Using Published Pathway Figures in Enrichment Analysis and Machine Learning

Min-Gyoung Shin[1] and Alexander R. Pico*[1]

[1] Institute of Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA, USA

*Corresponding author

Abstract

Pathway Figure OCR (PFOCR) is a novel kind of pathway database approaching the breadth and depth of Gene Ontology while providing rich, mechanistic diagrams and direct literature support. PFOCR content is extracted from published pathway figures currently emerging at a rate of 1000 new pathways each month. Here, we compare the pathway information contained in PFOCR against popular pathway databases with respect to overall and disease-specific coverage. In addition to common pathways analysis use cases, we present two advanced case studies demonstrating unique advantages of PFOCR in terms of cancer subtype and grade prediction analyses.

Keywords: pathway database, database comparison, enrichment analysis, machine learning, disease mechanism

Background

For the past few decades pathway databases have relied on the manual extraction of pathway knowledge from the literature by teams of biocurators ranging from small, centralized groups to communities of hundreds of contributors [1–6]. Individuals comb the literature and contribute their domain knowledge in order to model the vast diversity of biochemical reactions and cellular processes comprising biological systems. Pathway databases are commonly used in enrichment analysis (a.k.a., pathway analysis), where they are reduced to collections of gene sets and tested for statistical overrepresentation or enrichment with respect to a researcher-provided gene set [7,8]. The connections and mechanistic details present in pathway models are still relevant for interpretation and visualization of enrichment results and are a key advantage of pathway databases over typical gene set collections such as GO. However, pathway databases also suffer disadvantages relative to simpler forms of gene set annotation, namely limited breadth, depth and curation throughput.

While GO Biological Process terms annotate 62% of human genes, pathway databases only cover up to 44% [9]. This limited breadth means that a significant percentage of a researcher's genes of interest (e.g., from a differential expression dataset) would be essentially excluded from an enrichment analysis as they would yet be included in *any* pathway model. In terms of depth, pathway databases have traditionally focused on canonical pathways, for example having a single, generic representation of "apoptosis" or "hippo signaling", despite the actual diversity of these biological processes across cell types, developmental stages,

disease states and conditions. This oversimplification is understandable given the low throughput of pathway database content. Constructing a pathway model takes significant time and effort, including information gathering, synthesis, encoding and review [10,11]. Keeping up with the continuous flood of published findings spanning all aspects of cellular biology with manual curation is clearly a Sisyphean task. The WikiPathways project has had some success in addressing these challenges, sharing the burden of biocuration with any interested member of the research community [5,12]. Steadily acquiring ~90 new pathways a year and ~250 edits per month by a total of over 700 contributors, the WikiPathways project still pales in comparison to the volume of unique pathway diagrams routinely published in the literature as static images.

The Pathway Figure OCR (PFOCR) project takes a more direct approach to capturing pathway information in a pathway database [13]. Collecting 1,000 published pathway figures per month in recent years and a total of 79,949 pathway figures since 1995 from the indexes of PubMed Central, the project has extracted 1.5 million human genes (14,253 unique; 60% of all human genes), 218 thousand chemicals (11,100 unique), and 29 thousand disease names (1,204 unique) via a pipeline involving machine learning, optical character recognition (OCR), and named entity recognition (NER). The PFOCR database contains more unique genes than any other pathway database and is comparable in breadth to Gene Ontology. With remarkably little redundancy, PFOCR contains many dozens of unique instances of processes like apoptosis and hippo signaling that typically have only canonical representations in pathway databases. In terms of throughput, the algorithmic steps of the PFOCR project could feasibly generate an up-to-date database in lockstep with the indexing efforts of PubMed Central.

In this paper, we will provide a detailed characterization of the PFOCR collection of pathways with a particular focus on pathways and genes with disease associations. We will also demonstrate the utility of PFOCR in comparison with other pathway databases in enrichment analyses and machine learning. While PFOCR is applicable to any analysis involving pre-defined gene sets, we will highlight specific types of enrichment analyses and machine learning where its unique breadth and depth are crucial to fruitful results and insightful interpretation.

Results

Disease Coverage Comparison

A common goal of pathway analysis is to identify processes underlying a disease state. Thus, an important characteristic for a pathway database is its disease coverage, i.e., how many diseases are represented among its pathway-defined gene sets and how deep and diverse are the gene sets per disease. PFOCR has close to a hundred times as many pathways as a canonical pathway database, but how does their distribution across diseases compare to an intentionally curated database? In order to assess disease coverage, a standard set of 876 distinct disease names was compiled from the Comparative Toxicogenomics Database (CTD) [14]. The titles and descriptions associated with pathways from WikiPathways, Reactome, and KEGG were queried for disease name occurrences. Similarly, the titles and captions annotating PFOCR content were queried.

A total of 791 (90%) diseases were represented by at least one pathway in PFOCR. Reactome, WikiPathways and KEGG represented 153 (17%), 127 (14%), and 94 (11%) diseases, respectively (**Supplemental Table 1**). In terms of depth, we focused on the top 20 disease matches for each database; the union of these matches comprise the 37 diseases shown in **Table 1**. PFOCR includes 1954 pathways relevant to *breast cancer* and offers 649 pathways on average among its top 20 diseases. Reactome holds 23 pathways relevant to *leukemia* and 7 pathways on average among its top 20 diseases. WikiPathways has 18 pathways relevant to *SARS-CoV-2*—which aligns with its unique, community-driven approach to collecting new content on emerging topics—and also has 7 pathways on average among its top 20 diseases. KEGG has a four-way tie with 5 pathways for each of *lung cancer, insulin resistance, hepatitis,* and *cardiomyopathy,* and only 3 pathways on average among its top 20 diseases. By contrast, PFOCR has 5 or more pathways for 351 diseases and at least 1 pathway for 555 diseases that are missing from the other pathway databases (**Supplemental Table 1**).

Table 1. Disease Coverage Comparison. Union of top 20 disease-related pathways for each database by searching associated pathway titles and descriptions or captions. Disease names were compiled from the Comparative Toxicogenomics Database [14]. Average pathway counts among the top 20 disease matches per database are given in parentheses in column headers. The pathway count for the most represented disease per database is in bold italics. *Abbreviations: AML = acute myeloid leukemia, CDG = congenital disorders of glycosylation.

Disease	PFOCR (649)	Reactome (7)	WikiPathways (7)	KEGG (3)
Breast cancer	***1954***	20	6	3
Fibrosis	959	4	8	4
Alzheimer	874	6	12	2
Colorectal cancer	817	10	5	1
Prostate cancer	794	1	7	1
Lung cancer	792	7	7	***5***
Lymphoma	692	9	2	3
Leukemia	673	***23***	6	3
Melanoma	659	8	4	1
Insulin resistance	586	1	4	***5***
Obesity	529	4	8	4
Ischemia	483	2	2	0
Parkinson	464	4	6	2
Pancreatic cancer	435	1	6	1
Tuberous sclerosis	413	1	0	0
Hepatitis	397	4	3	***5***
Hypertrophy	380	0	6	4
Arthritis	372	6	4	1
Hypertension	358	2	4	2
Glioblastoma	343	8	2	1
Atherosclerosis	300	6	4	3
SARS-CoV-2	265	3	***18***	1
Retinoblastoma	204	3	5	2
Heart failure	185	0	2	3
Anemia	179	6	1	1
Diabetes mellitus	177	1	3	4
Schizophrenia	146	4	2	0
Renal cell carcinoma	142	0	5	1
Cardiomyopathy	133	1	1	***5***
Huntington	121	1	3	2
Multiple myeloma	120	4	0	0
Hyperplasia	100	2	2	3
AML*	96	6	1	1
Ulcer	67	1	0	3
Hyperlipidemia	33	0	6	0
Hemophilia A	3	4	0	0
CDG*	2	4	1	1

In order to assess gene coverage among disease pathways, we referenced the matching disease names in Jensen DISEASES, a database that provides gene-disease associations retrieved from text mining, literature, cancer mutation data, and genome-wide association studies [15]. **Table 2** shows these 17 matching diseases and the corresponding number of Jensen DISEASES genes covered by each pathway database. PFOCR covers 53 of cardiomyopathy genes (62%), which is the maximum count, and covers 63% of disease genes on average. Reactome has a maximum count of 12 (48%) for breast cancer and 9% on average. WikiPathways has a maximum count of 15 (19%) for diabetes mellitus and 21% on average. The maximum count by KEGG is for cardiomyopathy at 31 (36%) and the average coverage is 28%. For each of the diseases in **Table 2**, PFOCR covers a higher percentage of disease-associated genes than the other pathway databases. Interestingly, with the exception of diabetes mellitus, PFOCR includes over half of the genes associated with a given disease in pathway models labeled by those diseases.

