Five years of DisGeNET: Lessons learned and challenges ahead

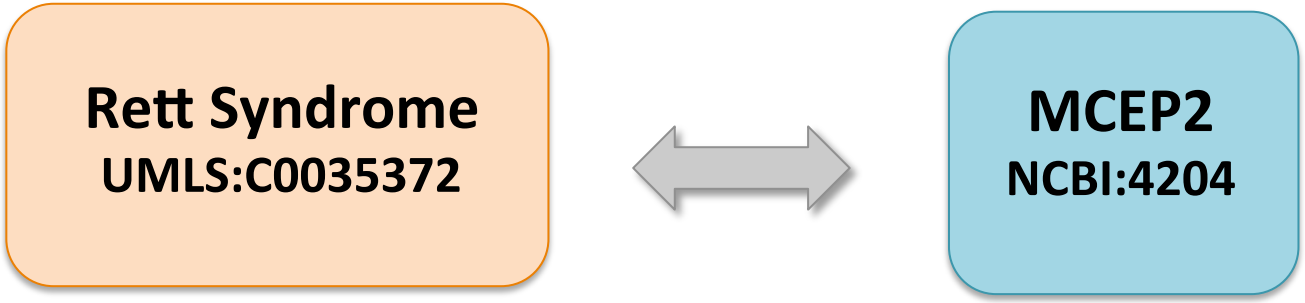
Laura I. Furlong
IMIM-UPF

Big Data in Biomedicine
Barcelona
November 11, 2014

GRIB RESEARCH PROGRAMME ON BIOMEDICAL INFORMATICS
upf. Universitat Pompeu Fabra Barcelona
IMIM Parc de Salut MAR Barcelona
Institut Hospital del Mar d'Investigacions Mèdiques

DisGeNET

- Knowledge platform on human diseases and their genes
- Aims to cover all disease therapeutic areas
- Developed by integration of data from expert-curated resources and from the literature by text mining
- Centered on gene-disease association (GDA) and its supporting evidence and provenance


