Hanspers *et al. Genome Biology* (2020) 21:273
https://doi.org/10.1186/s13059-020-02181-2

Genome Biology

DATABASE Open Access

Pathway information extracted from 25 years of pathway figures

Kristina Hanspers[1†], Anders Riutta[1†], Martina Summer-Kutmon[2,3] and Alexander R. Pico[1*]

*Correspondence:
alex.pico@gladstone.ucsf.edu
†Kristina Hanspers and Anders Riutta contributed equally to this work.
[1]Institute of Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA, USA
Full list of author information is available at the end of the article

Abstract

Thousands of pathway diagrams are published each year as static figures inaccessible to computational queries and analyses. Using a combination of machine learning, optical character recognition, and manual curation, we identified 64,643 pathway figures published between 1995 and 2019 and extracted 1,112,551 instances of human genes, comprising 13,464 unique NCBI genes, participating in a wide variety of biological processes. This collection represents an order of magnitude more genes than found in the text of the same papers, and thousands of genes missing from other pathway databases, thus presenting new opportunities for discovery and research.

Keywords: Pathways, Figures, Literature, OCR, Gene sets

Background

The molecular mechanisms underlying biology are often outlined as pathway diagrams. In textbooks and on whiteboards, these depictions are fundamental to a biologist's training. As mental models for practitioners, they serve as scaffolds for hypotheses and integrating new knowledge. And in the scientific literature, pathway figures are the pinnacle of communication for published work, synthesizing diverse sources and types of data spanning decades into a coherent model. Though often published only as static images, pathways express dynamic interactions. Common examples include metabolic cycles, gene regulation, and signaling cascades. Depicted interactions play out over a spectrum of electrochemical, enzymatic, and developmental timescales.

When properly modeled as an interaction network and annotated with standard identifiers, pathway knowledge can be conveyed with greater precision in formats amenable to computational analysis. Distinct from static images, pathway models can be used in enrichment analyses [1], enhanced data visualization [2, 3], knowledge graphs [4, 5], biomedical inference [6], and database queries [7, 8]. Over the past couple decades a number of pathway databases, including GenMAPP [9], MetaCyc [10, 11], KEGG [12] and Reactome [13, 14] took on the challenge of curating canonical pathway biology, each with their own unique focus and approach. A broader, community-curated approach was

undertaken by WikiPathways to allow any researcher to model and freely share their pathway knowledge [15–17]. And under an even broader umbrella, the NDEx database provides access to not only pathways, but also diverse types of network models, offering DOI minting for citation [18]. Despite the continued growth and active usage of these database efforts, the vast majority of pathway knowledge is still captured in static images submitted solely to publishers as figures. We estimate 1,000 pathway figures are indexed by PubMed Central (PMC) *each month* in recent years and less than 3% of these are sourced from a pathway database [19].

In this study, we have identified pathway figures published over the past 25 years and characterized their content in terms of recognized gene symbols by optical character recognition (OCR). While it is more common to process text from the abstract and body of papers in order to extract genes and other biological concepts including interactions, knowledge extraction from published pathway figures is relatively rare and incomplete [20–22]. In a pilot study of 4,000 pathway figures [19, 23], we developed a custom OCR pipeline and identified over twice as many unique human genes as detected in the text by PubTator [24]. The gene content extracted from this limited sample of pathway figures reproduced two thirds of the database content at WikiPathways and included over a thousand human genes not previously annotated in pathway models. Remarkably, no two pathways were identical among this set of 4000 figures. A wealth of novel and diverse pathway knowledge is essentially trapped in published pathway figures.

The goal of this work was to identify the human gene content in a comprehensive collection of published pathway figures, to characterize its biological relevance, and to increase meaningful, FAIR [25] access to this pathway knowledge resource. In the end, 65k pathway figures were found in publications from the past 25 years with over a million mentions of human genes identified in total. Of the 13.5k unique human genes identified, over a quarter had yet to be annotated in WikiPathways or Reactome databases. The biological relevance of the identified gene sets was assessed by performing enrichment analysis against annotated gene sets using Gene Ontology and an extensive disease ontology showing both diversity and depth. Finally, a series of usage examples demonstrate the potential of this content to enhance literature searches, elucidate the history of scientific discovery and support enrichment analyses.

Content

We identified and characterized 64,643 pathway figures published between the years 1995 and 2019. Starting with 235,081 figures from a PMC image query that specified the 25-year date range and keywords covering diverse types of pathways, machine learning was applied to more precisely distinguish figures containing molecular interaction diagrams from those depicting other types of pathways (e.g., neuronal pathways) or pathway-related content (e.g., pathway enrichment results).

Relying solely on the linearly ranked PMC results (i.e., without subsequent machine learning steps) would have resulted a relatively *diluted* set of figures containing a high proportion of non-pathways. Two rounds of machine learning effectively *concentrated* actual pathway figures. The second and final round relied on a set of 15,406 figures manually classified by a domain expert to train a model distinguishing pathway figures from other figures with 91.88% precision, 91.88% recall, and a Matthews Correlation Coefficient [26] of 0.82. The resulting set of 64,643 pathway figures is defined as our "65k set" used in this

study. The 65k set of pathway figures was ultimately assessed to consist of 94% pathways (±3% at 97% confidence) by manual classifying a random sample of 300 figures.

Papers containing pathway figures

Prior to any gene detection by OCR, the papers containing the 65k set of pathway figures were characterized by publicly available annotations. The pathway figures came from 56,095 papers authored by 216,542 unique authors and are published in 3,453 journals. Obviously, not all co-authors are involved in preparing a pathway figure in a given paper, but for comparison, the most successful effort to crowdsource pathway knowledge by the WikiPathways database has fewer than 800 unique authors.

The papers containing pathway figures can be characterized by paper-level annotations, for example disease ontology terms from Europe PMC and genes recognized in the text by PubTator. A subset of 29,187 (52%) pathway-figure containing papers had at least one disease ontology terms annotation from Europe PMC. The top 10 most frequent disease ontology terms annotating these papers are Cancer (39% of 29,187 papers), Infection (19%), Defects (15%), Tumor (9.3%), Diabetes (4.3%), Hypoxia (2.0%), Depression (1.3%), Obesity (0.8%), Ischemia (0.7%), Atherosclerosis (0.4%), and Other (9.2%). PubTator extracts genes and other concepts from the abstract and body text of PMC-indexed papers by natural language processing [24]. While figure captions may be included in text-based approaches, the figure images are not included in any way, as they require OCR and custom normalization prior to entity recognition. According to PubTator, 30,036 (53.5%) of pathway-figure containing papers had at least one gene found in the text with an average of 3.4 genes per paper. The top 10 genes found in the text of these papers are AKT1 (5.4% of 30,036 papers), MTOR (4.1%), TP53 (3.7%), MAPK1 (3.5%), TGFB1 (2.8%), PIK3CD (2.8%), EGFR (2.6%), TNF (2.4%), CTNNB1 (1.9%), and MAPK3 (1.7%). The majority of these genes match the dominant disease annotation for cancer-related biological processes. The prevalence of cancer in this set of papers is greater than that for PMC in general, which is estimated at 12% [27]. This bias could reflect the popularity and particular effectiveness of pathway diagrams in depicting cancer-related signaling and metabolic processes. It could also result from a bias in the pathway ontology [28] used to construct the original PMC query, which included "cancer pathways."

Genes in pathway figures

Having identified and characterized a set of papers containing pathway figures, the primary goal of this study was to extract their human gene content by an OCR pipeline customized for pathway figures [19, 23]. The approach targeted the most common ways authors refer to genes, proteins, complexes, and families in pathway figures, leveraging current, alias, and previous HGNC symbols as well as a curated collection of conventional bioentity names that have been mapped to official HGNC symbols. There is a wealth of information visually conveyed by pathway figures, including localization, reactions, cascades, cycles, co-factors, metabolites, drugs, and other biological concepts. The most tractable and widely used content—even from fully-annotated pathway models—is the gene content. The other components and concepts in pathway figures remain worthwhile pursuing in future work, but knowing the gene content on its own transforms a collection of static images into a resource with diverse research applications.