Table 2. Disease Gene Coverage Comparison. Comparison of the disease-related gene content in the top disease-annotated pathways from Table 1 based on matching reference gene sets available from Jensen DISEASES. Absolute gene counts together with percentage of DISEASES genes in parentheses are provided for each database-disease pair. The gene counts for the most represented disease per database is in bold italics. An average percentage of DISEASES genes per database is given in parentheses in column headers.

Disease	PFOCR (63%)	Reactome (9%)	WikiPathways (21%)	KEGG (28%)
Cardiomyopathy	***53 (62%)***	2 (2%)	14 (16%)	***31 (36%)***
Diabetes mellitus	31 (39%)	2 (3%)	***15 (19%)***	18 (23%)
Obesity	45 (71%)	4 (6%)	8 (13%)	12 (19%)
Parkinson	34 (64%)	3 (6%)	11 (21%)	15 (28%)
Prostate cancer	23 (64%)	2 (6%)	6 (17%)	5 (14%)
Alzheimer	18 (58%)	9 (29%)	9 (29%)	9 (29%)
Melanoma	21 (70%)	5 (17%)	6 (20%)	7 (23%)
Lung cancer	24 (86%)	7 (25%)	13 (46%)	17 (61%)
Breast cancer	23 (92%)	***12 (48%)***	14 (56%)	13 (52%)
Hypertension	7 (88%)	2 (25%)	3 (38%)	3 (38%)
Atherosclerosis	5 (63%)	1 (13%)	0 (0%)	3 (38%)
Retinoblastoma	2 (67%)	0 (0%)	1 (33%)	1 (33%)
Huntington	1 (50%)	0 (0%)	1 (50%)	1 (50%)
Renal cell carcinoma	1 (50%)	0 (0%)	0 (0%)	0 (0%)
Schizophrenia	8 (50%)	1 (6%)	1 (6%)	0 (0%)
Multiple myeloma	4 (50%)	1 (13%)	0 (0%)	0 (0%)
Tuberous sclerosis	2 (100%)	0 (0%)	0 (0%)	0 (0%)

General Pathway Analysis

In practice, PFOCR is available in a format commonly used by enrichment analysis algorithms (see GMT in Data Availability). Thus, PFOCR can fit into analytical pipelines involving any package supporting GMT files, including the top-ranked Bioconductor packages: fgsea [16], clusterProfiler [17], GSEABase [18], and GSVA [19]. The pathway figure database has already been incorporated into online tools such as Enrichr[20] and NDEx iQuery [21].

Enrichr (https://maayanlab.cloud/Enrichr) provides quick and easy enrichment analysis against over 200 gene set databases simultaneously. Among the 27 databases in the pathway category, PFOCR has the greatest number of gene sets and the second largest coverage of human genes (the kinase co-expression gene sets by ARCHS4 has the largest). PFOCR ranks fourth among all 209 Enrichr databases in terms of gene set size. Among the suite of plots provided by the Enrichment Analysis Visualizer "appyter" for Enrichr databases [22], a UMAP helps to visualize the density and diversity of gene sets per database (**Figure 1A**). The appyter analysis identified 35 clusters of PFOCR pathways. By contrast, GO Biological Process divides into only 18 clusters, Reactome 27, WikiPathways 8, and KEGG 11, thus characterizing the greater diversity of pathway-based gene sets in PFOCR.

NDEx Integrated Query (iQuery, https://www.ndexbio.org/iquery) provides network and pathway gene set analysis using multiple methods simultaneously against PFOCR, WikiPathways, INDRA-connected GO terms [23], and various interactomes [21]. Like Enrichr, the input is a simple list of gene names, and results are produced immediately. With a focus on networks and pathways, however, NDEx iQuery offers more detailed views of resulting gene sets ranked by similarity, p-value, or overlap (**Figure 1B**). In the case of PFOCR, the detailed view includes the original published figure next to its title, caption and other article metadata. Each pathway figure is associated with an interactive gene list and links to a dedicated page in the NDEx database. The PFOCR's pathway figure-based gene sets can also be imported as nodes into Cytoscape with a single click.

We also developed an R Shiny tool called Interactive Enrichment Analysis (https://github.com/gladstone-institutes/Interactive-Enrichment-Analysis) to perform two types of enrichment analyses for one or more datasets simultaneously against GO, WikiPathways, and PFOCR [24]. The tool supports interactive exploration of results with customizable plots (volcano, dot, bar, heatmap, emap,

and GSEA plots) and embedded pathway views (**Figure 1C**). In addition to views of original published figures, PFOCR results include links to dedicated web pages at the PFOCR database, which include a rich collection of metadata, crosslinks to PubMedCentral, NDEx and WikiPathways, and downloadable tables of extracted genes, chemicals and disease terms.

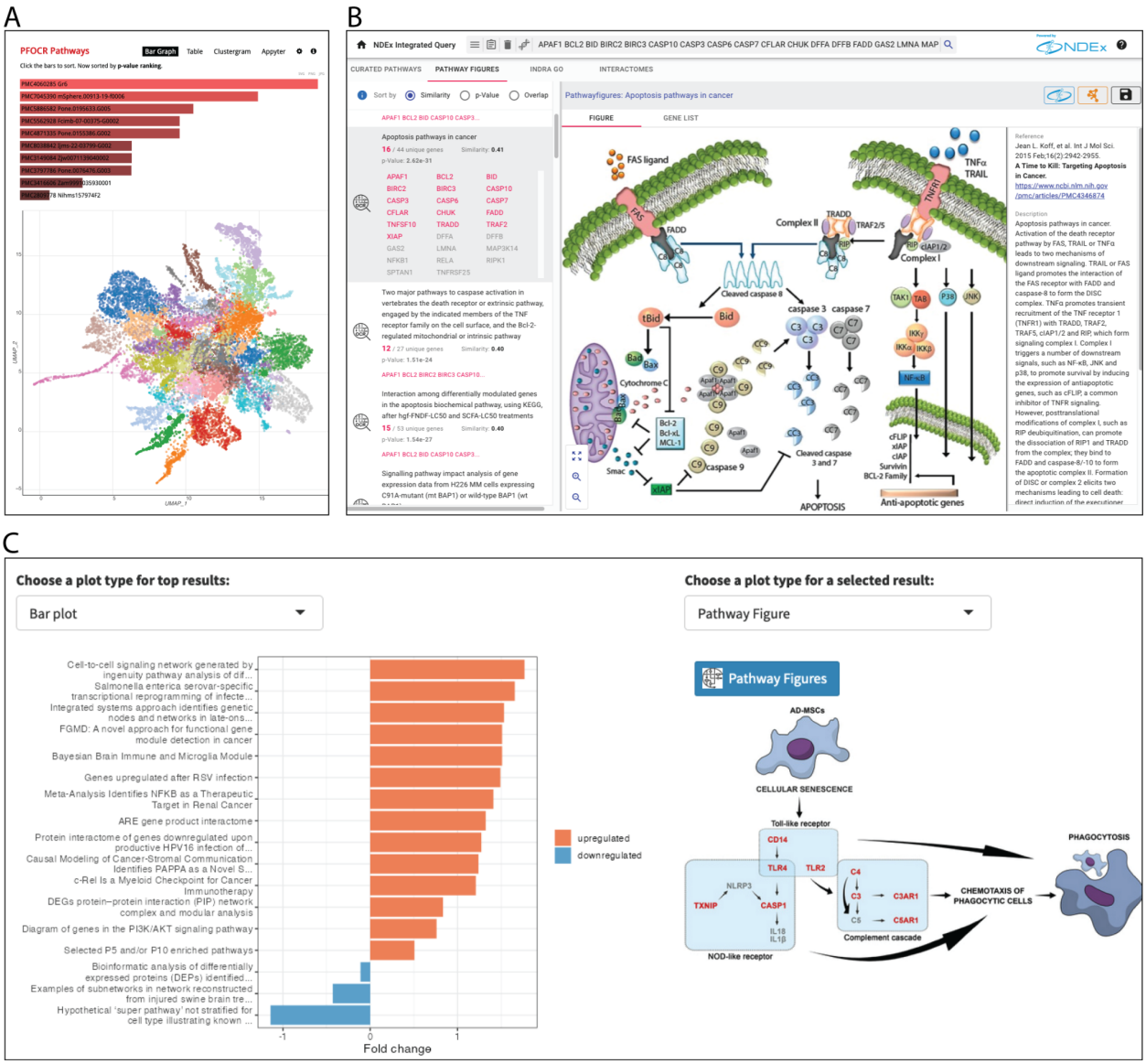

Figure 1. Pathway Analysis with PFOCR. (A) Typical bar graph of Enrichr results for PFOCR pathways, above. Appyter UMAP of PFOCR pathway clusters, below. The Enrichment Analysis Visualizer appyter computed term frequency-inverse document frequency (TF-IDF) values for the gene set corresponding to each pathway in PFOCR and plotted the first two dimensions of a UMAP applied to the resulting values. Generally, pathways with more similar gene sets are positioned closer together. Pathways are colored by Leiden algorithm-computed clusters. (B) NDEx iQuery screenshot showing ranked pathway figures (left panel), original pathway figure of selected result (middle), and article metadata (right). The top-right includes buttons to a dedicated NDEx page or to import the figure-extracted gene set into Cytoscape as nodes. (C) The plot options for ORA and GSEA results in the R Shiny tool called Interactive Enrichment Analysis. PFOCR results include a view of the original published figure and a link to a dedicated PFOCR website page (right).