Rett Syndrome
UMLS:C0035372
MCEP2
NCBI:4204

Rett Syndrome
UMLS:C0035372

MCEP2
NCBI:4204

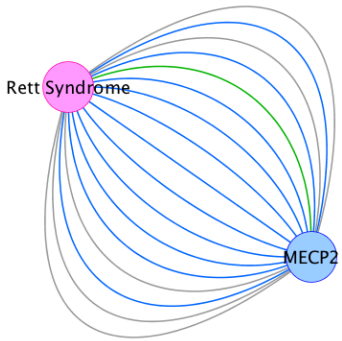
UniProt ✓

ctd ✓

MGI ✓

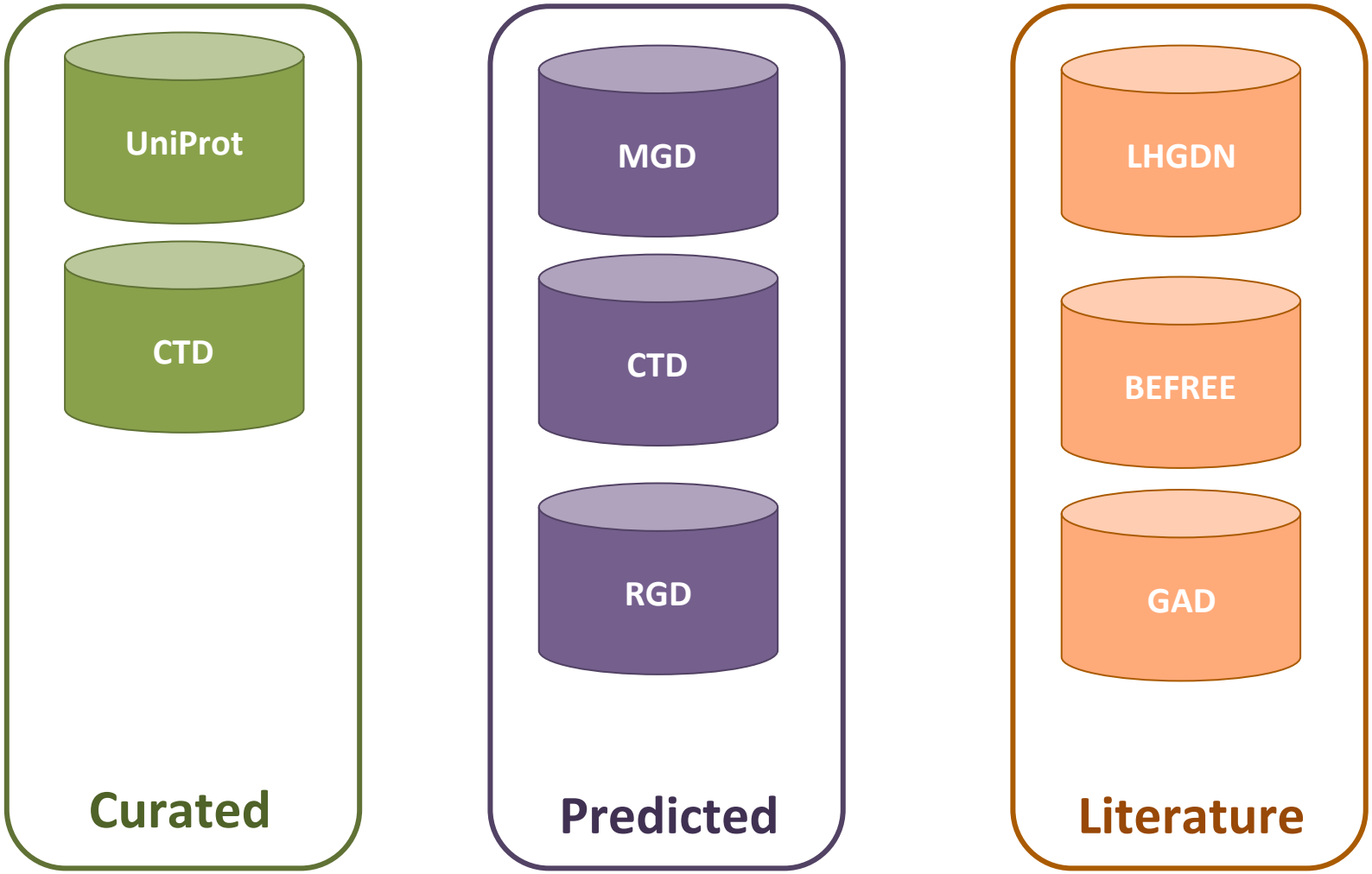
RGD X

MEDLINE ✓

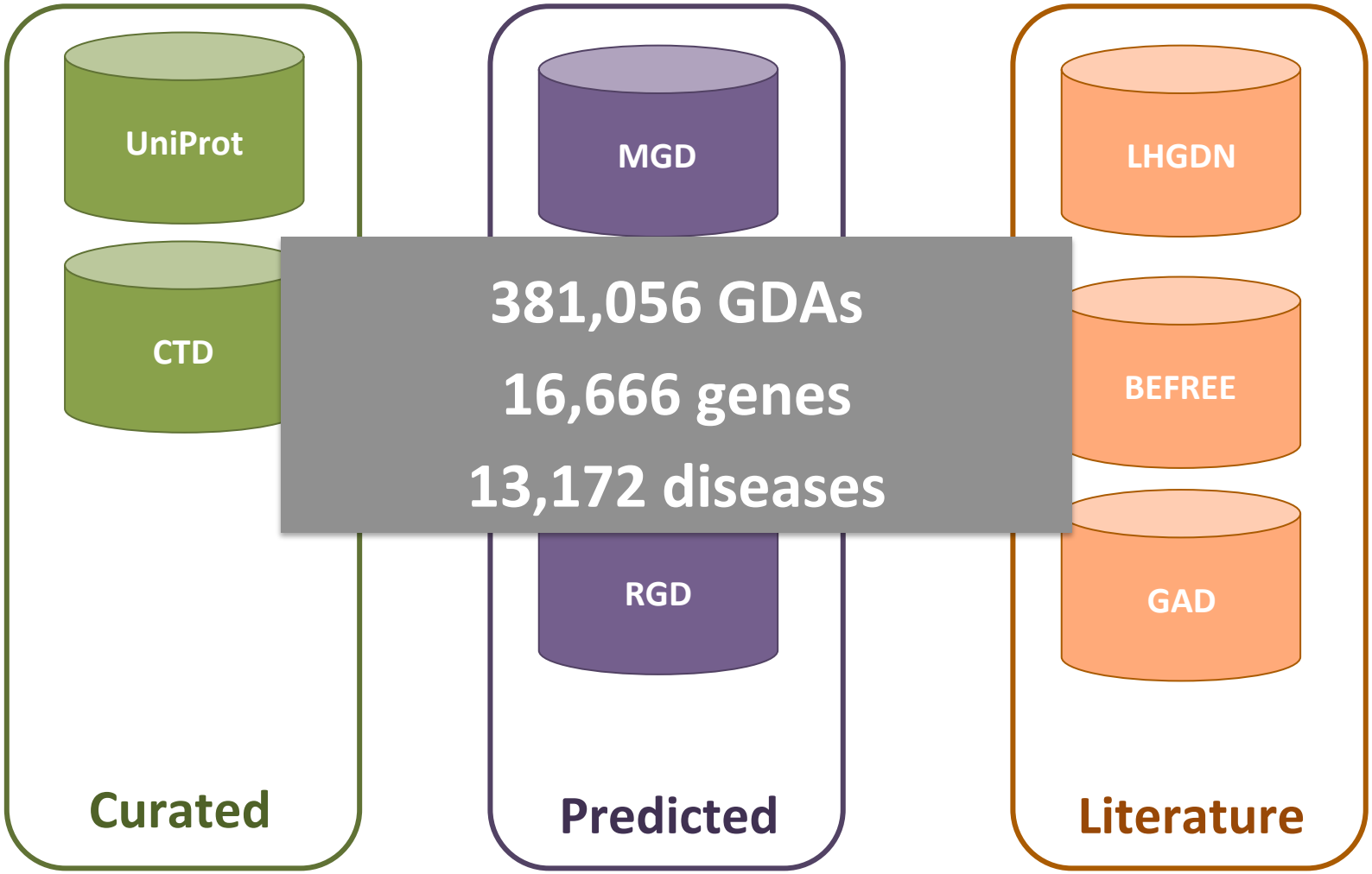


DisGeNET

DisGeNET

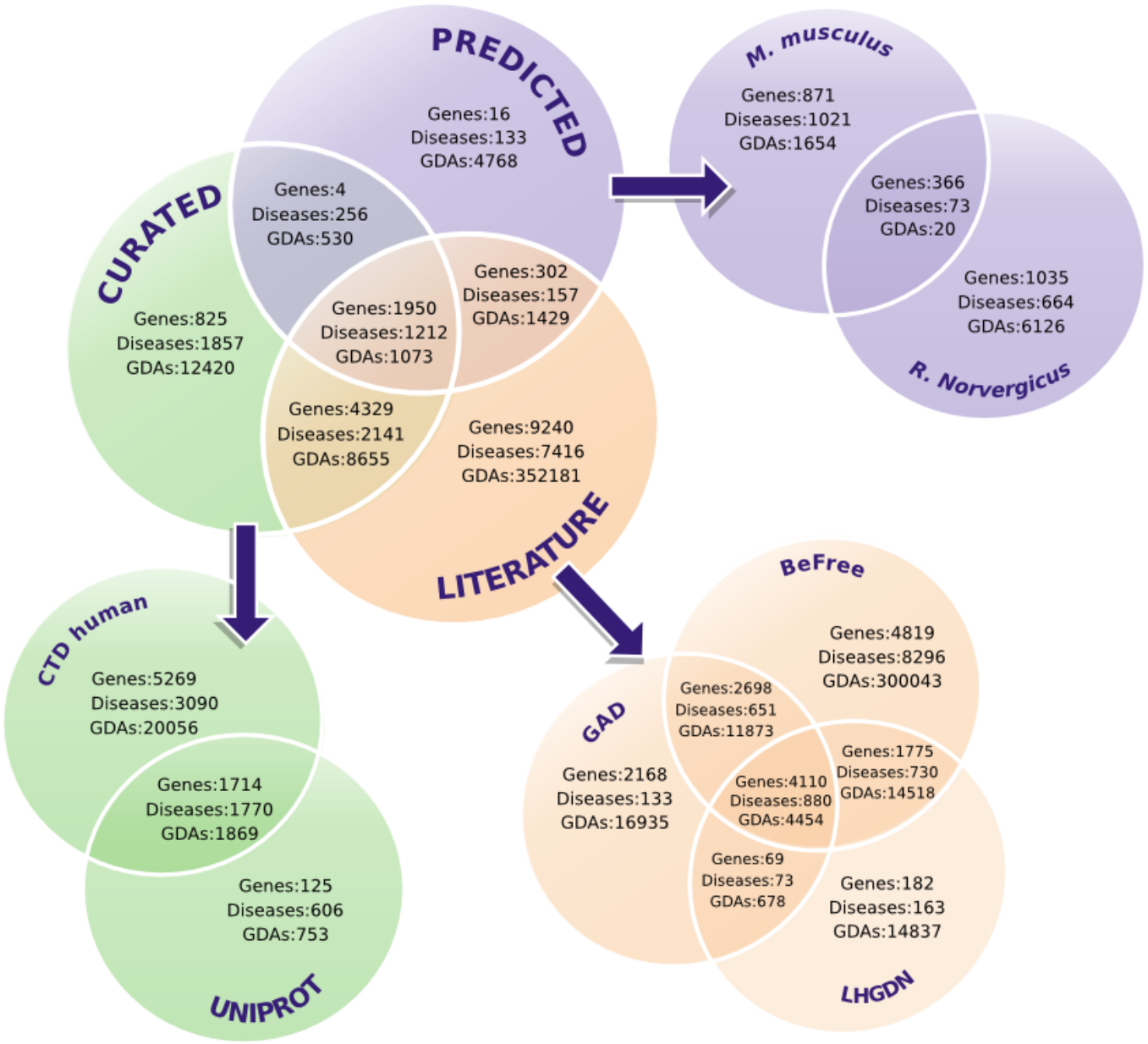

UniProt
CTD
Curated
MGD
CTD
RGD
Predicted
LHGDN
BEFREE
GAD
Literature

DisGeNET


UniProt
CTD
Curated
MGD
RGD
Predicted
LHGDN
BEFREE
GAD
Literature
381,056 GDAs
16,666 genes
13,172 diseases

Key points

1. **Fragmentation of information as a barrier to knowledge on the mechanisms of human diseases**

