Received: 1 March 2023 | Revised: 19 June 2023 | Accepted: 4 July 2023

DOI: 10.1002/ijc.34666

RESEARCH ARTICLE

Innovative Tools and Methods

IJC INTERNATIONAL JOURNAL of CANCER uicc

Comprehensive interrogation of gene lists from genome-scale cancer screens with oncoEnrichR

Sigve Nakken[1,2,3] | Sveinung Gundersen[3] | Fabian L. M. Bernal[4] | Dimitris Polychronopoulos[5] | Eivind Hovig[1,3] | Jørgen Wesche[1,2]

[1]Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway

[2]Centre for Cancer Cell Reprogramming, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway

[3]Centre for Bioinformatics, Department of Informatics, University of Oslo, Oslo, Norway

[4]University Center for Information Technology, University of Oslo, Oslo, Norway

[5]Ochre Bio Ltd, Oxford, UK

Correspondence

Sigve Nakken, Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway.
Email: sigven@ifi.uio.no

Funding information

Kreftforeningen, Grant/Award Number: 198093; Norges Forskningsråd, Grant/Award Number: 262652

Abstract

Genome-scale screening experiments in cancer produce long lists of candidate genes that require extensive interpretation for biological insight and prioritization for follow-up studies. Interrogation of gene lists frequently represents a significant and time-consuming undertaking, in which experimental biologists typically combine results from a variety of bioinformatics resources in an attempt to portray and understand cancer relevance. As a means to simplify and strengthen the support for this endeavor, we have developed oncoEnrichR, a flexible bioinformatics tool that allows cancer researchers to comprehensively interrogate a given gene list along multiple facets of cancer relevance. oncoEnrichR differs from general gene set analysis frameworks through the integration of an extensive set of prior knowledge specifically relevant for cancer, including ranked gene-tumor type associations, literature-supported proto-oncogene and tumor suppressor gene annotations, target druggability data, regulatory interactions, synthetic lethality predictions, as well as prognostic associations, gene aberrations and co-expression patterns across tumor types. The software produces a structured and user-friendly analysis report as its main output, where versions of all underlying data resources are explicitly logged, the latter being a critical component for reproducible science. We demonstrate the usefulness of oncoEnrichR through interrogation of two candidate lists from proteomic and CRISPR screens. oncoEnrichR is freely available as a web-based service hosted by the Galaxy platform (https://oncotools.elixir.no), and can also be accessed as a stand-alone R package (https://github.com/sigven/oncoEnrichR).

KEYWORDS

cancer relevance, gene set analysis, genome-scale screening, hit prioritization, target discovery

What's new?

Genome-scale screening experiments produce long lists of candidate genes that require extensive interpretation for biological insight and prioritization for follow-up studies. Interrogating these gene lists is a challenging undertaking. Building upon existing data integration frameworks

Abbreviations: CRISPR, clustered regularly interspaced short palindromic repeats; EFO, Experimental Factor Ontology; GO, Gene Ontology; HPA, Human Protein Atlas; OTP, Open Targets Platform; PPI, protein-protein interaction; TCGA, The Cancer Genome Atlas; TF, transcription factor; UMLS, Unified Medical Language System.

and multiple large-scale omics datasets, the authors developed a feature-rich gene set interpretation tool to systematically interpret and prioritize long lists of candidate genes. *oncoEnrichR* is a user-friendly reporting framework that portrays the cancer relevance of candidate hits more comprehensively than existing solutions, allowing researchers to efficiently gather evidence when picking candidates for in-depth follow-up experiments.

1 | INTRODUCTION

The search for novel cancer-implicated genes is currently fueled by data-driven high-throughput genome-scale screening studies. Gene perturbation experiments, possible with CRISPR/Cas9 technology and siRNA knockdown methodology, can detect survival-essential genes and genes related to drug resistance.[1] Proteomic screens, through affinity pull-down or proximity labeling approaches, can be used to reveal partners and functions of proto-oncogenes or tumor suppressors in a given cancer phenotype.[2] Moreover, genome-wide gene expression profiling is frequently employed to identify dysregulated genes in cancer cells.[3] Notably, all of the above-mentioned high-throughput screening technologies typically produce the same conceptual output, in the form of phenotype-associated gene hits ranked by effect size. Depending on the effect size thresholds adopted for a given experiment, the size of such hit lists can frequently be in the order of several hundred.[1] For a given set of hits, researchers need robust annotation tools that can systematically portray cancer relevance and guide the prioritization of genes for in-depth experimental validation. A considerable challenge in this respect is to mobilize sufficient prior knowledge from available and dispersed cancer data resources, not only when it comes to the functional nature of individual hits, but also system-wide properties of the total set of hits at the level of biological processes, pathways and networks.

