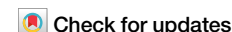


<https://doi.org/10.1038/s42003-024-06177-7>

# Rummagene: massive mining of gene sets from supporting materials of biomedical research publications



Daniel J. B. Clarke , Giacomo B. Marino , Eden Z. Deng, Zhuorui Xie, John Erol Evangelista & Avi Ma'ayan  

Many biomedical research publications contain gene sets in their supporting tables, and these sets are currently not available for search and reuse. By crawling PubMed Central, the Rummagene server provides access to hundreds of thousands of such mammalian gene sets. So far, we scanned 5,448,589 articles to find 121,237 articles that contain 642,389 gene sets. These sets are served for enrichment analysis, free text, and table title search. Investigating statistical patterns within the Rummagene database, we demonstrate that Rummagene can be used for transcription factor and kinase enrichment analyses, and for gene function predictions. By combining gene set similarity with abstract similarity, Rummagene can find surprising relationships between biological processes, concepts, and named entities. Overall, Rummagene brings to surface the ability to search a massive collection of published biomedical datasets that are currently buried and inaccessible. The Rummagene web application is available at <https://rummagene.com>.

The introduction of omics technologies has gradually moved biological and biomedical research from studying single genes and proteins towards studying gene sets, clusters of genes, molecular complexes, and gene expression modules<sup>1</sup>. Many biomedical and biological research studies produce and publish gene and protein sets. For example, differentially expressed genes and proteins from transcriptomics and proteomics assays, genes associated with genomic variants identified to be relevant to a phenotype, gene knockouts associated with a cellular or an organismal phenotype, target genes of transcription factors as determined by ChIP-seq experiments, proteins identified in differential phosphoproteomics, proteins identified in a complex from immunoprecipitation followed by mass-spectrometry studies, genes associated with a cellular phenotype from CRISPR screens, and many more types of sets can be generated. These gene sets are highly valuable but not often reused. This lack of reuse is partially because there are no standards for submitting gene sets in publications, and there are no centralized community repositories for depositing gene and protein sets. As a result, the potentially useful information about gene sets is buried in supporting material tables stored as PDF, Excel, CSV, or Word file formats. Since general and domain specific search engines do not index the contents of such supporting materials, there is no way to search through these tables. These supporting tables are not indexed by search engines because most search engines can only deal with free text and are not capable of parsing data tables.

Named entity recognition methods have been widely applied to biomedical and biological publication text, but not yet to extract gene sets from supporting tables. Manual gene set annotations and extraction of gene sets from publications has been achieved, but it is time consuming, labor intensive, and requires domain expertise. Most such efforts miss many relevant studies. For example, to create the ChIP-x Enrichment Analysis (ChEA) resource we manually extracted gene sets from supporting materials of ChIP-seq studies<sup>2,3</sup>. While the ChEA database achieved great success, it is difficult to maintain. Efforts such as ReMap<sup>4</sup>, Recount<sup>5</sup>, and ARCHS4<sup>6</sup> aim to address this challenge by uniformly reprocessing all the raw data available from community repositories to recompute gene sets from published studies, but such efforts rely on the existence of community repositories and uniform data collection standards. Another effort to automate the extraction of gene sets from publications is Pathway Figure Optical Character Recognition (PFOCR)<sup>7</sup>. PFOCR automatically extracts pathways from publications by scanning pathway diagrams. However, surprisingly, as far as we know there are no publications, databases, or community repositories that contain extracted gene sets from supporting materials of scientific biomedical research publications. Rummagene is a web-based software application that serves hundreds of thousands of gene sets extracted from publications listed on PubMed Central (PMC). It contains a softbot that scans supporting materials of publications listed on PMC to keep the

Department of Pharmacological Sciences, Mount Sinai Center for Bioinformatics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.

 e-mail: [avi.maayan@mssm.edu](mailto:avi.maayan@mssm.edu)

resource consistently updated. The Rummagene website provides the ability to search the corpus of gene sets by an input gene set query, a PMC free text search, and a table title search. To understand the statistical patterns within the Rummagene corpus, we performed various exploratory analyses, as well as demonstrate how this rich resource of organized biological knowledge can be used for specific applications.

## Results

### Descriptive statistics

The initial version of Rummagene contains 642,389 gene sets extracted from 121,237 articles. These 121,237 articles are identified as containing gene sets from 5,448,589 scanned PMC articles. The distribution of the occurrence of genes in gene sets is not even. Some genes are found in many sets, but most genes are members of few sets (Fig. 1a). At the same time, most identified gene sets have less than one hundred genes in each set (Fig. 1b). While most publications only contributed to the Rummagene collection one or two gene sets, there are few publications that contributed a few hundred sets (Fig. 1c). Over the years, more and more gene sets are found in publications (Fig. 1d). In fact, in the past four years, publications included many more sets compared to sets identified in the 30 years between 1988 and 2018. Since 2005, the average length of gene sets jumped from less than 20 genes in each set to ~150 genes in each set (Fig. 1e). This is likely due to the introduction of omics technologies and publications reporting gene sets identified from such studies. By projecting the gene set content into two dimensions with UMAP<sup>8</sup>, we see that, on average, short gene sets contain genes that are more commonly studied (Fig. 1f, g). While this is a general trend, some genes occur in many sets but are less commonly studied (Fig. 1h, i). Specifically, we identified 604 gene sets that are enriched in understudied genes. Understudied gene sets are defined as gene sets where the median citations per gene is less than 3 standard deviations from the median for citations observed for randomly assembled gene sets of similar size (Supplementary Data 1). These gene sets contain many sets that are made of orphan GPCRs and Znf family members. Other sets are mainly modules of differentially expressed genes. These modules are likely serving critical biological roles but are less explored. Next, we noticed that the Rummagene collection of gene sets has many duplicate entries. In fact, duplicated gene sets make up approximately 15% of the Rummagene gene sets. Many of these duplicate sets are found in the same publication. The publications having multiple tables often list the same sets but with different measurements or statistics, for example, measuring the expression of a set of genes under different conditions. We found fewer duplicate gene sets across multiple papers (Fig. 1j, k).

### Annotated collections of themed gene set libraries

For the collection of 642,389 gene sets we can identify subsets of gene sets associated with specific biological themes such as sets related to kinases, transcription factors, cell types, cell lines and tissues. Such themed gene sets can be used for specific enrichment analysis tasks such as kinase enrichment analysis<sup>9</sup>, transcription factor enrichment analysis<sup>2</sup>. Producing such subsets of gene sets can be done by simply searching the table titles for terms that match named entities such as protein kinases or transcription factors. Indeed, we identified 4525 gene sets that contain named human kinases, and 8078 gene sets that contain named transcription factors in the table titles. 444 kinase names and 1121 transcription factor names are unique in these collections of gene sets (Fig. 2a, Supplementary Data 2, 3). Similarly, we identified 4443 gene sets that contain named cell lines, and 6268 gene sets that contain cell types or tissues in table titles, with 450 and 670 unique terms, respectively (Fig. 2b). In addition, 5560 sets had the term “down” and 6677 had the term “up” in their table titles (Fig. 2c). These sets likely contain up- and down-regulated genes from gene expression signatures. A large portion of the identified gene sets contain gene names in their titles. Specifically, 97,478 table titles contain human gene symbols or synonyms (Fig. 2c). For the subset of gene sets containing known transcription factors in their titles, Uniform Manifold Approximation and Projection (UMAP) plots were generated from the inverse document frequency (IDF) vectors for all gene sets in the subset. Points representing different gene sets are colored

by both the PubMed Central ID (PMCID) of the original publication (Fig. 2f), and by the associated transcription factor (Fig. 2e). We found that these gene sets tend to cluster by transcription factor even when they are derived from different publications. This was further confirmed to be statistically significant (*T*-test;  $p < 0.0001$ ) by comparing the average and distribution of the Jaccard index similarities between gene sets mentioning the same transcription factor from different publications compared to those not mentioning the same TF (Fig. 2d). We also applied the same process to generate UMAP plots for the subset of terms containing known kinases, and similarly saw that these gene sets clustered by kinase (Fig. 2g) although originating from different PMCID (Fig. 2h). This trend was also confirmed statistically (Fig. 2d). Next, we aimed to assess whether kinase and transcription factor gene set libraries created from Rummagene contain useful information for performing gene set enrichment analysis. To achieve such an assessment, we queried each gene set from the Rummagene kinase and transcription factor libraries against corresponding kinase and transcription factor libraries created from multiple sources<sup>2,9</sup>. We observe a significant recovery of the correct kinases and transcription factors with all libraries, with best agreement observed for KEA<sup>9</sup> for kinases, and ChEA 2022<sup>10</sup> for transcription factors (Fig. 3a–f). This is likely because these two resources are manual efforts of extracting gene and protein sets from publications, including data from supporting tables. Comparing the kinase and transcription factor Rummagene libraries to KEA and ChEA, Rummagene is likely more comprehensive and updated, but less accurate.