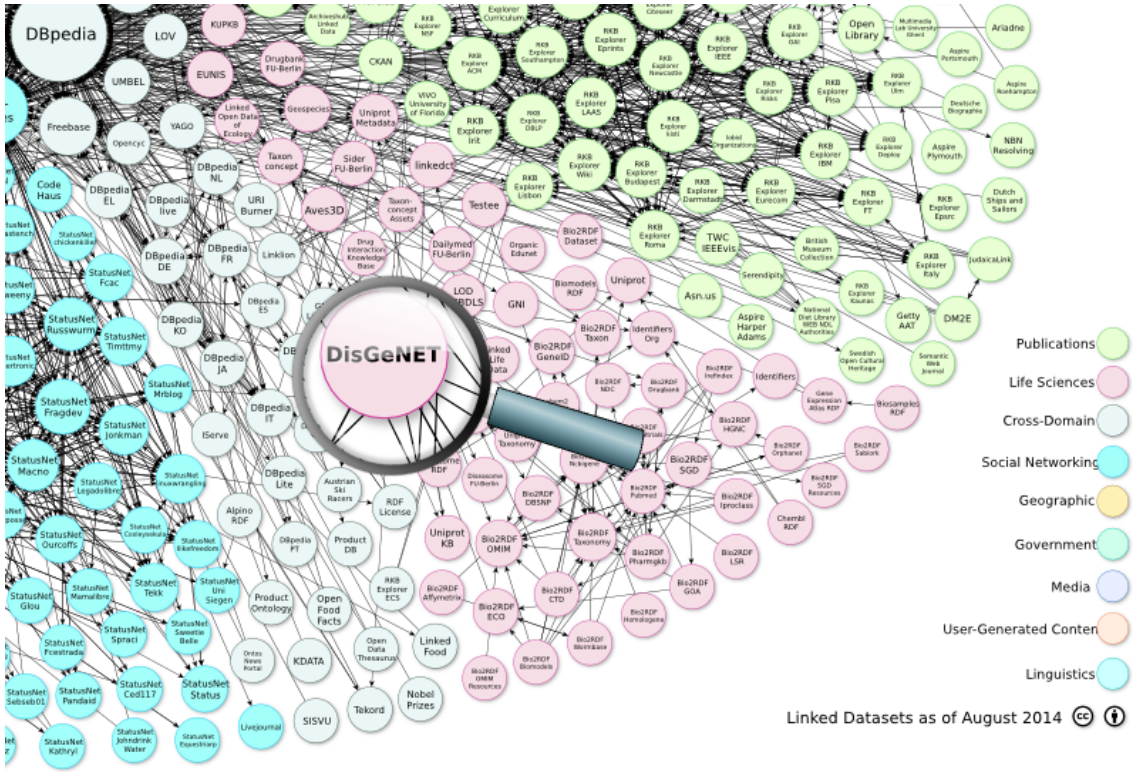

PREDICTED
Genes:16
Diseases:133
GDAs:4768
CURATED
Genes:825
Diseases:1857
GDAs:12420
Genes:4
Diseases:256
GDAs:530
Genes:1950
Diseases:1212
GDAs:1073
Genes:302
Diseases:157
GDAs:1429
Genes:4329
Diseases:2141
GDAs:8655
Genes:9240
Diseases:7416
GDAs:352181
LITERATURE
M. musculus
Genes:871
Diseases:1021
GDAs:1654
Genes:366
Diseases:73
GDAs:20
Genes:1035
Diseases:664
GDAs:6126
R. Norvergicus
CTD human
Genes:5269
Diseases:3090
GDAs:20056
Genes:1714
Diseases:1770
GDAs:1869
Genes:125
Diseases:606
GDAs:753
UNIPROT
BeFree
Genes:4819
Diseases:8296
GDAs:300043
GAD
Genes:2168
Diseases:133
GDAs:16935
Genes:2698
Diseases:651
GDAs:11873
Genes:4110
Diseases:880
GDAs:4454
Genes:1775
Diseases:730
GDAs:14518
Genes:69
Diseases:73
GDAs:678
Genes:182
Diseases:163
GDAs:14837
LHGDN

Standardization

- Controlled vocabularies and ontologies
- DisGeNET association type ontology

Open access

- RDF and nanopublications
- Data distributed under the Open database commons license



http://semanticscience.org

http://opendatacommons.org/

http://lod-cloud.net/

Key points

2. **High rate of data generation on GDAs poses challenges to biocuration pipelines**