In the next two sections, we demonstrate advanced pathway analyses on two different disease datasets, tailoring methodology for each and demonstrating the level of interpretive detail that is only possible with a pathway database having PFOCR's unique breadth and depth.

Case Study 1: Acute Myeloid Leukemia Subtype Analysis

Leukemia is a fatal disease with a 5-year overall survival rate of 24% and a long-term survival rate of less than 20% in adulthood [25–28]. Among different types of leukemia, acute myeloid leukemia (AML) is characterized by clonal disorders of the hematopoietic compartment, such as abnormal proliferation of undifferentiated myeloid progenitors, impaired hematopoiesis, bone marrow failure and variable response to therapy [28]. Leukemia became a treatable disease with the development of drugs such as midostaurin, gilteritinib, and ivosidenib [29]. Studies have shown that leukemia drug efficacy is highly dependent on a patient's genetic subtype profile. For example, midostaurin provides excellent treatment for patients with FLT3 mutations, whereas ivosidenib is particularly effective for patients with IDH1 mutations [30]. Additionally, Smoothened (SMO) inhibitors, such as Glasdegib, control the progression of acute leukemia by specifically targeting the Hedgehog (Hh) signaling pathway [31]. This subtype specificity can be effectively understood by investigating the effects of drugs on signaling cascades at the pathway level, rather than at a single gene level, assessing perturbations caused by different mutations [32]. Likewise, characterizing subtypes based on pathway-level transcriptomic profiles can help develop effective therapeutic strategies to optimize best survival outcomes of leukemia patients.

We characterized the perturbed transcriptomic profiles of AML subtypes using leukemia pathways from PFOCR, WikiPathways, Reactome, and KEGG. The purpose of the analysis was to compare the effectiveness of these pathway databases in characterizing leukemia subtypes based on gene expression. AML patient gene expression data with 8 mutations was retrieved from GEO (GSE3494 and GSE2990) and PCA-based quality control was performed. We then performed differentially expressed gene (DEG) analysis, where each mutation type was treated as a separate group and control samples were used as the reference, followed by gene set enrichment analysis (GSEA). Finally, hierarchical clustering was applied to the normalized enrichment scores (NES) from GSEA to cluster leukemia mutations in operational subtypes.

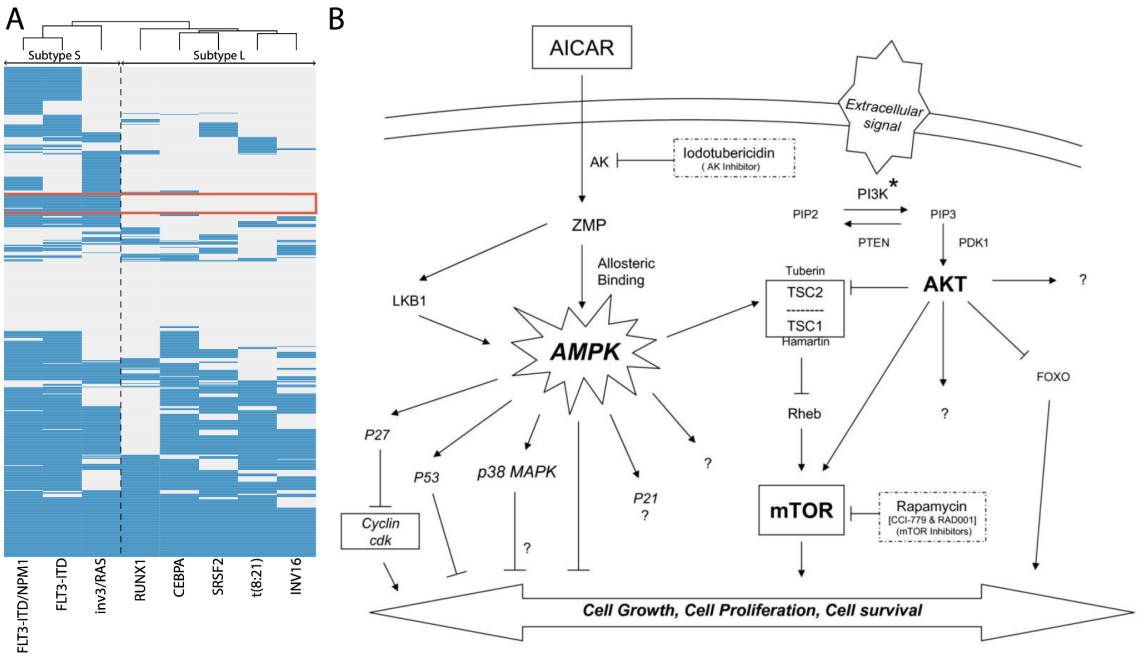

Figure 2. Heatmap of GSEA NES values from leukemia expression data. (A) Hierarchical clustering of GSEA NES values of leukemia mutations defines two top-level subtypes, denoted S and L. The cluster of 26 pathways that differentiates the two subtypes is highlighted with a red box. (B) PFOCR's PMC1948012__F10[33] is one of the 26 subtype-distinguishing pathways and contains the FOXO gene signaling pathway. This figure is reproduced from [28].

A total of 705 leukemia-enriched pathways were investigated by hierarchical clustering analysis, including 673 PFOCR, 6 WikiPathways, 23 Reactome, and 3 KEGG pathways (**Figure 2A**). At the top level of the hierarchical cluster, there were two core subclusters where the larger cluster (subtype L) was defined by five mutations, RUNX1, INV16, t(8:21), CEBPA, and SRSF2, and the smaller cluster (subtype S) was defined by

three mutations, FLT3-ITD, FLT3-ITD/NPM1, and inv3/RAS. Among all mutations of subtype S, FLT3 mutation is associated with the most unfavorable prognosis [29,34]. FLT3, a transmembrane ligand-activated receptor tyrosine kinase, is expressed in hematopoietic progenitor cells, and 25-30% of AML cases carry FLT3 mutations that result in abnormal cell growth and survival via mTOR and PI3K/AKT pathway [35]. In subtype L, RUNX1 is another well-known type of AML mutation associated with hematopoietic stem cell (HSC) growth, differentiation, and homeostasis. Its abnormalities are often associated with older age and male sex, and are found in 8–16% of AML patients [35].

Of the 705 leukemia pathways, there are 26 pathways that clearly distinguish subtypes L and S (**Figure 2A, boxed rows, Supplemental Table 2**). These 26 core leukemia pathways (CLGS) consist of 23 PFOCR pathways and 1 KEGG pathway. The predominance of PFOCR pathways in defining the CLGS highlights the utility of PFOCR in characterizing subtypes compared to other pathway databases. For example, one of the PFOCR pathways in the CLGS involves the regulation of forkhead box (FOX) genes through PI3K/AKT signaling cascade (**Figure 2B**). FOX genes are transcription factors involved in the regulation of multiple cellular functions, including development, differentiation, proliferation, and apoptosis [36–38], and may act as either tumor suppressors or oncogenes depending on the cellular and biological context [28,39]. In the context of leukemia, FOX genes have been reported to play different roles depending on the type of mutation. For example, upregulation of FOXO1 in AML with a RUNX1 mutation has been reported to help the growth of leukemia cells and inhibit the differentiation of CD34+ hematopoietic stem and progenitor cells [28]. In addition, the activity of FOXO1 has been reported to affect the antineoplastic drug sensitivity of AML cells, and has been proposed as a therapeutic for leukemia with RUNX1 mutations [40,41]. On the other hand, FOXO3 acts as a tumor suppressor, and phosphorylation of FOXO3 by FLT3-ITD results in inactivation of FOXO3-mediated apoptosis in leukemia with FLT3-ITD mutation [42].