### Topic modeling

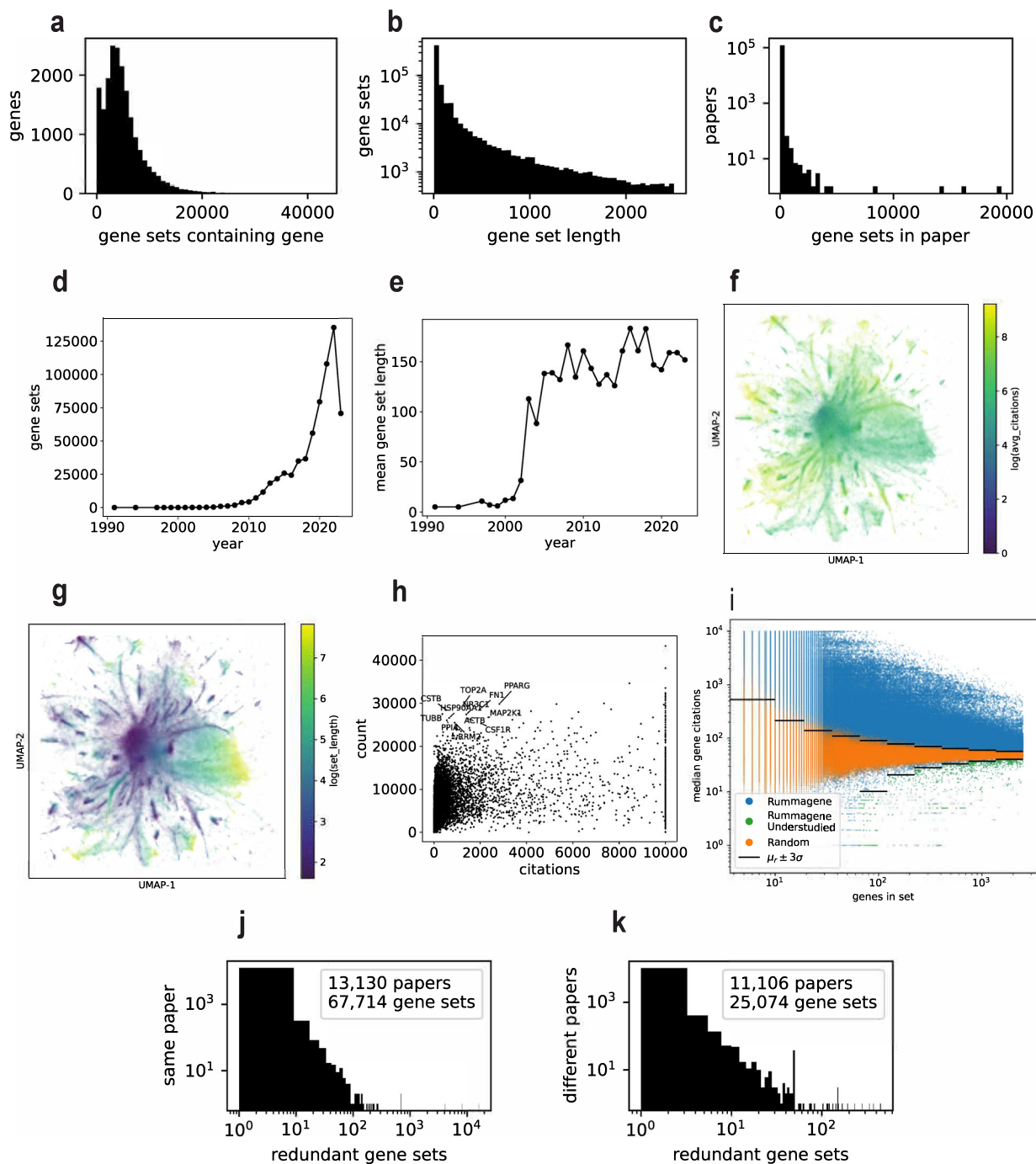
To obtain a global view of the contents of the gene sets in Rummagene, we performed latent Dirichlet allocation (LDA) analysis<sup>11</sup> on all abstracts from publications containing at least one extracted gene set. Nine topics were identified and subsequently manually labeled based on the most common terms and their relative weights (Fig. 4a). Some of the most frequently appearing terms across all topics included gene, cell, expression, DNA, patient, cancer, and analysis. The greatest portion of abstracts are relating to mutations and variants in diseases, protein-protein interactions, and mechanisms, while the topics with the least abstracts are related to immune functions and genome-wide associations and risks. The visualization of abstracts in topic space also reveals the relation and similarity between topics (Fig. 4b). For instance, the topic mutations and variants in disease borders DNA transcription and methylation. Additionally, the genome wide association and risk topic is isolated from the other topic clusters. The data and modeling topic is located adjacent to most of the other topics suggesting that abstracts with this topic may be related to a variety of other topics as expected. Overall, the topic analysis reveals the predominant categories of gene sets in Rummagene, specifically those concerning mutations and variants in diseases and those concerning protein interactions and functional mechanisms.

### Similar gene set pairs that are distant in abstract space

Next, we asked whether the knowledge embedded in Rummagene can lead to the construction of hypotheses by identifying gene sets with high similarity in gene set space while completely disjointed at the publication abstract text space. The rationale for this is that this way we can identify undiscovered associations between named entities such as genes and diseases. Surprisingly, we first observed that the pairs of gene sets with the highest similarity at the gene set level, with no similarity at the abstract level, are highly enriched in proteins that are commonly detected in mass-spectrometry proteomics studies (Fig. 5a), highly expressed in RNA-seq assays (Fig. 5b), but less widely studied (Fig. 5c). This is likely because proteomics studies tend to commonly report the same abundant, large-size, and “sticky” proteins, transcriptomics studies detected as differentially expressed highly expressed genes, and gene sets in publications commonly report overlapping genes in pathways and ontology terms containing highly studied genes. After filtering pairs of gene sets that are proteomics rich, or contain highly expressed genes, or composed of highly studied genes, we identified a few pairs of sets that contain a gene name in one table title of one

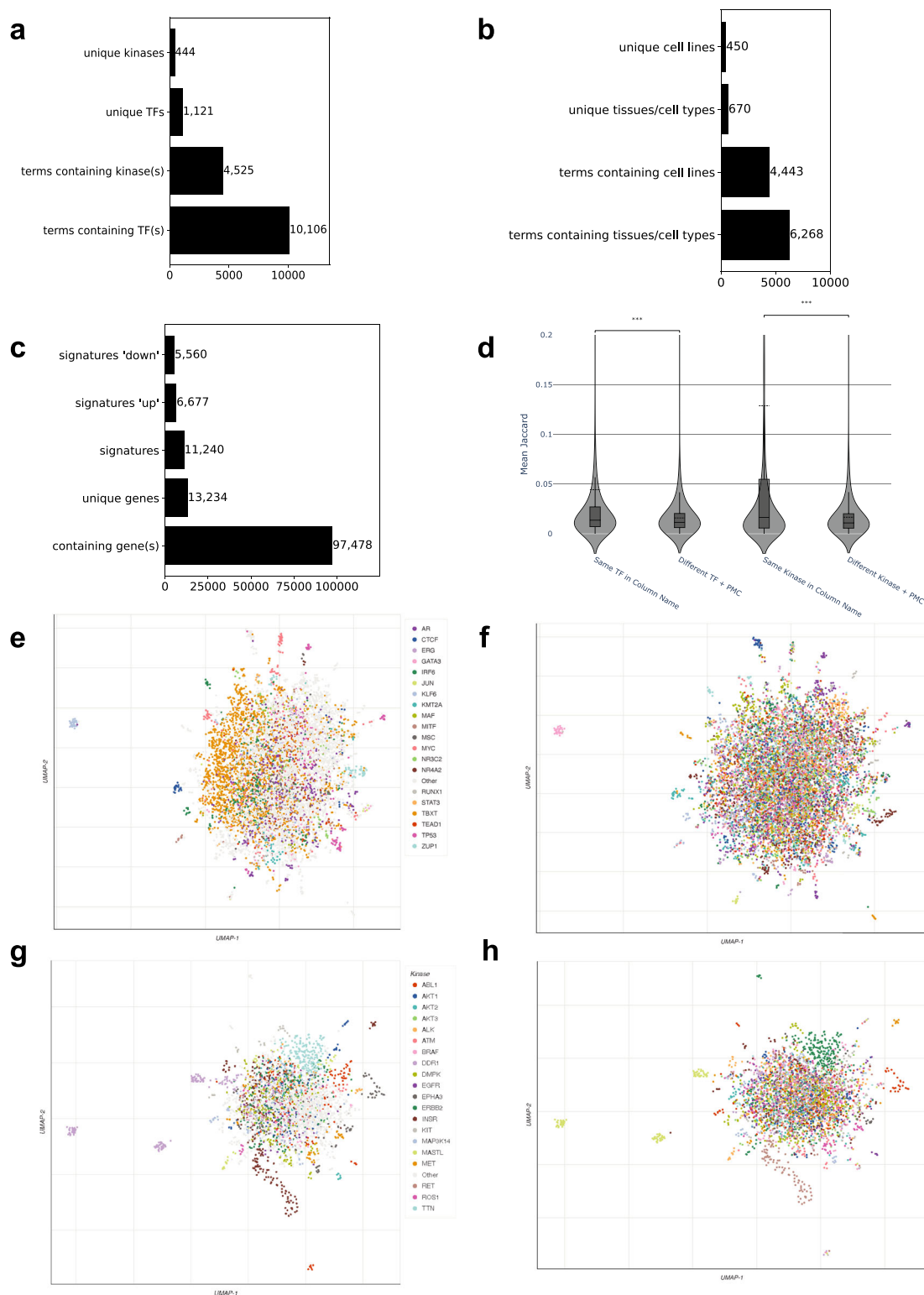
set, and a disease name in the table title of the second set (Supplementary Note). For example, some of the top identified pairs highlight a possible relationship between the proteins identified to interact with *CLUH*<sup>12</sup>, and gene sets identified in hypoxia<sup>13</sup>, melanoma<sup>14</sup>, and glioma<sup>15</sup>. This connection is logical because *CLUH* was found to be critical to mitochondrial function which is altered in these conditions. Similarly, other top overlapping pairs

include the *TOPBP1* interactome<sup>16</sup> and a potential relationship to melanoma<sup>14</sup>, hypoxia<sup>17</sup>, and teratomas<sup>18</sup>. To assist in possibly explaining these connections, we utilized the GPT-4 API, a large language model, to compose hypotheses that suggest how such seemingly unrelated named entities might be in fact related by giving GPT-4 the two abstracts. For example, when asked about the connection between the gene *CLUH* and the