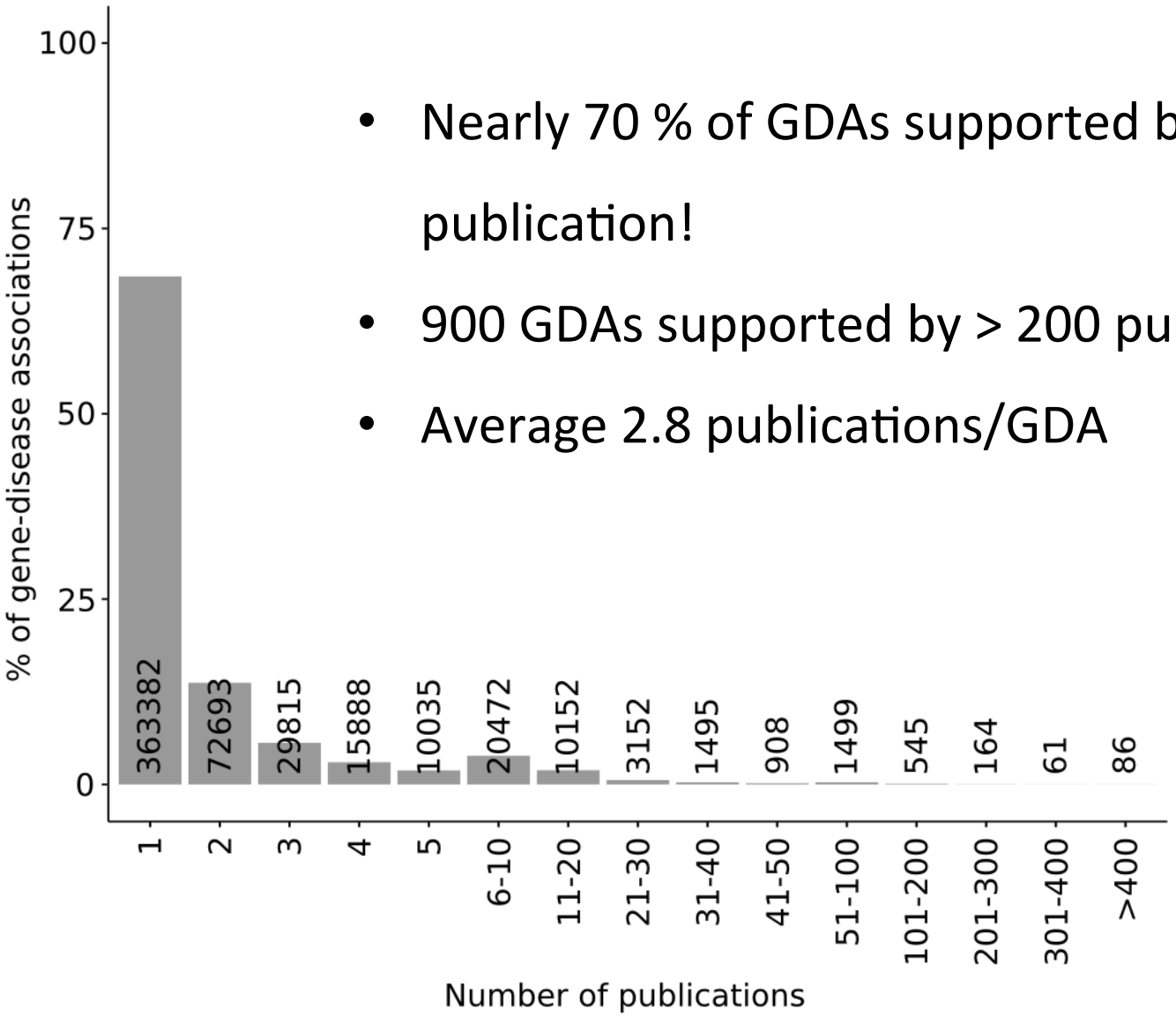
Publications on diseases and genes from 1980
(737,712 publications)

Text Mining

Supervised machine-learning approach for relation extraction between genes and diseases

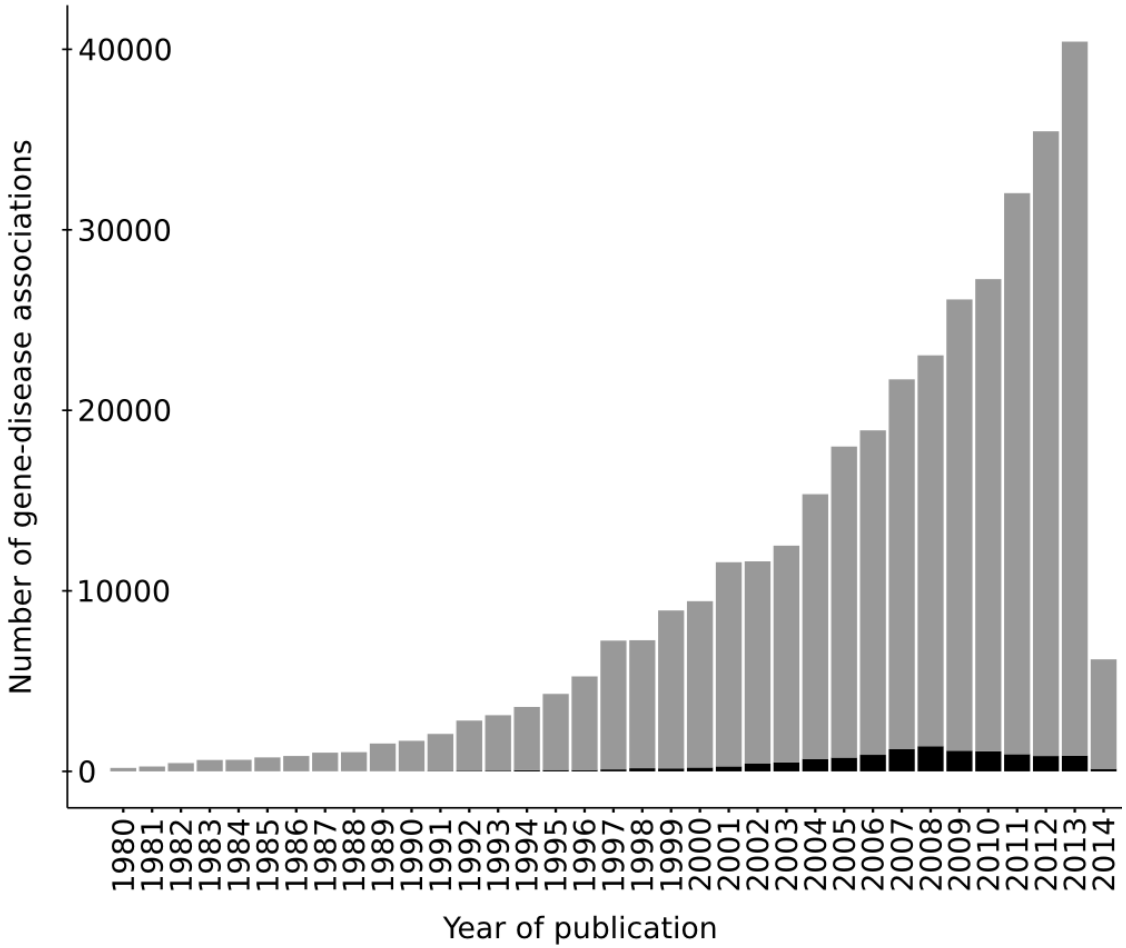
14,777 genes
12,650 diseases
355,976 publications

