ComPath: An ecosystem for exploring, analyzing, and curating pathway databases

Daniel Domingo-Fernández[1,2,*], Charles Tapley Hoyt[1,2], Carlos Bobis-Álvarez[3], Josep Marín-Llaó[1,4], and Martin Hofmann-Apitius[1,2]

1. Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin 53754, Germany
2. Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53115, Germany
3. Faculty of Medicine and Health Sciences, University of Oviedo, Oviedo 33006, Spain
4. Rovira i Virgili University, Tarragona 43003, Spain

***Corresponding Author**: Domingo-Fernández, D., Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53754, Germany. Telephone details: +49 2241 14-2269. Fax: +49 2241 14-2656

Keywords: bioinformatics, pathways, databases, data integration, software, research data

Abstract

Although pathways are widely used for the analysis and representation of biological systems, their lack of clear boundaries, their dispersion across numerous databases, and the lack of interoperability impedes the evaluation of the coverage, agreements, and discrepancies between them. Here, we present ComPath, an ecosystem that supports curation of pathway mappings between databases and fosters the exploration of pathway knowledge through several novel visualizations. We have curated mappings between three of the major pathway databases and present a case study focusing on Parkinson's disease that illustrates how ComPath can generate new biological insights by identifying pathway modules, clusters, and cross-talks with these mappings. The ComPath source code and resources are available at https://github.com/ComPath and the web application can be accessed at http://compath.scai.fraunhofer.de/.

Introduction

The notion of pathways enables the representation, formalization, and interpretation of biological events or series of interactions. Cataloging biological knowledge into pathways reduces complexity from all possible interacting molecular entities to a set of well studied and validated functional relationships between molecular entities culminating in biological processes. Several efforts have generated databases of pathways with varying specificity and granularity that comprise signaling cascades, metabolic routes, and regulatory networks from to precise

signatures with no more than a couple of acting players to general pathways involving thousands of molecular players (Kanehisa *et al.*, 2016; Fabregat *et al*., 2017; Slenter *et al*., 2017; Liberzon *et al*., 2011).

Representing biology as pathways inevitably results in a loss of information that entails the removal of spatiotemporal information, or even biological entity types from the equation, often leading to represent pathways as biological networks. The network abstraction facilitates pathway visualization and interpretation thanks to the harmony between biological networks and systems: nodes correspond to molecular entities and edges to types of interactions occurring between them (e.g., inhibition, phosphorylation, etc.). Although networks can comprise a broad range of molecular types (e.g., proteins, chemicals, small molecules, etc.), they are generally reduced to the most direct outcome of our genetic makeup - the genetic and protein levels - so that we can mechanistically understand their functionality. Thus, they are frequently viewed and simplified to "gene sets", the collection of all genes/proteins that constitute the pathway, due to the major challenges of incorporating network topology and translating the variety of relationships into pathway analysis methods.

While the majority of the efforts to construct and delineate biological knowledge into pathways come from research groups or commercial companies that employ experienced scientific curators (Fabregat *et al*., 2017; Krämer *et al*., 2013), other pathway databases have flourished completely maintained and developed by the scientific community (Slenter *et al*., 2017). The reason for this disparity of approaches to curate pathway knowledge is that each resource diverges conceptually depending on the scope (e.g., signaling pathways, gene regulatory networks, and metabolic processes). With such a disparity in areas of expertise, curation teams, and guidelines, every pathway community adopted a different syntax or format to contextualize knowledge such as BioPax and Systems Biology Markup Language (SBML) (Demir *et al*., 2010; Hucka *et al*., 2003). This motivated the development of centralized resources, such as Pathway Commons (Cerami *et al*. 2010), that integrate pathways from different databases in a consistent manners. However, the arbitrary and overlapping boundaries as well as the absence of a common nomenclature to name pathway representations hamper the linkage of pathways, enforcing manual inter-database pathway comparisons. Although pathway ontologies were generated in a first attempt towards a standardization of pathway naming (Petri *et al*., 2014; Iyappan *et al*., 2016), there has not been any effort dedicated to link different pathway databases, which in turn would enable the exploration of the coverage, agreements, or discrepancies in the pathway knowledge.

Though numerous algorithms (Khatri *et al*., 2012) and tools (Liberzon *et al*., 2011; Kuleshov *et al.*, 2016) have been successfully applied to interpret experimental data through the context of pathway databases (Cary *et al*., 2005; Subramanian *et al*., 2005), there has not yet been a systematic comparison between the contents of various pathway databases, an assessment of their overlaps and gaps, or an establishment of mappings. This is by no means trivial due to the heterogeneity of data formats and terminology used across databases, the biological scales represented, and the numerous types of relationships describing interactions. Previous studies have only focused on comparing a single or small set of well-established pathways across multiple resources (Bauer-Mehren *et al*., 2009; Chowdhury

and Sarkar, 2015). For example, a comparison focused on metabolic pathways revealed how a set of five databases only agreed in a minimum core of the biochemistry knowledge (Stobbe *et al*., 2011).

These studies demonstrate the need to connect insights provided by each pathway database to foster a greater understanding of the underlying biology. Here, we present ComPath, a web application that integrates content from publicly accessible pathway databases, generates comparisons, enables exploration, and facilitates curation of inter-database mappings.

Results

We developed an interactive web application that enables users to explore, analyze, and curate pathway knowledge. Below, we present three case studies illustrating how it can be used for each of these purposes. The figures for each were generated by interactive, dynamic views in the ComPath web application based on three major public pathway databases: KEGG, Reactome, and WikiPathways (**Figure 1**).