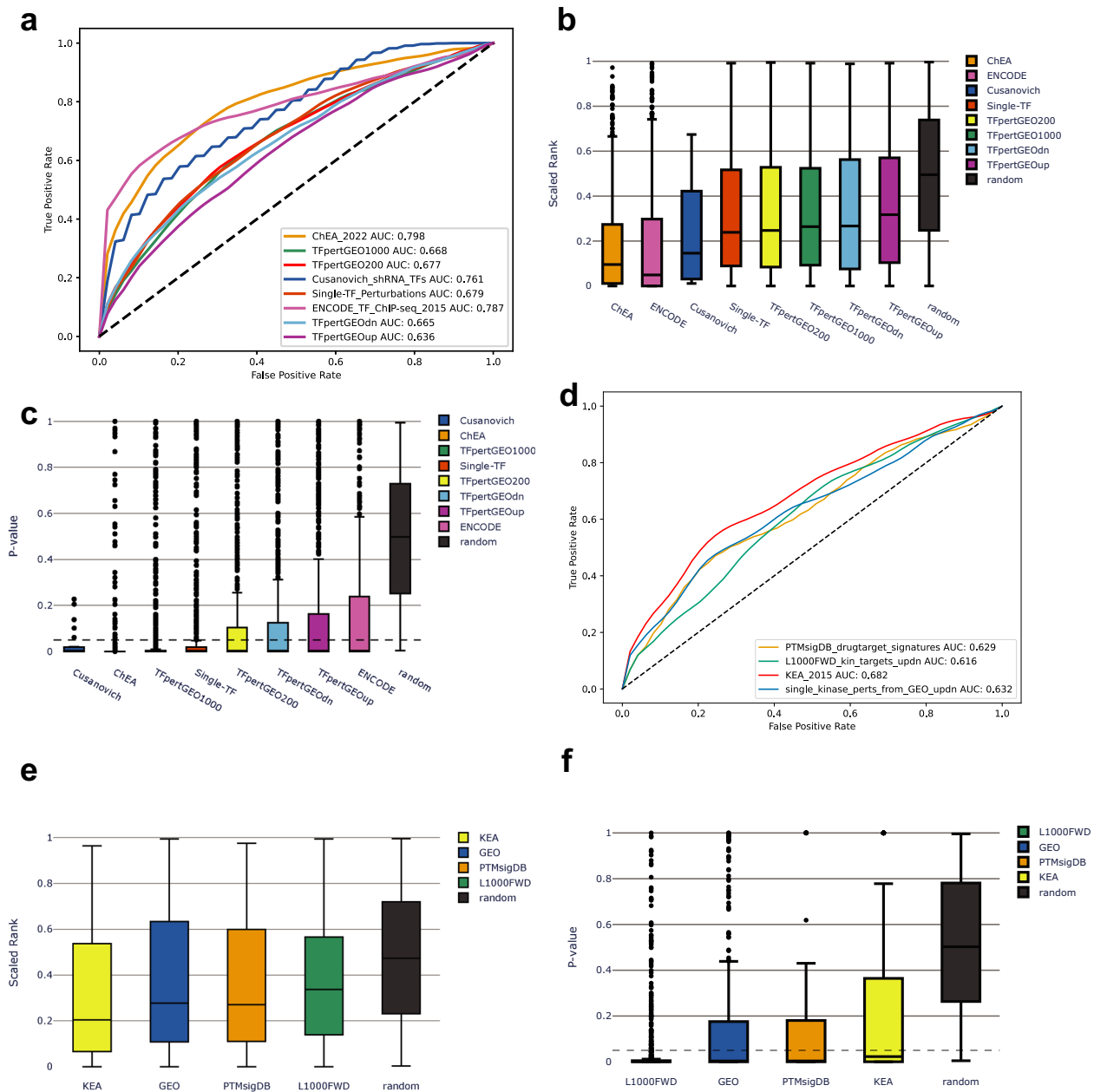
**Fig. 1 | Distributions of the genes and gene sets in the Rummagine database.** **a** Distribution of genes in gene sets. **b** Distribution of gene set lengths. **c** Distribution of gene sets per paper. **d** Gene set contribution per year. **e** Average gene set length per year. **f** Projection of the Rummagine gene sets into UMAP space where gene sets are colored by the natural logarithm of their average citations per gene. **g** Projection of the Rummagine gene sets into UMAP space where gene sets are colored the natural logarithm of by their length. **h** Scatter plot of all genes with their overall citations vs. membership in Rummagine sets. **i** Scatter plot of Rummagine sets with their median gene-citation vs. number of genes in the set compared with randomly

constructed gene sets' median gene-citation of similarly sized gene sets. The black lines show the  $\pm 3$  standard deviation from the mean median gene-citation boundary line for each binned range of gene set sizes. **j** Publications that contain at least one redundant gene set. The x-axis corresponds to the number of times the gene set is repeated, and the y-axis corresponds to the number of papers with such redundant gene sets. **k** Gene sets that are redundant across more than one publication. The x-axis corresponds to the number of times the gene set is repeated, and the y-axis corresponds to the number of different publications that contain the redundant gene set.



**Fig. 2 | Extracting kinases, transcription factors, tissues, cell types, and cell lines from Rummagene gene sets. a** The number of unique and non-unique kinases and transcription factors identified in table headers describing gene sets. **b** The number of unique and non-unique cell types and tissues, as well as cell lines identified in table headers describing gene sets. **c** Gene set table titles containing the terms “up”, “down”, or gene symbols. **d** Mean Jaccard similarity coefficient for gene sets containing a TF or a kinase in the column title and from different PMC articles with gene

sets containing the same TF or kinase compared to those with a different TF or kinase also from different PMC articles ( $p < 0.0001$ ). **e** UMAP projection of the transcription factors gene set library created from Rummagene where sets are colored by the top-most common transcription factors. **f** The same UMAP as **d** except that sets are colored by their PMCID. **g** UMAP projection of the kinases gene set library created from Rummagene where sets are colored by the top-most common kinases. **h** The same UMAP as **f** except that sets are colored by their PMCID.