As mentioned above, the success of FDA-approved leukemia drugs depends on a patient's tumor genetics. For example, a study tested the effect of midostaurin in leukemia patients with RUNX1 showed a limited role of the drug in leukemia control [43]. Midostaurin treatment in systemic mastocytosis patients has shown that patients with one or more mutations in the S/A/R (SRSF2, ASXL, or RUNX1) panel have a lower survival rate and a higher progression rate to AML or mast cell leukemia than patients without mutations. Additionally, the same study found that midostaurin treatment was not able to prevent the increase of RUNX1 mutations in patients which was associated with progression to secondary AML [43]. To date, there have been no officially approved drugs effective for leukemia with RUNX1 mutations [29,34,44].

The different roles of FOXO genes in RUNX1 leukemia and FLT3-ITD leukemia highlights the mechanistic differences between the two leukemia subtypes and why treatment of both subtypes with the same drug may fail. PFOCR-based gene set hierarchical clustering analysis is an effective methodology that can aid in understanding pathway-level mechanisms underlying the differences between cancer subtypes.

Case Study 2: Breast Cancer Prediction Analysis

To evaluate the disease prediction efficacy of these pathway databases, machine learning analysis was performed on breast cancer patient expression data. For this analysis, two independent breast cancer patient expression data sets were collected from GSE3494 and GSE2990, and samples with breast cancer grades 1 and 2 were selected for analysis. 51 samples with a grade value of 2 were randomly selected from GSE3494 to balance the number of samples in both grades. For the test data, GSE2990 was used, and there were 29 grade 1 and 31 grade 2 samples. Random forest was chosen to build predictive models, and tunned model parameters were selected based on out-of-sample error. On all pathways annotated as breast cancer, prediction accuracy was calculated from the test data and leave-one-out cross-validation accuracy was calculated from the training data. Gene importance (feature importance) was calculated as the average importance from the leave-one-out cross-validation iterations. The prediction accuracy rankings in **Table 3** shows 21 pathways with prediction accuracy and cross-validation accuracy larger than 0.65 (bold), which are all from PFOCR. **Supplemental Table 3** presents all results with prediction accuracy and cross-validation accuracy larger than 0.55 and min gene importance larger than 0.1, which include 695 PFOCR, 9 Reactome, and 1 KEGG pathways.

Table 3 Random forest breast cancer grade prediction accuracy with top important gene information. Pathways with cross-validation accuracy and prediction accuracy greater than 0.65 (in bold and all from PFOCR), plus the top cross-validation accuracy and prediction accuracy results from each of Reactome and KEGG (discontinuously below dashed line). The top results were determined by max(min(cross-validation accuracy, prediction accuracy)). The "Top Gene" corresponds to the gene with the highest feature importance score in each pathway. The "Ref." column provides paper and figure citations for each pathway.

Rank	Pathway	Ref.	Top Gene	Importance Score	CV Accuracy	Prediction Accuracy	Specificity	Sensitivity	F1
1	PMC2937358__F1	[45,46]	CXCL6	0.61	0.72	0.72	0.61	0.83	0.69
2	PMC2653381__F1	[47,48]	GSTM4	0.25	0.69	0.7	0.58	0.83	0.67
3	PMC5772637__F6	[49,50]	CASP9	0.91	0.69	0.68	0.84	0.52	0.73
4	PMC5772637__F7	[49,51]	CASP9	1.54	0.68	0.72	0.77	0.66	0.74
5	PMC8040471__F7	[52,53]	CASP9	0.45	0.68	0.67	0.81	0.52	0.71
6	PMC6759650__F5	[54,55]	CYCS	0.37	0.67	0.78	0.71	0.86	0.77
7	PMC7811378__F9	[56,57]	CASP9	0.85	0.67	0.78	0.84	0.72	0.8
8	PMC5715135__F7	[58,59]	CASP9	0.15	0.67	0.75	0.84	0.66	0.78
9	PMC2673236__F1	[60,61]	ACSM3	1.32	0.67	0.72	0.74	0.69	0.73
10	PMC387764__F8	[62,63]	CASP9	1.78	0.67	0.68	0.74	0.62	0.71
11	PMC6947643__F2	[64,65]	HEY1	1.23	0.67	0.67	0.84	0.48	0.72
12	PMC7409684__F1	[66,67]	WNT11	0.18	0.67	0.67	0.68	0.66	0.68
13	PMC8023395__F2	68,69]	EGFR	0.17	0.66	0.73	0.84	0.62	0.76
14	PMC4336604__F9	[70,71]	MAPK9	0.33	0.66	0.72	0.74	0.69	0.73
15	PMC6499473__F1	[72,73]	PARP1	0.23	0.66	0.72	0.61	0.83	0.69
16	PMC3219187__F5	[74,75]	ZBTB7A	2.29	0.66	0.7	0.81	0.59	0.74
17	PMC6024909__F4	[76,77]	CDK1	0.26	0.66	0.7	0.52	0.9	0.64
18	PMC2694962__F1	[78,79]	ACSM3	1.52	0.66	0.68	0.48	0.9	0.61
19	PMC5256616__F6	[80,81]	CASP9	0.36	0.66	0.68	0.48	0.9	0.61
20	PMC4407294__F2	[82,83]	NOTCH2	0.7	0.66	0.67	0.52	0.83	0.62
21	PMC6305585__F2	[84,85]	CPEB1	0.77	0.66	0.67	0.71	0.62	0.69
63	R-HSA-8864260	[86]	HSPD1	0.25	0.64	0.68	0.81	0.55	0.72
115	KEGG_hsa01522	[87]	BAX	0.11	0.62	0.85	0.9	0.79	0.86

To assess the overlap of gene content among the top 21 pathways from PFOCR and any of the results from the other databases that passed the minimum accuracy 0.55 threshold, we plotted the genes per pathway, ordered (and sized) by their importance scores (**Figure 3**). Remarkably, the majority of genes, including top scoring "important" genes, from the highest accuracy pathways are unique to PFOCR results.

CASP9, the gene with the second highest importance score in 21 PFOCR pathways, is an initiator of apoptosis in the mitochondrial apoptosis pathway. A study by Sharifi and Moridnia found an association of CASP9 expression with miR-182-5p [88]. The breast cancer cell line MCF-7 had poor viability when miR-182-5p was inhibited, suggesting CASP9 upregulation is related to MCF-7 cell viability. In addition, an independent study investigated SNPs in CASP9 to find an increased breast cancer risk in patients with CASP9 mutations. In particular, CASP9 SNPs rs4645978 and rs4645981 were associated with high breast cancer risk, suggesting that CASP9 contributes to breast cancer development [89].

Among the overlapping genes of PFOCR and REACTOME, HEY1, one of the well-known Notch target genes, had the highest shared gene importance score. Chen et al. investigated the expression of HEY1 in breast cancer cells and found that the HEY1 expression level increased under hypoxic conditions [90]. In addition, increased Notch4-Hey1 mRNA expression and decreased patient survival were found to be correlated in another study, confirming Hey1 as a marker for breast cancer development [91].

On the other hand, NOTCH2 was the gene with the highest importance score overlapping between PFOCR and KEGG. The role of NOTCH2 was studied by Fu et al. by investigating NOTCH2 expression and polymorphisms of SNP rs11249433 in breast cancer patient data [92]. That study suggested that increased expression of NOTCH2 with the rs11249433 polymorphism may contribute to the development of ER+ luminal tumors.