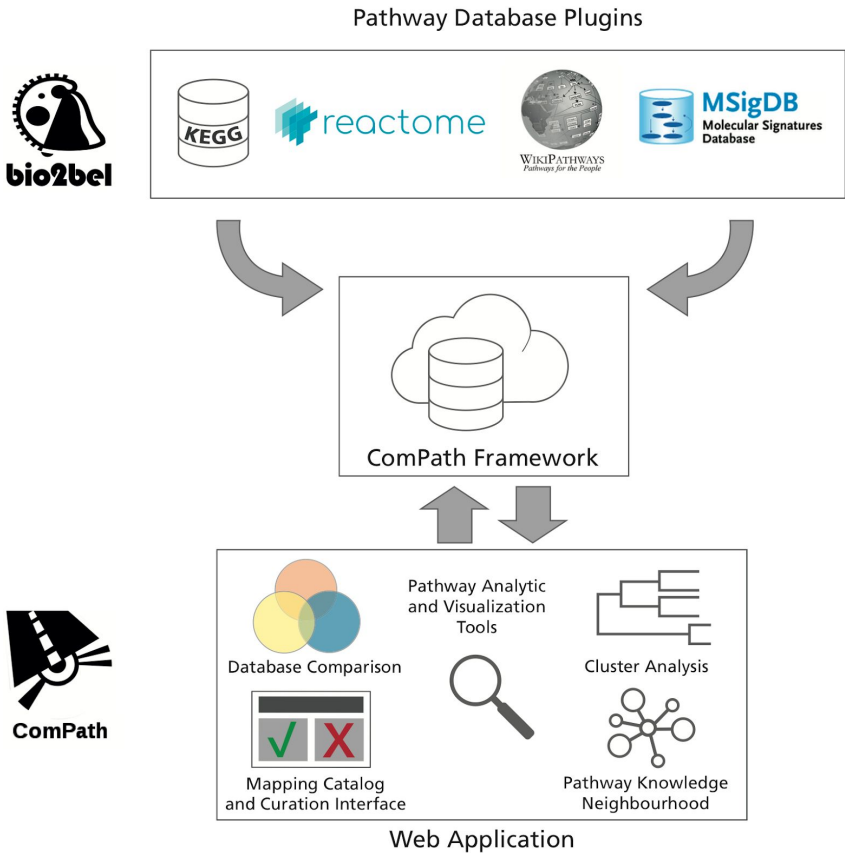

Figure 1. The ComPath ecosystem has three main components: the pathway database plugins, the ComPath framework, and the ComPath web application. The ComPath framework mediates the communication between the plugins containing the pathway database information and the web application.

Case Study I: Comparison of Pathway Databases

Assessment of Gene Coverage

Analysis of the overlaps between Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, and WikiPathways revealed that there are approximately 3800 common human genes shared between the three databases (**Figure 2A**). While at least one common human gene was present in almost every pathway across each database, the number of pathways with more common human genes diminishes much more quickly in WikiPathways and Reactome (**Supplementary Figure S1**). This difference might be due to database properties such as pathway size (e.g., on average, pathways contain 90 genes in KEGG, 50 in Reactome, and 42 in WikiPathways) or gene promiscuity (i.e. genes functionally linked to many pathways) (**Supplementary Table 1**). For further investigation, the ComPath web application generates summary tables and creates several visualizations to enable exploration of the distributions of pathway size and gene memberships for each database (**Figure 2B**).

Exploration of Pathways

While the previous views produced gene-centric summaries of the contents of pathway databases, ComPath also enables the exploration of pathway similarity landscape using Clustergrammer.js (Fernandez *et al*., 2017). **Figure 2C** illustrates how this view can elucidate the hierarchical relationships between the *Metabolic* pathway, the largest KEGG pathway, and other more high-granular KEGG metabolic pathways (e.g., *alpha-Linolenic acid metabolism, Lipoic acid metabolism, and ether lipid metabolism*).

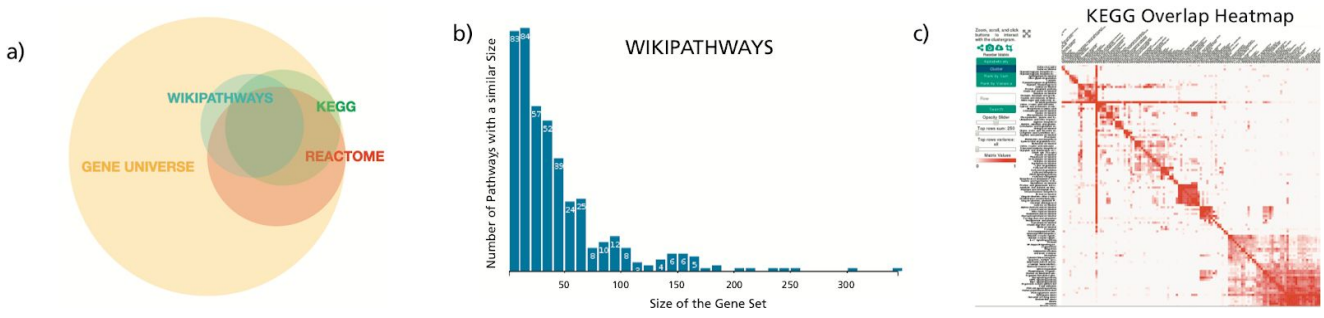

Figure 2. **A)** An Euler diagram summarizing the human gene-centric coverage of KEGG, Reactome, and WikiPathways compared to the universe of all genes from HGNC. **B)** Histogram views present gene promiscuity or pathway size distributions. **C)** The pathway similarity landscape of KEGG visualized as a heatmap.

Case Study II: Identification of pathway modules, overlaps, and interplays using pathway enrichment

ComPath couples classic pathway enrichment analysis (Cavalli *et al*., 2017) with pathway-centric visualizations to identify modules, investigate overlaps, and cluster pathways. This case study demonstrates their use to investigate the roles of the pathways related to established genetic associations in the context of Parkinson's disease (PD).

Pathway enrichment with Fisher's exact test using a gene panel associated with PD reviewed by Brás *et al*. (the gene set will be referenced as PDgset) yielded over 300 pathways containing at least one of the panel's genes (**Figure 3A**). We discarded pathways with fewer than two genes from PDgset, that were larger than 300 genes, or that were not found to be statistically significant (false discovery rate > 5%) after applying multiple hypothesis testing correction with the Benjamini-Yekutieli method under dependency (Benjamini and Yekutieli, 2001).

Three views were used to assist in the interpretation of the remaining 29 enriched pathways: a pathway network view was used to identify pathway modules, a pathway overlap view was used to explore the intersections and cross-talks between pathways, and a pathway dendrogram view was used for clustering.

We used the pathway network view, in which nodes represent pathways and weighted edges represent their gene-based similarities (see Methods), to identify six different modules (i.e., groups of pathways) by filtering out the edges with a similarity lower than 0.2 (**Figure 3B**). The largest module (labelled as M_1) contained pathways related to the processes of endocytosis and vesicle transport; both of which are putatively disrupted in PD (Perrett *et al.*, 2015). M_2 comprised pathways related to PTK6 signaling (e.g., Reactome pathway *PTK6 promotes HIF1A stabilization;* q-value=0.0005) whose high pathway enrichment significance as well as the fact that it regulates another PDgset gene, ATP13A2, (Rajagopalan *et al.*, 2016) suggests that might be linked to PD. ATP13A2 is directly responsible for Kufor-Rakeb syndrome (Gusdon *et al.*, 2012), a rare juvenile form of PD, and participates in two other PD mechanisms: lysosomal iron storage and mitochondrial stress. Because their related pathways (i.e., *Lysosome pathway* from KEGG, *Pink/Parkin mediated mitophagy* from Reactome, and *Mitophagy pathway* from both KEGG and Reactome) were also enriched by pathway enrichment analysis, we further investigated the role of ATP13A2 in PD. It is activated by phosphatidylinositol(3,5)bisphosphate, a particular phosphatidylinositol involved in M_3 pathways (phosphatidylinositol metabolism and signaling pathways). Because this activation leads to a reduction in mitochondrial stress and α-synuclein toxicity, two hallmarks of PD, ATP13A2 has been proposed as a therapeutic target (Holemans *et al.*, 2015). Ultimately, this exploration of the similarities and cross-talks between these three modules suggests further investigation of the candidate PD gene ATP13A2.