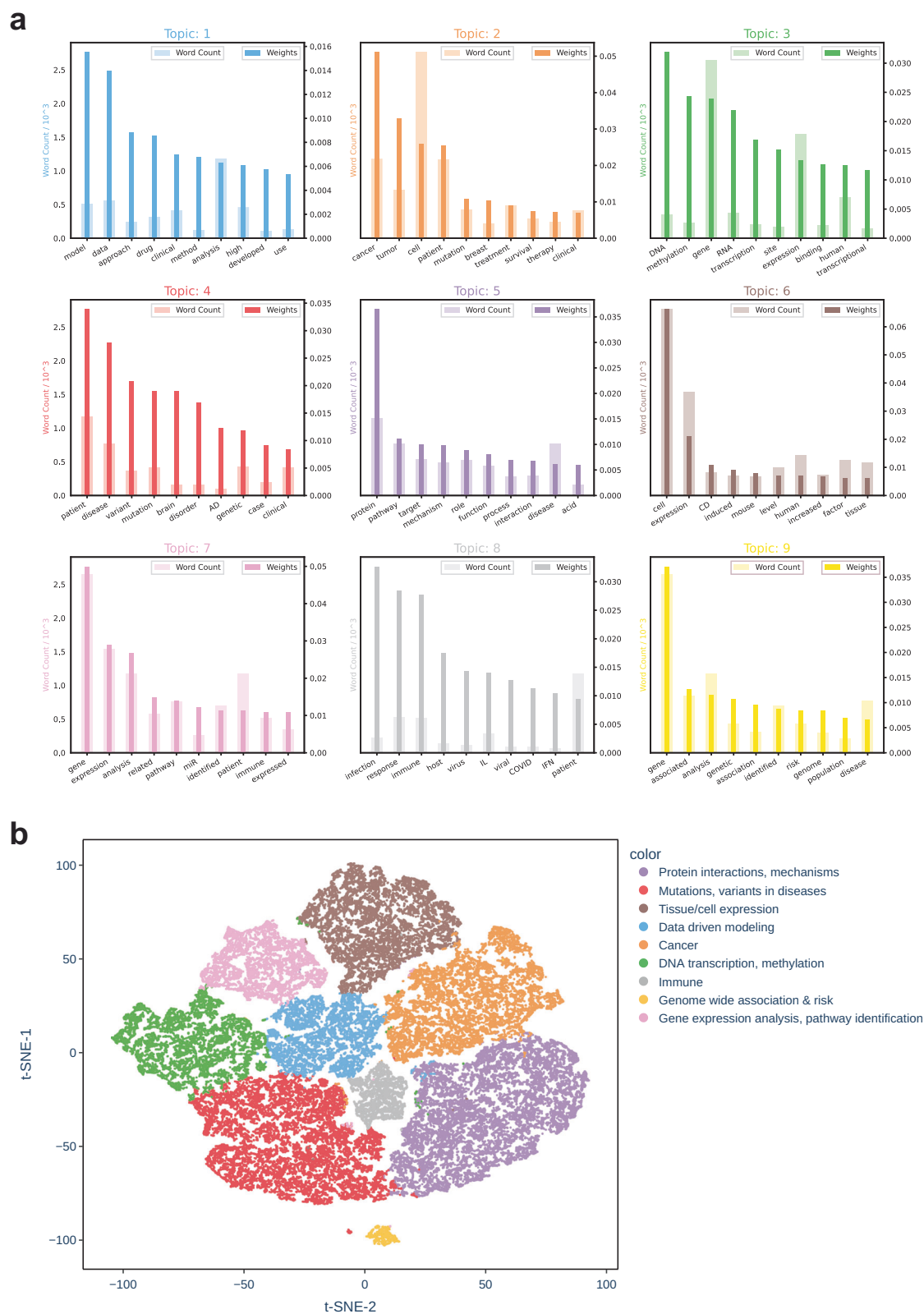


**Fig. 3 | Benchmarking the consensus transcription factor gene and kinase set libraries created from Rummagene. a** Mean receiver operating characteristic (ROC) curves and mean area under the ROC curves (AUC) generated from 5000 bootstrapped curves. **b** Scaled ranks of each TF in the benchmarking datasets when enriched against the consensus TF gene set library. **c** Mean significance of overlap (Fisher’s exact test) for Rummagene generated consensus TF gene sets with each

benchmarking library. **d** Mean ROC curves and AUCs generated from 5000 bootstrapped curves. **e** Scaled ranks of each kinase in the benchmarking datasets when enriched against the consensus kinase gene set library. **f** Mean significance of overlap (Fisher’s exact test) for Rummagene generated consensus kinase gene sets with each benchmarking library.

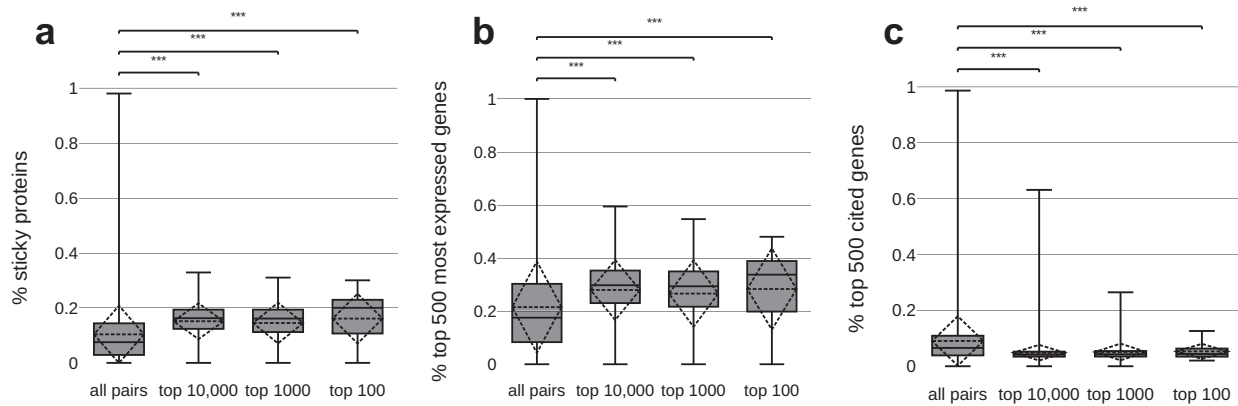
disease hypoxia, prompted with the abstracts and gene set terms, the LLM responded with a plausible explanation concerning mitochondrial function, specifically: “Therefore, it is plausible that the *CLUH* gene may be involved in the adaptive response of SKOV-3 ovarian cancer cells to hypoxia, possibly by regulating the translation and stability of mitochondrial proteins. This could explain the high overlap between the two gene sets. Further experimental studies would be needed to confirm this hypothesis.” The model successfully determined the cell line used to produce the gene set concerning hypoxia from the abstract provided and it made a reasonable hypothesis about the relationship between the two gene sets given the dissimilar context

of the abstracts. Additionally, when asking the LLM about the connection between the gene sets with *TOPBP1* and teratomas in their column names, using the two abstracts associated with these gene sets, the LLM produced a plausible explanation about their similarity after stating a hypothesis and reiterating information from the abstracts: “Given the role of *TOPBP1* in DNA repair and the importance of gene mutations in the development of teratomas, it is plausible that mutations or dysregulation of *TOPBP1* could contribute to the development or progression of teratomas. This could explain the high overlap between the two gene sets. Further research would be needed to confirm this hypothesis and elucidate the exact mechanisms



**Fig. 4 | Topic modeling with LDA from the abstracts of 121,043 papers from which gene sets in Rummagene were extracted. a** Based on the word counts and importance of each word from the LDA model, topics were manually labeled: Topic 1: Data driven modeling; Topic 2: Cancer; Topic 3: DNA transcription, methylation; Topic 4: Mutations, variants in diseases; Topic 5: Protein interactions, mechanisms;

Topic 6: Tissue/cell expression; Topic 7: Gene expression analysis, pathway identification; Topic 8: Immune; Topic 9: Genome wide associations and risk. **b** t-SNE projection of each of the 121,043 papers from which gene sets were extracted labeled by their topic identified through LDA.



**Fig. 5 | Distribution of percent of sticky proteins.** The fraction of the top 500 most cited genes from GeneRIF, and the fraction of the top 500 most highly expressed gene sourced from ARCHS4, in the overlapping genes of the gene set pairs. “All pairs” refers to all gene set pairs from different publications with significant overlapping genes and no abstract similarity. Top 100 and top 10,000 refer to the most significantly overlapping pairs ranked by  $p$ -value. Boxes represent the 25th% percentile, 50th% percentile, and 75th% percentile, while whiskers show the upper and lower fence. The dashed diamonds represent the mean and standard deviation. \*\*\*

indicates  $P$ -values < 0.001.  $P$ -values computed using the Welch’s  $t$ -test: (a) fraction of sticky protein: all pairs vs. top 10,000  $P$ -value = 0.00E + 00; all pairs vs. top 1000  $P$ -value = 1.36E-60; all pairs vs. top 100  $P$ -value = 8.21E-09; (b) fraction of top 500 most expressed genes: all pairs vs. top 10,000  $P$ -value = 0.00E + 00; all pairs vs. top 1000  $P$ -value = 3.73E-35; all pairs vs. top 100  $P$ -value = 2.44E-05; (c) fraction of top 500 most cited genes: all pairs vs. top 10,000  $P$ -value = 0.00E + 00; all pairs vs. top 1000  $P$ -value = 5.87E-220; all pairs vs. top 100  $P$ -value = 1.02E-24.

involved”. The summaries produced by the GPT-4 LLM are mostly helpful and logical but should be manually verified as the model states on its own.

### Gene function predictions

Large collections of gene sets can be used to effectively predict gene functions with semi-supervised learning<sup>19</sup>. The first step to produce such predictions is to construct a gene-gene similarity matrix from the Rummagine database of gene sets. This can be done with different algorithms. Here we tested the ability of three previously published co-occurrence algorithms<sup>20</sup> to make such predictions, and compare the quality of such predictions to predictions made with a similar method that utilizes gene-gene co-expression correlations from thousands of RNA-seq samples<sup>6</sup>. The gene-gene similarity matrices from Rummagine were able to predict with high accuracy and precision the gene membership for functional terms created from the Gene Ontology (GO) Biological Process<sup>21</sup>, GWAS Catalog<sup>22</sup>, Mouse Genome Informatics (MGI) Mammalian Phenotypes (MP)<sup>23</sup>, and WikiPathways<sup>24</sup> (Fig. 6a). To illustrate an example for one term, the term “Fasting Plasma Glucose” from GWAS Catalog was selected. The top 10 genes that are closest to the genes known to be associated with this phenotype are *SLCO1B3*, *SLCO1B7*, *P3R3URF-PIK3R3*, *SLC30A8*, *FAM240B*, *MTNR1B*, *PERCC1*, *EEF1AKMT4-ECE2*, *KLF14*, *CCDC201*, and *PAX4*; and the ROC curve to assess the quality of the predictions has a 0.75 area under the curve (Fig. 6b). The top 10 predicted genes for each term from these three gene set libraries are provided as a supporting table (Supplementary Data 4).

