Supplementary Information

Quantifying the impact of public omics data

Yasset Perez-Riverol [1, *], Andrey Zorin [1], Gaurhari Dass [1], Manh-Tu Vu [1], Pan Xu [2], Mihai Glont [1], Juan Antonio Vizcaíno [1], Andrew F. Jarnuczak [1], Robert Petryszak [1], Peipei Ping [3,4], Henning Hermjakob [1,2]

[1] European Molecular Biology Laboratory, EMBL-European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, UK.

[2] State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Lifeomics; National Center for Protein Sciences (The PHOENIX Center, Beijing), Beijing 102206, China

[3] Department of Physiology, Division of Cardiology, David Geffen School of Medicine at UCLA, University of California, Los Angeles, California, USA.

[4] Department of Medicine, Division of Cardiology, David Geffen School of Medicine at UCLA, University of California, Los Angeles, California, USA.

Table of Contents

Supplementary Note 1: Biological connection metric

The *connections* metric is calculated by retrieving the references to the dataset from knowledgebases. For example, if a UniProt protein entry references a proteomics dataset, this counts as a connection to the dataset. The main idea is to keep track of all the biological knowledge generated by one omics dataset. This metric promotes the idea that the value of a dataset is related not only to the number of times it has been downloaded/reused/cited, but also how much biological knowledge (e.g. UniProt proteins, Ensembl genes, interactions or pathways) are supported by that dataset. **Table 1** shows all the knowledgebases data resources used by OmicsDI to compute the connections to omics datasets. The first column of the table includes the name of the data resource, the second column is the type of data and the last column is the number of entities it contains (by March 2019).

Table 1: Knowledge bases data resources included in OmicsDI.

Name of the Database	Type of Biological Knowledge	Number of Biological Entities
Assembly	Nucleotide sequences	144,952
Assembly scaffold (Release)	Nucleotide sequences	41,759,024
Baseline Expression Atlas Genes	Gene expression	775,676
ChEMBL Molecule	Bioactive molecules	1,735,442
ChEMBL Target	Bioactive molecules	11,538
Coding (Release)	Nucleotide sequences	356,332,997
DGVa	Genomes & metagenomes	42,992,846
Differential Expression Atlas Genes	Gene expression	376,636
Ensembl Gene	Genomes & metagenomes	3,113,803
Ensembl Genomes Gene	Genomes & metagenomes	175,370,460
Enzyme Portal	Enzymes	7,161
HGNC	Genomes & metagenomes	45,739
IMGT/HLA	Nucleotide sequences	13,641
IntAct Complexes	Molecular interactions	2,122
IntAct Interactions	Molecular interactions	528,593
IntAct Interactors	Molecular interactions	102,685
IntEnz	Enzymes	7,686
InterPro	Protein families	32,568
JPO	Protein sequences	2,795,178
KIPO	Protein sequences	518,927

Large assembly	Nucleotide sequences	44,032,859
Ligands	Bioactive molecules	19,672
Non-coding (Release)	Nucleotide sequences	19,045,545
PDBe	Macromolecular structures	107,958
PfamEntry	Protein families	16,712
PomBase	Genomes & metagenomes	6,995
RNAcentral	Nucleotide sequences	13,353,531
Reactome	Reactions & pathways	686,345
Rfam	Nucleotide sequences	2,309,950
Rhea	Reactions & pathways	43,456
Sequence (Release)	Nucleotide sequences	205,714,494
TreeFam	Protein families	15,736
USPTO	Protein sequences	4,077,368
UniProtKB	Protein sequences	108,184,003
WormBase ParaSite	Genomes & metagenomes	2,550,351

All these knowledge bases are indexed by the EBI Search resource (www.ebi.ac.uk/ebisearch/) [1]. The EBI search API is used to retrieve all the biological entities that reference the OmicsDI dataset or its corresponding publications (some databases use the publication identifier instead of the dataset accession). In addition, we add to the previous list the biological entities that have been reported by the dataset. By adding the list of referenced biological entities within the dataset we mitigate the time effect between the dataset publication and the use of the data in a knowledgebase.

For example, the dataset **E-MTAB-599** has 21,741 connections in EBI SEARCH (www.ebi.ac.uk/ebisearch/search.ebi?db=allebi&query=E-MTAB-599&requestFrom=searchBox); 115 Nucleotide sequences, 21,572 gene expression profiles, 54 Samples & ontologies mentions (lets called this **group A**). Then, we query the EBI Search for the manuscript associated with the dataset (https://www.ebi.ac.uk/ebisearch/crossrefsearch.ebi?db=europepmc&id=21921910&ref=genomes&requestFrom=relatedData-xref). The results showed 1,689,177 new connections from the Database from Genomics Variation Archive (**lets called this group B**). In addition, the 21,572 reported in the dataset considered (**lets called this group C**). The final connection metric number is the union of the three sets (**A** U **B** U **C**). By using the union of the three groups we avoid duplications of connections among providers (**Fig. 1**).