Of the 64,643 pathway figures identified by image classification, 58,962 (91%) figures had at least one human gene recognized by our pathway figure OCR pipeline. A total of 1,112,551 instances of human genes were recognized, consisting of 13,464 unique human NCBI Genes. On average, there were 18.9 genes recognized per figure, compared to only 3.4 genes recognized in the text of the same papers by PubTator. PubTator found a tenth as many genes (101,617) in the text of these same papers overall; only half of the papers (53.5%) mentioning one or more genes in the text. In our pathway figure OCR results, there were over 600 figures with more than 100 genes each. While many of the largest figures are interaction networks, the largest with 385 recognized genes is an augmented KEGG pathway where the authors properly listed the individual gene family and paralog members referenced generically in the original [29]. At the other end, there was a long tail of just over 20k (37%) figures that had fewer than seven recognized genes.

The top 10 human genes identified in pathway figures representing unique families are MAPK1 (15% of 58,962 figures), AKT1 (14%), PIK3CA (10%), NFKB1 (8.9%), KRAS (7.6%), MTOR (7.5%), MAP2K1 (6.2%), TNF (5.6%), RAF1 (5.3%), and TP53 (5.1%). Compared to genes extracted from the text of the same set of papers containing pathway figures, the same trend of cancer-related biology dominates the content. In more detail, this top 10 set includes four of the top five genes found in the text of the same papers by PubTator: MAPK1, AKT1, MTOR, and TP53. Figure 1 presents the results of a set analysis performed on these overlapping genes found in text and figures, showing approximately 50% overlap with respect to text occurrences (i.e., half of the occurrences in text were also found in the pathway figures of the same papers) and overall a far greater number of occurrences in figures; the occurrence of combinations of these genes being almost exclusive to figures. Of the 13,464 unique human genes identified in the figures, half (6564 or 49%) were not identified in the text of any of the papers by PubTator. Compared to pathway databases, over a quarter of the unique genes recognized in pathway figures (3710 or 28%) were not present in either WikiPathways nor Reactome collections (as of January 2020). Clearly, pathway figures represent biological models that are not fully described in the text nor captured in curated pathway databases.

Sets of genes in pathway figures

Seeking to optimize across coverage, performance, and interpretability, we defined a subset of figures with at least seven distinct NCBI Genes. Compared to the overall set, these 28,836 pathway figures contained 13,216 (98%) unique genes, thus retaining the coverage and novelty of the collection. Among these 28,836 pathway figures, 28,520 (99%) were significantly associated with at least one Gene Ontology (Biological Process) term by enrichment analysis, indicating general biological relevance. In terms of disease relevance, we found 20,227 (70%) pathway figures significantly associated with at least one disease ontology term [30], and 98% of disease terms in the ontology (157/160) were represented by one or more figures. Not surprisingly, the top disease term was Cancer (42% of 20,227 figures). Manual inspection of paper and figure titles verified 64 different disease terms as accurate annotations for as many as 8,419 pathway figures. In addition to Cancer, prevalent terms included Cardiomyopathy, Lung cancer, Melanoma, Breast cancer, Rheumatoid arthritis, Diabetes mellitus, and Neurodegenerative disease. The gene information extracted from pathway figures allows for a new way to annotate the literature by enrichment analysis against any gene set-based ontology or resource, such as

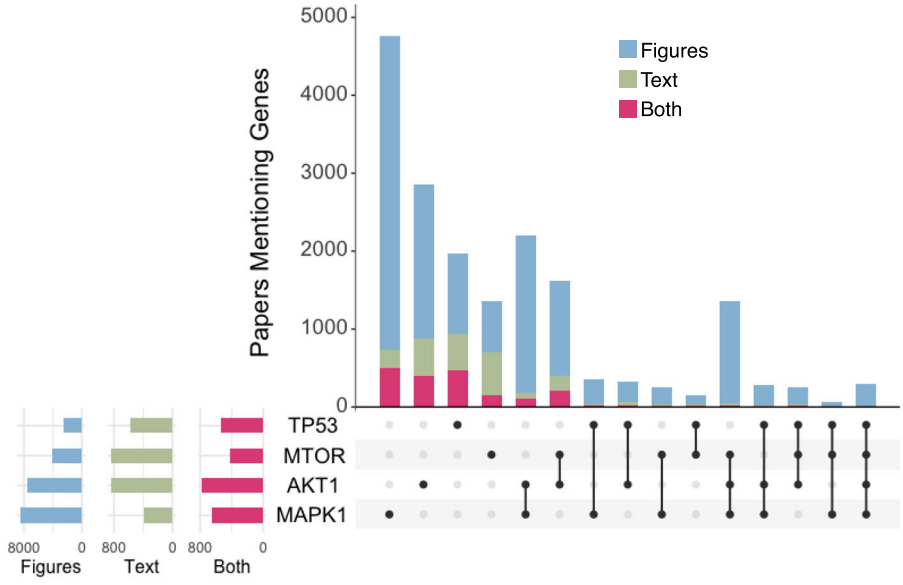

Fig. 1 UpSet plot of the most common genes found in the text of pathway figure-containing papers. The stacked bar plot (top) shows the number of papers that include one or more of these genes in either a pathway figure (blue), the text via PubTator (green), or both (red). The matrix below the bar plot indicates which sets of genes are represented by each bar. The bar plots on the left show the total number of genes in either pathway figures, text, or both using independent scales

Gene Ontology, OMIM, MSigDB, or even other pathway databases like WikiPathways and Reactome.

Utility

The initial characterization of published pathway figures and their gene content revealed a novel resource of relevant pathway knowledge that is practically inaccessible to researchers. The following examples demonstrate how this resource could be utilized in a variety of applications.

Searching scientific literature

With over 600,000 papers added to PMC in the last year, researchers are resigned to merely sampling the work most relevant to them via search engines, feeds, subscriptions, and recommendations. While figure captions are accessible to text-based processing and indexing, the actual contents of figures remain hidden to any commonly available search engine (e.g., PubMed, PMC, Europe PMC or even Google). The systematic identification of genes in pathway figures enables access to this content through new and existing tools.

Literature search tools

Search engines commonly index papers by the genes found in the abstract, body, and caption text. Europe PMC goes further by supporting community-contributed mappings between genes and papers (https://europepmc.org/annotations). While these typically derive from text-based processing, the same deposition and integration system could be used to accept gene-paper mappings based on figures. Querying one or more genes at

Europe PMC would then return paper results containing both text and figure references to those genes.

As another example, the Chan Zuckerberg Initiative is working on a new literature feed service called Meta (https://meta.org; in open beta) that processes the latest publications and preprints on a daily basis. The gene-paper mappings from pathway figures would be a natural fit for an indexing system that links papers via their contents. Furthermore, the characterization of gene sets as demonstrated above could provide mappings from papers to disease ontology terms and other gene-based and pathway-based annotations. A feed service that could produce a regularly updated set of papers that contained relevant pathway figures would be a welcome innovation to researchers attempting to stay abreast of scientific literature.

Interactive pathway figure app

We produced an online tool using R Shiny (https://gladstone-bioinformatics.shinyapps.io/shiny-25years) to enable filtering, searching, and viewing the full collection of 65k pathway figures by enriched disease terms, genes, date, and various publication metadata fields (Fig. 2). Organized into three stages, the first stage offers auto-complete fields to define OR-based filters for disease annotations, gene content and publication years, and displays bar plots of the top 40 disease ontology terms, top 40 human genes, and publication dates represented by the currently filtered set of figures. The second stage displays a paginated table view of the currently filtered set of figures, each row representing a pathway figure and its parent paper. The columns can be used to sort and query within the table to further refine the current set. Selecting a row in the table will update the third stage, which displays the pathway figure, a link to PMC and table of recognized genes. The gene table includes the symbol found in the figure and the source of the lexicon it matched along with the official HGNC symbol and NCBI Gene identifier. This tool provides a way to easily query figures of interest given a set of genes or topic. For example, in the preparation of this manuscript the tool was used for the "History of scientific discovery" section on relating to the Hippo Signaling Pathway.