### The knowledge space that is covered by Rummagine compared with Enrichr

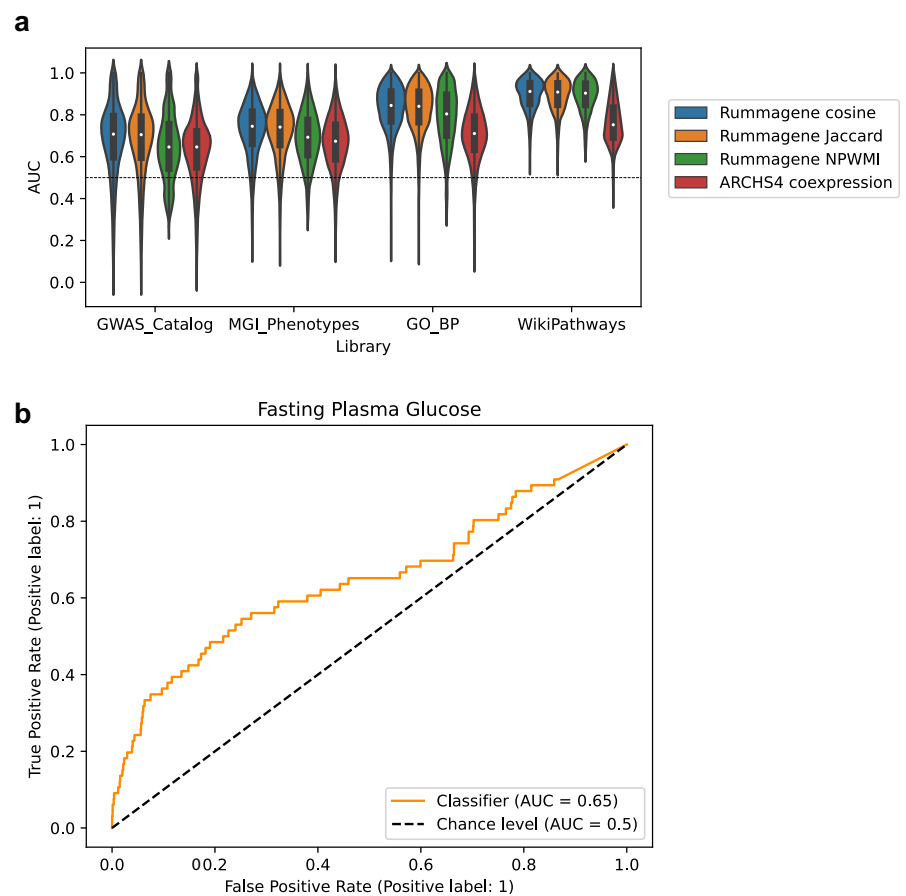
To assess the breadth and coverage of the automatically curated Rummagine gene set space, we contrasted it against the Enrichr<sup>10</sup> gene set space. Enrichr is a large-scale curated database of gene sets of similar size when compared to Rummagine. UMAP<sup>25</sup> was applied to project over 1 million gene sets into two dimensions for the purpose of data visualization where each point represents a gene set from either Rummagine or Enrichr. Gene sets are colored by whether they originate from Rummagine or Enrichr’s gene set library categorization: Transcription, Pathways, Ontologies, Diseases/Drugs, Cell Types, Miscellaneous, Legacy, Crowd (Fig. 7a, b). We observe that Rummagine gene sets cluster into many punctate clusters that likely represent themed gene sets (Fig. 7A). Also, Enrichr’s gene sets are

clustered by category (Fig. 7b). When overlaying the Rummagine gene sets on the Enrichr gene sets, most categories are covered with some few exceptions. We observe that some gene set libraries are not covered by Rummagine, while few areas in gene set space are much more common in Rummagine compared with Enrichr. To quantitatively verify the presence of these unique clusters, UMAP enhanced clustering was employed with a UMAP projection with  $\text{min\_dist}$  of 0 followed by HDBSCAN clustering<sup>26</sup>. Clusters were assigned labels based on whether 25% of the gene sets within that cluster were from a given Enrichr gene set library, or otherwise they were labeled by a cluster number. Mostly Enrichr and mostly Rummagine clusters, making up 90% of gene sets in the cluster across the projection are visible (Fig. 7c). The largest clusters that are mostly from Enrichr are from gene set libraries that were created from unique sources, for example, the LINCS L1000 data<sup>27,28</sup>, single cell transcriptomics<sup>29</sup>, virus-host protein-protein interactions<sup>30</sup>, pathways extracted from figures<sup>31</sup>, and gene sets related to NIH funded investigators<sup>32</sup> (Fig. 7d). On the other hand, several clusters were unique to Rummagine (Supplementary Data 5–8). One of these clusters, namely cluster 81, contains gene sets that are exclusively transcription factors. This is likely because there are specific assays and studies that focus on profiling these genes exclusively.

### The Rummagine website

The Rummagine data is served on the website <https://rummagine.com> with three search engines. The first search engine accepts gene sets as the input query and then returns matching gene sets based on the overlap between the input gene set and the unique sets in the Rummagine database. The results are ranked by the Fisher’s exact test, and to optimize responsiveness, a fast in-memory algorithm is implemented. The results are presented to the user in paginated tables with hyperlinks to the original publication and the supplemental material from which the gene sets were extracted, the genes in the matching sets, the overlapping genes, the  $p$ -values and the Benjamini-Hochberg corrected  $p$ -values of the overlap, and the odds ratios. When clicking the overlap numbers, a popup screen shows the overlapping genes with the ability to copy them to the clipboard, submit them to Rummagine, or submit them to Enrichr<sup>10</sup>. Similarly, the original gene set can be accessed by clicking on the column name. The second search engine facilitates a free-text PMC search. This search engine queries PMC with the entered terms to receive PMCID that match the query. It then compares the returned

**Fig. 6 | Benchmarking gene function prediction using Rummagene gene sets.** **a** The area under the receiver operating characteristic (AUROC) curve distributions for predicting genes associated with terms from four different gene-set libraries. Predictions were made using three gene-gene similarity matrices (cosine, Jaccard, and NPWMI) and the ARCHS4 gene-gene co-expression matrix. **b** The AUROC curve for the GWAS Catalog term “Fasting Plasma Glucose”, produced from the average NPWMI of each gene to the 66 genes included in the “Fasting Plasma Glucose” gene set.



PMCID to the PMCID in Rummagene to identify matching PMCIDs. Once such matches are detected, the gene sets in the Rummagene database are returned to the user as a paginated table with hyperlinks to the original publications and the matching gene sets. The third search engine queries the table titles from which gene sets were extracted. Table titles that match the inputted search terms are displayed in a paginated table with hyperlinks to the matching publications and gene sets. All search engine results can be filtered, shared by URL, and downloaded. The entire database is available for download as a text file, and access to the data is provided via a GraphQL API. Importantly, the Rummagene resource is updated automatically once a week.

## Discussion

By crawling through full articles and supporting materials from over five million research publications available from PMC, we were able to identify over 150,000 publications that contain over 600,000 mammalian gene sets of various lengths. Smaller gene sets are enriched for widely studied genes while longer lists contain less studied genes likely due to their origin from omics studies. Interestingly, in the past five years, the publication of gene sets in articles has been increasing exponentially. Hence, most gene sets in the Rummagene database are from this period. Here we demonstrated how the Rummagene resource can be used for various applications. Specifically, we showed how a subset of the extracted sets can be used for transcription factor and kinase enrichment analyses. We also showed how the rich knowledge in Rummagene can be used for gene function predictions. In addition, we demonstrated how we can form hypotheses by identifying gene set pairs with high similarity in gene set space and low similarity in abstract space. However, many additional applications are possible. For example, Rummagene can be used to produce textual descriptions for gene sets using large language models (LLMs). Given the large collection of Rummagene gene sets, as well as the fast enrichment search engine that is implemented,