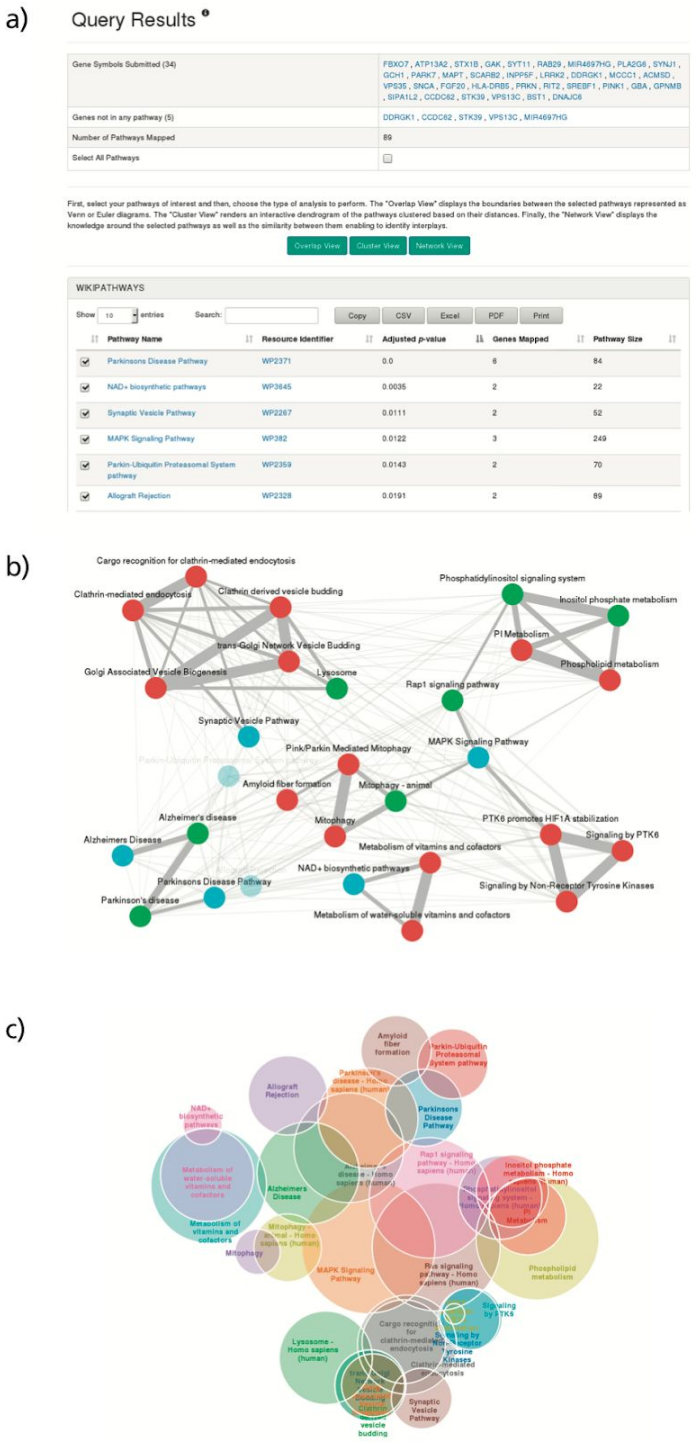

Figure 3. A) Results of pathway enrichment using the PDgset as input using the ComPath pathway enrichment wizard **B)** The Pathway Network Viewer displays the similarity around a selection of pathways. **C)** The Pathway Overlap View depicts the overlaps and intersection of pathways enriched from the PDgset.

While the pathway network viewer provides an overview of the different modules and their cross-talks, it does not reveal information about their contained pathways' boundaries and intersections. Therefore, we implemented the pathway overlap view; an interactive Euler diagram that allows exploration of pathway demarcations (**Figure 3C**). We employed this view to identify the set of genes common to all pathways in M_5, a module comprising the two Alzheimer's disease (AD) and two PD pathways from KEGG and WikiPathways. Subsequently, we used the ComPath pathway enrichment wizard to investigate in which pathways the common five genes identified (APAF1, CASP3, CASP9, CYCS, and SNCA) participate. The analysis revealed that they are predominantly involved in apoptosis, an important process in both AD and PD pathophysiology (Obulesu *et al*., 2014; Tatton *et al*., 2003).

To confirm the modules described in this case scenario, we used ComPath to conduct the clustering approach described in Chen *et al*. (**Supplementary Figure S2**). While this deterministic clustering method validated the modules previously defined in the Pathway Network Viewer, it also suggested that the three pathways that were not assigned to any module (i.e., *Allograft Rejection*, *MAPK Signaling pathway, and Rasp1 signaling pathway)* could be merged into M_2. Additionally, the resulting dendrogram revealed hierarchical relationships between pathways (e.g., *Pink/Parkin Mediated Mitophagy* is a subset of the Reactome *Mitophagy* pathway), information that can be used to establish pathway mappings, as we show in the following case study.

Case Study III: Establishing mappings between pathway databases

ComPath, as well as other tools, have demonstrated the benefits of integrating pathway knowledge from diverse resources to improve biological functional analysis (Cerami *et al*. 2010; Belinky *et al*., 2015; Kuleshov *et al.* 2016). However, even after overcoming the technical hurdle of harmonizing different formats used by different databases, these resources do not enable the interoperability necessary to investigate of pathways' interplays, cross-talks, consistency of boundaries, and gaps because inter-database pathway mappings have not yet been established.

In order to address this, ComPath introduces a curation environment in which users from the scientific community can propose and maintain a collection of established mappings between pathways from various databases. This laborious task is facilitated by the interactive visualizations (i.e., a dendrogram view and a similarity landscape heatmap) presented in the previous case studies as well as dedicated pathway pages where the content, descriptions, references, and the established mappings can be examined (**Figure 4A**). Furthermore, ComPath suggests the most similar pathways based on this information so users can propose new mappings . This new mappings are included into the mapping catalog that serves as a search interface as well as a distribution platform for mappings (**Figure 4B**). In addition, the mapping catalog promotes community engaging incorporating a voting system where authenticated users can agree or disagree on mappings; this way, proposed mappings with a net sum of votes greater than 3 are automatically registered as accepted.