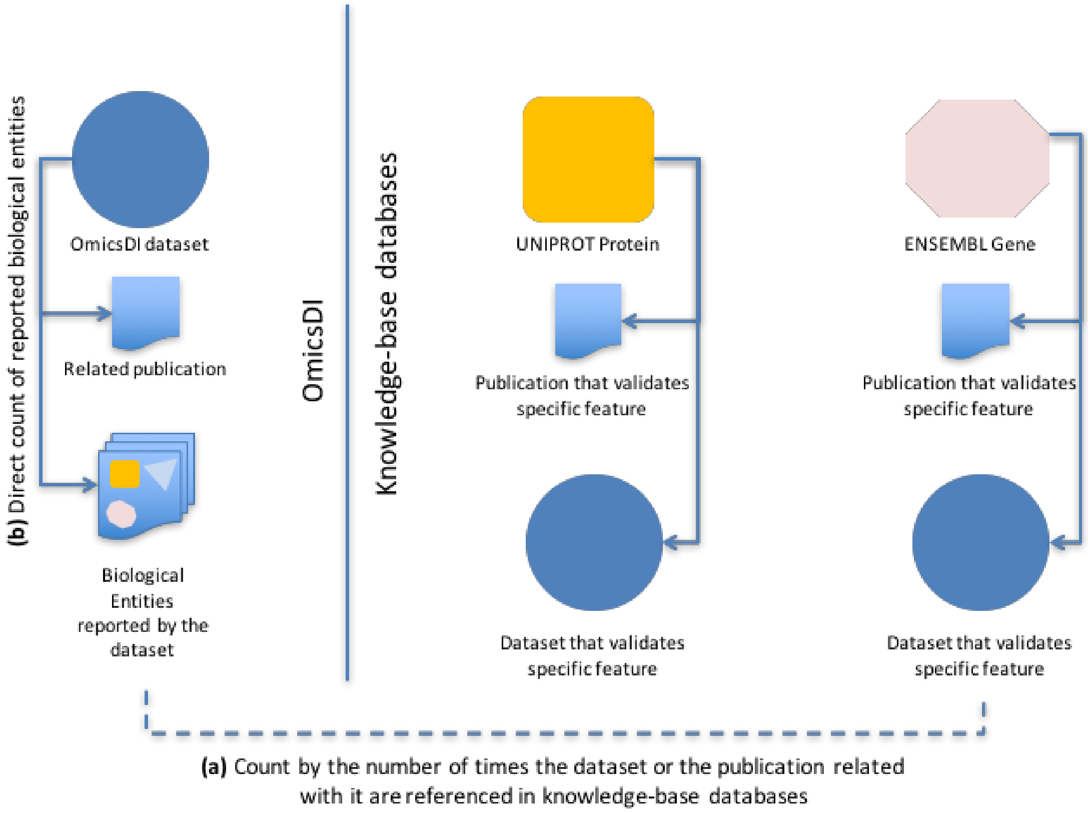

(b) Direct count of reported biological entities
OmicsDI dataset
Related publication
Biological Entities reported by the dataset
OmicsDI
Knowledge-base databases
UNIPROT Protein
Publication that validates specific feature
Dataset that validates specific feature
ENSEMBL Gene
Publication that validates specific feature
Dataset that validates specific feature
(a) Count by the number of times the dataset or the publication related with it are referenced in knowledge-base databases

Figure 1: Schema of the reference system used in OmicsDI based on the EBI Search architecture. Each OmicsDI dataset is supported by a publication and a list of biological entities in knowledgebases (UniProt, Ensembl, etc). In the same way, entities in knowledgebases are supported by publications or datasets. (**a**) The connections pipeline in OmicsDI generates counts of connections to a dataset or its related publication, if referenced in a knowledgebase. In addition, (**b**) the pipeline uses the list of biological entities reported by the dataset to complement the previous lists in case the dataset hasn't been added to the knowledgebase. The final connection metric is the **union** of all the collected lists.

Supplementary Note 2: Clustering based on metrics

We have combined the data *citation* and *reanalyses* metrics using a simple *k-means* clustering algorithm to identify which are the datasets that are highly-cited, highly-reanalysed, moderately-cited, and moderately-reanalysed (**Fig. 2**). This simple classification enables users and services to search and find relevant datasets in a resource such as OmicsDI, with contains more than 100,000 datasets. The current results show that there are 113 (highly-cited), 682 (moderately-cited), 245 (moderately-reanalysed), 8 (highly-reanalysed) datasets.