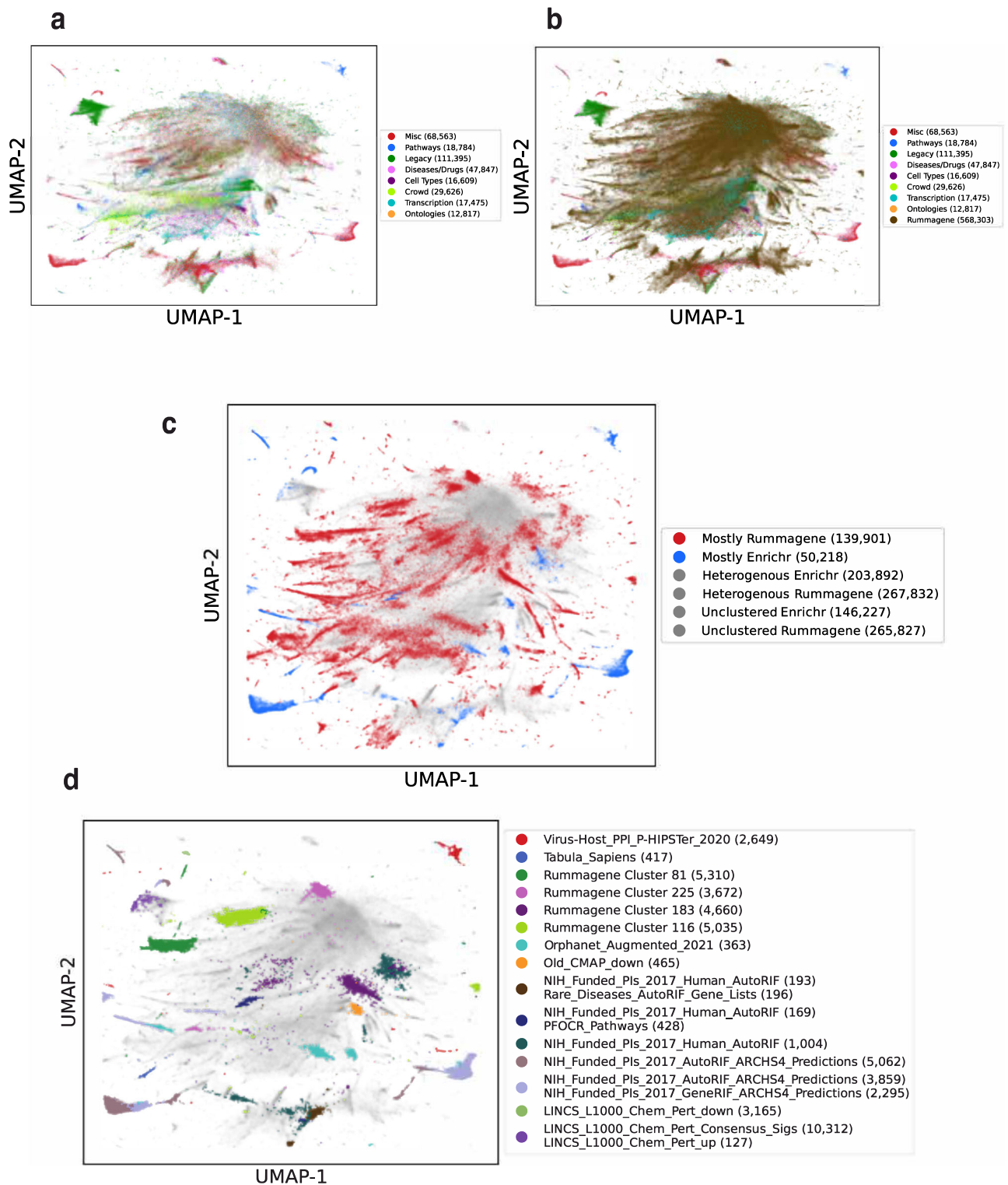
we could provide the Rummagene API to an LLM to act as a chatbot that searches for relevant papers that are related to a given gene set and then summarize the collective functions identified in these papers. This is different from just giving an LLM a gene set because it adds focus to the search by utilizing the Rummagene API. The LLM use case currently implemented in Rummagene is forming hypotheses about two highly overlapping gene sets with dissimilar abstracts. We show how when submitting the two abstracts to an LLM to provide an explanation about why the seemingly unrelated abstracts might have highly overlapping gene sets, the LLM is constrained to provide a plausible explanation. Although such an explanation is at times trivial, in all cases that we tested, it was based on correct facts. Hence, the prompt is detailed and constrained enough to produce high-quality responses from the LLM. One of the opportunities provided by Rummagene is its integration with other resources that contain large collections of gene sets and signatures, for example, Enrichr<sup>10</sup>, ARCHS4<sup>6</sup>, and SigCom LINCS<sup>27</sup>. Biomedical research has been traditionally communicated via hardcopy printed paper journals. The transition into fully digital research communication, and with the introduction of omics technologies, increased efforts are placed on better annotation and standardization of published research data including the publication of gene sets and data tables. During this transition period toward such improved annotations, Rummagene plays an important role in making previously published data, buried in supplemental materials of publications, more findable, accessible, interoperable, and reusable (FAIR)<sup>33</sup>.

## Methods

### Crawler to extract gene sets from publications listed in PMC

The PMC Open Access Subset<sup>34</sup> contains millions of journal articles available under license terms that permit reuse. Additionally, PMC provides uniformly structured bundles that can be retrieved in bulk over FTP. An index file contains a tabular listing of all PMCIDs represented with a pointer





**Fig. 7 | Visualizing the global space of the gene sets contained within Rummagene and Enrichr.** Vectorization by IDF followed by Truncated SVD to 50 dimensions is applied to the combined gene sets from both Rummagene and Enrichr. The Enrichr gene sets are colored by library categories. UMAP visualization of the gene sets in Enrichr without (a) and with (b) the Rummagene sets. 10% of extreme points are omitted for zooming into the area with the most sets/points. HDBSCAN was used to assign cluster labels to the gene set clusters that are over-represented with gene sets

from unique Enrichr libraries or were uniquely labeled in Rummagene. The gene sets are colored by whether they were assigned to a cluster with other gene sets from the same source, Enrichr or Rummagene in (c). Gene sets that are colored in gray include those that are unclustered or are members of a heterogeneous cluster. The gene sets are colored by the top representative Enrichr libraries, or by the Rummagene cluster number, or otherwise gray in (d), including gene sets that are unclustered or belong to a heterogeneous cluster.

to the compressed bundle corresponding to that PMCID. Each bundle has a PDF of the paper, an XML document containing structured metadata about the paper, figures, and supplemental material files. First, the index file is downloaded, a job is then submitted for each paper. The job downloads and extracts the archive and processes the XML structured paper by loading the tables from the paper and all supplemental files. Both the tables from the main paper, and the tables in the supplemental files may have captions or labels. These captions or labels are saved. Additionally, places in the text that mention the table, or the supplemental file, are identified when they are linked in the markup; at most, 15 words before such a call to the tables are saved. Every supplemental file is processed by one of several table-extractor-functions, selected based on the file extension. These extractor functions include support for Excel, CSV, TSV, and inferred separator loading of TXT files, as well as a PDF table extractor based on Tabula-Py. For each supporting materials table that is extracted, every column in the table is considered. The extractor function attempts to map all unique strings to gene symbols. Mapping may be direct, through some synonym, or identifier. Any column where more than half of the strings can be successfully mapped to a valid human gene symbol using NCBI's Gene Info<sup>35</sup> file for *Homo sapiens* are retained. In other words, all columns passing this filter become a gene set in the Rummagene gene set library. This approach aims to capture human gene sets, but also captures gene sets from other mammalian organisms such as mouse or rat because of the high overlap in gene symbols. Hence, we consider the overall collection of gene sets in Rummagene as mammalian. The term describing the gene set is made of the PMCID, the file name in the bundle, the Excel spreadsheet name or the XML table label, the column's first cell, and additional sequential numbers that are added to the term to make it unique if needed. The description field is constructed by concatenating any available caption, label, and text mention. The original items in each table column that pass the filter are preserved, but genes are included only if they can be mapped to official symbols. In addition to filtering out columns with too few mapped genes (< 5), columns with too many mapped genes (> 2500) are ignored. This is because these are likely to contain gene sets that cover all measured genes and not a subset of identified genes with a potentially unique function. This pipeline produces a large gene matrix transpose (GMT) file which can be added to incrementally. The pipeline is designed to continue where it left off when it is re-run. It is set to run weekly to extend the database with any new publications that are added to the PMC Open Access database. The new entries to the GMT are stored in the Rummagene database to be accessed from the web-based application. By extracting gene sets from supporting material of published research articles we can make these more accessible for search and reuse.

### Search engine implementation

The large size of the Rummagene gene set library requires special implementation of an algorithm that can quickly compare the input gene set to all the gene sets in the Rummagene database. Besides a fast algorithm that can compare the input set to all other sets, efficient storage of the gene sets is needed as well as sufficient hardware. To enable a fast gene set search, a Rust-powered REST API was implemented. The algorithm first initializes several in-memory data structures: 1) a background sorted set of all genes across all gene sets in the database; 2) the index of each gene saved in a hashmap mapping where each gene is mapped to a 32 bit unsigned integer (U32) index; 3) the gene set IDs and unique hashes stored as UUIDs; and 4) a hashset of mapped genes using the Fowler–Noll–Vo (FNV) hash function on each gene for each unique gene set. FNV is known to perform well when dealing with small keys. This is the case in our implementation which uses 32-bit unsigned integer keys. In our tests, FNV performed much faster than the default hasher. These data structures are created by querying the database with Rust. When the user presses the search button, the queried gene sets are forwarded to the API. After ensuring that the index is initialized, the code maps the user submitted gene set to a U32 hash set. It then computes the intersections between the user's gene set and the gene sets in memory and performs the Fisher's exact test using the identified overlap. Parallel processing with Rayon<sup>36</sup> is employed to further speed up this process. Once

completed, Benjamini-Hochberg adjusted *p*-values are computed. Next, the results are sorted by *p*-value, temporarily cached, and returned. The gene sets in Rummagene are stored in a Postgres database<sup>37</sup>. A function in the Postgres database is responsible for mapping the gene symbols to UUIDs before passing them to the Rust API to obtain results. These returned results can be joined by ID with the gene sets and genes in the database to facilitate further filtering. In this way, the use of an API is transparent to the front-end which queries the database with PostGraphile powered GraphQL. By implementing an advanced fast search engine, we can offer an interactive real-time service to users of the Rummagene application. The Rummagene database is automatically updated once a week by processing all the new articles added to PMC in the past week to identify new gene sets in the supporting materials of these articles. When a batch of new gene sets are added to the database, a new reference of valid gene names is constructed with the complete set of genes in the database. At that time, the API is called to prepare the new gene name reference prior to removing the old reference. By automatically updating the database, we ensure that it will remain relevant and current long term with minimal effort.