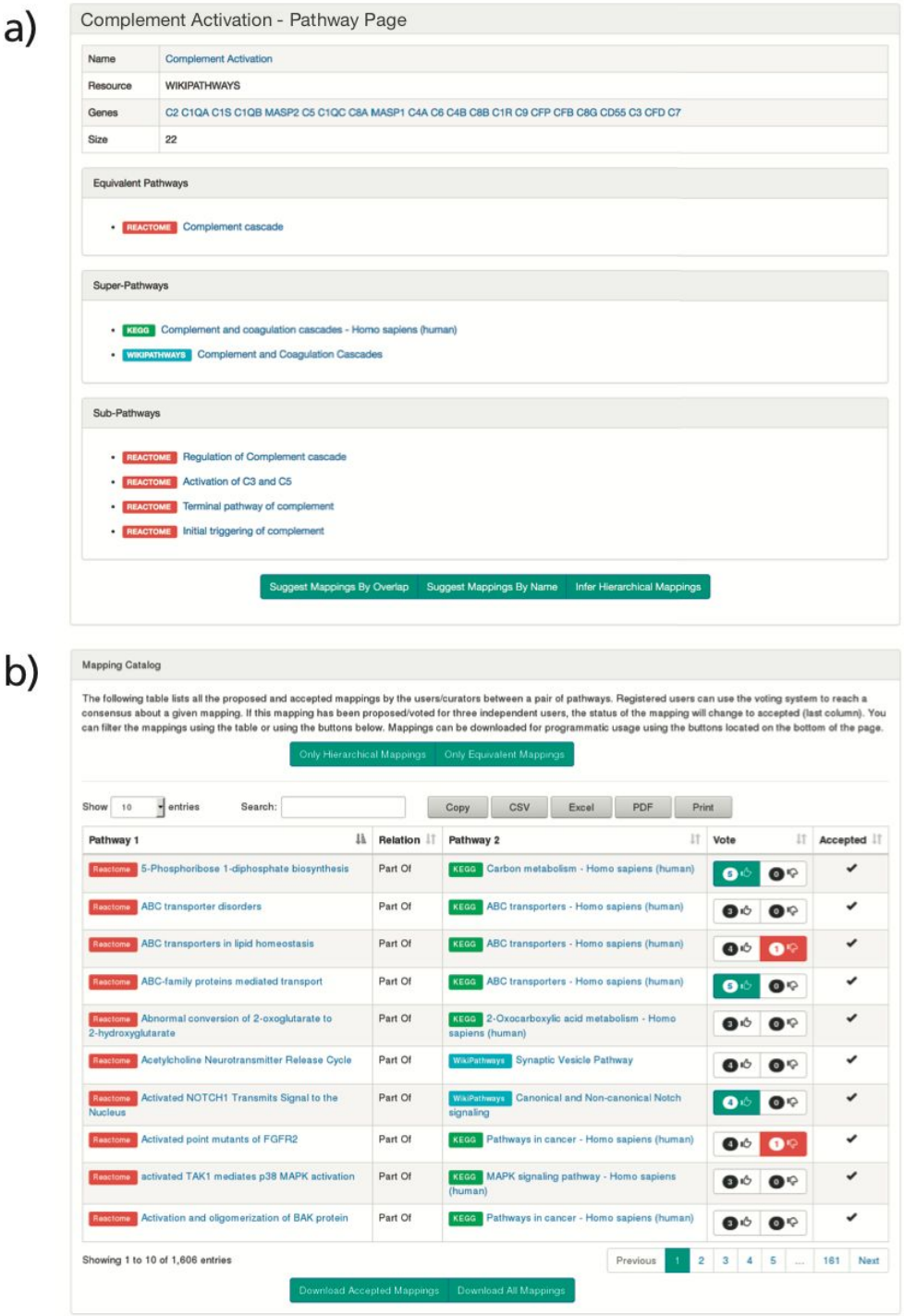

Figure 4. A) The pathway info view introduces basic pathway information such as its participating molecular entities, references, or mappings and enables automatic mapping suggestions based on different similarity metrics. **B)** The mappings view allows users to browse established mappings, propose new mappings, and give feedback on putative mappings.

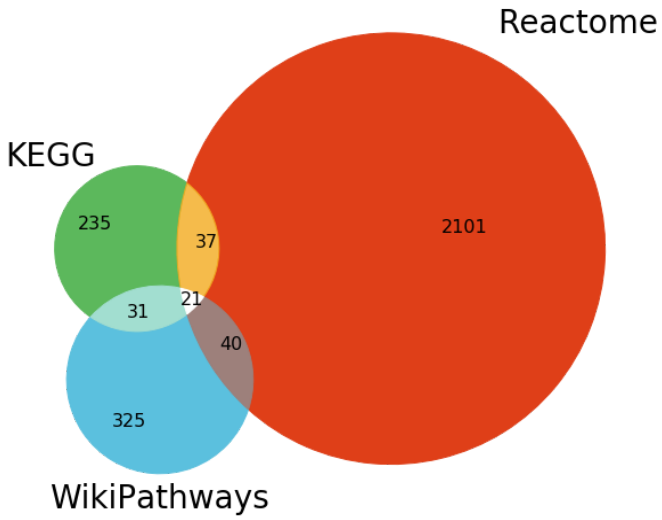

Figure 5. Venn diagram illustrating the overlaps of equivalent pathways between KEGG, Reactome, WikiPathways resulting from the curation exercise. Note: the number of overlapping pathways in the Venn diagram do not exactly match the number of equivalent mappings since there are equivalent pathways within WikiPathways that, when mapped to another database, could have more than one equivalent pathway. For example, there are two equivalent *Wnt signaling pathways* in WikiPathways that are both mapped to their corresponding Reactome pathway. This is resolved to a unique in the Venn diagram. A list of intra-database equivalent pathways is presented in the **Supplementary Table 2.**

After an exhaustive investigation of all possible mappings between pathways in KEGG, Reactome, and WikiPathways (see Methods), we identified 58 equivalencies between KEGG and Reactome, 64 between Reactome and WikiPathways, and 55 between KEGG and WikiPathways. Of these equivalent pathways, 21 are shared between the three resources (**Figure 5 and Supplementary Table 3**). We also identified 247 hierarchical relationships between KEGG and Reactome, 597 between KEGG and WikiPathways, and 564 between Reactome and WikiPathways. After considering these, approximately 26% of KEGG, 70% of Reactome, and 35% of WikiPathways did not share any mappings with any other database (**Supplementary Figure S3**). The high uniqueness observed in Reactome could be attributed to several factors: its small pathway sizes, its high granularity, and its high coverage of HGNC (**Figure 2A**).

The results of this curation effort are distributed at https://github.com/ComPath/resources and http://compath.scai.fraunhofer.de/ so they can be revised, updated, and exploited by the research community hoping that this work serves as a first endeavor towards unifying pathway knowledge.