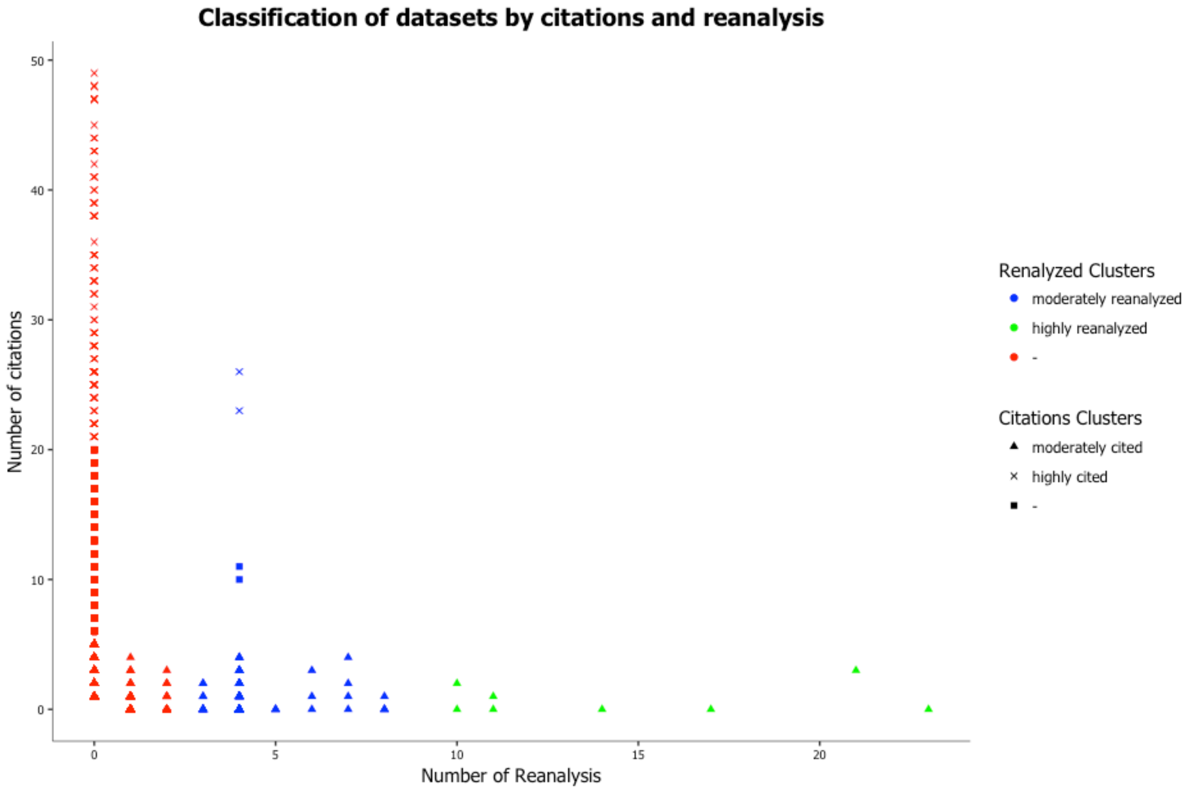
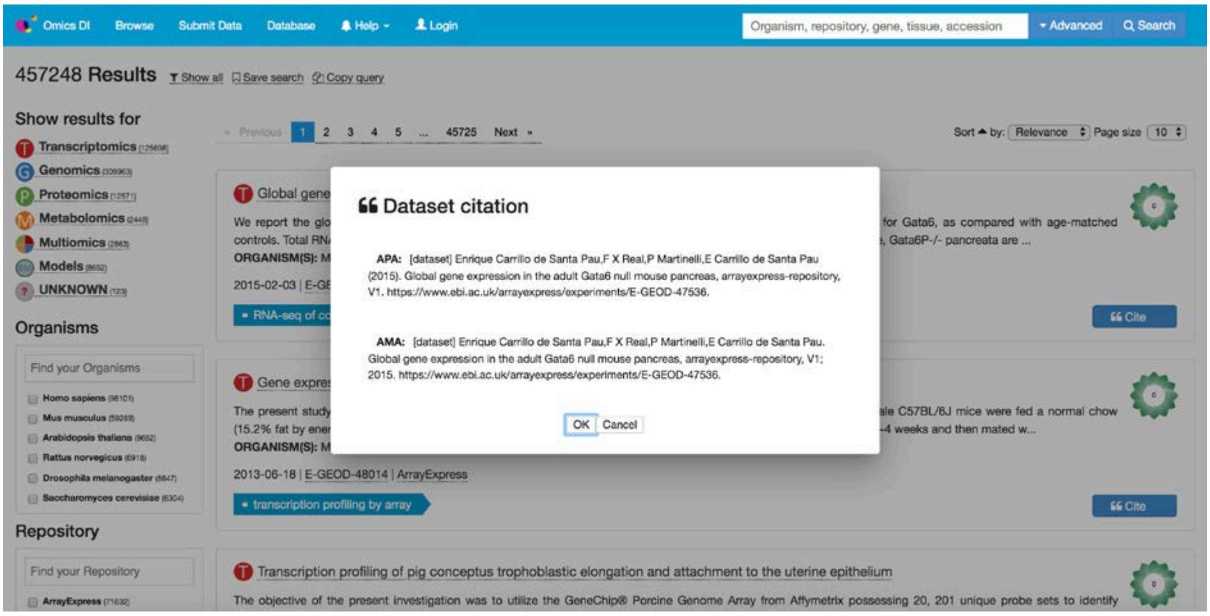


Figure 2: Classification of datasets into four different categories (highly-cited, highly-reanalysed, moderately-cited, moderately-reanalysed) using the *k-means* cluster algorithm.

Supplementary Note 3: Citing datasets

OmicsDI has implemented a simple visualization component (**Fig. 3**) that allows users to cite the corresponding dataset using the FORCE11 Data Citation Synthesis Group recommendations (http://www.dcc.ac.uk/resources/how-guides/cite-datasets)[3].



Omics DI
Browse
Submit Data
Database
Help
Login
Organism, repository, gene, tissue, accession
Advanced
Search
457248 Results
Show all
Save search
Copy query
Show results for
Transcriptomics (125698)
Genomics (309963)
Proteomics (12571)
Metabolomics (2448)
Multiomics (2863)
Models (8652)
UNKNOWN (123)
Organisms
Find your Organisms
Homo sapiens (98101)
Mus musculus (59289)
Arabidopsis thaliana (9662)
Rattus norvegicus (6918)
Drosophila melanogaster (6847)
Saccharomyces cerevisiae (6304)
Repository
Find your Repository
ArrayExpress (71632)
« Previous
1
2
3
4
5
...
45725
Next »
Sort by: Relevance
Page size 10
Global gene
We report the glo
controls. Total RN
ORGANISM(S): M
for Gata6, as compared with age-matched
, Gata6P-/- pancreata are ...
2015-02-03 | E-GE
RNA-seq of co
Cite
Gene expres
The present study
(15.2% fat by ener
ORGANISM(S): M
ale C57BL/6J mice were fed a normal chow
-4 weeks and then mated w...
2013-06-18 | E-GEOD-48014 | ArrayExpress
transcription profiling by array
Cite
Transcription profiling of pig conceptus trophoblastic elongation and attachment to the uterine epithelium
The objective of the present investigation was to utilize the GeneChip® Porcine Genome Array from Affymetrix possessing 20, 201 unique probe sets to identify
OK
Cancel