### Extracting functional terms from column titles

To assess the contents of the extracted gene sets, the column titles for each table were examined to identify a variety of functional terms. Supplementary table titles often include DOI and other identification information, thus these were ignored when conducting this analysis. After separating column titles in each gene set, column titles were split on dashes, underscores, and periods. To identify gene sets in each column, each resulting string was examined to assess if it was an NCBI Entrez<sup>38</sup> approved human gene symbol or a listed synonym. All gene synonyms were subsequently converted to their official symbol. Although genes can be represented with integer identifiers, strings only containing numbers were ignored because after manual examination, we discovered that many of these as artifacts. Additionally, strings containing S succeeded by an integer were ignored considering the vast majority of these refer to the supplemental table number. Transcription factors and kinases were subsequently identified from the extracted gene symbols. To identify gene sets that may represent signatures, the strings 'up', 'down', and 'dn' were searched for in the split column titles. To identify tissues, cell types and cell lines present in the column titles, the Brenda Tissue Ontology (BTO)<sup>39</sup> official terms and synonyms were extracted, and exact matches were identified. For gene sets containing multiple BTO terms, they were hyphenated to capture, for instance, a cell type from a specific tissue.

### Visualization of the kinase and TF gene set libraries

For each extracted gene set, IDF vectors were computed using the Scikit-learn<sup>40</sup> Python package using the set of all included genes as the corpus. Using the Scanpy<sup>41</sup> Python package, Uniform Manifold Approximation and Projection (UMAP)<sup>8</sup> plots for different categories of gene sets were then generated from the IDF vectors and clusters were automatically computed using the Leiden algorithm<sup>42</sup>. To visualize broad patterns across the data, each point representing a gene set was colored based on the cluster, associated PMCID, and associated kinase or transcription factor, if applicable. By visualizing the kinase and TF gene set libraries we can observe higher level functional clusters of related kinases and TFs.

### Benchmarking transcription factor and kinase enrichment analyses

Consensus transcription factor and kinase gene set libraries were created by performing a metadata search of the Rummagene database by submitting the kinase or transcription factor named entities as the search term. Returned entries are matches where the transcription factor or kinase terms appear in the gene set's table title, table legend, or column legend. The gene set for each transcription factor and kinase is composed from the union of all identified gene sets corresponding to the given transcription factor or kinase. Benchmarking datasets were sourced from ChEA3<sup>2</sup> for transcription factors and from KEA3<sup>9</sup> for kinases. To benchmark enrichment analysis performed

with the constructed consensus gene set libraries, the rank of each transcription factor/kinase was identified using the Fisher's exact test  $p$ -value for each matching gene set in each benchmarking dataset. To generate ROC curves, we downsampled the negative class to the same size as the positive class to achieve class balance. ROC curves were then bootstrapped over 5000 iterations and the mean ROC and AUCs were reported. Since we are randomly downsampling the negative class, bootstrapping the curve over several thousand iterations ensures a more accurate depiction of the ability of the Rummagene transcription factor and kinases gene set libraries to accurately predict the perturbed transcription factor or kinase. The numpy interp function was used to linearly interpolate between all points from the 5000 ROC curves to generate composite ROC for each benchmarking library.

### Topic modeling

To identify the predominant topics associated with gene sets in the Rummagene database, the abstracts of each paper contributing at least one gene set were assembled from the PMC bulk download. The text contained within the <abstract> tags was concatenated. Papers containing no abstracts were excluded from the analysis. Each abstract was then tokenized, stop words were removed, and lemmatized using the Python package Natural Language Toolkit (NLTK)<sup>43</sup>. The LdaModel class of Python package Gensim<sup>44</sup> was then used to identify nine topics with a chunksize of 100 over 10 passes. The number of topics was chosen manually by observing the separation of topics given different sets of parameters. Word counts and word importance were extracted from the model for each of the nine topics. The abstracts were visualized in topic space using the vectors produced by the latent Dirichlet allocation (LDA) model<sup>11</sup> for adherence of each paper to each topic using t-SNE<sup>25</sup>.

### Similar gene set pairs that are distant in abstract space

The preprocessing of publications' abstracts followed the same procedure as in topic modeling where abstracts were first extracted from the PMC bulk download, then cleaned of stopwords and lemmatized using the NLTK<sup>43</sup> Python package. Abstracts were then converted to word counts using the count vectorizer and subsequently fit to term frequency - inverse document frequency (TF-IDF) vectors using the Scikit-learn<sup>40</sup> Python package. The cosine similarity of each paper abstract to all other abstracts was then assessed using the Scikit-learn pairwise linear kernel metric based on the computed TF-IDF vectors. Only pairs of gene sets from different publications with zero cosine similarity of their abstracts were retained. For each pair of such gene sets, Fisher's exact test was performed to assess the significance of the overlap among the genes within these two sets. Only pairs with  $p < 0.05$  were retained for further analysis. Pairs with identical gene sets were excluded. Pairs were further filtered to only include those with overlaps of more than 50 genes. Additionally, to assess novelty of the recovered pairs, the percentage of their overlapping genes with 'sticky proteins' identified in analysis of protein-protein interactions<sup>45</sup> were used (Supplementary Data 9). In the analysis of gene set pairs including a gene or a disease in the table or column title and legend, only the top 10,000 most significant pairs with < 10% 'sticky proteins' were included. To assess the amount of highly cited genes, present in the overlapping genes of gene set pairs, the top 500 most cited genes according to GeneRIF<sup>38</sup> were used (Supplementary Data 9). Additionally, to determine the amount of highly expressed genes present in the overlapping genes of gene set pairs, the top 500 most highly expressed protein coding genes were sourced based on mean expression across 5000 random samples from ARCHS4<sup>6</sup> (Supplementary Data 9). To identify disease names in column titles of the gene set pairs, DisGeNet<sup>46</sup> disease terms were used and gene names were identified using NCBI gene<sup>38</sup> mappings. The OpenAI API chat completion module using the GPT-4 model was utilized to hypothesize about the connection between the remaining top pairs of gene sets from the subset of filtered genes sets based on the filtering steps described above. When prompting the model, we provide it with the gene set terms, the abstracts of both papers, as well as any identified disease or gene extracted from the gene set term column title in following format:

"Based on the pair of extracted gene sets from two research publications, hypothesize why there might be a connection between these gene sets based on the two abstracts, and the provided gene and disease terms: Gene set term 1: [term1], disease from gene set 1 term: [disease], abstract of publication for gene set term 1: [term1\_abstract], Gene set term 2: [term2], gene(s) from gene set 2 term: [term2\_abstract], abstract of publication for gene set term 2: [term2\_abstract]." Additionally, the system message explains the task as follows: "You are a biologist who attempts to generate a hypothesis about why two gene sets, which are lists of genes, may have a high overlap despite being extracted from two publications that have dissimilar abstracts. The gene set/paper pairs you will be given have one gene set with a disease term and the other with a gene name, so you should include reasoning as to a possible connection between the disease and the gene and explain this possible connection. Such a connection should be related to the abstracts." The response from the model along with statistics about the significance of the overlap and a PubMed query with the disease and the gene is provided to help uncover if this association is already published in literature.

### Gene function predictions

50,000 gene sets were randomly selected from Rummagene and filtered for sets with less than 2000 genes. For all human genes, we formed a matrix  $A$  where  $A(i, j) = 1$  if gene  $i$  is a member of gene set  $j$  and 0 otherwise. Then the co-occurrence matrix  $\Phi = A \cdot A^T$ . As previously described<sup>20</sup>, the co-occurrence probability between two genes:

$$P(\alpha, \beta) = \frac{\Phi(\alpha, \beta)}{\phi_0},$$

where  $\phi_0$  is the total number of co-occurrences, and the marginal probability  $P(\alpha) = \frac{1}{\phi_0} \sum \Phi(\alpha, \beta)$ .

The cosine similarity, Jaccard index, and normalized pointwise mutual information (NPWMI) for each pair of genes were then calculated as follows:

$$\text{Cosine}(\alpha, \beta) = \frac{P(\alpha, \beta)}{\sqrt{P(\alpha)P(\beta)}}$$

$$\text{Jaccard}(\alpha, \beta) = \frac{P(\alpha, \beta)}{P(\alpha) + P(\beta) - P(\alpha, \beta)}$$

$$\text{NPWMI}(\alpha, \beta) = \frac{-1}{\ln(P(\alpha, \beta))} \cdot \max \left\{ 0, \ln \left( \frac{P(\alpha, \beta)}{P(\alpha)P(\beta)} \right) \right\}$$

The NPWMI is a value between 0 and 1, where a larger value indicates the two genes co-occur with greater probability than expected by random chance<sup>48</sup>. Four gene set libraries were used to benchmark gene function prediction: GO Biological Process (2023), GWAS Catalog (2023), MGI Mammalian Phenotypes (2021), and Human WikiPathways (2021). To perform the predictions of the likelihood that a gene belongs to a gene set, we measured the distance of each gene to each gene set in each library by computing the average distance of the gene to each gene in each gene set. Suppose  $L$  is a matrix where  $L(i, j) = 1$  if gene  $i$  is a member of gene set  $j$  in the library  $L$ , and 0 otherwise. Let  $D$  be the similarity matrix as described above, where the diagonal is set to 0. The gene/gene-set association matrix  $G = \frac{D \cdot L}{L \cdot 1^T}$  where the division is elementwise. Each entry  $G(i, j)$  is then the mean similarity of gene  $i$  to all the genes in gene set  $j$ . The matrix  $G$  can then be used to predict membership of gene  $i$  in any gene set. ROC curves and AUC values for each term in the library were computed using the Python sklearn.metrics module<sup>40</sup>.