Gene list interpretation is frequently dominated by generic geneset or pathway enrichment resources, such as MetaScape,[4] DAVID[5] or Enrichr,[6] all offering systems-level, functional views for a given set of genes. Web-based services for interrogating the function and potential disease relevance of individual genes are also available, as exemplified by GeneCards.[7] A functional network view for a given gene set can further be provided by tools such as STRING[8] or GeneMania,[9] considering, for example, the presence of physical protein-protein interactions. Nevertheless, users often end up with a fragmented approach, where they need to consolidate results from a diverse set of annotation tools.[4] More importantly, existing tools for gene list interpretation are generally limited in their portrayal and prioritization of cancer relevance. Although cross-examination of genes in relation to large-scale cancer omics data can be performed, using portals such as cBioPortal[10] or GEPIA2,[11] there is a scarcity of dedicated gene set interrogation tools that comprehensively integrate prior molecular knowledge on cancer, the latter being a necessity when prioritizing large lists of hits from screens conducted in a cancer setting.

We have developed *oncoEnrichR*, a user-friendly gene set interpretation workflow in which a diverse suite of annotation and prioritization modules highlight the cancer relevance of candidate hits, both quantitatively through global and tumor-type specific rank scores, but also qualitatively through multiple annotations that underscore potential tumorigenic roles and drug-target opportunities. The gene-centric modules of oncoEnrichR are supplemented with system-level features that offer insights from a candidate list at the level of signaling pathways, molecular gene signatures, as well as regulatory interactions and the protein interactome. A structured and interactive gene set report, highly configurable by the user, is readily available as the main output per analysis.

Here, we describe the various analysis modules available in oncoEnrichR, and demonstrate how it can function as a valuable platform for candidate hit interpretation through its application on hits from CRISPR and proteomic screens.

2 | MATERIALS AND METHODS

2.1 | Data collection

The backend of oncoEnrichR is based on the integration of cancer-relevant properties of human genes and their interrelationships from 28 publicly available knowledge resources (a complete overview is provided in Table S2). The main objective is to provide a breadth of large-scale molecular datasets that cover key dimensions in our current comprehension of cancer relevance. Conceptually, the datasets can be broadly categorized as (a) *interactions*, such as protein-protein or transcription factor-target interactions, (b) *collections*, like protein complexes and biological pathways and (c) basic *annotations*, including descriptive types, for instance tumor suppressor gene or cancer hallmark annotations, and quantitative types, such as those derived from large-scale omics datasets, for example, percent tumor samples in The Cancer Genome Atlas (TCGA,[12]) with a homozygous deletion for a given gene. The preparation of all data sources is outlined in detail in Methods S1.

2.2 | Gene-cancer association rank

We utilized data from the powerful Open Targets Platform (OTP, https://targetvalidation.org) to create a quantitative ranking of genes with respect to cancer relevance.[13] OTP provides an aggregated association score (range 0-1) between distinct disease phenotypes and human genes based on a range of evidence sources, for example, genetic associations, text mining or data from animal models. The disease phenotypes coming from OTP are formatted as terms from the Experimental Factor Ontology (EFO), an established vocabulary of human disease phenotypes.[14] To estimate the strength of association