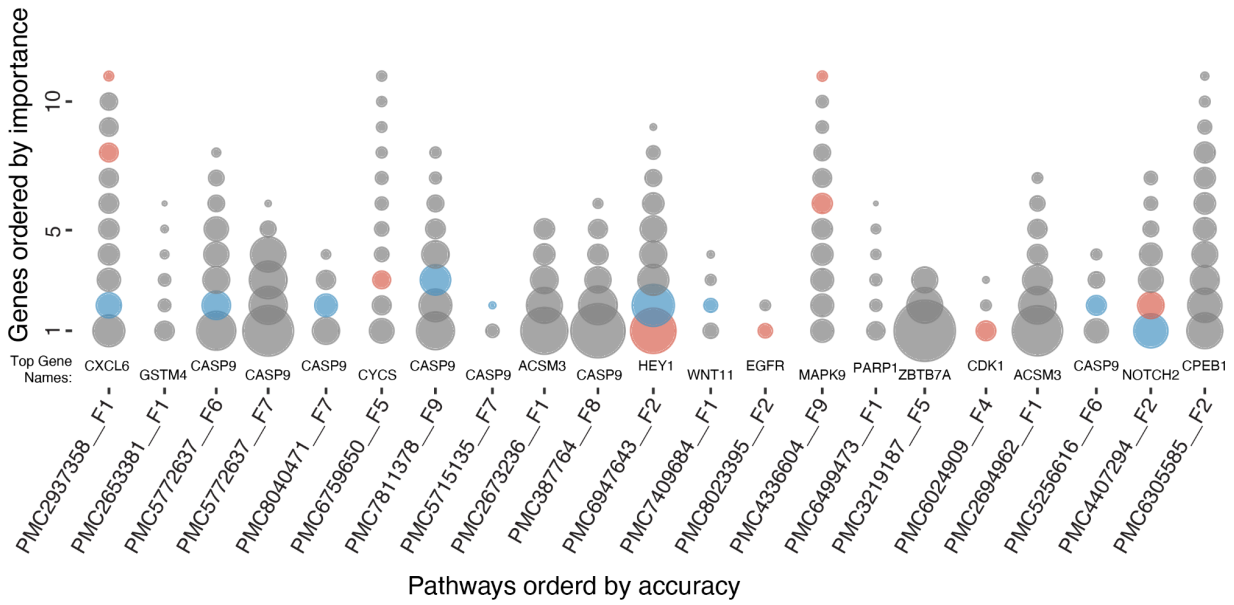

Figure 3. Gene importance from random forest breast cancer models built on pathway genes. Represented as circles, the gene content is sized and ordered by importance (large-to-small, bottom-to-top) for the top twenty-one pathways ordered by prediction accuracy (left-to-right). The "Top Gene" from Table 3 is represented by the largest, bottom-most circle and is labeled for each pathway. The vast majority of circles (gray) denote genes uniquely found in PFOCR pathways among all significant pathways with "important" genes. The genes that are also present within Reactome and KEGG results are highlighted in red and blue, respectively. There were no significant pathways with gene importance scores above 0.1 from WikiPathways.

Discussion

Using the Comparative Toxicogenomics Database (CTD) [14] for a comprehensive, independent set of 876 disease names, we chose the most comparable sources of metadata (titles and descriptions or captions) to assess the disease coverage by PFOCR and popular pathway databases: Reactome, WikiPathways, and KEGG. In terms of both breadth and depth, PFOCR surpassed the other pathway databases by a considerable margin. We assessed the most common diseases represented by each database (e.g., PFOCR has 1954 breast cancer pathways), the average number of pathways among their top 20 diseases (e.g., PFOCR has 649 pathways on average), and the minimal coverage (e.g., PFOCR has 5 or more pathways for 351 diseases and at least 1 pathway for 555 diseases that are lacking representation from any of the other pathway databases). In terms of actual disease gene content in these pathways, we made use of Jensen DISEASE [15] annotations and focused on 17 diseases including the best representatives from each database and spanning neurological, cancer, heart, lung and metabolic categories. Again, the contrast was striking, with PFOCR covering more disease genes in every case and over half of the genes in all diseases except diabetes mellitus (39%), while the other pathway databases averaged between 9% and 28% coverage. Researchers interested in associating their genes of interest (e.g., from differential gene expression analyses) with mechanisms of disease will find broader coverage and more diverse instances in the PFOCR database.

Like other pathway databases, PFOCR content is available in formats amenable to gene set enrichment analyses, such as GMT (see Data Availability), allowing bioinformaticians to import PFOCR into practically any pathway analysis workflow. PFOCR pathways have also been integrated into commonly used online tools for enrichment analyses, including Enrichr [20] and iQuery by NDEx [21], and we also have introduced an R Shiny app called Interactive Enrichment Analysis, which uses PFOCR, WikiPathways and GO by default [24].

AML subtype clustering analysis showed how PFOCR pathway information can be used to gain insights to understand differential regulatory mechanisms of disease subtypes at the pathway level. One example of a more general approach of enrichment analysis using canonical gene sets was demonstrated in the study performed by Asi et al. [93]. In this study, enrichment analysis was performed using KEGG and GO on the

combined data of DNase I-seq and RNA-seq data from FLT3-ITD and t(8;21) AML patient samples. Integration was performed by finding DNase I hypersensitivity site (DHS) peaks specific to two mutation types and combining the result with the DEG analysis result. Subsequently, GO and KEGG enrichment analysis performed for each subtype and enriched gene sets for each AML subtype were investigated. This study provided a list of gene sets and compared the presence/absence of gene set signals in each subtype. While this approach is the most common way gene set enrichment is used in research, utilizing PFOCR with larger extensions of more diverse and detailed gene sets can help to investigate pathway level signals in depth, while clustering approaches can make gene set comparisons more systematic and interpretable by categorizing pathway-level signals specific to each subtype.

Each published pathway figure offers not only unique content, but also unique context based on the specific experiments and insights described in its parent article. As a collection, PFOCR offers a distinct advantage to researchers in providing diverse examples of a given biological process linked to specific experimental designs and analytical methods. For example, our case study on breast cancer grade prediction showed that PFOCR outperformed other pathway databases in terms of both rank and total number of results. In our breast cancer analysis results, the top 62 pathways that passed the accuracy threshold were all PFOCR pathways, showing the relative superiority of PFOCR compared to Reactome, WikiPathways, and KEGG. The fact that the majority of genes—including those with the highest importance scores—are unique to PFOCR suggests that the unique pathway information in PFOCR can support research into disease mechanisms in a way that other pathway databases cannot.

Conclusions

PFOCR is a novel kind of pathway database approaching the breadth and depth of GO while providing rich, mechanistic diagrams and direct literature support. PFOCR content is extracted from the pathway figures embedded in published articles dating as far back as 1995 and at a current pace of one thousand new pathways a month. Compared to other popular pathway databases, PFOCR contains an unprecedented number and diversity of pathways, genes, and chemicals. Another 23,000 pathways figures will be added to the database soon, covering 2021-2023, with plans to regularly update as new figures are published. With a focus on disease-associated content, PFOCR leads not only in terms of the number of diseases covered, but also the number of pathways per disease, and the number of unique genes per disease. In the context of pathway analysis, PFOCR offers more avenues to connect a researcher's gene set of interest to biological processes related to disease. Available in GMT format and pre-integrated into user-friendly online tools, PFOCR is easy to include in a researcher's analysis plan. More advanced pathway analyses, for example investigating cancer subtypes and grade prediction, can leverage the unique depth of pathway content in PFOCR that lends support to explorations into possible mechanistic models and machine learning applications where typical pathway databases are not particularly useful.

Methods

The PFOCR gmt (pfocr-20210515-gmt-Homo_sapiens.gmt) includes figures published between 1995 and 2021 and was downloaded from the PFOCR data archive [94]. The WikiPathways gmt file (wikipathways-20211210-gmt-Homo_sapiens.gmt) includes curated pathway models up until 2021 and was downloaded from the WikiPathways data archive [95]. KEGG pathways were retrieved from json files downloaded from TogoWS [96]. TogoWS only supports KEGG data uploaded prior to their proprietary licensing in 2012. Reactome pathways were retrieved as JSON files using the Reactome REST API on 10 January 2022, and were parsed using R jsonlite library [97].

PFOCR, Reactome, WikiPathways, and KEGG pathways are filtered according to the criteria of a minimum number of genes of 3 and a maximum number of genes of 500 in each pathway. As a result of this filter, the number of pathways from their sources decreased from 2029 to 1663 in Reactome, from 703 to 686 in WikiPathways, and from 345 to 345 in KEGG. The PFOCR gmt is provided with these constraints already applied.

Disease coverage comparison

Disease information was downloaded from the Comparative Toxicogenomics Database (CTD) [14]. Diseases were screened according to the following criteria in order to compile a distinct set of names amenable to making unambiguous occurrence counts in text annotations in pathway databases: 1. A disease name is not an extension of another disease name from the same disease, 2. A disease name is not related to a psychological condition, 3. A disease is not a category for multiple diseases otherwise included (e.g., neurodegenerative disease), 4. A disease is not related to an environmental condition (e.g., mite infestations), 5. A disease is not an alias for another disease, 6. A disease is not a symptom (e.g., abdominal pain), 7. A disease name is not ambiguous relative to included disease names (e.g., cancer). The final number of filtered disease names was 876. Text titles and descriptions (or captions) were collected for each of the human pathways from PFOCR, Reactome, WikiPathways, and KEGG. Case-insensitive string matching functions were used to identify disease name occurrences in the collected text samples. A match was only counted once per pathway even if the disease name occurred multiple times within or across text samples for that pathway. The resulting pathway counts per disease and per database are shown in **Supplement Table 1** and a subset in Table 1. Reactome and WikiPathways provide additional sources for disease annotation, including ontology tags, gene descriptions, and bibliography titles that we did not include in this accounting in order to make a fair comparison across all four resources.