As a topical demonstration of the tool, a second version was produced focusing on COVID-19 related pathways as defined by the COVID-19 Open Research Dataset (https://gladstone-bioinformatics.shinyapps.io/shiny-covidpathways). There are 221 pathway figures in this collection that can be rapidly queried and viewed by the same three-stage procedure described above and in Fig. 2. This tool has already proven useful in building SARS-CoV-2 pathways at WikiPathways (http://covid.wikipathways.org) as part of the COVID-19 Disease Map initiative (https://covid.pages.uni.lu/map_curation) [31].

Knowledge graph query paths

This source of annotated pathway information is also finding utility in an advanced platform for distributed knowledge integration. The BioThings Explorer platform includes a collection of APIs semantically defining inputs and outputs that comprise a knowledge graph (https://biothings-explorer.readthedocs.io) [5]. The platform also includes an engine that supports queries that traverse paths through the graph, e.g., *drugs* that bind *components of pathways* associated with a *disease of interest*. By defining an API that recognizes standard paper and gene identifiers in a JSON file export of pathway figure-based gene sets, the 1.1M gene-paper links from this study can be used to bridge query paths

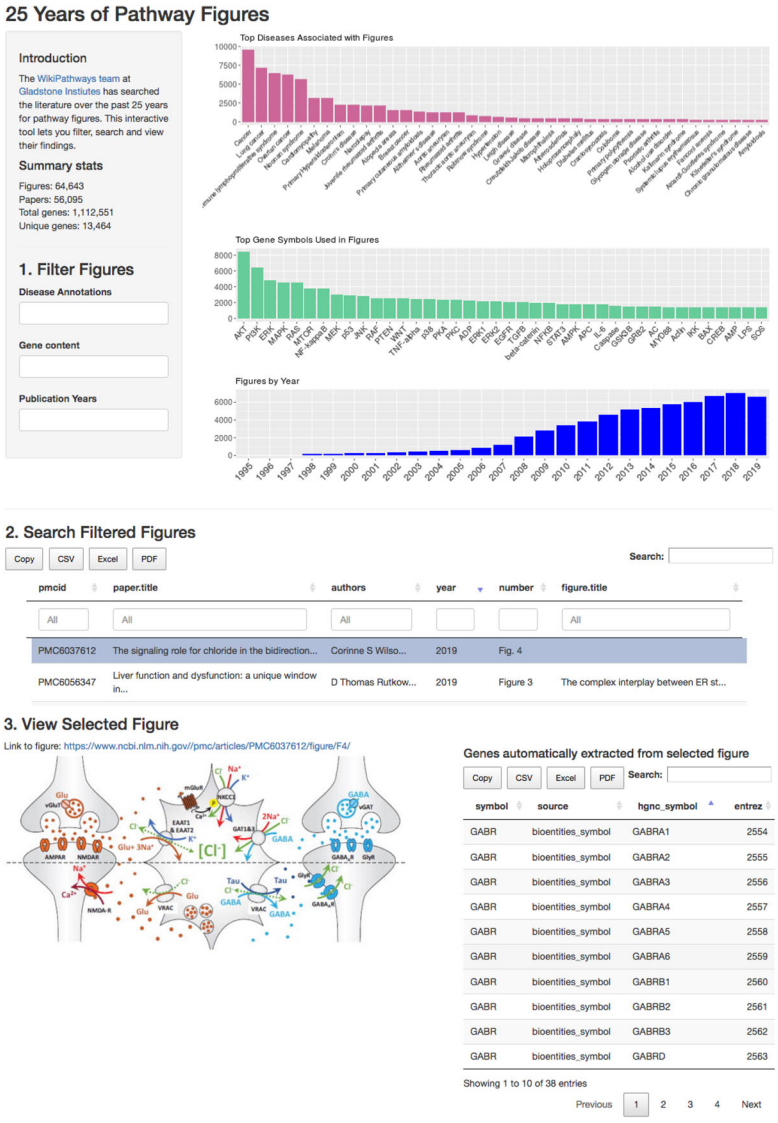

Fig. 2 Screenshot of interactive query tool for pathway figures. The tool allows for easy querying of the full set of 65k pathway figures by any combination of gene content, publication year, disease annotations, figure and paper title, authors, etc.

to other content in the knowledge graph. The addition of disease annotations and future extraction of metabolites, drugs, and concepts from the OCR results will establish additional bridges and reinforce paths inferred by computational reasoning. This work is in active development as part of the NCATS Biomedical Data Translator program, targeting critical use cases, such as drug repurposing (https://ncats.nih.gov/tidbit/tidbit_04.html).

History of scientific discovery

Given the 25-year span of this new pathway resource, it was natural to reflect on the role of pathway figures in scientific discovery. The R Shiny app previously described can be used by any researcher or historian to investigate the story of particular diseases and genes

from a pathway perspective. Tracing the development of the Hippo signaling pathway as one example demonstrates this potential (Fig. 3).

The Hippo signaling pathway controls organ size in animals by regulation of cell proliferation and apoptosis. The components of the Hippo signaling pathway are highly conserved [32], and many of the early discoveries were made via Drosophila genetic screens. The Hippo signaling pathway includes a central kinase signaling cascade, where MST1/2 (Hpo) phosphorylates LATS1/2 (Wts), which activates it. Activated LATS1/2 phosphorylates YAP/TAZ (Yorkie/Yki), leading to its inactivation and degradation in the cytoplasm. When activated, YAP/TAZ translocates to the nucleus and binds to several transcription factors, including TEADs, leading to transcription of proliferation and survival genes. Phosphorylation of LATS1/2 is facilitated by binding to SAV and MOB1. The Hippo pathway can be activated by many different stimuli including cell density and polarity, mechanical sensation and soluble factors via upstream regulators WWC1 (Kibra), NF2, etc. Cross-talk with multiple pathways is known, including TGF-β, Notch, and Wnt signaling. The Wts gene (LATS1 in humans) was discovered in Drosophila in 1995 [33, 34], and the first mention of a pathway involving Wts, Sav, and Hpo was in 2003 [35, 36], initially termed the Salvador/Warts pathway. From 2003 on, several discoveries were made which further defined the pathway components and process [37].

In our 65k pathway figure collection, the first published figures representing the Hippo signaling pathway appeared in 2007 [38, 39], more than a decade after the initial discovery of the central Wts gene, and after the core components were described in the literature (Fig. 3). The early published pathway figures are sparse and some even include question marks for components that are not yet known [38]. Discoveries of specific components