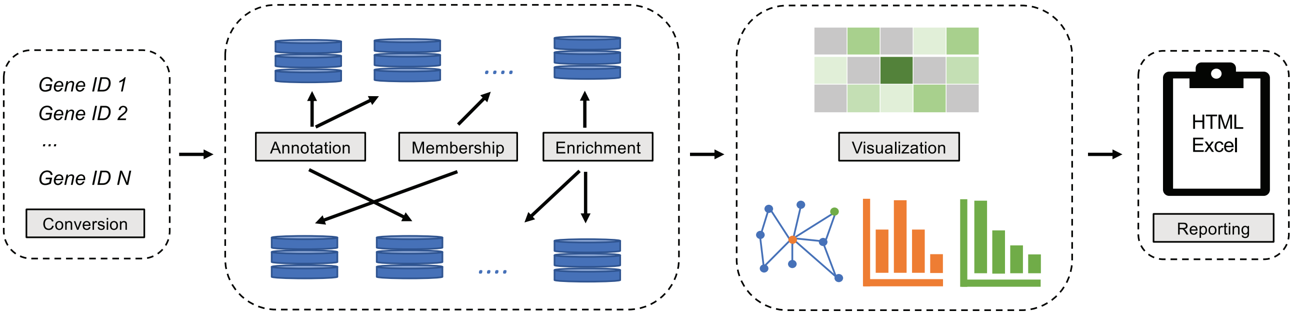


FIGURE 1 Conceptual overview of the oncoEnrichR workflow. The tool consists of four key processing steps: (a) conversion and harmonization of gene/protein query identifiers, (b) annotation, membership and enrichment analyses of the query set against a comprehensive collection of cancer-relevant databases, (c) visualization of analysis results and (d) report generation, either as HTML or Excel.

between human genes and cancers at a broader level (ie, at the level of tissues), we initially created a curated data resource coined *phenOncoX* that maps individual ontology terms from multiple vocabularies toward their respective human tissues (https://github.com/sigven/phenOncoX). This mapping was facilitated by OncoTree,[15] a classification system for precision oncology which has defined a primary set of tumor types/tissues and associated cancer subtypes in the form of ontology terms from the NCI Thesaurus or the Unified Medical Language System (UMLS, see http://oncotree.mskcc.org/#/home).

The cancer-dedicated ontology mapping created by *phenOncoX* enabled a consolidation of disparate phenotype sources found in the underlying knowledge resources of oncoEnrichR. Furthermore, considering the raw EFO-tied association scores from OTP, our cancer phenotype map was essential in the computation of a scaled rank with respect to the association between any given gene and a primary tumor type, and also globally, that is, pan-cancer (see Method S1 and Figure S1).

2.3 | Architecture

The oncoEnrichR gene set interpretation tool has been developed as an R package, and functions conceptually as a workflow of four main steps (Figure 1). The web interface to oncoEnrichR is facilitated by the Galaxy platform (https://oncotools.elixir.no), which ensures a reproducible and collaborative analysis framework.[16]

3 | RESULTS

The main user input to oncoEnrichR is a basic list of human gene identifiers, with an upper limit of 1000 entries. To cater for ease of use, multiple commonly used identifier types are permitted, for example, official gene symbols, UniProt accessions, as well as transcript, gene or protein identifiers (RefSeq/Ensembl). Importantly, oncoEnrichR can also map historical gene aliases toward their up-to-date primary symbol designations, a feature that is often necessary in the processing of user-generated gene lists.

An optional specification of a background gene set for use in gene set enrichment analysis is further available, which ensures that enrichment results can be interpreted correctly with respect to the nature of the underlying screen. The user can also configure multiple parameters of the individual analysis modules, exemplified by significance thresholds in enrichment analysis or confidence thresholds for queried protein-protein interactions. Finally, the user can flexibly choose any combination of modules to be run and included as output in the final report, in that sense allowing the user to control the focus and scope of the gene set analysis.

The output of the oncoEnrichR workflow is two-fold. A standalone, structured HTML report is provided as the main output, enabling user interaction with the resulting visualizations and ranked tables provided for each individual analysis module (outlined below). Moreover, a multisheet Excel workbook with all annotations, interactions and enrichment results is readily available. To cater to transparency, all versions of underlying software and databases are provided in the output files.