Discussion

The lack of a lingua franca in systems biology hampers the harmonization that would enable the exploration of the coverage, agreements, or discrepancies in the pathway knowledge. Harmonizing this information is an important step to better comprehend and model biology as well as improve the bioinformatics pipelines that utilize this knowledge to elucidate biological insights. As a first step towards closing this gap, we have implemented an environment capable of accomodating the pathway knowledge from multiple databases in order to facilitate its exploration and analysis through a web application. Additionally, an embedded curation interface allow users to curate and establish mappings between pathways. Subsequently, we used ComPath to conduct an extensive curation work to link the pathways from three of the major pathway databases in order to evaluate their similarities and differences.

The common genes between KEGG, Reactome, and WikiPathways covered the majority of pathways, indicating that their pathway knowledge is partially biased towards this shared gene set, even while there are still thousands of genes that have not yet been functionally annotated to pathways. Furthermore, our curation effort revealed that a surprisingly low number of pathways (21) were equivalent between KEGG, Reactome, and WikiPathways. On the other hand, the number of mapped pathways increased significantly when the hierarchical mappings were considered, revealing the inconsistent granularity employed to delineate pathway boundaries.

Although the absence of topological pathway information in ComPath is an irrefutable limitation in this study, gene-centric approaches enable a reduction of complexity in pathway comparison as well as integration of resources which do not provide topology information (Belinky *et al*., 2015). Furthermore, recent studies revealed significant differences across a large sample of topology-based pathway analysis methods (Ihnatova, *et al*., 2018), and highlighted that gene sets alone might be sufficient to detect an enriched pathway under realistic circumstances (Bayerlová *et al*., 2015). Hence, even if the abstraction of pathways as gene sets might not exploit all the existing pathway information, it is sufficient to drive an investigation of the pathway knowledge.

The established inter-database mappings allowed to link pathways from three major databases, opening the door towards a better integration of the pathway knowledge. In the future, these links can be used to complement and fill pathway knowledge as well as to conduct a precise evaluation of equivalent or related pathways by exploiting the available format converters such as the converter from Reactome to WikiPathways (Bohler *et al.*, 2016). Furthermore, ComPath have been designed to accommodate multiple types of molecular entities participating in pathways (i.e. Reactome chemical information); thus, enabling to replicate the analyses presented with lipid or metabolite databases such as LIPEA (Acevedo *et al*., 2018) or HMDB (Wishart *et al*., 2017).

In summary, we demonstrated that ComPath serves as an exploratory, analytic, and curation framework for pathway databases. Furthemore, we showed how the ComPath web application can complement enrichment approaches to elucidate and prioritize pathways and genes related to interesting biological phenomenon. Finally, we hope that the

implementation of a curation ecosystem and the first mapping efforts conducted in this work pave the way towards unifying the pathway knowledge.

Methods

ComPath Framework

At its core, ComPath is a framework for integrating pathway and gene set databases. We defined a set of guidelines for implementing wrappers around the processes of downloading data, transforming it into a common data model, and making queries. These guidelines are encoded in an abstract class with the Python programming language such that new plugins can be quickly implemented for new resources. Each implementation must have a mapping between genes and pathways as well as functions for exporting pathways as gene sets, performing pathway enrichment analysis, and performing reasoning/inference over pathway hierarchies.

Compath Plugins

We implemented plugins for four major public pathway databases: KEGG, Reactome, WikiPathways, and MSigDB (Kanehisa *et al*., 2016; Fabregat *et al*., 2017; Slenter *et al*., 2017; Liberzon *et al*., 2011). They can be used individually as a way of extracting and exploring the pathways contained within the database. Additionally, they can be used jointly in the ComPath web application where the pathways from multiple databases are integrated for their exploration, analysis, and curation.

ComPath Web Application

The web application was implemented in the Python programming language using the Flask microframework and a suite of its extensions. The compatibility between Flask and the data models defined in all pathway plugins allows the integration and harmonization of the pathway knowledge in an extensible manner. To illustrate the flexibility of ComPath, we have included plugins for the Alzheimer's disease and Parkinson's disease gene sets associated with disease-specific mechanisms from NeuroMMSig (Domingo-Fernández *et al*., 2017) in the public version of the ComPath web (http://compath.scai.fraunhofer.de/).

ComPath leverages a variety of state-of-the-art libraries for visualization and exploration of pathway knowledge. We chose Bootstrap for the design of the website since its responsive design retains full compatibility across all devices. Interactive visualizations are generated using several Javascript libraries, including D3.js, Clustergrammer.js (Fernandez *et al*., 2017), and Cytoscape.js (Franz *et al*., 2015).

We implemented a RESTful API documented with an OpenAPI specification that can be accessed through the ComPath instance released at http://compath.scai.fraunhofer.de/apidocs. The API enables users to programmatically extract mapping information and perform queries using different genes or pathways identifiers.

Code Availability

The source code for ComPath and its plugins can be found on GitHub (https://github.com/ComPath and https://github.com/Bio2BEL) under the MIT license. Both the plugins and the web application can be installed with PyPI (https://pypi.org), the main packaging system for Python. Furthermore, we have included a Dockerfile to enable reproducing the ComPath environment with Docker (https://www.docker.com/). Finally, documentation is included in each GitHub repository and it is also accessible at Read the Docs (https://readthedocs.org).

Estimating Pathway Similarity

While a variety of indices (e.g., Jaccard, Sørensen–Dice, Tversky) have been used to assess the similarity between sets, the Szymkiewicz-Simpson coefficient is most appropriate for comparing sets widely varying in size. Similarly to previous studies, we have chosen this index to not only calculate pathway similarity but also reveal *contained* pathways (i.e., when most of the nodes from a small pathway are in a larger pathway) to indicate potential hierarchical relationships (Chen *et al.* 2014, Pita-Juarez *et al*., 2018; Belinky *et al*., 2015; Katiyar *et al*., 2018).

Curation of Pathway Mappings

Here, we describe the curation procedure we used in order to systematically generate equivalency and hierarchical mappings between pathways in KEGG, Reactome, and WikiPathways. First, we define two types of mappings:

1. **equivalentTo**. An undirected relationship denoting both pathways refer to the same biological process. The requirements for this relationship are:
    - *Scope*: both pathways represent the same biological pathway information.
    - *Similarity*: both pathways must share at minimum of one overlapping gene.
    - *Context*: both pathways should take place in the same context (e.g., cell line, physiology).
2. **isPartOf**. A directed relationship denoting the hierarchical relationship between the pathway 1 (child) and 2 (parent). The requirements are:
    - *Subset Scope*: The subject (pathway 1) is a subset of pathway 2 (e.g., Reactome pathway hierarchy).
    - *Similarity*: same as above.
    - *Context:* same as above.

We generated all possible mappings between pathways in each database (KEGG-WikiPathways, KEGG-Reactome, and WikiPathways-Reactome) and prioritized them based on the follow two independent metrics that have been proposed to calculate pathway similarity (Belinky *et al*., 2015):

1. Lexical similarity between each pair of pathways' names was calculated using the Levenshtein distance (Levenshtein, 1966).