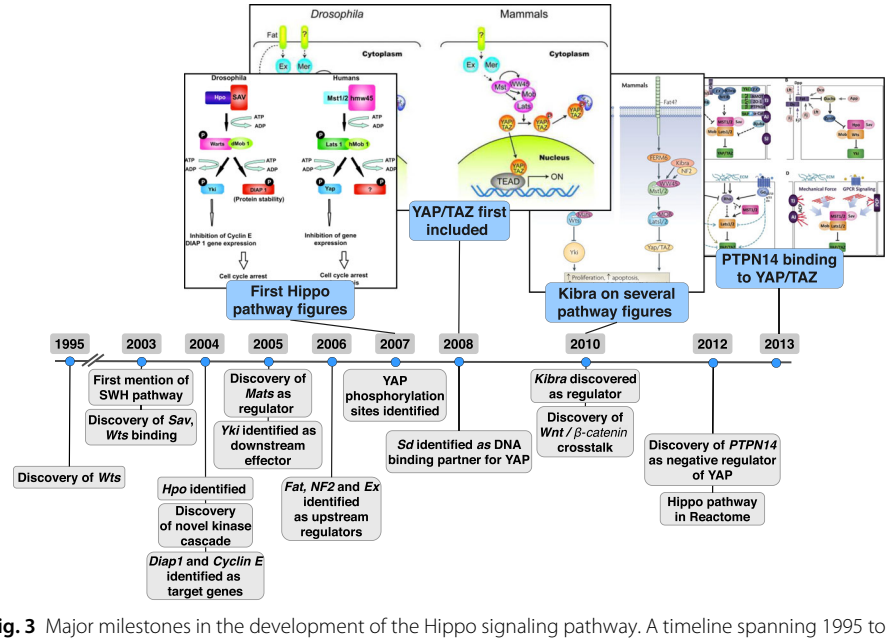

Fig. 3 Major milestones in the development of the Hippo signaling pathway. A timeline spanning 1995 to 2013 focuses on the discovery of the components and processes of the Hippo signaling pathway (below timeline), illustrated by representative pathway figures that grow in complexity and detail (above timeline). The events depicted in gray below the timeline were collected mostly from Kim et al. [37]. Drosophila and human orthologs are Wts:LATS1/2, Sav:SAV1, Hpo:MST1/MST2(STK3), Mer:NF2, Fat: FAT1, Ex: FRMD6, Yorkie/Yki:YAP(YAP1)/TAZ(WWTR1), Sd:TEAD1-4, Kibra:WWC1

are in some cases followed by a pathway figure from an independent publication, adding that component to the pathway. For example, discovery of the binding of PTPN14 to YAP in 2012 [40] is followed in 2013 by a pathway figure showing the interaction [41]. Another interesting observation is related to the Kibra gene (WWC1 in humans), which was first characterized in a yeast two hybrid screen in 2003 [42]. It was investigated in a variety of contexts (cytoskeleton, memory function, etc.), and in 2010, Kibra was shown to be an upstream regulator of the Hippo pathway [43]. Interestingly, there were no pathway figure hits for Kibra before 2010, but from 2010 on, the number of pathway figures including Kibra has grown steadily, with the vast majority of them representing the Hippo pathway [44]. Yorkie (YAP/TAZ in humans) was first indicated as the transcriptional activator of the Hippo pathway in 2005 [45] and it was subsequently found that phosphorylation of YAP at S127 inhibits transcriptional activity by retaining YAP in the cytoplasm [46]. These critical findings were followed in 2008 by the first pathway figure showing the details of YAP/TAZ signal transduction [47].

Access to pathway diagrams and their gene contents provides a new way to track the discovery and depiction of key molecules and interactions for a given pathway over time; many of these pathway figures predate any pathway database. For the Hippo pathway example, there were 31 representations of the pathway in published literature spanning the 13 years prior to its first entry in a database in 2012 (https://reactome.org/content/detail/R-HSA-2028269). Even with the advent of pathway databases, evolving and emerging pathway knowledge can be accessed via our pathway figure OCR pipeline once published and publicly indexed, whereas traditional database curation might lag by months, by years or indefinitely. In the next section, this example is also used to demonstrate the utility of having access to this history in a computational format for the first time, enabling clustering, visualization, and other applications.

Trends for all pathways and their gene content can also be explored. For example, by using data gathered from OMIM [48], the time spanning the initial cloning of a gene and its first appearance on a pathway can be determined. A decade in the case of the Wts (LATS1/2 in human) gene and the Hippo signaling pathway, the cloning-to-pathway time spans for all 13,464 genes has an overall median of 12 years. By comparison, the time spanning the initial cloning of a gene and its first biochemical feature characterization also has a median of 12 years.

Pathway figure enrichment analysis

The 65k set contains more unique human genes and greater contextual depth than any pathway database, providing an interesting new resource for pathway analysis. However, to make the set usable for enrichment analysis like overrepresentation and gene set enrichment analysis, there are aspects of gene content and redundancy to consider.

Even though certain methods normalize enrichment scores for gene set size, the process is not accurate for extremely small or extremely large gene sets. While large sets are not an issue in our pathway figure-based gene sets, approximately half of the figures have fewer than 10 genes. Filtering the 65k set with a cutoff of at least 10 unique genes, a set of 32,277 (49%) pathway figures can be defined for use in enrichment analysis. By contrast, only 1072 (1.9%) papers had 10 or more genes identified in the text by PubTator. Also of note, this set of 49% of the largest pathway figures retains 97% (13,153) of the unique human genes found in the overall 65k collection. Among the set of 32,277 pathway figures

identified for use in enrichment analysis, there were 878 figures that shared the exact same gene content with at least one other figure and 4,937 figures entirely contained by one or more other pathway figures.

In order to assess the redundancy and hierarchical structure among pathway figures in more detail, a sample subset of 55 Hippo signaling pathway figures was clustered by gene overlaps (Fig. 4). The overlap (intersection/number of genes in gene set) and Jaccard index (intersection/union) [49] were calculated between each pair of gene sets. No two pathways in this set were identical, but many pathway figures contained the contents of smaller pathway figures, analogous to the nesting of Biological Process terms in Gene Ontology. The "Core" cluster (red), for example, contains four small figures each with the defining set of Hippo signaling pathway genes, which are an essential component of all the figures in this set; thus, the high scores across top four rows. The "Meta" cluster (purple) has 25 large figures associating multiple pathways with Hippo signaling. Though there is much lower similarity for this cluster, a few bright subclusters along the diagonal indicate sets of pathway figures with high mutual overlap. The "Plus" cluster (green) has 26 small-to-medium figures that contain just a few additional genes interacting with the core Hippo signaling pathway. The dendrogram and heatmap for this example set of pathways illustrate a novel way to conceive of a pathway in general, that is, as a hierarchically organized collections of core and peripheral components rather than as a singular summary diagram.

For the purpose of an enrichment analysis, any of these pathways have the potential to provide a highly specific result with greater context and interpretability than a single, so-called canonical, Hippo signaling pathway. This potential argues against additional prefiltering of the pathway figure set. Furthermore, many enrichment tools already employ

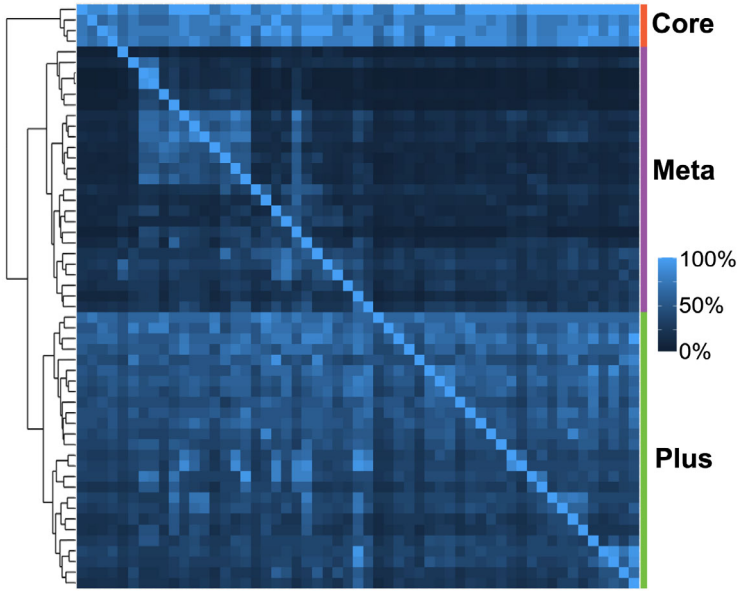

Fig. 4 Overlap among Hippo signaling pathway figures. A matrix view of the directional overlap among 55 pathway figures containing at least five core genes involved in hippo signaling. The dendrogram (left) identifies various levels of clustering. The top three clusters are annotated by labels and a color bar for reference (right)

post-filtering of results by, for example, a Jaccard distance measure to handle ontologies with even greater redundancy and higher degree of nesting. The same approach could optionally be applied to enrichment results from these pathway figure-based gene sets as part of a researcher's exploratory analysis.