Dataset citation

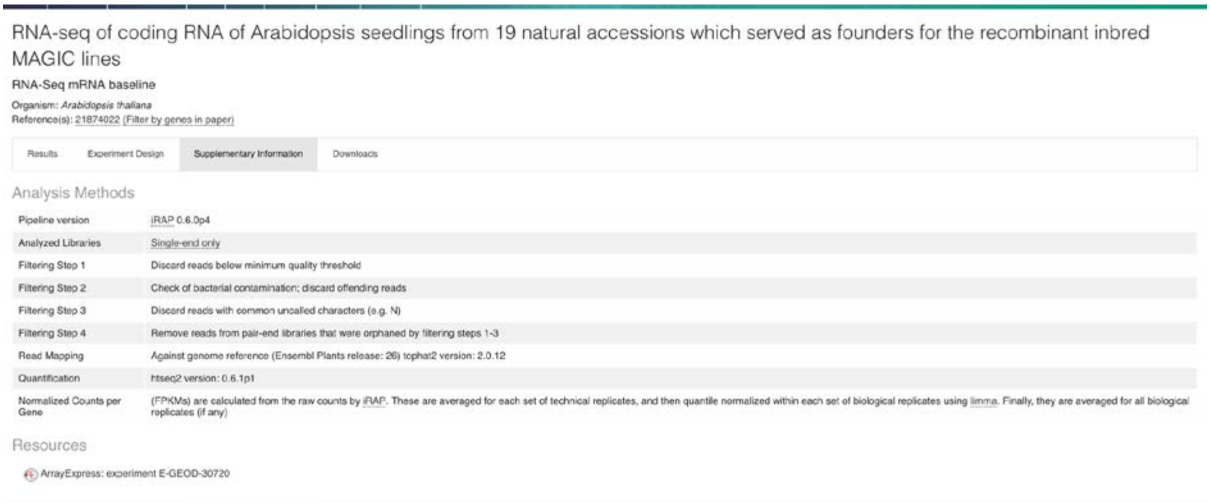
APA: [dataset] Enrique Carrillo de Santa Pau,F X Real,P Martinelli,E Carrillo de Santa Pau (2015). Global gene expression in the adult Gata6 null mouse pancreas, arrayexpress-repository, V1. https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-47536.

AMA: [dataset] Enrique Carrillo de Santa Pau,F X Real,P Martinelli,E Carrillo de Santa Pau. Global gene expression in the adult Gata6 null mouse pancreas, arrayexpress-repository, V1; 2015. https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-47536.

Figure 3: Dataset citations in OmicsDI can be exported in APA and AMA styles.

Supplementary Note 4: References between databases

Expression Atlas (www.ebi.ac.uk/gxa) adds to each reanalysed dataset the accession number and URL of the original ArrayExpress dataset. The reference to the original dataset in ArrayExpress enables OmicsDI to easily trace the number of reanalyses for each dataset.



RNA-seq of coding RNA of Arabidopsis seedlings from 19 natural accessions which served as founders for the recombinant inbred MAGIC lines

RNA-Seq mRNA baseline

Organism: *Arabidopsis thaliana*
Reference(s): 21874022 (Filter by genes in paper)

Analysis Methods

Pipeline version	iRAP 0.6.0p4
Analyzed Libraries	Single-end only
Filtering Step 1	Discard reads below minimum quality threshold
Filtering Step 2	Check of bacterial contamination; discard offending reads
Filtering Step 3	Discard reads with common uncalled characters (e.g. N)
Filtering Step 4	Remove reads from pair-end libraries that were orphaned by filtering steps 1-3
Read Mapping	Against genome reference (Ensembl Plants release: 26) tophat2 version: 2.0.12
Quantification	htseq2 version: 0.6.1p1
Normalized Counts per Gene	(FPKMs) are calculated from the raw counts by iRAP. These are averaged for each set of technical replicates, and then quantile normalized within each set of biological replicates using limma. Finally, they are averaged for all biological replicates (if any)

Resources

ArrayExpress: experiment E-GEOD-30720

Figure 4: Expression Atlas' direct reference to the original re-used transcriptomics dataset originally deposited in ArrayExpress.

Supplementary Note 5: Metrics Pipeline

The metrics estimation pipeline is based on the Omics XML file format that is used to transfer datasets from each provider into OmicsDI (https://github.com/OmicsDI/specifications) (**Fig. 5**). **Fig. 5** shows how the data is imported from each provider into a central mongoDB database. An automatic pipeline is run to detect duplication and data replication across the resource between databases. For example, since ArrayExpress replicates all the datasets from GEO, the ArrayExpress dataset https://www.omicsdi.org/dataset/arrayexpress-repository/E-GEOD-46807 (accession number E-GEOD-46807) is a replicate of the GEO dataset GSE46807 and BioProject PRJNA202371. Some other types of replication are more complex to detect automatically and for this reason, a manual curation tool is also provided to merge datasets. Finally, the pipelines use the central mongoDB database, the EuropePMC API and the knowledgebases (e.g. Ensembl, UniProt, see Table 1) to compute/estimate the different metrics (**Fig. 9**).

The *reanalysis* count is computed using the reanalysis link provided on each OmicsDI XML file. If the dataset constitutes a *reanalysis* of a previous one, the XML file contains a reference to the original dataset. In the same way, each dataset contains the list of molecules reported in the analysis (e.g. UniProt protein, ChEBI or Ensembl gene identifiers) that are used to crosslink the dataset with the knowledgebases and compute the *connections* metric.

The *citations* metric is computed using the EuropePMC API (https://europepmc.org/RestfulWebService). Each dataset accession is searched using the API and the number of manuscripts that mentions the manuscript is retrieved (*citations* count). A precondition of this method is that the manuscript content needs to be accessible on EuropePMC and/or PubMed Central. The number of views is estimated by counting the number of times users access a dataset in the OmicsDI web interface but also the number of times the dataset is accessed using the API. By March 2019, for calculating the direct downloads of each dataset, major providers from EMBL-EBI including PRIDE, ArrayExpress, ENA, MetaboLigths and BioModels agreed to provide the number of downloads via the different file transfer methods: FTP, HTTP and Aspera. The number of downloads can be provided also using the OmicsDI XML file schema using the additional field *download_count*.