2. Content similarity between each pair of pathways' genes was calculated using the previously described Szymkiewicz-Simpson coefficient.

After prioritization, our three curators from different areas of expertise (neuroscience, medicine, and biology) independently evaluated both similarities and the scope and context included in the pathway descriptions to assign the mapping types and to remove false positives. Furthermore, we investigated possible intra-database mappings within KEGG and WikiPathways since these resources do not yet contain hierarchical relationships. Finally, our curators combined the results and re-evaluated them to generate a consensus mapping file. It is available at https://github.com/ComPath/resources under the MIT License.

References

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, *45*(D1), D353-D361 (2016).

Fabregat, A., *et al*. The reactome pathway knowledgebase. *Nucleic acids research*, *46*(D1), D649-D655 (2017).

Slenter, D. N., *et al*. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic acids research*, *46*(D1), D661-D667 (2017)

Liberzon, A., *et al*. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, *27*(12), 1739-1740 (2011).

Krämer, A., Green, J., Pollard Jr, J., and Tugendreich, S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, 30(4), 523-530 (2013).

Demir, E., *et al.* The BioPAX community standard for pathway data sharing. *Nature biotechnology*, *28*(9), 935 (2010).

Hucka, M., *et al*. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, *19*(4), 524-531 (2003).

Cerami, E. G., *et al*. Pathway Commons, a web resource for biological pathway data. *Nucleic acids research*, *39*, D685-D690 (2010).

Petri, V., *et al*. The pathway ontology–updates and applications. *Journal of biomedical semantics*, *5*(1), 7. (2014).

Iyappan, A., *et al*. Towards a pathway inventory of the human brain for modeling disease mechanisms underlying neurodegeneration. *Journal of Alzheimer's Disease*, *52*(4), 1343-1360 (2016).

Khatri, P., Sirota, M., and Butte, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, *8*(2), e1002375 (2012).

Kuleshov, M. V., *et al*. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, *44*(W1), W90-W97 (2016).

Cary, M. P., Bader, G. D., and Sander, C. Pathway information for systems biology. *FEBS letters*, *579*(8), 1815-1820 (2005).

Subramanian, et *al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545-15550 (2005).

Bauer-Mehren, A., Furlong, L .I. and Sanz, F. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Mol. Syst. Biol*., 5, 290 (2009).

Chowdhury, S., and Sarkar, R. R. Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges. *Database*, Volume 2015, bau126 (2015)

Stobbe, M. D., Houten, S. M., Jansen, G. A., van Kampen, A. H., and Moerland, P. D. Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC systems biology*, *5*(1), 165 (2011).

Fernández, N. F., *et al*. Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Scientific data*, *4*, 170151 (2017).

Cavalli, F. M. G., *et al*. Intertumoral Heterogeneity within Medulloblastoma Subgroups. *Cancer Cel*l, 31(6), 737–754.e6 (2017).

Brás, J., Guerreiro, R., and Hardy, J. SnapShot: genetics of Parkinson's disease. *Cell*, 160(3), 570-570 (2015).

Benjamini, Y., and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165-1188 (2001).

Perrett, R. M., Alexopoulou, Z., and Tofaris, G. K. The endosomal pathway in Parkinson's disease. *Molecular and Cellular Neuroscience*, *66*, 21-28 (2015).

Rajagopalan, S., Rane, A., Chinta, S. J., and Andersen, J. K. Regulation of ATP13A2 via PHD2-HIF1α signaling is critical for cellular iron homeostasis: implications for Parkinson's disease. *Journal of Neuroscience*, *36*(4), 1086-1095 (2016).

Gusdon, A. M., Zhu, J., Van Houten, B., and Chu, C. T. ATP13A2 regulates mitochondrial bioenergetics through macroautophagy. *Neurobiology of disease*, *45*(3), 962-972 (2012).

Holemans, T., *et al*. A lipid switch unlocks Parkinson's disease-associated ATP13A2. *Proceedings of the National Academy of Sciences*, *112*(29), 9040-9045 (2015).

Obulesu, M., and Lakshmi, M. J. Apoptosis in Alzheimer's disease: an understanding of the physiology, pathology and therapeutic avenues. *Neurochemical research*, *39*(12), 2301-2312 (2014).

Tatton, W. G., Chalmers‐Redman, R., Brown, D., and Tatton, N. Apoptosis in Parkinson's disease: signals for neuronal degradation. *Annals of neurology*, *53*(S3) (2003).

Chen, Y. A., Tripathi, L. P., Dessailly, B. H., Nyström-Persson, J., Ahmad, S., and Mizuguchi, K. Integrated pathway clusters with coherent biological themes for target prioritisation. *PloS one*, *9*(6), e99030 (2014).

Belinky, F., *et al*. PathCards: multi-source consolidation of human biological pathways. *Database*, *2015* (2015).

Ihnatova, I., Popovici, V., and Budinska, E. A critical comparison of topology-based pathway analysis methods. *PloS one*, *13*(1), e0191154 (2018).

Bayerlová, M., Jung, K., Kramer, F., Klemm, F., Bleckmann, A., and Beißbarth, T. Comparative study on gene set and pathway topology-based enrichment methods. *BMC bioinformatics*, *16*(1), 334 (2015).

Bohler, A., *et al.* Reactome from a WikiPathways perspective. *PLoS computational biology*, 12(5), e1004941 (2016).

Acevedo, A., Duran, C., Ciucci, S., Gerl, M., and Cannistraci, C. V. LIPEA: Lipid Pathway Enrichment Analysis. bioRxiv, 274969 (2018).

Wishart, D. S., *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic acids research,* 46(D1), D608-D617 (2017).

Domingo-Fernández, D., *et al*,. Multimodal Mechanistic Signatures for Neurodegenerative Diseases (NeuroMMSig): a web server for mechanism enrichment. *Bioinformatics*, *33*(22), 3679-3681 (2017).

Franz, M., *et al*. Cytoscape. js: a graph theory library for visualisation and analysis. *Bioinformatics*, *32*(2), 309-311 (2015).

Pita-Juarez, Y., *et al*. The Pathway Coexpression Network: Revealing pathway relationships. *PLoS computational biology*, *14*(3), e1006042 (2018).

Katiyar, A., Sharma, S., Singh, T. P., and Kaur, P. Identification of Shared Molecular Signatures Indicate the Susceptibility of Endometriosis to Multiple Sclerosis. *Frontiers in Genetics*, *9*, 42 (2018).

Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*. 10 (8): 707–710 (1966).

Acknowledgements

This work was supported by the EU/EFPIA Innovative Medicines Initiative Joint Undertaking under AETIONOMY [grant number 115568], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies in kind contribution.