Aprox. 30,000 GDAs



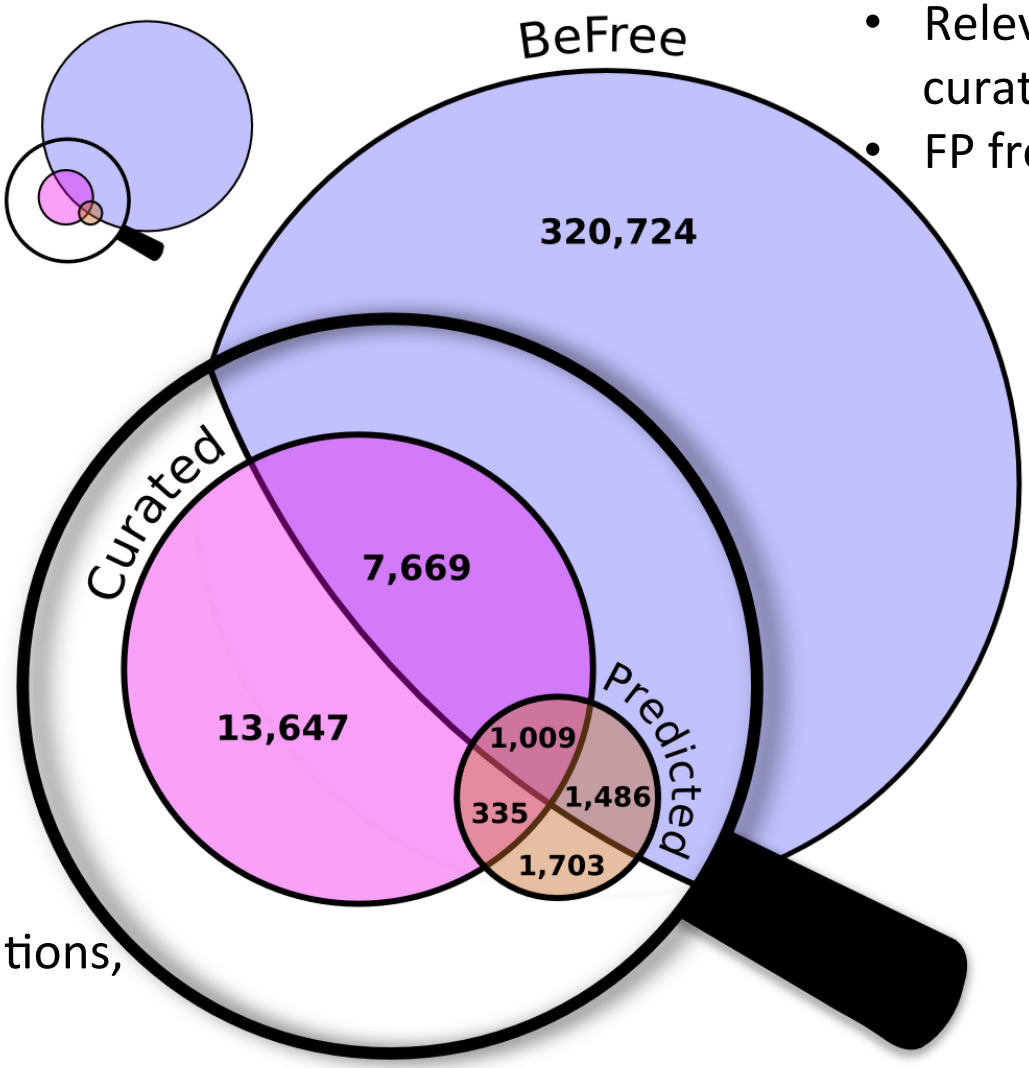
- Nearly 70 % of GDAs supported by only **one** publication!
- 900 GDAs supported by > 200 publications
- Average 2.8 publications/GDA

GDAs supported by only one publication



Little overlap between text mined data and data present in curated repositories

- Non-relevant data for DBs
- Relevant data but still not curated
- FP from TM



- Other publications, full text, etc
- FN from TM

- This is not raw data (e.g. from NGS analysis), is data already collated and filtered that have passed peer review before publication

- Text mined from abstracts:

  - Not mining full text, nor supplementary material

  - Not mining tables, figures

- Focus on relations stated on sentences, not handling anaphoras

- We are just looking into a *small fraction* of all the available data!!!