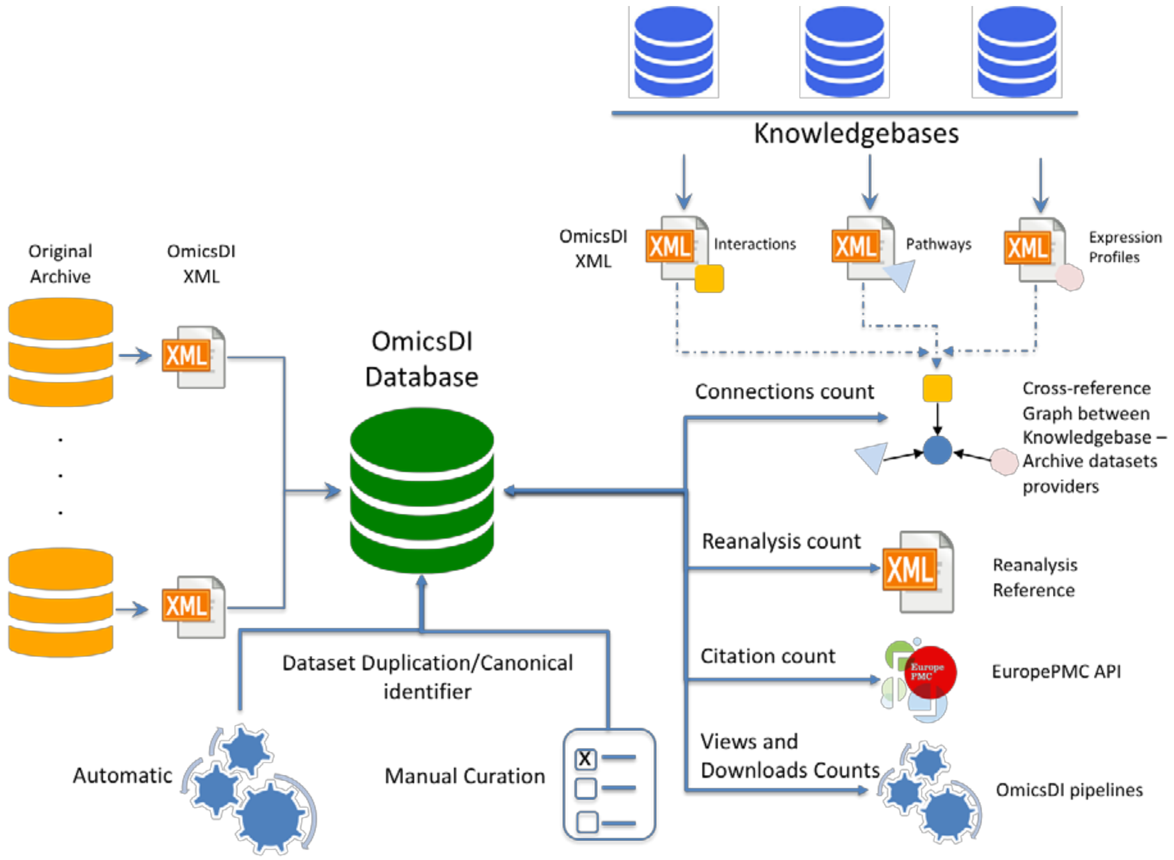
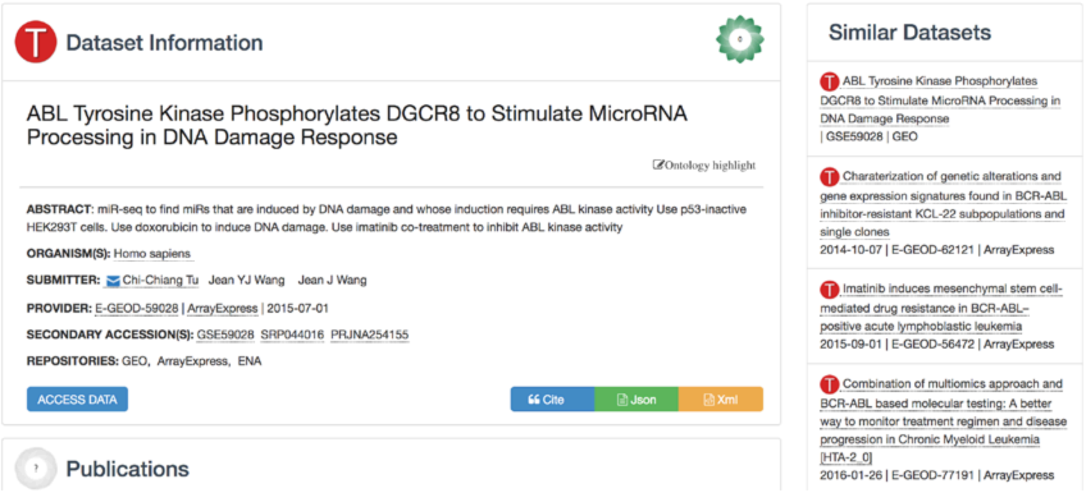


Figure 9: The metrics estimation pipeline is based on the Omics XML file format that is used to transfer datasets from each provider into OmicsDI. Data is imported from each provider into a central MongoDB database. An automatic pipeline is run to detect duplication and data replication across the resource. The pipelines use the central MongoDB database, the EuropePMC API and the knowledgebases (e.g. Ensembl, UniProt) to compute/estimate the different metrics.

Supplementary Note 6: Dataset replicate detection system

Dataset replication is considered a good practice because it guarantees that the dataset will be available even if one of the repositories needs to be closed down or runs out of funding. However, for indexing resources it is desirable to detect that the dataset is the same in the different providers to avoid redundancy. This is a complex task, the identifier of the dataset in the original provider is not always kept in the “replicate” versions. An automatic pipeline and a manual annotation system enable OmicsDI to remove duplicated datasets with potentially different identifiers (e.g. transcriptomics datasets available in ArrayExpress and Gene Expression Omnibus (GEO)). The pipeline designates one of the datasets as the canonical representation and annotates the rest of identifiers as additional secondary ones.