Author Contributions

M.H.A and D.D.F conceived and designed the study. D.D.F implemented ComPath and the pathway database plugins with help from C.T.H. D.D.F, C.B.A, and J.M.L curated the pathway mappings. D.D.F and C.T.H wrote the paper. M.H.A. reviewed the content.

Competing Interests

The authors declare no competing interests.

Supplement

Outline

All the links in this supplement point to resources located at https://github.com/ComPath/resources.

The supplementary information is divided into three sections based on the content: Figures, Tables, and Supplementary Text.

- Figures:
    - Genetic-centric Coverage of Pathway Databases
    - Case Study II: Dendrogram View
    - Pathways without mappings in each Database
    - Similarity Landscape of the Curated Mappings
- Tables:
    - Statistics Summary of Pathway Sizes across KEGG, Reactome, and WikiPathways
    - Equivalent Pathways within the same Database
    - Equivalent pathways across KEGG, Reactome, and WikiPathways
    - Curation examples
- Supplementary Text:
    - Case Study II: Additional Findings
    - Software Installation

Figures

Genetic-centric Coverage of Pathway Databases

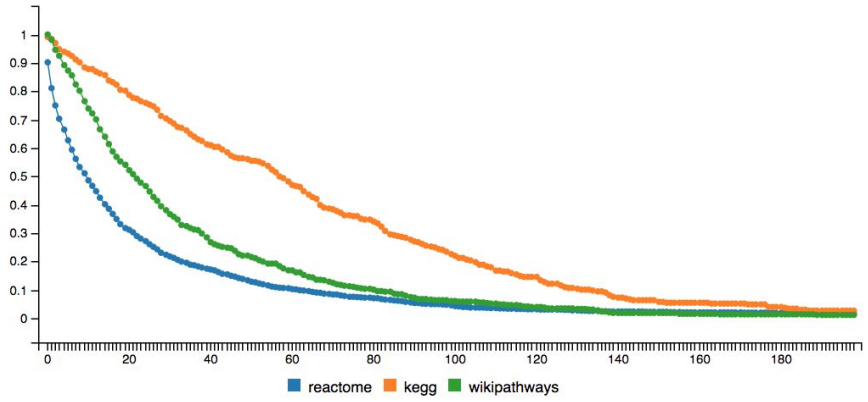

Figure 1. Relative number of pathways in KEGG, Reactome, and WikiPathways with at least X common genes must be present in a pathway. This figure can be interactively visualized at http://compath.scai.fraunhofer.de/simulation and it is described step by stept in this following Jupyter notebook: https://github.com/ComPath/resources/blob/master/notebooks/Gene%20centric%20coverage%20of%20Pathway%20Databases.ipynb

Case Study II: Dendrogram View

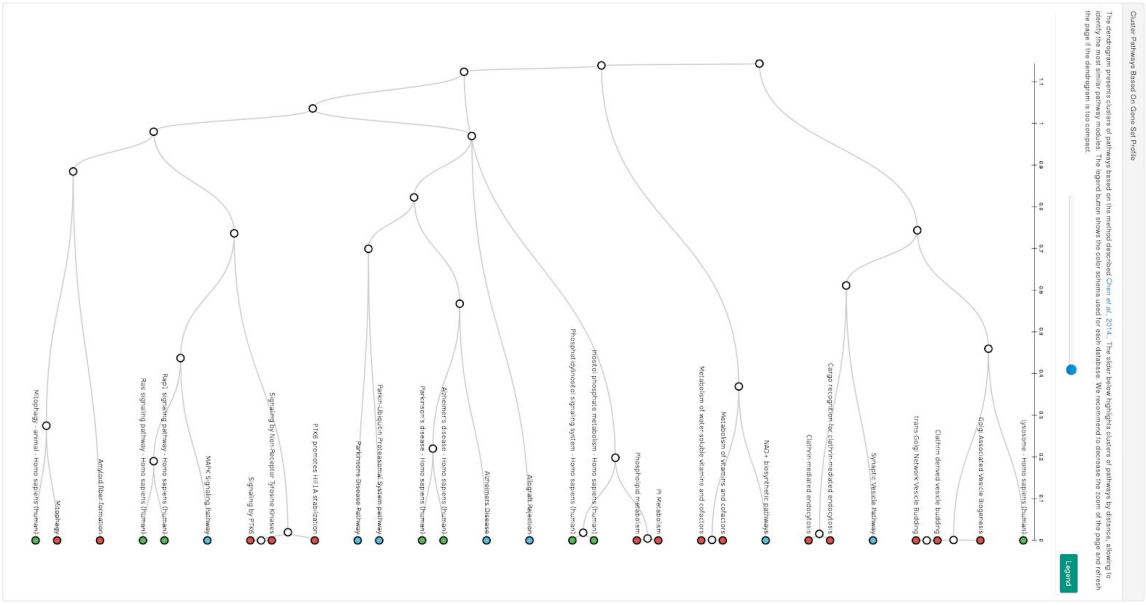

Figure 2. Dendrogram representation of the enriched PDgset pathways hierarchically clustered by their similarity. Link to the interactive visualization here.

Pathways without mappings in each Database

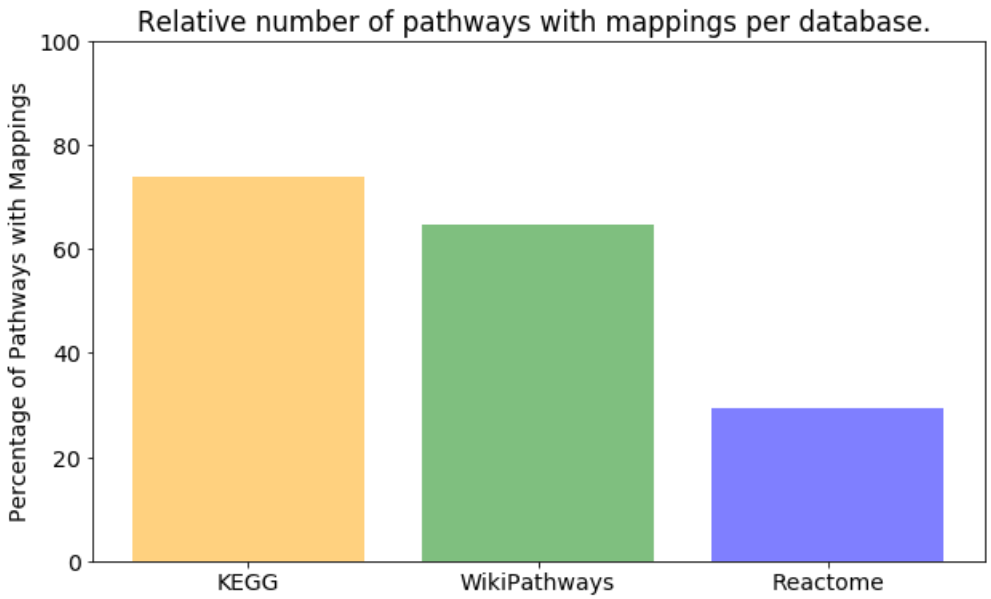

Figure 3. Number of pathways with at least one mapping with respect to the total number of pathways in each database: KEGG, Wikipathways, and Reactome. The Jupyter notebook that outlines this analysis is located at https://github.com/ComPath/resources/blob/master/notebooks/Pathways%20without%20mappings.ipynb

