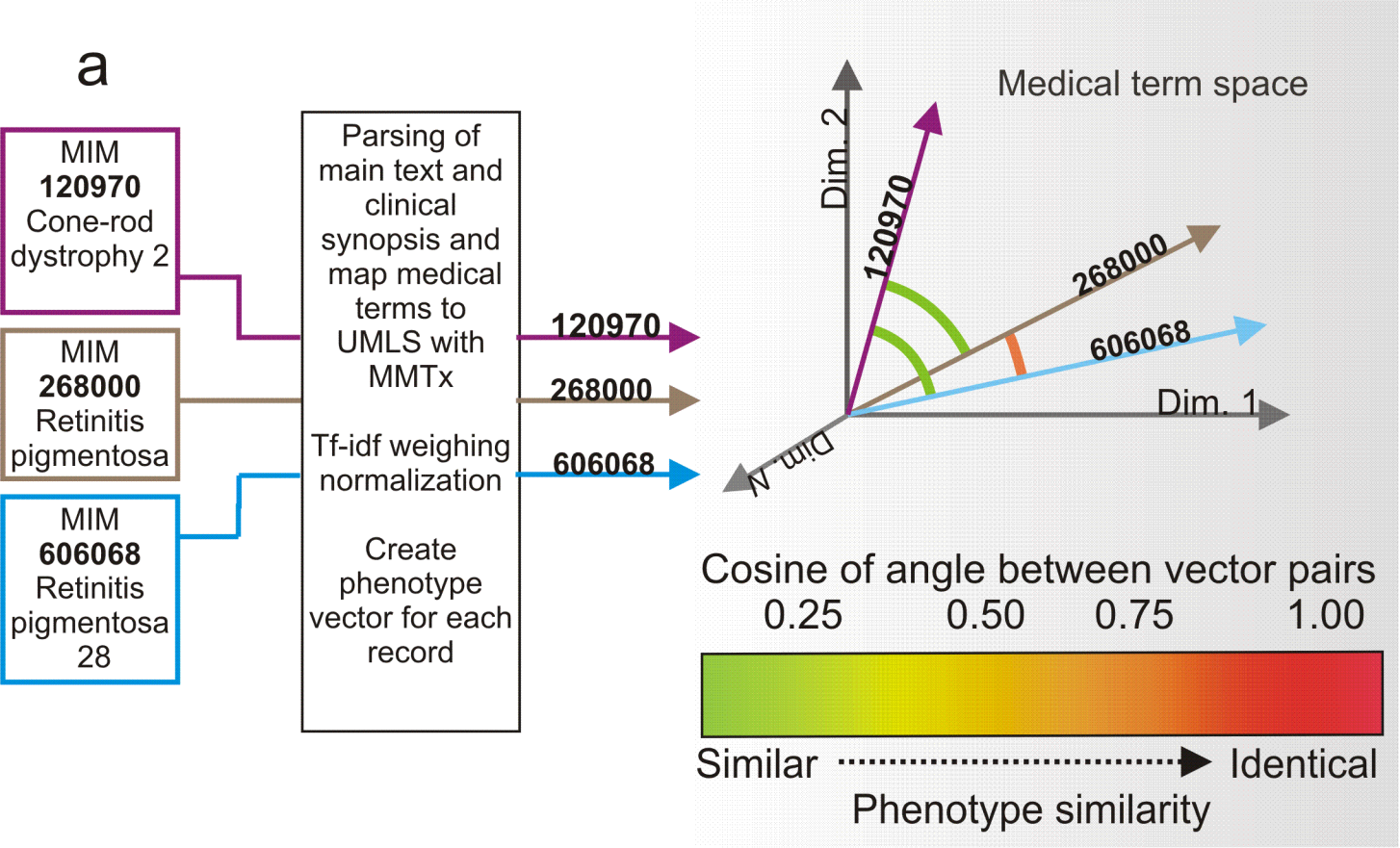
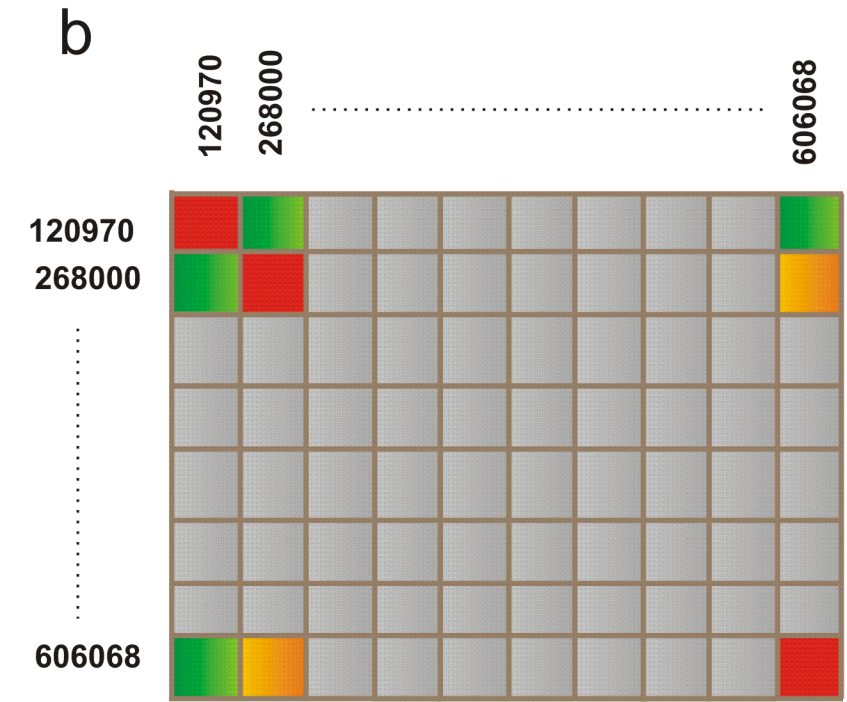

a
MIM
120970
Cone-rod
dystrophy 2
MIM
268000
Retinitis
pigmentosa
MIM
606068
Retinitis
pigmentosa
28
Parsing of
main text and
clinical
synopsis and
map medical
terms to
UMLS with
MMTx
Tf-idf weighing
normalization
Create
phenotype
vector for each
record
120970
268000
606068
Medical term space
Dim. 2
Dim. 1
Dim. N
120970
268000
606068
Cosine of angle between vector pairs
0.25
0.50
0.75
1.00
Similar
Identical
Phenotype similarity


b
120970
268000
606068
120970
268000
606068

Supplementary Figure 1 Measuring phenotype association scores between OMIM records. For every record we created a phenotype vector **a**). A phenotype vector consists of weighted medical terms present in the record, and represents the phenotype described in that particular record. For a single record a vector was made by parsing the clinical synopsis and main text of a record with MMTx. MMtx maps the words to medical terms in a subset of the UMLS vocabulary. Every term in the vector was weighted in relation to the record it was describing, and the weight normalizes the term scores in relation to the length of the records. Based on these medical terms a vector representing every OMIM record was then plotted in medical term space. The pairwise phenotypic overlap between records was quantified by the cosine of the angle between normalized vector pairs. Doing this for all combinations of vector pairs, enabled us to make a matrix of pairwise phenotypic similarity scores between all OMIM records **b**). For clarity we only show the procedure for three OMIM records (MIM 120970, MIM 268000, MIM 606068).