The following dataset (**Fig. 6**) was originally deposited in the Gene Expression Omnibus (GEO) (www.ncbi.nlm.nih.gov/geo/), replicated in the European Nucleotide Archive (ENA) (www.ebi.ac.uk/ena) and indexed by ArrayExpress (www.ebi.ac.uk/arrayexpress/) and BioProjects (www.ncbi.nlm.nih.gov/bioproject/). All those databases assign different identifiers to the same dataset, GSE59028, SRP044016, E-GEOD-59028 and PRJNA254155, respectively. The current OmicsDI pipeline automatically annotates all the identifiers for a dataset with replicates (ArrayExpress - E-GEOD-59028, **Fig. 6**). It also provides a web-interface to manually annotate those datasets that cannot be automatically “merged”.



Dataset Information

ABL Tyrosine Kinase Phosphorylates DGCR8 to Stimulate MicroRNA Processing in DNA Damage Response

ABSTRACT: miR-seq to find miRs that are induced by DNA damage and whose induction requires ABL kinase activity Use p53-inactive HEK293T cells. Use doxorubicin to induce DNA damage. Use imatinib co-treatment to inhibit ABL kinase activity

ORGANISM(S): Homo sapiens

SUBMITTER: Chi-Chiang Tu Jean YJ Wang Jean J Wang

PROVIDER: E-GEOD-59028 | ArrayExpress | 2015-07-01

SECONDARY ACCESSION(S): GSE59028 SRP044016 PRJNA254155

REPOSITORIES: GEO, ArrayExpress, ENA

Publications

Similar Datasets

ABL Tyrosine Kinase Phosphorylates DGCR8 to Stimulate MicroRNA Processing in DNA Damage Response
| GSE59028 | GEO

Charaterization of genetic alterations and gene expression signatures found in BCR-ABL inhibitor-resistant KCL-22 subpopulations and single clones
2014-10-07 | E-GEOD-62121 | ArrayExpress

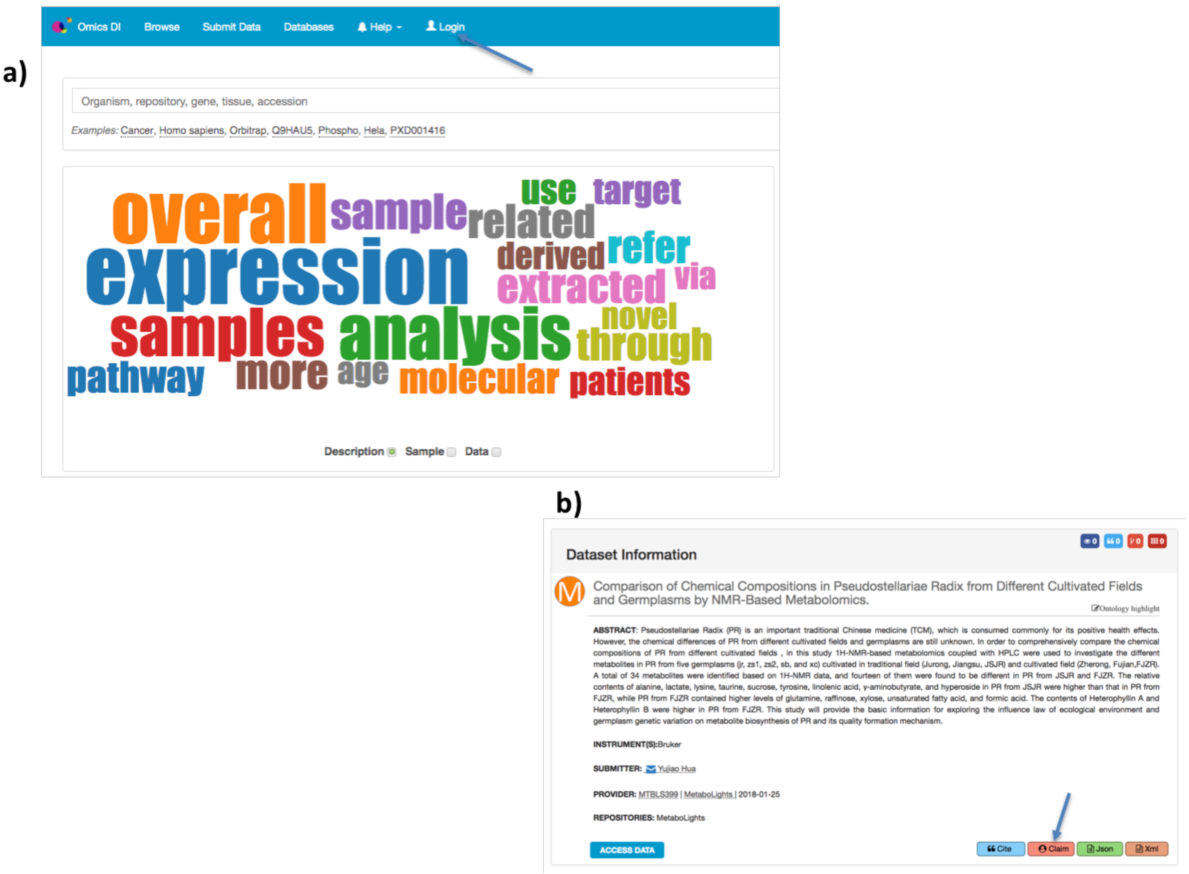
Imatinib induces mesenchymal stem cell-mediated drug resistance in BCR-ABL– positive acute lymphoblastic leukemia
2015-09-01 | E-GEOD-56472 | ArrayExpress

Combination of multiomics approach and BCR-ABL based molecular testing: A better way to monitor treatment regimen and disease progression in Chronic Myeloid Leukemia [HTA-2_0]
2016-01-26 | E-GEOD-77191 | ArrayExpress

Figure 6: ArrayExpress dataset E-GEOD-59028. This dataset can be also found in the original database provider GEO (dataset identifier GSE59028), in the ENA database (identifier SRP044016) and in NCBI BioProjects (identifier PRJNA254155). The OmicsDI canonical identifier pipeline only represents one entry in the OmicsDI resource (in this case the ArrayExpress one) for the corresponding dataset but also provides all the alternative/additional identifiers in other data resources.

