Guide for Downloading Data

Introduction

The eRAM provides enriched annotations for 15,942 rare diseases, yielding 6,147 human disease related phenotype terms, 31,661 mammalians phenotype terms, 10,202 symptoms from UMLS, 18,815 genes and 92,580 genotypes.

Public Data

Users can download these data through http://www.unimd.org/eram/download.php
The public data contains:

Disease-Gene associations

eRAM provides disease-gene associations from several existing databases including Orphanet, OMIM, UniProtKB (1-3), ClinVar (4), DISEASES (including text-mined data) (5) and DisGeNET (CTD data) (6,7), as well as disease-gene associations inferred by the disease comorbidity-based network approach (8) using data in ClinVar.

All this data can be easily get through eRAM Gene.zip.

Disease-Manifestation associations

eRAM provides disease-manifestation associations from the HPO, Orphanet and UMLS (MRREL.RRF, 2017AA release). We provide two files (eRAM Integrated Phenotype.txt and eRAM Integrated Symtoms.txt) for users to get these data.

Other Data

Besides the above data, eRAM also contains the following data:

Disease-Manifestation associations

eRAM provides disease-manifestation associations by text mining abstracts/full-text articles from MEDLINE.

Disease-Genotype associations

eRAM contains genotype information from the following resources:

1) Existing databases: DisGeNET, GWASdb (9), LOVD (10) and PharmGKB (11).
2) Data from Beijing Children's hospital.

Disease-Comorbidity associations

The comorbidity information in eRAM was mainly collected in two ways:

1) Extracted from electronic health records (EHRs). We extracted the disease comorbidity information from Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II) database, from which we collected 34,261 unique disease comorbidity pairs. This information has been presented in eRAM in the

comorbidity section.

2) Text-mined disease comorbidity information from MEDLINE.

Users can get the above data or any other information via email.

- ***Contact information:***

Jinmeng Jia (jiajm.ecnu@gmail.com)

Reference:

1. The UniProt, C. (2017) UniProt: the universal protein knowledgebase. *Nucleic acids research*, **45**, D158-D169.
2. Magrane, M. and UniProt, C. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database : the journal of biological databases and curation*, **2011**, bar009.
3. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A.J., Poux, S., Bougueleret, L. and Xenarios, I. (2016) UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Methods in molecular biology*, **1374**, 23-54.
4. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J. *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, **44**, D862-868.
5. Pletscher-Frankild, S., Palleja, A., Tsafou, K., Binder, J.X. and Jensen, L.J. (2015) DISEASES: text mining and data integration of disease-gene associations. *Methods*, **74**, 83-89.
6. Pinero, J., Bravo, A., Queralt-Rosinach, N., Gutierrez-Sacristan, A., Deu-Pons, J., Centeno, E., Garcia-Garcia, J., Sanz, F. and Furlong, L.I. (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, **45**, D833-D839.
7. Pinero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F. and Furlong, L.I. (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database : the journal of biological databases and curation*, **2015**, bav028.
8. Wu, X., Jiang, R., Zhang, M.Q. and Li, S. (2008) Network-based global inference of human disease genes. *Molecular systems biology*, **4**, 189.
9. Li, M.J., Liu, Z., Wang, P., Wong, M.P., Nelson, M.R., Kocher, J.P., Yeager, M., Sham, P.C., Chanock, S.J., Xia, Z. *et al.* (2016) GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic acids research*, **44**, D869-876.
10. Pan, M., Cong, P., Wang, Y., Lin, C., Yuan, Y., Dong, J., Banerjee, S., Zhang, T., Chen, Y., Zhang, T. *et al.* (2011) Novel LOVD databases for hereditary breast cancer and colorectal cancer genes in the Chinese population. *Hum Mutat*, **32**, 1335-1340.
11. Thorn, C.F., Klein, T.E. and Altman, R.B. (2013) PharmGKB: the Pharmacogenomics Knowledge Base. *Methods in molecular biology*, **1015**, 311-320.