3.1 | Analysis modules

A candidate gene set can currently be interrogated with oncoEnrichR through 17 different analysis modules (Table S1). In general, each module considers the candidate set through a particular perspective on gene function or cancer relevance, and makes up an individual section in the output report. A small, educational user guide is populated for each section in the report, giving the user a basic introduction to the type of analysis or annotation performed and underlying databases used. Here, we briefly outline the various types of modules that are offered by oncoEnrichR, with a focus on what they provide regarding functional insights and cancer relevance for the gene set in question.

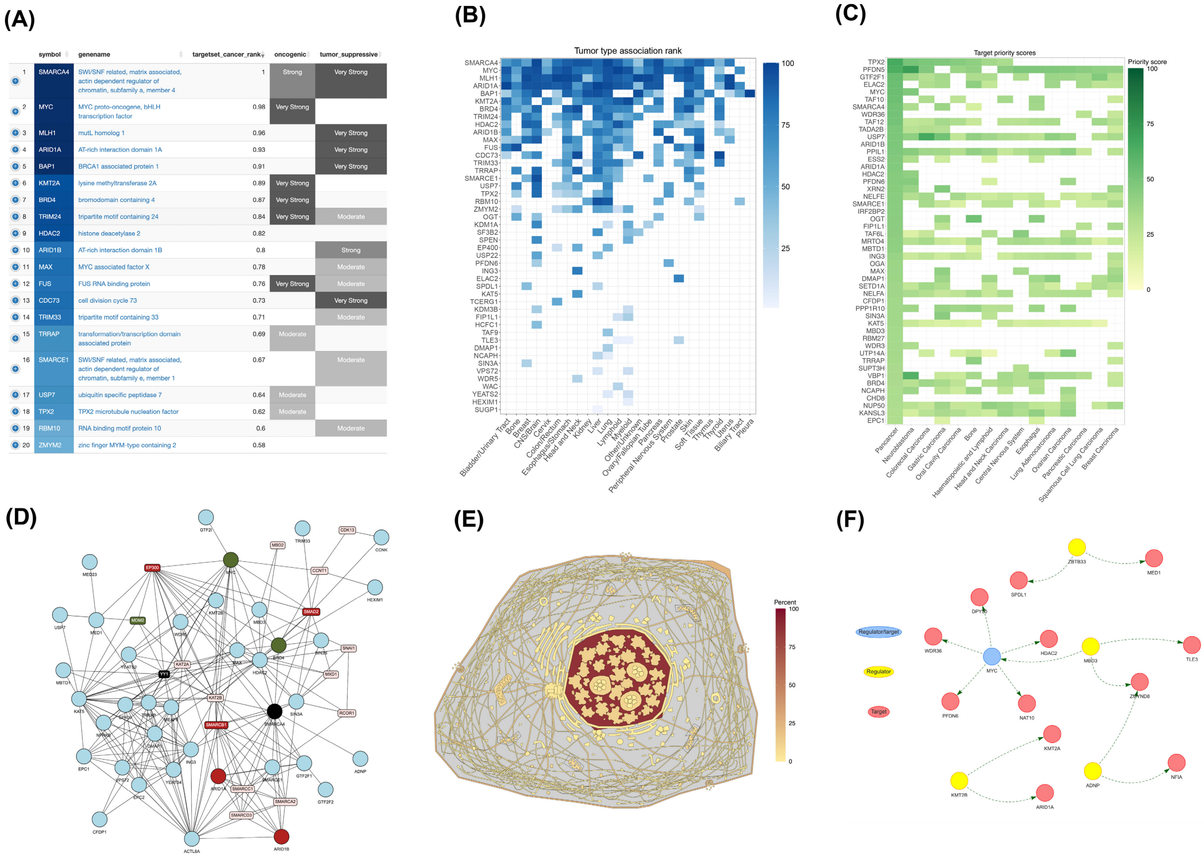
3.1.1 | Gene function, tumor-type associations and drug-target opportunities

A significant number of genes in the human genome are still poorly characterized with respect to function, and members of this set frequently show up in candidate lists from high-throughput screens.

Given that such uncharacterized genes provide a level of novelty for potential follow-up experiments, oncoEnrichR features a dedicated *Poorly characterized genes* module that highlights and ranks such entries, essentially considering genes with a lack of curated Gene Ontology (GO) annotations or gene summary descriptions.