To investigate disease gene coverage of the pathway databases, the human disease gene file 'human_disease_knowledge_filtered.tsv' was downloaded from Jensen DISEASES. Jensen disease names that exactly matched the CTD disease names were selected for investigation. The number of genes present in integrated pathways for each disease was determined for each pathway database and also expressed as a percentage of the number of genes defined by Jensen DISEASES for each disease.

Case Study 1: Acute Myeloid Leukemia Subtype Analysis

AML patient gene expression data was downloaded from GEO under accession number GSE108316. Samples with mutations in RUNX1, inv(16), t(8;21), CEBPA, SRSF2, FLT3-ITD, FLT3-ITD/NMP1, and inv(3)/RAS were selected for this study. Differentially expressed gene (DEG) analysis was performed using genes with 2 or more read counts in at least five samples using the limma pipeline [98]. Gene counts were transformed to log2-counts per million (logCPM), and the mean-variance relationship and weights were estimated using voom in the R library limma [98]. Then, limma's lmFit was used to fit linear models and limma's eBayes to compute statistics of the fitted models. Control samples were used as reference samples for each hypothesis comparing leukemia patients and normal gene expression levels.

Based on the DEG results, GSEA was performed for each mutation type. First, leukemia related pathways were retrieved from PFOCR, WikiPathways, KEGG, and Reactome. GSEA was performed using GSEA in the R library ClusterProfiler [17]. Normalized enrichment scores (NES) were retrieved from GSEA results, and scores for each mutation type were subjected to hierarchical clustering using the R function *heatmap.2*. At the top level of the sample hierarchy, there are two main clusters. The core leukemia gene set (CLGS) was defined by the hierarchical clustering and included 26 pathways with the greatest average NES differences.

Case Study 2: Breast Cancer Prediction Analysis

GSE3494 and GSE2990 breast cancer patient gene expression data measured by Affymetrix Human Genome U133A Array were downloaded for breast cancer analysis. Robust Multi-array Average (RMA) [99] was calculated using rma in the R library affy. Principal component analysis (PCA) was performed to confirm that there was no batch effect in data sets.

To build random forest models for predicting breast cancer grade in patient data, first, breast cancer pathways were retrieved from PFOCR, Reactome, WikiPathways, and KEGG. Breast cancer pathways compiled from the disease coverage comparison analysis resulted in 1954 PFOCR, 20 Reactomes, 6 WikiPathways and 3 KEGG pathways. The gene expression values of genes in each pathway were used as feature values in random forest models. For best results, hyperparameters (number of features, minimum node size, and fraction of

observations to sample) were tuned using the tune function in the R library randomForestSRC [100] to ensure that the prediction error is the minimum. The number of trees was set to 1000 times the number of features. Then, ranger [101] was used to make random forest models and measure feature importance given the optimal hyperparameters. The overall training performance of each model and the feature importance was measured as the average of the one-out cross-validation results. The pathways with cross validation accuracy and prediction accuracy higher than 0.65 were selected as top pathways and pathways with cross validation accuracy and prediction accuracy higher than 0.55 were selected and used to find top PFOCR pathway unique genes.

Declarations

Availability of data and materials

- Filtered GMT with pathway figure ID aliases convenient for enrichment analyses: https://doi.org/10.5281/zenodo.8118381

Funding

This work was supported by NIH/NIGMS P41GM103504 and R01GM100039.

Authors' contributions

All the analyses were performed by M-GS. Figures and tables were prepared by M-GS and ARP. The manuscript was written by M-GS and ARP.

Acknowledgements

We would like to acknowledge Reuben Thomas, Martina Summer-Kutmon, Kristina Hanspers, Anders Riutta, and Laurent Winckers for early discussions and brainstorming leading to the conception of this manuscript.

References

1. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.

2. 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;31:19–20.

3. Karp PD, Paley S, Romero P. The Pathway Tools software. Bioinformatics. 2002;18 Suppl 1:S225–32.

4. Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, de Bono B, et al. Reactome: a knowledge base of biologic pathways and processes. Genome Biol. 2007;8:R39.

5. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: pathway editing for the people. PLoS Biol. 2008;6:e184.

6. Bauer-Mehren A, Furlong LI, Sanz F. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. Mol Syst Biol. 2009;5:290.

7. Mubeen S, Hoyt CT, Gemünd A, Hofmann-Apitius M, Fröhlich H, Domingo-Fernández D. The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. Front Genet. 2019;10:1203.

8. García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway Analysis: State of the Art. Front Physiol. 2015;6:383.

9. Chen EY. Enrichr [Internet]. [cited 2023 Feb 15]. Available from: https://maayanlab.cloud/Enrichr/

10. Waagmeester AS, Kelder T, Evelo CTA. The role of bioinformatics in pathway curation. Genes Nutr. 2008;3:139–42.

11. Hanspers K, Kutmon M, Coort SL, Digles D, Dupuis LJ, Ehrhart F, et al. Ten simple rules for creating reusable pathway models for computational analysis and visualization. PLoS Comput Biol. 2021;17:e1009226.

12. Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, et al. WikiPathways: connecting communities. Nucleic Acids Res. 2021;49:D613–21.

13. Hanspers K, Riutta A, Summer-Kutmon M, Pico AR. Pathway information extracted from 25 years of pathway figures. Genome Biol. 2020;21:273.

14. Davis AP, Wiegers TC, Johnson RJ, Sciaky D, Wiegers J, Mattingly CJ. Comparative Toxicogenomics Database (CTD): update 2023. Nucleic Acids Res. 2023;51:D1257–62.

15. Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: text mining and data integration of disease-gene associations. Methods. 2015;74:83–9.

16. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis [Internet]. bioRxiv. 2021 [cited 2023 Jun 14]. p. 060012. Available from: https://www.biorxiv.org/content/10.1101/060012v3

17. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb). 2021;2:100141.

18. Morgan M, Falcon S, Gentleman R. GSEABase: Gene set enrichment data structures and methods. R package version.

19. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.

20. Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, et al. Gene Set Knowledge Discovery with Enrichr. Curr Protoc. 2021;1:e90.

21. Pillich RT, Chen J, Churas C, Fong D, Ideker T, Liu SN, et al. NDEx IQuery: a multi-method network gene set analysis leveraging the Network Data Exchange [Internet]. bioRxiv. 2022 [cited 2023 Feb 14]. p. 2022.10.24.513552. Available from: https://www.biorxiv.org/content/biorxiv/early/2022/10/25/2022.10.24.513552

22. Clarke DJB, Jeon M, Stein DJ, Moiseyev N, Kropiwnicki E, Dai C, et al. Appyters: Turning Jupyter Notebooks into data-driven web apps. Patterns (N Y). 2021;2:100213.

23. Gyori BM, Bachman JA, Subramanian K, Muhlich JL, Galescu L, Sorger PK. From word models to executable models of signaling networks using automated assembly. Mol Syst Biol. 2017;13:954.

24. Interactive-Enrichment-Analysis: A set of Shiny apps to provide interactive enrichment analysis and

exploration of results [Internet]. Github; [cited 2023 Feb 14]. Available from: https://github.com/gladstone-institutes/Interactive-Enrichment-Analysis

25. Zwaan CM, Kolb EA, Reinhardt D, Abrahamsson J, Adachi S, Aplenc R, et al. Collaborative Efforts Driving Progress in Pediatric Acute Myeloid Leukemia. J Clin Oncol. 2015;33:2949–62.

26. Rubnitz JE. Current Management of Childhood Acute Myeloid Leukemia. Paediatr Drugs. 2017;19:1–10.

27. Shallis RM, Wang R, Davidoff A, Ma X, Zeidan AM. Epidemiology of acute myeloid leukemia: Recent progress and enduring challenges. Blood Rev. 2019;36:70–87.

28. Sengupta TK, Leclerc GM, Hsieh-Kinser TT, Leclerc GJ, Singh I, Barredo JC. Cytotoxic effect of 5-aminoimidazole-4-carboxamide-1-beta-4-ribofuranoside (AICAR) on childhood acute lymphoblastic leukemia (ALL) cells: implication for targeted therapy. Mol Cancer. 2007;6:46.

29. Yu J, Li Y, Zhang D, Wan D, Jiang Z. Clinical implications of recurrent gene mutations in acute myeloid leukemia. Exp Hematol Oncol. 2020;9:4.