Key points

2. **High rate of data generation on GDAs poses challenges to biocuration pipelines**

- Need to find alternative strategies to expert curation
- "wisdom of the crowds" approaches (e.g. crowdsourcing)
- ?

Key points

3. **Data prioritization to support interpretation of data on the genetic determinants of human diseases**

GDAs prioritization

Ranks gene-disease associations based on the supporting **evidence**

DisGeNET score = $S_{CURATED} + S_{PREDICTED} + S_{LITERATURE}$

$S_{CURATED} = W_{UNIPROT} + W_{CTD}$

$S_{PREDICTED} = W_{Rat} + W_{Mouse}$

$S_{LITERATURE} = W_{GAD} + W_{LHGDN} + W_{BeFree}$

Disease	Gene	Score	UniProt	CTD	Rat	Mouse	Number of publications		
							BeFree	GAD	LHGDN
Wilson's Disease	ATP7B	0.99	0.3	0.3	0.1	0.1	174	31	23
Rett Syndrome	MECP2	0.9	0.3	0.3	0	0.1	438	27	43
Cystic Fibrosis	CFTR	0.9	0.3	0.3	0	0.1	1429	150	78
Obesity	MC4R	0.94	0.3	0.3	0.1	0.1	220	46	0
Alzheimer Disease	APP	0.88	0.3	0.3	0	0.1	1096	18	81

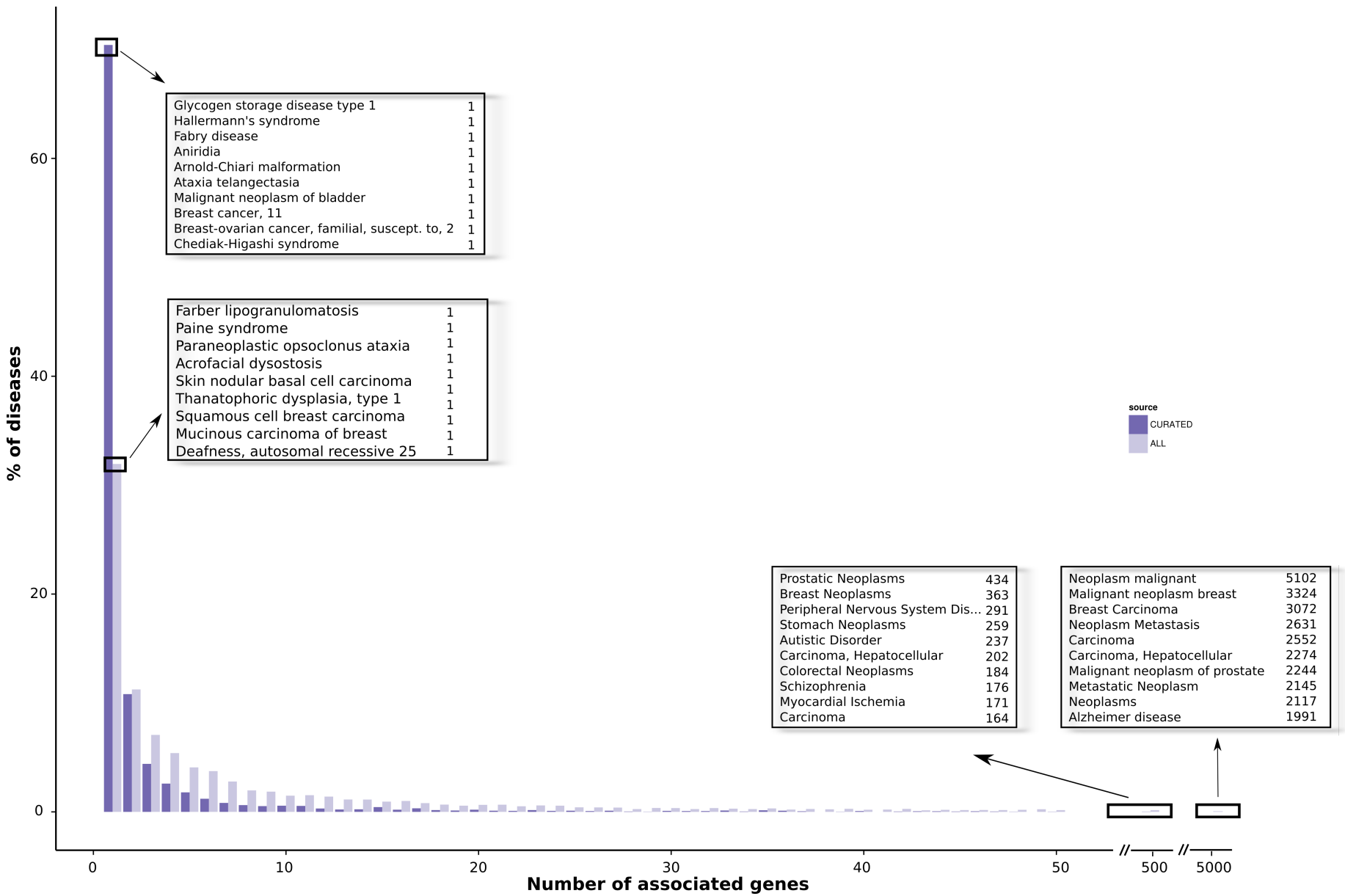
Key points

3. **Data prioritization to support interpretation of data on the genetic determinants of human diseases**

- Still we have a large dataset with low score (350,000 GDAs)
- Other approaches based on:
  - type of association of the gene to the disease
  - experimental evidence for the GDAs
  - network-based gene-prioritization algorithms
- Take into account contradictory findings
- Different prioritization approaches for different purposes?