In the *Cancer associations* module, the user can interrogate the candidate hits with respect to current knowledge on genes with tumorigenic roles. Genes are assigned tumor suppressive and oncogenic roles with varying levels of confidence, utilizing support from curated resources and text mining results of the biomedical literature (Methods S1[17-20]). All genes are ranked within the candidate set according to their overall strength of association to cancer, and where the association metric is based upon aggregated evidence from multiple data types, such as text mining, genetic associations or animal models (Figure 2A). A dedicated heatmap is furthermore showing the relative strength of association of candidate genes toward distinct tumor types (Figure 2B). A separate *Cancer hallmarks* module indicates whether the different hallmarks of cancer are promoted or suppressed by members of the candidate set, including links to associated literature. Targeted drugs, both in early and late clinical development, that are specifically indicated for one or more cancer conditions, are listed in a *Drug association* module. This module also allows the user to prioritize the complete set of candidates with respect to their drug targeting potential, based on comprehensive target tractability data.

The workflow contains two additional modules that can aid the identification of therapeutic targets within the candidate set. In the *Gene fitness effects* module, the tool shows which candidate targets are required for cellular fitness in different molecular contexts, lending



	symbol	genename	targetset_cancer_rank	oncogenic	tumor_suppressive
1	SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4	1	Strong	Very Strong
2	MYC	MYC proto-oncogene, bHLH transcription factor	0.98	Very Strong	
3	MLH1	mutL homolog 1	0.96		Very Strong
4	ARID1A	AT-rich interaction domain 1A	0.93		Very Strong
5	BAP1	BRCA1 associated protein 1	0.91		Very Strong
6	KMT2A	lysine methyltransferase 2A	0.89	Very Strong	
7	BRD4	bromodomain containing 4	0.87	Very Strong	
8	TRIM24	tripartite motif containing 24	0.84	Very Strong	Moderate
9	HDAC2	histone deacetylase 2	0.82		
10	ARID1B	AT-rich interaction domain 1B	0.8		Strong
11	MAX	MYC associated factor X	0.78		Moderate
12	FUS	FUS RNA binding protein	0.76	Very Strong	Moderate
13	CDC73	cell division cycle 73	0.73		Very Strong
14	TRIM33	tripartite motif containing 33	0.71		Moderate
15	TRRAP	transformation/transcription domain associated protein	0.69	Moderate	
16	SMARCE1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1	0.67		Moderate
17	USP7	ubiquitin specific peptidase 7	0.64	Moderate	
18	TPX2	TPX2 microtubule nucleation factor	0.62	Moderate	
19	RBM10	RNA binding motif protein 10	0.6		Moderate
20	ZMYM2	zinc finger MYM-type containing 2	0.58		

FIGURE 2 Examples of analysis output from the oncoEnrichR workflow. Output views from oncoEnrichR analysis modules when analyzing a set of n = 134 c-MYC interacting partners.[39] The complete oncoEnrichR report for this example gene set is available for download at https://doi.org/10.5281/zenodo.8051153. (A) Overall rank of query set proteins with respect to cancer relevance, considering the built-in pan-cancer gene rank score of oncoEnrichR. Users can also interrogate/filter the query set with respect to classifications as proto-oncogene/tumor suppressor genes and view supporting literature. (B) A heatmap of cancer relevance among query set members, measured on a per tumor-type basis according to oncoEnrichR's rank score. (C) Target priority scores of query set members from Project Score, considering the combination of gene fitness scores in cancer cell lines and the availability of known biomarkers. (D) An interactive protein-protein interaction network of the query set, here shown with underlying data from the STRING database. Query set members are shaped as circles, and the network is expanded with the most important interacting non-query set proteins (shaped as squares). Cancer gene roles are color-coded (tumor suppressors in red, proto-oncogenes in green, dual roles in black), and targeted cancer drugs can also be attached to the network. (E) A heatmap of subcellular localization for query set members, facilitated by the subcellular template provided by the *gganatogram* R package (https://github.com/jespermaag/gganatogram), showcasing the most abundant subcellular compartments for the hits analyzed. (F) Network of known regulatory interactions (transcription factor-target) for members of the query set.

upon data from large-scale CRISPR-Cas9 whole-genome dropout screens in human cancer cell lines.[21] Notably, the intersection of candidate hits with such genetic screen data also indicates relevant cell lines to be used in experimental follow-up studies. The integration of fitness effects with target tractability data and known molecular biomarkers allows for further prioritization of candidate hits as therapeutic targets (Figure 2C). On top of that, oncoEnrichR incorporates recent predictions on synthetic lethality in cancer cell lines from a machine learning model,[22] effectively allowing the prioritization of candidate hits that are involved in such interactions, either as internal interactions within the candidate set or through interactions with other genes.