30. Benard B, Gentles AJ, Köhnke T, Majeti R, Thomas D. Data mining for mutation-specific targets in acute myeloid leukemia. Leukemia. 2019;33:826–43.

31. Yu J, Jiang PYZ, Sun H, Zhang X, Jiang Z, Li Y, et al. Advances in targeted therapy for acute myeloid leukemia. Biomark Res. 2020;8:17.

32. Sinkala M, Mulder N, Martin D. Machine Learning and Network Analyses Reveal Disease Subtypes of Pancreatic Cancer and their Molecular Characteristics. Sci Rep. 2020;10:1212.

33. Pathway Figure OCR. Action for AICAR in human leukemia ALL cells [Internet]. 2021 [cited 2023 Mar 20]. Available from: https://pfocr.wikipathways.org/figures/PMC1948012__1476-4598-6-46-10.html

34. Padmakumar D, Chandraprabha VR, Gopinath P, Vimala Devi ART, Anitha GRJ, Sreelatha MM, et al. A concise review on the molecular genetics of acute myeloid leukemia. Leuk Res. 2021;111:106727.

35. Doucette K, Karp J, Lai C. Advances in therapeutic options for newly diagnosed, high-risk AML patients. Ther Adv Hematol. 2021;12:20406207211001138.

36. Golson ML, Kaestner KH. Fox transcription factors: from development to disease. Development. 2016;143:4558–70.

37. He L, Gomes AP, Wang X, Yoon SO, Lee G, Nagiec MJ, et al. mTORC1 Promotes Metabolic Reprogramming by the Suppression of GSK3-Dependent Foxk1 Phosphorylation. Mol Cell. 2018;70:949–60.e4.

38. Alvarez-Fernández M, Medema RH. Novel functions of FoxM1: from molecular mechanisms to cancer therapy. Front Oncol. 2013;3:30.

39. Lam EW-F, Brosens JJ, Gomes AR, Koo C-Y. Forkhead box proteins: tuning forks for transcriptional harmony. Nat Rev Cancer. 2013;13:482–95.

40. Lin S, Ptasinska A, Chen X, Shrestha M, Assi SA, Chin PS, et al. A FOXO1-induced oncogenic network defines the AML1-ETO preleukemic program. Blood. 2017;130:1213–22.

41. Han C-Y, Cho K-B, Choi H-S, Han H-K, Kang K-W. Role of FoxO1 activation in MDR1 expression in adriamycin-resistant breast cancer cells. Carcinogenesis. 2008;29:1837–44.

42. Scheijen B, Ngo HT, Kang H, Griffin JD. FLT3 receptors with internal tandem duplications promote cell viability and proliferation by signaling through Foxo proteins. Oncogene. 2004;23:3338–49.

43. Jawhar M, Schwaab J, Naumann N, Horny H-P, Sotlar K, Haferlach T, et al. Response and progression on midostaurin in advanced systemic mastocytosis: KIT D816V and other molecular markers. Blood. 2017;130:137–45.

44. Short NJ, Konopleva M, Kadia TM, Borthakur G, Ravandi F, DiNardo CD, et al. Advances in the Treatment of Acute Myeloid Leukemia: New Drugs and New Challenges. Cancer Discov. 2020;10:506–25.

45. Chen J-Q, Russo J. ERalpha-negative and triple negative breast cancer: molecular features and potential therapeutic approaches. Biochim Biophys Acta. 2009;1796:162–75.

46. WikiPathways. Pathway Figure OCR [Internet]. WikiPathways; [cited 2023 Jun 13]. Available from: https://pfocr.wikipathways.org/figures/PMC2937358__nihms192921f1.html

47. Macias RIR, Marin JJG, Serrano MA. Excretion of biliary compounds during intrauterine life. World J Gastroenterol. 2009;15:817–28.

48. Foetal liver-placenta-maternal liver excretory pathway [Internet]. Pathway Figure OCR. WikiPathways; 2009 [cited 2023 Jun 13]. Available from: https://pfocr.wikipathways.org/figures/PMC2653381__WJG-15-817-g001.html

49. Kopp S, Sahana J, Islam T, Petersen AG, Bauer J, Corydon TJ, et al. The role of NFκB in spheroid formation of human breast cancer cells cultured on the Random Positioning Machine. Sci Rep. 2018;8:921.

50. Pathway studio analysis of genes analyzed in the study by the gene array analysis and qPCR [Internet]. Pathway Figure OCR. WikiPathways; 2018 [cited 2023 Jun 13]. Available from: https://pfocr.wikipathways.org/figures/PMC5772637__41598_2017_18556_Fig6_HTML.html

51. Pathway studio analysis of proteins whose genes were analysed in the study by the gene array analysis and qPCR [Internet]. Pathway Figure OCR. WikiPathways; 2018 [cited 2023 Jun 13]. Available from: https://pfocr.wikipathways.org/figures/PMC5772637__41598_2017_18556_Fig7_HTML.html

52. Proposed model of PAK5-AIF signaling pathway in breast cancer proliferation [Internet]. Pathway Figure OCR. WikiPathways; 2021 [cited 2023 Jun 13]. Available from: https://pfocr.wikipathways.org/figures/PMC8040471__ijbsv17p1315g007.html

53. Xing Y, Li Y, Hu B, Han F, Zhao X, Zhang H, et al. PAK5-mediated AIF phosphorylation inhibits its nuclear translocation and promotes breast cancer tumorigenesis. Int J Biol Sci. 2021;17:1315–27.

54. Enrichment in multiple cancer-associated networks in breast cancer [Internet]. Pathway Figure OCR. WikiPathways; 2019 [cited 2023 Jun 13]. Available from: https://pfocr.wikipathways.org/figures/PMC6759650__fonc-09-00910-g0005.html

55. Vishnubalaji R, Sasidharan Nair V, Ouararhni K, Elkord E, Alajez NM. Integrated Transcriptome and Pathway Analyses Revealed Multiple Activated Pathways in Breast Cancer. Front Oncol. 2019;9:910.

56. Possible action mechanism of capsaicin on proliferation and apoptosis in breast cancer [Internet]. Pathway Figure OCR. WikiPathways; 2021 [cited 2023 Jun 13]. Available from: https://pfocr.wikipathways.org/figures/PMC7811378__DDDT-15-125-g0009.html

57. Chen M, Xiao C, Jiang W, Yang W, Qin Q, Tan Q, et al. Capsaicin Inhibits Proliferation and Induces Apoptosis in Breast Cancer by Down-Regulating FBI-1-Mediated NF-κB Pathway. Drug Des Devel Ther. 2021;15:125–40.

58. Canonical pathway of molecular mechanisms of cancer from IPA showing dysregulated mRNA targets along with their potential differentially expressed miRNA regulators found in Lebanese samples [Internet]. Pathway Figure OCR. WikiPathways; 2017 [cited 2023 Jun 14]. Available from: https://pfocr.wikipathways.org/figures/PMC5715135__41598_2017_16978_Fig7_HTML.html

59. Nassar FJ, Talhouk R, Zgheib NK, Tfayli A, El Sabban M, El Saghir NS, et al. microRNA Expression in Ethnic Specific Early Stage Breast Cancer: an Integration and Comparative Analysis. Sci Rep. 2017;7:16829.

60. Xu X, Gammon MD, Wetmur JG, Bradshaw PT, Teitelbaum SL, Neugut AI, et al. B-vitamin intake, one-carbon metabolism, and survival in a population-based study of women with breast cancer. Cancer Epidemiol Biomarkers Prev. 2008;17:2109–16.

61. Schematic illustration of one-carbon metabolism pathway [Internet]. Pathway Figure OCR. WikiPathways; 2009 [cited 2023 Jul 5]. Available from: https://pfocr.wikipathways.org/figures/PMC2673236__nihms-107145-f0001.html

62. NRIF3- or DD1-induced apoptosis [Internet]. Pathway Figure OCR. WikiPathways; 2004 [cited 2023 Jul 5]. Available from: https://pfocr.wikipathways.org/figures/PMC387764__zmb0090417790008.html

63. Li D, Das S, Yamada T, Samuels HH. The NRIF3 family of transcriptional coregulators induces rapid and profound apoptosis in breast cancer cells. Mol Cell Biol. 2004;24:3838–48.