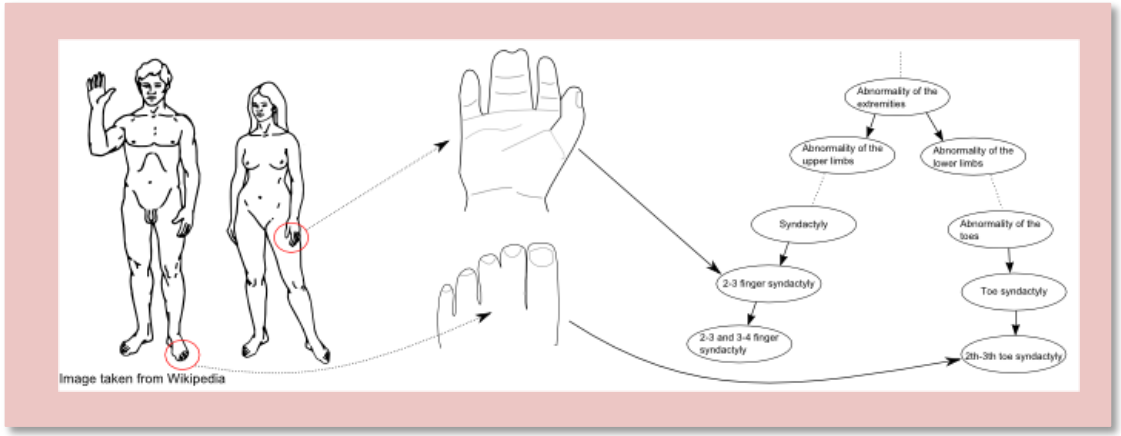
Key points

4. **Large number of genes associated to some diseases might reflect phenotypic diversity**


% of diseases
Number of associated genes
0
20
40
60
0
10
20
30
40
50
500
5000
Glycogen storage disease type 1 1
Hallermann's syndrome 1
Fabry disease 1
Aniridia 1
Arnold-Chiari malformation 1
Ataxia telangectasia 1
Malignant neoplasm of bladder 1
Breast cancer, 11 1
Breast-ovarian cancer, familial, suscept. to, 2 1
Chediak-Higashi syndrome 1
Farber lipogranulomatosis 1
Paine syndrome 1
Paraneoplastic opsoclonus ataxia 1
Acrofacial dysostosis 1
Skin nodular basal cell carcinoma 1
Thanatophoric dysplasia, type 1 1
Squamous cell breast carcinoma 1
Mucinous carcinoma of breast 1
Deafness, autosomal recessive 25 1
source
CURATED
ALL
Prostatic Neoplasms 434
Breast Neoplasms 363
Peripheral Nervous System Dis... 291
Stomach Neoplasms 259
Autistic Disorder 237
Carcinoma, Hepatocellular 202
Colorectal Neoplasms 184
Schizophrenia 176
Myocardial Ischemia 171
Carcinoma 164
Neoplasm malignant 5102
Malignant neoplasm breast 3324
Breast Carcinoma 3072
Neoplasm Metastasis 2631
Carcinoma 2552
Carcinoma, Hepatocellular 2274
Malignant neoplasm of prostate 2244
Metastatic Neoplasm 2145
Neoplasms 2117
Alzheimer disease 1991

Need for deep *phenotyping* of diseases to identify disease subtypes associated to different gene networks

Human Phenotype Ontology (HPO)



http://www.human-phenotype-ontology.org/

- Ontology of phenotypic abnormalities of human diseases
- Based on OMIM, Orphanet and DECIPHER

DisGeNET-HPO annotations

33 % of DisGeNET diseases annotated with HPO terms

Mainly come from OMIM diseases (28 % of DisGeNET diseases come from OMIM)

Need for other approaches to annotate the full spectrum of diseases at phenotypic level

Key points

1. Fragmentation of information as a barrier to knowledge on the mechanisms of human diseases

2. High rate of data generation on GDAs poses challenges to biocuration pipelines

3. Data prioritization to support interpretation of data on the genetic determinants of human diseases

4. Large number of genes associated to some diseases might reflect phenotypic diversity

IBI Group

Alba Gutiérrez
Àlex Bravo
Janet Piñero
Núria Queralt-Rosiñach
Miguel A. Mayer
Pablo Carbonell
Laura I. Furlong
Ferran Sanz

IBI Past members

Montserrat Cases
Michael Rautschka
Anna Bauer-Mehren
Solène Grosdidier

Unión Europea
Fondo Europeo de Desarrollo Regional

Instituto de Salud Carlos III

inb

TECNIO
Be tech. Be competitive

Open PHACTS
Open Pharmacological Space

eTOX

GRIB
RESEARCH PROGRAMME ON BIOMEDICAL INFORMATICS

upf. Universitat Pompeu Fabra Barcelona

IMIM
Parc de Salut MAR Barcelona
Institut Hospital del Mar d'Investigacions Mèdiques