Supplementary Note 7: Claiming datasets system

OmicsDI has implemented a dataset claiming system that allows researchers to get credit for generating and sharing their datasets. The platform is built on top of three main components: i) a log-in system based on ORCID (**Fig. 7a**); ii) a dataset search-based claiming system that allows users to search and add the relevant datasets to their OmicsDI profiles (**Fig. 7b**); and iii) a user profile that enables users to update datasets in their OmicsDI profile (**Fig. 7a**) and synchronize them with ORCID (**Fig. 8b**). Additionally, as a key point, OmicsDI claimed datasets can be synchronized to the researcher's own ORCID profile, highlighting datasets there as a research product as well [2]. When the profile is synchronized with ORCID, the user's datasets are added to their ORCID profile. The claiming step is based on the *principle of trust* as in other popular analogous resources for publications such as Google Scholar or ResearchGate.



Organism, repository, gene, tissue, accession

Examples: Cancer, Homo sapiens, Orbitrap, Q9HAU5, Phospho, Hela, PXD001416

Dataset Information

Comparison of Chemical Compositions in Pseudostellariae Radix from Different Cultivated Fields and Germplasms by NMR-Based Metabolomics.

ABSTRACT: Pseudostellariae Radix (PR) is an important traditional Chinese medicine (TCM), which is consumed commonly for its positive health effects. However, the chemical differences of PR from different cultivated fields and germplasms are still unknown. In order to comprehensively compare the chemical compositions of PR from different cultivated fields , in this study 1H-NMR-based metabolomics coupled with HPLC were used to investigate the different metabolites in PR from five germplasms (jr, zs1, zs2, sb, and xc) cultivated in traditional field (Jurong, Jiangsu, JSJR) and cultivated field (Zherong, Fujian,FJZR). A total of 34 metabolites were identified based on 1H-NMR data, and fourteen of them were found to be different in PR from JSJR and FJZR. The relative contents of alanine, lactate, lysine, taurine, sucrose, tyrosine, linolenic acid, γ-aminobutyrate, and hyperoside in PR from JSJR were higher than that in PR from FJZR, while PR from FJZR contained higher levels of glutamine, raffinose, xylose, unsaturated fatty acid, and formic acid. The contents of Heterophyllin A and Heterophyllin B were higher in PR from FJZR. This study will provide the basic information for exploring the influence law of ecological environment and germplasm genetic variation on metabolite biosynthesis of PR and its quality formation mechanism.

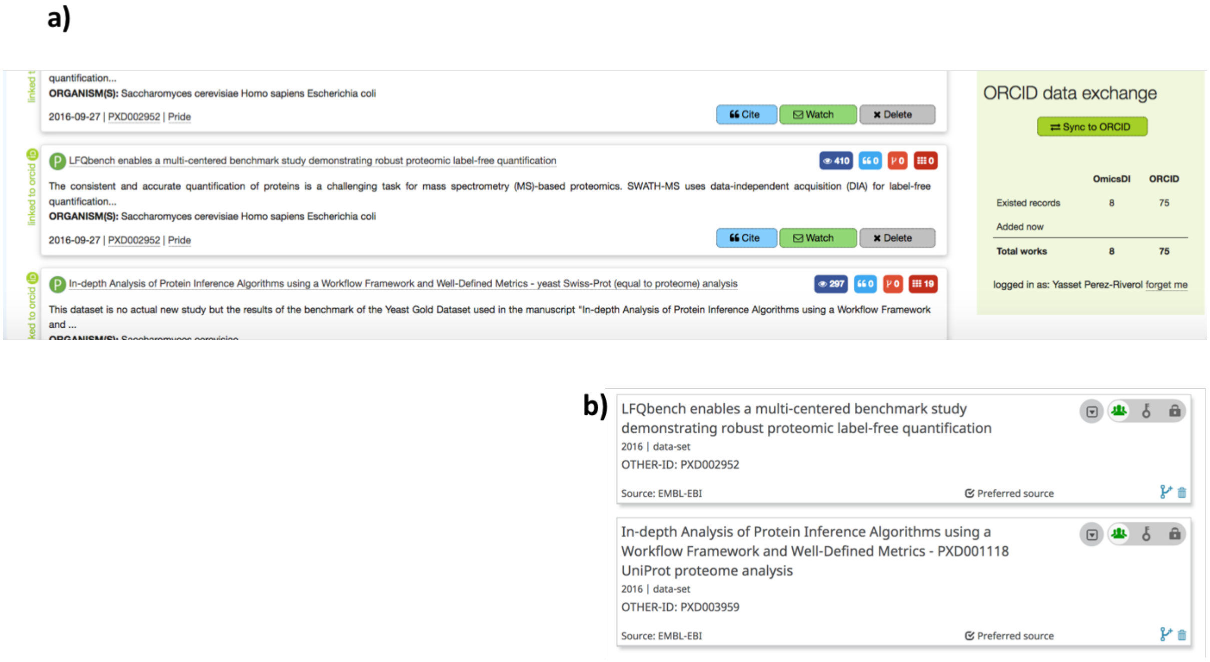
INSTRUMENT(S): Bruker

SUBMITTER: Yujiao Hua

PROVIDER: MTBLS399 | MetaboLights | 2018-01-25

REPOSITORIES: MetaboLights

Figure 7: **(a)** OmicsDI provides a simple log-in and personal profile by using the ORCID authentication system. **(b)** When the researchers are logged in, a new option appears at the bottom of each dataset page, which allows them to claim the corresponding dataset.



quantification...
ORGANISM(S): Saccharomyces cerevisiae Homo sapiens Escherichia coli
2016-09-27 | PXD002952 | Pride