Discussion

There is a vast resource of pathway knowledge trapped in the form of published pathway figures. We have identified these figures and extracted their gene content to make this resource accessible and to demonstrate its potential to enhance biomedical research. We found 64,643 pathway figures in publications dating from 1995 to 2019 and identified 1,112,551 occurrences of human genes in total (13,464 unique NCBI Genes). Much of this content represents novel pathway knowledge that is present neither in the text of papers, nor in pathway databases. The extraction of drugs, metabolites, and disease terms from these same pathway figures is now a more tractable project, as is the identification of genes from other species, by expanding the lexicon used to match against the OCR results. A greater challenge lies in the extraction of interactions, subcellular compartments, and other graphical representations of biology beyond just the molecules. While the OCR of text from images is a well-studied problem with widely available solutions, a more customized machine learning approach will be required for these non-standard graphics. Adding to the challenge, the process of modeling interactions in a pathway involves more than simply retracing what is typically provided in a pathway figure. A certain level of domain expertise is required for curators of pathway databases as their task requires independent interpretation and investigation. Nevertheless, the refined collection of gene-characterized pathway figures presented here can be used as a reference dataset by researchers interested in this challenge.

Despite unlocking the basic contents of published pathway figures, this work only partially mitigates gross deficiencies in current pathway knowledge representation and communication. *Oh, the figures we have seen!* Having scanned many thousands of pathway figures as a collection, the most obvious point to make is that standardization is desperately needed. Standards for pathway models have been around for decades [50–52] and have been implemented in a variety of freely available software tools [53–57]. Likewise, standard practices for the deposition and sharing of scientific models are well established (e.g., sequences, structures and ontologies). We recommend that authors make use of these standards and we implore reviewers, editors, journals, and funders to encourage and enforce the application of good scientific publishing practices to pathway knowledge. This recommendation applies generally to the publication of pathways, networks, and other models of system biology consisting of identifiable entities and their relationships. By using proper modeling tools, pathway knowledge can be databased, indexed, shared, and used more effectively and FAIRly [25].

In the meantime, the *post hoc* extraction of knowledge from published pathway figures can serve to make this content more findable, accessible, interoperable, and reusable. The gene sets extracted from these figures can be indexed to enhance literature searches, they can create and reinforce links in knowledge graphs, they can inform the historiography of gene and pathway discovery, and they can enable pathway figure-based enrichment analysis. We are currently working with NDEx to host the initial set of gene-annotated pathway figures for enrichment analysis (https://home.ndexbio.org). The annotated

pathway figures are also being prioritized by their novel gene content and disease associations for manual curation as proper models in pathway databases. Finally, the interactive pathway figure tool allows anyone to explore the complete set of 65k figures by various metrics and metadata (https://gladstone-bioinformatics.shinyapps.io/shiny-25years).

Construction

Collection of figures

A PMC image query URL was composed, specifying a date range starting at 1995-01-01. Exploratory queries previously identified 1995 as the first year with accurately indexed pathway figures. Keywords in the query were specified as an OR set of the following pathway types together with the word *pathway*: *signaling, signaling, regulatory, disease, drug, metabolic, biosynthetic, synthesis, cancer, response*, and *cycle*. These were determined based on exploratory queries and referencing the first two levels of pathway ontology terms [28]. The query returned just over 235,000 figure images, which were retrieved by an HTML-scraping script, along with metadata pertaining to the figure and parent paper: PMCID, paper title, paper citation, publication year, figure filename, figure URL, figure number, figure title, and figure caption. The collection was filtered for unique entries and publication dates spanning the 25-year period of 1995 January 01 to 2019 December 31. It is worth noting that while the query was performed on 2020 January 31, the results for 2019 are not expected to be complete since many journals unfortunately wait 6 months to a year to make their content openly accessible.

Limitations

Web scraping the results of a PMC image query is inefficient and imprecise. It is likely that many pathway figures are missing from the results due to incomplete keyword listing and database indexing. At the same time, the query results included many non-pathway figures. Given the ranked order of images provided by provided by PMC, we manually checked the percentage of *actual* pathway figures at three points: the first thousand figures (of 235k) contained 66.3% pathways, the middle thousand contained 35%, and the final thousand contained fewer than 10% (Fig. 5a). The order of query results from PMC was thus informative, but not sufficient to distinguish pathway and non-pathway figures.

Classification of figures

To properly identify pathway figures among the PMC image query results, two rounds of machine learning were performed using Google Cloud AutoML Vision. The AutoML service was accessed and controlled via a REST API and a web-based dashboard. The first model was trained on 2000 manually classified figures selected from high, middle, and low relevance ranges based on PMC query result ranking (Fig. 5a, dots). The service randomly split the provided figures into three sets: 80% training, 10% testing, and 10% validation. The manual classification was performed by a domain expert relying on their own organic neural network (a.k.a. brain) trained on 15 years of experience creating, curating, and using pathway diagrams in biomedical research. For the purposes of this OCR-based project, figures were considered pathways if they described a biological process or set of interactions involving identifiable genes and proteins. Molecular interaction networks and developmental processes were thus included, while cellular diagrams that did not name genes or proteins were excluded. The AutoML service evaluated the

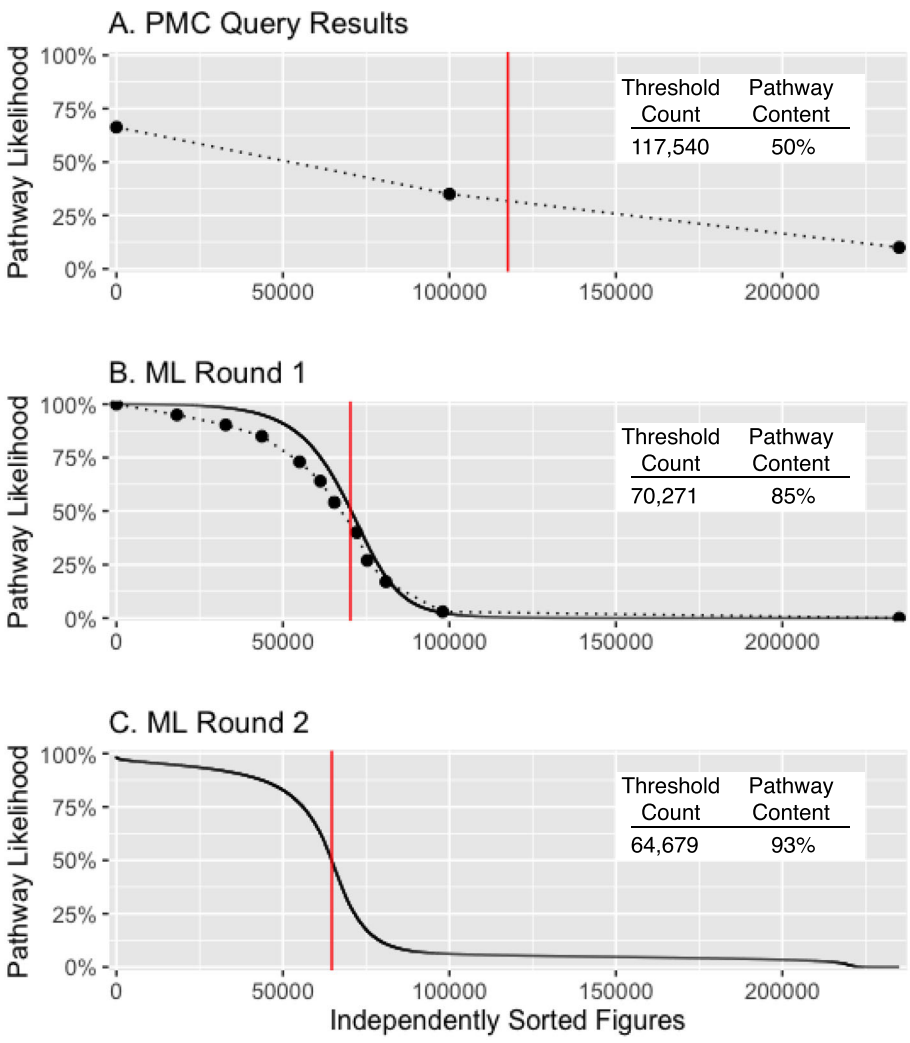

Fig. 5 Machine learning performance. Classification of pathway figures assessed in **a** original PMC-sorted query results and **b** first and **c** second rounds of machine learning results independently sorted by estimated pathway likelihood (*y*-axis). Dots in **a** and **b** (connected by dotted lines) indicate where figures were sampled along their sorted indices for manual classification. These data were used in training subsequent rounds of ML. The solid black curves in **b** and **c** indicate the prediction of the learned model from each round of ML. The red vertical lines indicate the threshold used to define a set of pathway figures at each stage and assess their count and actual pathway content (inset tables). The final set of 64,643 pathway figures replaced 1.2% of known false positives and false negatives based on manual classification and was estimated to contain 94% pathway content