### Comparing the Rummagene gene set space to the Enrichr gene set space

All the gene set libraries in Enrichr were assembled and processed together with the Rummagene gene sets so they can be projected into the same two-

dimensional space. First, all genes were mapped to their official NCBI gene symbols for *Homo sapiens* or filtered out. Gene sets were then converted into vectors with values corresponding to the inverse document frequency (IDF)<sup>49</sup>. Truncated Singular Value Decomposition (Truncated SVD)<sup>50</sup> was then used to reduce the dimensionality of the IDF vectors to the 50 largest singular values. A UMAP<sup>25</sup> with the default settings was then used to embed all samples into two dimensions. Finally, to better position the visualization, we computed the mean and standard deviation of the embedding dimension axes and show the bulk of the samples that are within 1.68 standard deviations from the mean.

### Data availability

The Rummagene dataset version analyzed here is available for download from: <https://rummagene.com/download> and from Figshare<sup>51</sup>. The most recent updated version of the Rummagene dataset is also available from <https://rummagene.com/download>. This dataset is updated weekly on Mondays. Additional files needed to reproduce the results are provided as Supplementary Data files.

### Code availability

The Rummagene web server application is available from: <https://rummagene.com/>. The Rummagene source code is available from: <https://github.com/MaayanLab/rummagene> and a snapshot of the source code was deposited in Figshare<sup>52</sup>. The code and files needed to reproduce the figures are available from: <https://github.com/MaayanLab/rummagene/tree/main/figures>.

Received: 11 October 2023; Accepted: 10 April 2024;

Published online: 20 April 2024

### References

- Manzoni, C. et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief. Bioinform.* **19**, 286–302 (2018).
- Keenan, A. B. et al. ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Res.* **47**, W212–W224 (2019).
- Lachmann, A. et al. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438–2444 (2010).
- Hammal, F., de Langen, P., Bergon, A., Lopez, F. & Ballester, B. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res.* **50**, D316–D325 (2022).
- Wilks, C. et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol.* **22**, 323 (2021).
- Lachmann, A. et al. Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **9**, 1366 (2018).
- Shin, M.-G. & Pico, A. Using Published Pathway Figures in Enrichment Analysis and Machine Learning. *bioRxiv*. <https://doi.org/10.1101/2023.07.06.548037>. (2023).
- McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).
- Kuleshov, M. V. et al. KEA3: improved kinase enrichment analysis via data integration. *Nucleic Acids Res.* **49**, W304–W316 (2021).
- Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
- Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
- Hémono, M., Haller, A., Chicher, J., Duchêne, A.-M. & Ngondo, R. P. The interactome of CLUH reveals its association to SPAG5 and its co-translational proximity to mitochondrial proteins. *BMC Biol.* **20**, 13 (2022).
- Bileck, A. et al. Inward Outward Signaling in Ovarian Cancer: Morpho-Phospho-Proteomic Profiling Upon Application of Hypoxia and Shear Stress Characterizes the Adaptive Plasticity of OVCAR-3 and SKOV-3 Cells. *Front. Oncol.* **11**, 746411 (2021).
- Rolfs, F., Piersma, S. R., Dias, M. P., Jonkers, J. & Jimenez, C. R. Feasibility of Phosphoproteomics on Leftover Samples After RNA Extraction With Guanidinium Thiocyanate. *Mol. Cell. Proteom.* **20**, 100078 (2021).
- Monsivais, D. et al. Mass-spectrometry-based proteomic correlates of grade and stage reveal pathways and kinases associated with aggressive human cancers. *Oncogene* **40**, 2081–2095 (2021).
- Mooser, C. et al. Treacle controls the nucleolar response to rDNA breaks via TOPBP1 recruitment and ATR activation. *Nat. Commun.* **11**, 123 (2020).
- Salaverry, L. S. et al. Metabolic plasticity in blast crisis-chronic myeloid leukaemia cells under hypoxia reduces the cytotoxic potency of drugs targeting mitochondria. *Discov. Oncol.* **13**, 60 (2022).
- Shen, H. et al. Integrated Molecular Characterization of Testicular Germ Cell Tumors. *Cell Rep.* **23**, 3392–3406 (2018).
- Lachmann, A. et al. Geneshot: search engine for ranking genes from arbitrary text queries. *Nucleic Acids Res.* **47**, W571–W577 (2019).
- Ma’ayan, A. & Clark, N. R. Large Collection of Diverse Gene Set Search Queries Recapitulate Known Protein-Protein Interactions and Gene-Gene Functional Associations. *arXiv [q-bio.MN]* (2016).
- Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
- MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
- Smith, C. L., Goldsmith, C.-A. W. & Eppig, J. T. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.* **6**, R7 (2005).
- Pico, A. R. et al. WikiPathways: pathway editing for the people. *PLoS Biol.* **6**, e184 (2008).
- Van Der Maaten, L., Postma, E. O., van den Herik, H. J. & Others. Dimensionality reduction: A comparative review. *J. Mach. Learn. Res.* **10**, 13 (2009).
- Campello, R. J. G. B., Moulavi, D. & Sander, J. Density-based clustering based on hierarchical density estimates. in *Advances in Knowledge Discovery and Data Mining* 160–172 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013).
- Evangelista, J. E. et al. SigCom LINCS: data and metadata search engine for a million gene expression signatures. *Nucleic Acids Res.* **50**, W697–W709 (2022).
- Subramanian, A. et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437–1452.e17 (2017).
- Tabula Sapiens Consortium\*. et al. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
- Lasso, G. et al. A Structure-Informed Atlas of Human-Virus Interactions. *Cell* **178**, 1526–1541.e16 (2019).
- Hanspers, K., Riutta, A., Summer-Kutmon, M. & Pico, A. R. Pathway information extracted from 25 years of pathway figures. *Genome Biol.* **21**, 273 (2020).
- Talley, E. M. et al. Database of NIH grants using machine-learned categories and graphical clustering. *Nat. Methods* **8**, 443–444 (2011).
- Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
- Gamble, A. PubMed Central (PMC). *Charlest. Advisor* **19**, 48–54 (2017).
- Brown, G. R. et al. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* **43**, D36–D42 (2015).

36. Pieper, R., Löff, J., Hoffmann, R. B., Griebler, D. & Fernandes, L. G. High-level and efficient structured stream parallelism for rust on multi-cores. *J. Computer Lang.* **65**, 101054 (2021).
37. Obe, R. O. & Hsu, L. S. *PostgreSQL: Up and Running: A Practical Guide to the Advanced Open Source Database*. (“O’Reilly Media, Inc.,” 2017).
38. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* **33**, D54–D58 (2005).
39. Gremse, M. et al. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res* **39**, D507–D513 (2011).
40. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
41. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
42. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
43. Bird, S., Klein, E. & Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. (“O’Reilly Media, Inc.,” 2009).
44. Rehurek, R. & Sojka, P. Gensim—python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University* (2011).
45. Mazloom, A. R. et al. Recovering protein-protein and domain-domain interactions from aggregation of IP-MS proteomics of coregulator complexes. *PLoS Comput. Biol.* **7**, e1002319 (2011).
46. Piñero, J. et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* **45**, D833–D839 (2017).
47. Sun, B. B. et al. Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. *bioRxiv* 2022.06.17.496443 (2022) <https://doi.org/10.1101/2022.06.17.496443>.
48. Chiarcos, C., de Castilho, R. E. & Stede, M. *Von Der Form Zur Bedeutung: Texte Automatisch Verarbeiten: From Form to Meaning: Processing Texts Automatically. Proceedings of the Biennial GSCL Conference 2009*. (Narr Francke Attempto Verlag, 2009).
49. Karen, S. J. A statistical interpretation of term specificity and its application in retrieval. *J. Documentation* **28**, 11–21 (1972).
50. Chicco, D. & Masseroli, M. Software Suite for Gene and Protein Annotation Prediction and Similarity Search. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **12**, 837–843 (2015).
51. Clarke, D. J. B. et al. Rummagene gene sets with descriptions 01172024. figshare. Dataset. <https://doi.org/10.6084/m9.figshare.25017023.v3> (2024).
52. Clarke, D. J. B. et al. Rummagene source code snapshot from 03132024. figshare. Software. <https://doi.org/10.6084/m9.figshare.25404637.v1> (2024).

## Acknowledgements

This study is partially supported by NIH grants OT2OD030160, U24CA264250, RC2DK131995, and U24CA224260.

## Author contributions

D.J.B.C. and G.B.M. developed the website, wrote the manuscript, performed data analysis, and produced figures. D.J.B.C. wrote the crawler and developed the code for the fast search engine. Z.X. and E.Z.D. wrote the manuscript and performed data analysis. J.E.E. contributed to the data analysis. A.M. conceived the project, wrote the manuscript, managed the project, and was responsible for funding the project.

## Competing interests

The authors declare competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s42003-024-06177-7>.

**Correspondence** and requests for materials should be addressed to Avi Ma’ayan.

**Peer review information** *Communications Biology* thanks Alexander R. Pico and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Chien-Yu Chen and Tobias Goris.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024