Tables

Statistics Summary of Pathway Sizes across KEGG, Reactome, and WikiPathways

Descriptor	**KEGG**	**Reactome**	**WikiPathways**
Mean	89.63	50.48	42.66
Unbiased variance	8915.80	20841.04	2231.94
Maximum	1273	2729	345
Minimum	1	1	1

Table 1. Statistics of the Pathway Sizes across KEGG, Reactome, and WikiPathways. The Jupyter notebook used to generate this table is available at https://github.com/ComPath/resources/blob/master/notebooks/Statistics%20Pathway%20Sizes.ipynb

Equivalent Pathways within the same Database

Database	Pathway I	Identifier I	Pathway II	Identifier II
WikiPathways	Wnt signaling pathway	WP428	Wnt signaling pathway	WP363
WikiPathways	Notch Signaling Pathway	WP61	Canonical and Non-canonical Notch signaling	WP3845
WikiPathways	Toll-like Receptor Signaling Pathway	WP75	Toll-like Receptor Signaling	WP3858
WikiPathways	Aryl Hydrocarbon Receptor Pathway	WP2873	Aryl Hydrocarbon Receptor	WP2586

Table 2. Equivalent pathways in the same database.

Equivalent pathways across KEGG, Reactome, and WikiPathways

The Jupyter notebook that outlines this analysis is located at https://github.com/ComPath/resources/blob/master/notebooks/Common%20pathways%20across%20databases.ipynb

KEGG	Reactome	WikiPathways
Mismatch repair	Mismatch Repair	Mismatch repair
Thyroid hormone synthesis	Thyroxine biosynthesis	Thyroxine (Thyroid Hormone) Production
Wnt signaling pathway	Signaling by WNT	Wnt Signaling Pathway
MAPK signaling pathway	MAPK family signaling cascades	MAPK Signaling Pathway
Pentose phosphate pathway	Pentose phosphate pathway (hexose monophosphate shunt)	Pentose Phosphate Pathway
PI3K-Akt signaling pathway	PI3K/AKT activation	PI3K-Akt Signaling Pathway
Toll-like receptor signaling pathway	Toll-Like Receptors Cascades	Toll-like Receptor Signaling Pathway
B cell receptor signaling pathway	B Cell Receptor Signaling Pathway	Signaling by the B Cell Receptor (BCR)
Notch signaling pathway	Signaling by NOTCH	Notch Signaling Pathway
Apoptosis	Apoptosis	Apoptosis
Sphingolipid metabolism	Sphingolipid metabolism	Sphingolipid Metabolism
Hedgehog signaling pathway	Signaling by Hedgehog	Hedgehog Signaling Pathway
Citrate cycle (TCA cycle)	Citric acid cycle (TCA cycle)	TCA Cycle
DNA replication	DNA Replication	DNA Replication
Non-homologous end-joining	Nonhomologous End-Joining (NHEJ)	Non-homologous end joining
Cell cycle	Cell Cycle	Cell Cycle
TGF-beta signaling pathway	Signaling by TGF-beta family members	TGF-beta Signaling Pathway

mTOR signaling pathway	mTOR signalling	Target Of Rapamycin (TOR) Signaling
IL-17 signaling pathway	Interleukin-17 signaling	IL17 signaling pathway
Synthesis and degradation of ketone bodies	Ketone body metabolism	Synthesis and Degradation of Ketone Bodies
Prolactin signaling pathway	Prolactin receptor signaling	Prolactin receptor signaling

Table 3. Equivalent pathways across KEGG, Reactome, and WikiPathways.

Curation examples

Example	Pathway 1	Pathway 2	Mapping
Same pathway with different context	AGE-RAGE signaling pathway in diabetic complications (KEGG)	AGE/RAGE pathway (WikiPathways)	No Mapping
Description provides additional information	Circadian rhythm related genes (WikiPathways) -	Circadian rhythm (KEGG)	Circadian rythm related genes (WikiPathways) isPartOf Circadian rhythm (KEGG)
Valid Hierarchical relationship	Alanine, aspartate and glutamate metabolism (KEGG)	Amino Acid metabolism (WikiPathways)	Alanine, aspartate and glutamate metabolism isPartOf Amino Acid metabolism
Valid Equivalent mapping	Mismatch repair (KEGG)	Mismatch Repair (Reactome)	Mismatch repair (KEGG) equivalentTo Mismatch Repair (Reactome)

Table 4. Curation exercise real examples.

Supplementary Text

Case Study II: Additional Findings

M_4 is composed of mitophagy-related pathways and *amyloid fiber formation*, one of the main hypothesis in Alzheimer's disease (AD), in concordance with M_5 that comprises the two pairs of *AD and PD disease pathways* from KEGG and WikiPathways. In contrast to what one could expect, the most similar pathways in this module are the AD and PD KEGG pathways, and not the more natural and expected higher similarity between the two pairs of AD and PD equivalent pathways. Furthermore, the analysis also spotlighted that the WikiPathways *MAPK signaling pathway* from WikiPathways is central between M_2, M_3, M_4; thus, indicating that the three modules might shared cross-talks with MAPK signaling. Finally, the M_6 comprised pathways related to vitamin and cofactors metabolic pathways that also implicated in PD (Etminan *et al.*, 2005; Fariss *et al.*, 2003; De Lau *et al.*, 2006).

Software Installation

All packages described in the manuscript are available through GitHub (https://github.com/) or PyPI (https://pypi.org), the main packaging system for Python 3, under the MIT license. All relevant information for installation is bundled in the package, so it can be easily and quickly installed independently of the operating system, running any modern version of the Python programming language. The documentation for all packages was built using the Python documenting tool Sphinx and is accessible at Read The Docs (https://readthedocs.org).

References

Etminan, M., Gill, S. S., and Samii, A. Intake of vitamin E, vitamin C, and carotenoids and the risk of Parkinson's disease: a meta-analysis. *The Lancet Neurology*, *4*(6), 362-365 (2005).

Fariss, M. W., and Zhang, J. G. Vitamin E therapy in Parkinson's disease. *Toxicology*, *189*(1-2), 129-146 (2003).

De Lau, L. M. L., Koudstaal, P. J., Witteman, J. C. M., Hofman, A., and Breteler, M. M. B. Dietary folate, vitamin B12, and vitamin B6 and the risk of Parkinson disease. *Neurology*, *67*(2), 315-318 (2006).