performance of the first model at 88.42% precision and 91.3% recall at a 50% confidence threshold. The model was then applied to the complete set of 235k figures to obtain pathway likelihood scores (Fig. 5b, solid line). Additional figures were sampled along the full distribution of scores and manually classified (Fig. 5b, dots). The total actual pathway content above the threshold (red line) was estimated to be 85%. Given these results, a second model was trained on combined set of 15,406 manually classified figures and assessed to perform at 91.88% precision and 91.88% recall at a 50% confidence threshold, resulting in a steeper transition (i.e., fewer uncertain calls) and increased accuracy at the extremes (Fig. 5c). A Matthews correlation coefficient of 0.82 was calculated [26]. 64,679

figures had a predicted pathway likelihood of 50% or greater from the second round of machine learning (red line). Finally, 383 false negative pathways were added back and 419 false positive non-pathways were removed based on prior manual classification, resulting in the set of 64,643 pathway figures used in this study. We have included our manual classification calls for the 15,406 pathways and non-pathways in the supplemental data files [58].

Limitations

A random sample of 300 figures from the final set was manually classified to estimate the proportion of actual pathway figures at 94% (±3% at 97% confidence). Of the 18 figures classified as non-pathways, two were composite figures that included a pathway as a minor panel in figure that contained a lot of non-pathway content. Composite figures were typically excluded from the "pathway" training sets in our manual classification in order to avoid recognizing gene occurrences in figure elements unrelated to pathways, so these were conservatively included in the false positive count. Of the remaining 16 non-pathways, only three had three or more genes subsequently detected by our pathway figure OCR pipeline (see next section), suggesting the majority of false positives could be effectively ignored.

Identification of genes in pathway figures

The set of pathway figures was then fed into our pathway figure OCR pipeline [19, 23]. Briefly summarized here, the pipeline's main components are word isolation, transformation, and lexicon matching. A series of custom transforms were applied to newline- and space-delimited words provided by the OCR output. Transforms included normalization of characters, substitutions, and expansions, correcting for common OCR mistakes and common habits of pathway authors to use non-standard gene names in pathway diagrams. For example, a transform would strip away extraneous annotations prefixed or suffixed to gene symbols (e.g., p-AKT or CDK1-FLAG) and would expand numerical ranges into individual gene symbols (e.g., WNT1-5). The "transforms" directory in our public GitHub repository contains all transforms applied in this work. After each round of transformation, a match is attempted against a lexicon of human genes including HGNC symbols (official, aliases and previous) and bioentities (https://github.com/wikipathways/bioentities). The pipeline output is a table of pathway figures associated with sets of recognized genes with human NCBI Gene identifiers. Paper and figure metadata including titles and captions are maintained along with additional information such as the raw OCR results, the normalized and transformed symbols, and the lexicon source that was matched.

Limitations

There were three non-composite false positive cases that contained three or more human genes: a diagram of nerve fibers with gene markers [59], a meta-network of gene-named pathways (e.g., IL-6 signaling) [60], and a figure containing three-letter amino acid codes (e.g., Tyr, His, Met) that happen to match gene aliases [61]. The first two cases are not egregious in that relevant biology is still being detected; it is just not in the context of a pathway diagram as defined here. The last case is the only one that is a problematic false positive (i.e., mistakenly identifying an amino acid as a gene). The pathway figure OCR

pipeline was also limited by the lexicon of human genes. Many of the pathway figures with zero or small numbers of recognized genes were for other species, e.g., Drosophila signaling pathways, yeast networks, and microbial metabolism. The lexicon can be expanded in the future to include other species, other types of molecules and other biological concepts.

Characterization of pathway figures

Annotations on the 56,095 papers containing pathway figures were retrieved from the PMC query site and web services, including authors, journal titles, paper titles, paper identifiers, publication dates, figure titles, hyperlinks, and captions.

Disease annotations for 29,187 papers were available from Europe PMC and collected using the *europepmc* R package [62]. A less redundant list of top 10 terms was made by excluding previously counted papers and re-sorting by disease term frequency. Singular and plural forms of the same disease term were combined, e.g., "Tumor" and "Tumors". After identifying the top ten, remaining publications were counted as "other."

Gene associations for 30,036 papers were available from PubTator and downloaded as NCBI Gene-to-PMID mappings from the PubTator FTP Service. A PMID-to-PMCID mapping file from the PMC FTP Service enabled comparison with our PMCID-indexed pathway figures and extracted gene sets.

Gene Ontology and disease annotations for pathway figures were determined by performing enrichment analysis on the sets of figure-extracted genes against ontology-associated gene sets. The source of disease annotations was the "knowledge" channel of the DISEASES database [30] filtered for disease terms with seven or more associated genes, resulting in a set of 160 disease terms in total with 5088 gene associations. Top 10 disease term lists were made less redundant by excluding previously counted figures (i.e., figures associated with more than one disease term) and re-sorting by disease term frequency.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-020-02181-2.

Additional file 1: Review history.

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history
The review history is available as Additional file 1.

Authors' contributions
KH, AR, and ARP conceived of the project. AR implemented the majority of the AutoML and OCR pipelines. All authors conducted analyses presented in the paper. All authors contributed to writing and editing of the manuscript. The author(s) read and approved the final manuscript.

Authors' information
Twitter handles: @xanderpico (Alexander R. Pico).

Funding
This work was supported by an R01 grant from NIH/NIGMS (GM100039) and promotional credit from Google subsidizing usage of their cloud platform.

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Institute of Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA, USA. [2]Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, The Netherlands. [3]Department of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, Maastricht, The Netherlands.

Received: 16 June 2020 Accepted: 16 October 2020
Published online: 09 November 2020