64. The gamma secretase (Gamma-secretase) complex: It is comprised of presenilin enhancer (PEN-2), anterior pharynx-defective 1 (APH1) and nicastrin [Internet]. Pathway Figure OCR. WikiPathways; 2019 [cited 2023 Jul 6]. Available from: https://pfocr.wikipathways.org/figures/PMC6947643__genes-10-00961-g002.html

65. Kar R, Jha NK, Jha SK, Sharma A, Dholpuria S, Asthana N, et al. A “NOTCH” Deeper into the Epithelial-To-Mesenchymal Transition (EMT) Program in Breast Cancer. Genes [Internet]. 2019;10. Available from: http://dx.doi.org/10.3390/genes10120961

66. Pathological roles of ion channels and transporters in triple-negative breast cancer cells [Internet]. Pathway Figure OCR. WikiPathways; 2020 [cited 2023 Jul 6]. Available from: https://pfocr.wikipathways.org/figures/PMC7409684__12935_2020_1464_Fig1_HTML.html

67. Lu C, Ma Z, Cheng X, Wu H, Tuo B, Liu X, et al. Pathological role of ion channels and transporters in the development and progression of triple-negative breast cancer. Cancer Cell Int. 2020;20:377.

68. Towards personalized treatment for early stage HER2-positive breast cancer [Internet]. Pathway Figure OCR. WikiPathways; 2019 [cited 2023 Jul 6]. Available from: https://pfocr.wikipathways.org/figures/PMC8023395__nihms-1688996-f0002.html

69. Goutsouliak K, Veeraraghavan J, Sethunath V, De Angelis C, Osborne CK, Rimawi MF, et al. Towards personalized treatment for early stage HER2-positive breast cancer. Nat Rev Clin Oncol. 2020;17:233–50.

70. How p-CREB acts as a common downstream effector for multiple signaling pathways that regulate BCRP mRNA expression [Internet]. Pathway Figure OCR. WikiPathways; 2015 [cited 2023 Jul 6]. Available from: https://pfocr.wikipathways.org/figures/PMC4336604__nihms657356f9.html

71. Xie Y, Nakanishi T, Natarajan K, Safren L, Hamburger AW, Hussain A, et al. Functional cyclic AMP response element in the breast cancer resistance protein (BCRP/ABCG2) promoter modulates epidermal growth factor receptor pathway- or androgen withdrawal-mediated BCRP/ABCG2 transcription in human cancer cells. Biochim Biophys Acta. 2015;1849:317–27.

72. A Developing Therapeutic Decision Tree for Ductal Breast Cancer with Emphasis on the TNBC Subtype [Internet]. Pathway Figure OCR. WikiPathways; 2019 [cited 2023 Jul 6]. Available from: https://pfocr.wikipathways.org/figures/PMC6499473__nihms-1025532-f0001.html

73. Heiser LM, Mills GB, Gray JW. Therapeutic Clues from an Integrated Omic Assessment of East Asian Triple Negative Breast Cancers. Cancer Cell. 2019. p. 341–3.

74. Hypothetic pathways by which Pokemon regulates survivin expression [Internet]. Pathway Figure OCR. WikiPathways; 2011 [cited 2023 Jul 6]. Available from: https://pfocr.wikipathways.org/figures/PMC3219187__bcr2843-5.html

75. Zu X, Ma J, Liu H, Liu F, Tan C, Yu L, et al. Pro-oncogene Pokemon promotes breast cancer progression by upregulating survivin expression. Breast Cancer Res. 2011;13:R26.

76. P14ARF-p53 pathway [Internet]. Pathway Figure OCR. WikiPathways; 2018 [cited 2023 Jul 6]. Available from: https://pfocr.wikipathways.org/figures/PMC6024909__cancers-10-00189-g004.html

77. Moulder DE, Hatoum D, Tay E, Lin Y, McGowan EM. The Roles of p53 in Mitochondrial Dynamics and Cancer Metabolism: The Pendulum between Survival and Death in Breast Cancer? Cancers [Internet]. 2018;10. Available from: http://dx.doi.org/10.3390/cancers10060189

78. Schematic illustration of overview of one-carbon metabolism pathway, linking to methylation reactions and nucleotide synthesis [Internet]. Pathway Figure OCR. WikiPathways; 2009 [cited 2023 Jul 6]. Available from: https://pfocr.wikipathways.org/figures/PMC2694962__nihms107224f1.html

79. Xu X, Chen J. One-carbon metabolism and breast cancer: an epidemiological perspective. J Genet Genomics. 2009;36:203–14.

80. Schematic illustrating the G protein-independent role of RGS6 in doxorubicin-induced apoptosis and antiproliferative signaling in the breast [Internet]. Pathway Figure OCR. WikiPathways; 2016 [cited 2023 Jul 6]. Available from: https://pfocr.wikipathways.org/figures/PMC5256616__12248_2016_9899_Fig6_HTML.html

81. Ahlers KE, Chakravarti B, Fisher RA. RGS6 as a Novel Therapeutic Target in CNS Diseases and Cancer. AAPS J. 2016;18:560–72.

82. Molecular pathways regulating breast cancer stem cells (CSCs) [Internet]. Pathway Figure OCR. WikiPathways; 2015 [cited 2023 Jul 6]. Available from: https://pfocr.wikipathways.org/figures/PMC4407294__13058_2015_560_Fig2_HTML.html

83. Toss A, Cristofanilli M. Molecular characterization and targeted therapeutic approaches in breast cancer. Breast Cancer Res. 2015;17:60.

84. ADAR1-mediated RNA editing in cancer development [Internet]. Pathway Figure OCR. WikiPathways; 2018 [cited 2023 Jul 6]. Available from: https://pfocr.wikipathways.org/figures/PMC6305585__fendo-09-00762-g0002.html

85. Kung C-P, Maggi LB Jr, Weber JD. The Role of RNA Editing in Cancer Development and Metabolic Disorders. Front Endocrinol . 2018;9:762.

86. Orlic-Milacic M. Transcriptional regulation by the AP-2 (TFAP2) family of transcription factors [Internet]. Reactome. 2016 [cited 2023 Jul 6]. Available from: https://reactome.org/content/detail/R-HSA-8864260

87. Endocrine resistance - Homo sapiens (hsa01522) [Internet]. KEGG. 2011 [cited 2023 Jul 6]. Available from: https://www.genome.jp/dbget-bin/www_bget?pathway:hsa01522

88. Sharifi M, Moridnia A. Apoptosis-inducing and antiproliferative effect by inhibition of miR-182-5p through the regulation of CASP9 expression in human breast cancer. Cancer Gene Ther. 2017;24:75–82.

89. Theodoropoulos GE, Michalopoulos NV, Pantou MP, Kontogianni P, Gazouli M, Karantanos T, et al. Caspase 9 promoter polymorphisms confer increased susceptibility to breast cancer. Cancer Genet. 2012;205:508–12.

90. Chen J, Imanaka N, Chen J, Griffin JD. Hypoxia potentiates Notch signaling in breast cancer leading to decreased E-cadherin expression and increased cell migration and invasion. Br J Cancer. 2010;102:351–60.

91. D'Angelo RC, Ouzounova M, Davis A, Choi D, Tchuenkam SM, Kim G, et al. Notch reporter activity in breast cancer cell lines identifies a subset of cells with stem cell activity. Mol Cancer Ther. 2015;14:779–87.

92. Fu Y-P, Edvardsen H, Kaushiva A, Arhancet JP, Howe TM, Kohaar I, et al. NOTCH2 in breast cancer: association of SNP rs11249433 with gene expression in ER-positive breast tumors without TP53 mutations. Mol Cancer. 2010;9:113.

93. Assi SA, Imperato MR, Coleman DJL, Pickin A, Potluri S, Ptasinska A, et al. Subtype-specific regulatory network rewiring in acute myeloid leukemia. Nat Genet. 2019;51:151–62.

94. PFOCR Data Archive [Internet]. [cited 2023 Feb 15]. Available from: https://wikipathways-data.wmcloud.org/pfocr/

95. WikiPathways Data [Internet]. [cited 2023 Feb 15]. Available from: https://data.wikipathways.org/20211210/gmt/

96. Katayama T, Nakao M, Takagi T. TogoWS: integrated SOAP and REST APIs for interoperable bioinformatics Web services. Nucleic Acids Res. 2010;38:W706–11.

97. Ooms J. The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects [Internet]. arXiv [stat.CO]. 2014. Available from: http://arxiv.org/abs/1403.2805

98. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression

analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.

99. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4:249–64.

100. Hemant Ishwaran UBK. RandomForestSRC: Fast unified random forests for survival, regression, and classification (RF-SRC) [Internet]. Comprehensive R Archive Network (CRAN). [cited 2023 Feb 10]. Available from: https://cran.r-project.org/package=randomForestSRC

101. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw. 2017;77:1–17.