P LFQbench enables a multi-centered benchmark study demonstrating robust proteomic label-free quantification 410 0 0 0

The consistent and accurate quantification of proteins is a challenging task for mass spectrometry (MS)-based proteomics. SWATH-MS uses data-independent acquisition (DIA) for label-free quantification...
ORGANISM(S): Saccharomyces cerevisiae Homo sapiens Escherichia coli
2016-09-27 | PXD002952 | Pride

P In-depth Analysis of Protein Inference Algorithms using a Workflow Framework and Well-Defined Metrics - yeast Swiss-Prot (equal to proteome) analysis 297 0 0 19

This dataset is no actual new study but the results of the benchmark of the Yeast Gold Dataset used in the manuscript "In-depth Analysis of Protein Inference Algorithms using a Workflow Framework and ...
ORGANISM(S): Saccharomyces cerevisiae

ORCID data exchange

	OmicsDI	ORCID
Existed records	8	75
Added now		
Total works	**8**	**75**

logged in as: Yasset Perez-Riverol forget me

LFQbench enables a multi-centered benchmark study demonstrating robust proteomic label-free quantification
2016 | data-set
OTHER-ID: PXD002952
Source: EMBL-EBI ✓ Preferred source

In-depth Analysis of Protein Inference Algorithms using a Workflow Framework and Well-Defined Metrics - PXD001118 UniProt proteome analysis
2016 | data-set
OTHER-ID: PXD003959
Source: EMBL-EBI ✓ Preferred source

Figure 8: (**a**) Datasets claimed by a specific user are listed in their OmicsDI profile. OmicsDI enables users to update their profile. A profile can be made public and the URL can be shared in the public domain to demonstrate the impact of the author's datasets. **(b)** Each researcher's dataset list can be synchronized with their own ORCID profile.

Supplementary Note 8: Metrics per species.

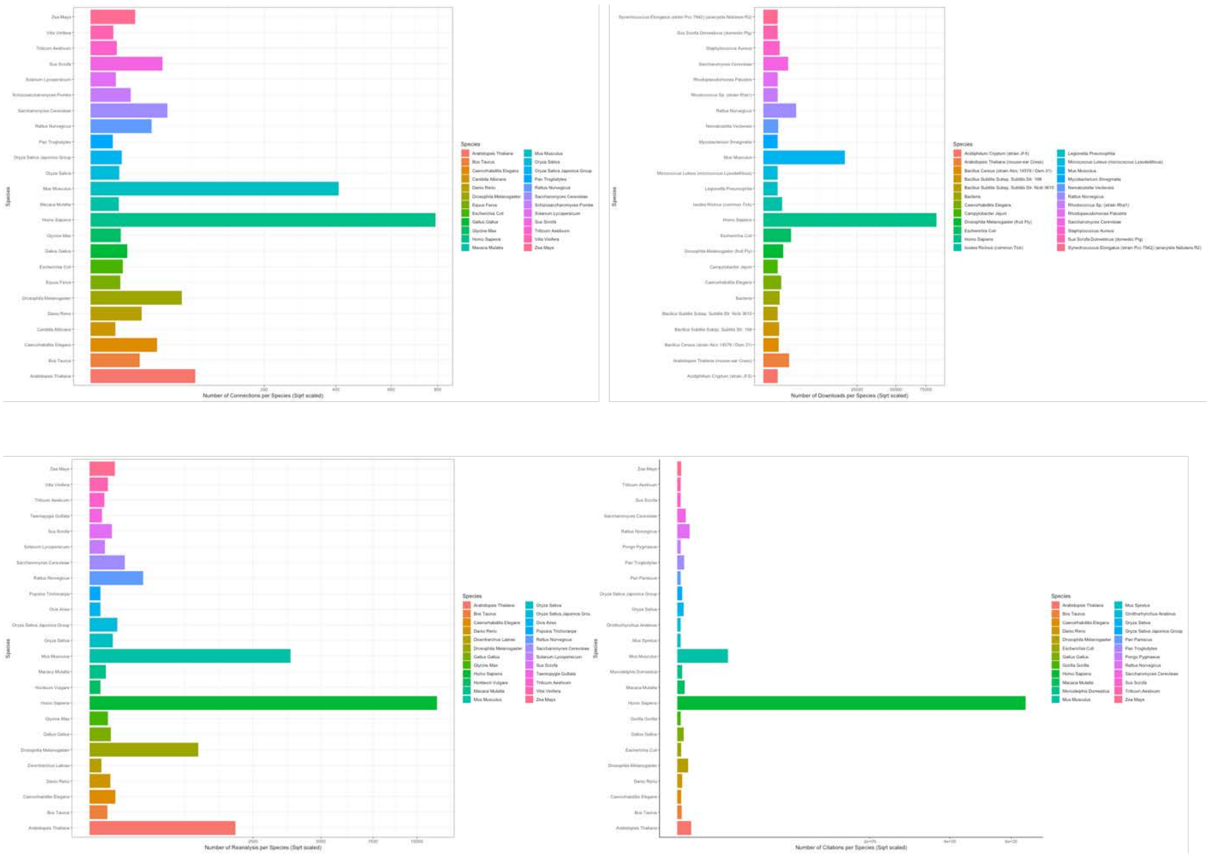


Figure 9: Distribution of metrics (*connections, downloads, reanalyses* and *citations*) per species.

References

1. Park, Y.M., Squizzato, S., Buso, N., Gur, T. & Lopez, R. The EBI search engine: EBI search as a service-making biological data accessible for all. *Nucleic Acids Res* **45**, W545-W549 (2017).
2. Haak, L.L., Fenner, M., Paglione, L., Pentz, E. & Ratner, H. ORCID: a system to uniquely identify researchers. *Learned Publishing* **25**, 259-264 (2012).
3. Ball, A. & Duke, M. How to cite datasets and link to publications. (Digital Curation Centre, 2011).