References

1. Nguyen T-M, Shafi A, Nguyen T, Draghici S. Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biol. 2019;20(1):1–15. https://doi.org/10.1186/s13059-019-1790-4. Accessed 20 May 2020.
2. Kutmon M, Lotia S, Evelo CT, Pico AR. WikiPathways App for Cytoscape: making biological pathways amenable to network analysis and visualization. F1000Research. 2014;3:152. https://doi.org/10.12688/f1000research.4254.2.
3. Cirillo E, Parnell LD, Evelo CT. A review of pathway-based analysis tools that visualize genetic variants. Front Genet. 2017;8:174. https://doi.org/10.3389/fgene.2017.00174. Accessed 20 May 2020.
4. Waagmeester A, Stupp G, Burgstaller-Muehlbacher S, Good BM, Griffith M, Griffith OL, Hanspers K, Hermjakob H, Hudson TS, Hybiske K, Keating SM, Manske M, Mayers M, Mietchen D, Mitraka E, Pico AR, Putman T, Riutta A, Queralt-Rosinach N, Schriml LM, Shafee T, Slenter D, Stephan R, Thornton K, Tsueng G, Tu R, Ul-Hasan S, Willighagen E, Wu C, Su AI. Wikidata as a knowledge graph for the life sciences. eLife. 2020;9. https://doi.org/10.7554/eLife.52614. Accessed 20 May 2020.
5. Xin J, Afrasiabi C, Lelong S, Adesara J, Tsueng G, Su AI, Wu C. Cross-linking BioThings APIs through JSON-LD to facilitate knowledge exploration. BMC Bioinformatics. 2018;19:30. https://doi.org/10.1186/s12859-018-2041-5. Accessed 20 May 2020.
6. Hunter LE. Knowledge-based biomedical data science. Data Sci. 2017;1(1-2):19–25. https://doi.org/10.3233/DS-170001. Accessed 20 May 2020.
7. Bader GD, Cary MP, Sander C. Pathguide: a pathway resource list. Nucleic Acids Res. 2006;34(suppl_1):504–6. https://doi.org/10.1093/nar/gkj126. Accessed 20 May 2020.
8. Rodchenkov I, Babur O, Luna A, Aksoy BA, Wong JV, Fong D, Franz M, Siper MC, Cheung M, Wrana M, Mistry H, Mosier L, Dlin J, Wen Q, O'Callaghan C, Li W, Elder G, Smith PT, Dallago C, Cerami E, Gross B, Dogrusoz U, Demir E, Bader GD, Sander C. Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. Nucleic Acids Res. 2020;48(D1):489–97. https://doi.org/10.1093/nar/gkz946. Accessed 20 May 2020.
9. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet. 2002;31(1):19–20. https://doi.org/10.1038/ng0502-19.
10. Karp PD. Pathway databases: a case study in computational symbolic theories. Science (New York, N.Y.) 2001;293(5537):2040–4. https://doi.org/10.1126/science.1064621.
11. Karp PD, Caspi R. A survey of metabolic databases emphasizing the MetaCyc family. Arch Toxicol. 2011;85:1015–33. https://doi.org/10.1007/s00204-011-0705-2. Accessed 20 May 2020.
12. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30. https://doi.org/10.1093/nar/28.1.27.
13. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, Stein L. Reactome: a knowledge base of biologic pathways and processes. Genome Biol. 2007;8(3):1–13. https://doi.org/10.1186/gb-2007-8-3-r39. Accessed 20 May 2020.
14. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, Loney F, May B, Milacic M, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. The reactome pathway knowledgebase. Nucleic Acids Res. 2020;48(D1):498–503. https://doi.org/10.1093/nar/gkz1031. Accessed 20 May 2020.
15. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: pathway editing for the people. PLoS Biol. 2008;6(7):184. https://doi.org/10.1371/journal.pbio.0060184. Accessed 20 May 2020.
16. Kutmon M, Riutta A, Nunes N, Hanspers K, Willighagen E, Bohler A, Mélius J, Waagmeester A, Sinha S, Miller R, Coort SL, Cirillo E, Smeets B, Evelo C, Pico AR. WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Res. 2016;44(D1):488–94. https://doi.org/10.1093/nar/gkv1024. Accessed 20 May 2020.
17. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D, Ehrhart F, Giesbertz P, Kalafati M, Martens M, Miller R, Nishida K, Rieswijk L, Waagmeester A, Eijssen LMT, Evelo CT, Pico AR, Willighagen EL. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. 2018;46(D1):661–7. https://doi.org/10.1093/nar/gkx1064. Accessed 20 May 2020.
18. Pratt D, Chen J, Pillich R, Rynkov V, Gary A, Demchak B, Ideker T. NDEx 2.0: a clearinghouse for research on cancer pathways. Cancer Res. 2017;77(21):58–61. https://doi.org/10.1158/0008-5472.CAN-17-0606.
19. Riutta A, Hanspers K, Pico AR. Identifying genes in published pathway figure images. BioRxiv. 2018. https://doi.org/10.1101/379446. Accessed 20 May 2020.
20. Hearst MA, Divoli A, Guturu H, Ksikes A, Nakov P, Wooldridge MA, Ye J. BioText Search Engine: beyond abstract search. Bioinformatics (Oxford, England). 2007;23(16):2196–7. https://doi.org/10.1093/bioinformatics/btm301.
21. Kozhenkov S, Baitaluk M. Mining and integration of pathway diagrams from imaging data. Bioinformatics (Oxford, England). 2012;28(5):739–42. https://doi.org/10.1093/bioinformatics/bts018.
22. Rodriguez-Esteban R, Iossifov I. Figure mining for biomedical research. Bioinformatics. 2009;25(16):2082–4. https://doi.org/10.1093/bioinformatics/btp318.
23. Pico A, Riutta A, Hanspers K. wikipathways/pathway-figure-ocr: 25 years of pathway figures. 2020. https://doi.org/10.5281/zenodo.3880094.
24. Wei CH, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. 2019;47(W1):587–93.

25. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:1–9. https://doi.org/10.1038/sdata.2016.18.
26. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6. https://doi.org/10.1186/s12864-019-6413-7.
27. Reyes-Aldasoro CC. The proportion of cancer-related entries in PubMed has increased considerably; is cancer truly "The Emperor of All Maladies"?. PLoS ONE. 2017;12(3):0173671.
28. Petri V, Jayaraman P, Tutaj M, Hayman GT, Smith JR, De Pons J, Laulederkind SJ, Lowry TF, Nigam R, Wang S-J, Shimoyama M, Dwinell MR, Munzenmaier DH, Worthey EA, Jacob HJ. The pathway ontology - updates and applications. J Biomed Semant. 2014;5(1):7. https://doi.org/10.1186/2041-1480-5-7.
29. Ryu D, Lee C. Expression quantitative trait loci for PI3K/AKT pathway. Medicine. 2017;96(1):5817. https://doi.org/10.1097/MD.0000000000005817.
30. Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: text mining and data integration of disease-gene associations. Methods (San Diego, Calif.) 2015;74:83–9. https://doi.org/10.1016/j.ymeth.2014.11.020.
31. Ostaszewski M, Mazein A, Gillespie ME, Kuperstein I, Niarakis A, Hermjakob H, Pico AR, Willighagen EL, Evelo CT, Hasenauer J, Schreiber F, Dräger A, Demir E, Wolkenhauer O, Furlong LI, Barillot E, Dopazo J, Orta-Resendiz A, Messina F, Valencia A, Funahashi A, Kitano H, Auffray C, Balling R, Schneider R. COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. Sci Data. 2020;7(1):1–4. https://doi.org/10.1038/s41597-020-0477-8.
32. Hilman D, Gat U. The evolutionary history of YAP and the Hippo/YAP pathway. Mol Biol Evol. 2011;28(8):2403–17. https://doi.org/10.1093/molbev/msr065. Accessed 20 May 2020.
33. Justice RW, Zilian O, Woods DF, Noll M, Bryant PJ. The Drosophila tumor suppressor gene warts encodes a homolog of human myotonic dystrophy kinase and is required for the control of cell shape and proliferation. Genes Dev. 1995;9(5):534–46. https://doi.org/10.1101/gad.9.5.534. Accessed 20 May 2020.
34. Xu T, Wang W, Zhang S, Stewart RA, Yu W. Identifying tumor suppressors in genetic mosaics: the Drosophila lats gene encodes a putative protein kinase. Development. 1995;121(4):1053–63. Accessed 20 May 2020.
35. Wu S, Huang J, Dong J, Pan D. hippo encodes a Ste-20 family protein kinase that restricts cell proliferation and promotes apoptosis in Conjunction with salvador and warts. Cell. 2003;114(4):445–56. https://doi.org/10.1016/S0092-8674(03)00549-X. Accessed 20 May 2020.
36. Udan RS, Kango-Singh M, Nolo R, Tao C, Halder G. Hippo promotes proliferation arrest and apoptosis in the Salvador/Warts pathway. Nat Cell Biol. 2003;5(10):914–20. https://doi.org/10.1038/ncb1050.
37. Kim W, Jho E-H. The history and regulatory mechanism of the Hippo pathway. BMB reports. 2018;51(3):106–18. https://doi.org/10.5483/bmbrep.2018.51.3.022.
38. Vitulo N, Vezzi A, Galla G, Citterio S, Marino G, Ruperti B, Zermiani M, Albertini E, Valle G, Barcaccia G. Characterization and evolution of the cell cycle-associated Mob domain-containing proteins in eukaryotes. Evol Bioinforma. 2007;3:121–58. Accessed 20 May 2020.
39. Andl T. miRNAs: miracle or mirage? Organogenesis. 2007;3(1):25–33. Accessed 20 May 2020.
40. Wang W, Huang J, Wang X, Yuan J, Li X, Feng L, Park J-I, Chen J. PTPN14 is required for the density-dependent control of YAP1. Genes Dev. 2012;26(17):1959–71. https://doi.org/10.1101/gad.192955.112. Accessed 20 May 2020.
41. Yu F-X, Guan K-L. The Hippo pathway: regulators and regulations. Genes Dev. 2013;27(4):355–71. https://doi.org/10.1101/gad.210773.112. Accessed 20 May 2020.
42. Kremerskothen J, Plaas C, Büther K, Finger I, Veltel S, Matanis T, Liedtke T, Barnekow A. Characterization of KIBRA, a novel WW domain-containing protein. Biochem Biophys Res Commun. 2003;300(4):862–7. https://doi.org/10.1016/S0006-291X(02)02945-5. Accessed 20 May 2020.
43. Yu J, Zheng Y, Dong J, Klusza S, Deng W-M, Pan D. Kibra functions as a tumor suppressor protein that regulates Hippo signaling in conjunction with Merlin and Expanded. Dev Cell. 2010;18(2):288–99. https://doi.org/10.1016/j.devcel.2009.12.012. Accessed 20 May 2020.
44. McNeill H, Woodgett JR. When pathways collide: collaboration and connivance among signalling proteins in development. Nat Rev Mol Cell Biol. 2010;11(6):404–13.
45. Huang J, Wu S, Barrera J, Matthews K, Pan D. The Hippo signaling pathway coordinately regulates cell proliferation and apoptosis by inactivating Yorkie, the Drosophila Homolog of YAP. Cell. 2005;122(3):421–34. https://doi.org/10.1016/j.cell.2005.06.007. Accessed 20 May 2020.
46. Zhao B, Wei X, Li W, Udan RS, Yang Q, Kim J, Xie J, Ikenoue T, Yu J, Li L, Zheng P, Ye K, Chinnaiyan A, Halder G, Lai Z-C, Guan K-L. Inactivation of YAP oncoprotein by the Hippo pathway is involved in cell contact inhibition and tissue growth control. Genes Dev. 2007;21(21):2747–61. https://doi.org/10.1101/gad.1602907. Accessed 20 May 2020.
47. Zhao B, Lei QY, Guan KL. The Hippo-YAP pathway: new connections between regulation of organ size and cancer. Curr Opin Cell Biol. 2008;20(6):638–46.
48. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43(D1):789–98. https://doi.org/10.1093/nar/gku1205.
49. Levandowsky M, Winter D. Distance between sets. Nature. 1971;234(5323):34–5. https://doi.org/10.1038/234034a0. Accessed 20 May 2020.
50. Kohn KW. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. Mol Biol Cell. 1999;10(8):2703–34. Accessed 20 May 2020.
51. Luna A, Karac EI, Sunshine M, Chang L, Nussinov R, Aladjem MI, Kohn KW. A formal MIM specification and tools for the common exchange of MIM diagrams: an XML-based format, an API, and a validation method. BMC Bioinformatics. 2011;12:167. https://doi.org/10.1186/1471-2105-12-167. Accessed 20 May 2020.

52. Rougny A, Touré V, Moodie S, Balaur I, Czauderna T, Borlinghaus H, Dogrusoz U, Mazein A, Dräger A, Blinov ML, Villéger A, Haw R, Demir E, Mi H, Sorokin A, Schreiber F, Luna A. Systems biology graphical notation: process description language level 1 version 2.0. J Integr Bioinforma. 2019;16(2). https://doi.org/10.1515/jib-2019-0022. https://www.degruyter.com/view/journals/jib/16/2/article-20190022.xml. Accessed 20 May 2020.
53. Salomonis N, Hanspers K, Zambon AC, Vranizan K, Lawlor SC, Dahlquist KD, Doniger SW, Stuart J, Conklin BR, Pico AR. GenMAPP 2: new features and resources for pathway analysis. BMC Bioinformatics. 2007;8:217. https://doi.org/10.1186/1471-2105-8-217. Accessed 20 May 2020.
54. Mi H, Muruganujan A, Demir E, Matsuoka Y, Funahashi A, Kitano H, Thomas PD. BioPAX support in CellDesigner. Bioinformatics. 2011;27(24):3437–8. https://doi.org/10.1093/bioinformatics/btr586. Accessed 20 May 2020.
55. Kutmon M, van Iersel MP, Bohler A, Kelder T, Nunes N, Pico AR, Evelo CT. PathVisio 3: an extendable pathway analysis toolbox. PLoS Comput Biol. 2015;11(2):1004085. https://doi.org/10.1371/journal.pcbi.1004085. Accessed 20 May 2020.
56. Karp PD, Latendresse M, Paley SM, Krummenacker M, Ong QD, Billington R, Kothari A, Weaver D, Lee T, Subhraveti P, Spaulding A, Fulcher C, Keseler IM, Caspi R. Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. Brief Bioinform. 2016;17(5):877–90. https://doi.org/10.1093/bib/bbv079. Accessed 20 May 2020.
57. Kondratova M, Sompairac N, Barillot E, Zinovyev A, Kuperstein I. Signalling maps in cancer research: construction and data analysis. Database. 2018;2018. https://doi.org/10.1093/database/bay036. https://academic.oup.com/database/article/doi/10.1093/database/bay036/4964960. Accessed 20 May 2020.
58. Pico A, Riutta A, Hanspers K, Kutmon M. Supplementary materials for 25 years of pathway figures. The NIH Figshare Archive. 2020. https://doi.org/10.35092/yhjc.c.5005697.v1. https://nih.figshare.com/collections/Supplementary_Materials_for_25_Years_of_Pathway_Figures/5005697/1.
59. Drescher MJ, Cho WJ, Folbe AJ, Selvakumar D, Kewson DT, Abu-Hamdan MD, Oh CK, Ramakrishnan NA, Hatfield JS, Khan KM, Anne S, Harpool EC, Drescher DG. An adenylyl cyclase signaling pathway predicts direct dopaminergic input to vestibular hair cells. Neuroscience. 2010;171(4):1054–74. https://doi.org/10.1016/j.neuroscience.2010.09.051.
60. Huang Y, Ma S-F, Espindola MS, Vij R, Oldham JM, Huffnagle GB, Erb-Downward JR, Flaherty KR, Moore BB, White ES, Zhou T, Li J, Lussier YA, Han MK, Kaminski N, Garcia JGN, Hogaboam CM, Martinez FJ, Noth I, COMET-IPF Investigators. Microbes are associated with host innate immune response in idiopathic pulmonary fibrosis. Am J Respir Crit Care Med. 2017;196(2):208–19. https://doi.org/10.1164/rccm.201607-1525OC.
61. Zhang Q, Yang X, Wang H, van der Donk WA. High divergence of the precursor peptides in combinatorial lanthipeptide biosynthesis. ACS Chem Biol. 2014;9(11):2686–94. https://doi.org/10.1021/cb500622c.
62. Levchenko M, Gou Y, Graef F, Hamelers A, Huang Z, Ide-Smith M, Iyer A, Kilian O, Katuri J, Kim J-H, Marinos N, Nambiar R, Parkin M, Pi X, Rogers F, Talo F, Vartak V, Venkatesan A, McEntyre J. Europe PMC in 2017. Nucleic Acids Res. 2017;46(D1):1254–60. https://doi.org/10.1093/nar/gkx1005. https://academic.oup.com/nar/article-pdf/46/D1/D1254/23161868/gkx1005.pdf.

Publisher's Note