3.1.2 | Functional enrichment, subcellular localization and molecular interactions

Exploring the candidate set through a systems perspective can often give important biological insights, and has become a common approach for the interpretation of gene lists, particularly within complex diseases such as cancer. In oncoEnrichR, system level analyses of the candidate set are provided through a series of modules. A *Functional enrichment* module offers standard enrichment/over-representation analysis of the candidate gene set toward GO, multiple pathway databases and established cancer-relevant gene signatures (Table S2), all supported by the clusterProfiler enrichment analysis tool.[23] As mentioned above, the user can supply a dedicated background gene set to be used in the enrichment analysis, and configure enrichment significance thresholds.

Interactive protein-protein interaction (PPI) networks of the candidate set are provided based on available data from the STRING (https://string-db.org) and BioGRID (http://thebiogrid.org) databases.[8,24] While BioGRID data here contains curated physical interactions only, STRING allows for exploring additional functional interactions predicted from other types of data (eg, text mining). oncoEnrichR shows both internal interactions among the candidate hits, and the most significant interactions with non-candidate set proteins. The user can further configure the confidence threshold of retrieved interactions. Importantly, as an aid to portray cancer relevance in the network, the different roles of cancer genes are explicitly color-coded, and the user may also opt to attach targeted cancer drugs to the nodes in the network (Figure 2D). The PPI module further allows local community structures in the network to be interrogated, and all proteins in the candidate set are ranked according to their level of centrality (ie, number of interactions) in the network.

A *protein complex* module further considers both curated and predicted protein complexes from multiple databases (CORUM,[25] ComplexPortal,[26] COMPLEAT[27] and hu.MAP[28]), catering for a ranking of the most cancer-relevant protein complexes with respect to candidate set members, including links to supporting literature. A *subcellular compartment localization* module offers annotation of the subcellular localization patterns for proteins in the query set, in which a subcellular "anatogram" provides an effective heatmap that indicates the most common subcellular compartments among proteins found in the candidate list (Figure 2E).

Considering the importance of inter- and intracellular signaling in cancer, oncoEnrichR integrates data on curated ligand-receptor interactions, as well as data on genes involved in transcriptional regulation.[29,30] The latter set of interactions can be shown through a network, indicating the presence of directed transcription-factor (TF)-target relationships among members of the candidate set (Figure 2F).

3.1.3 | Gene mutation frequencies, co-expression patterns and prognostic associations

Through the use of large-scale genomics data from The Cancer Genome Atlas (TCGA), candidate hits can be interrogated and prioritized for aberration frequencies in different tumor types (somatic copy number events and point mutations/indels), a well-known indication of cancer relevance.[12] The user can further examine the presence of known somatic mutation hotspots within candidate genes, and identify the amino acid sites that harbor loss-of-function variants in tumor samples. Genes in the candidate set that are found significantly co-expressed with known cancer genes in various tumor types are further highlighted in a separate section.

Expression profiling data from healthy tissues (Human Protein Atlas [HPA] and Genotype-Tissue Expression Project) are included in oncoEnrichR to portray tissue- and cell-type specific expression patterns for the candidate genes, which in turn may reveal enrichment of particular tissues/cell-types.[31,32]

It is well known that expression levels and genomic aberrations of particular protein-coding genes are associated with overall patient survival in a number of tumor types.[33] In the *Prognostic associations* module, oncoEnrichR shows both favorable and unfavorable prognostic associations of hits in the candidate set, considering either expression or methylation levels or mutation or copy-number status of these genes across tumor samples.[33]

3.2 | Applications of oncoEnrichR

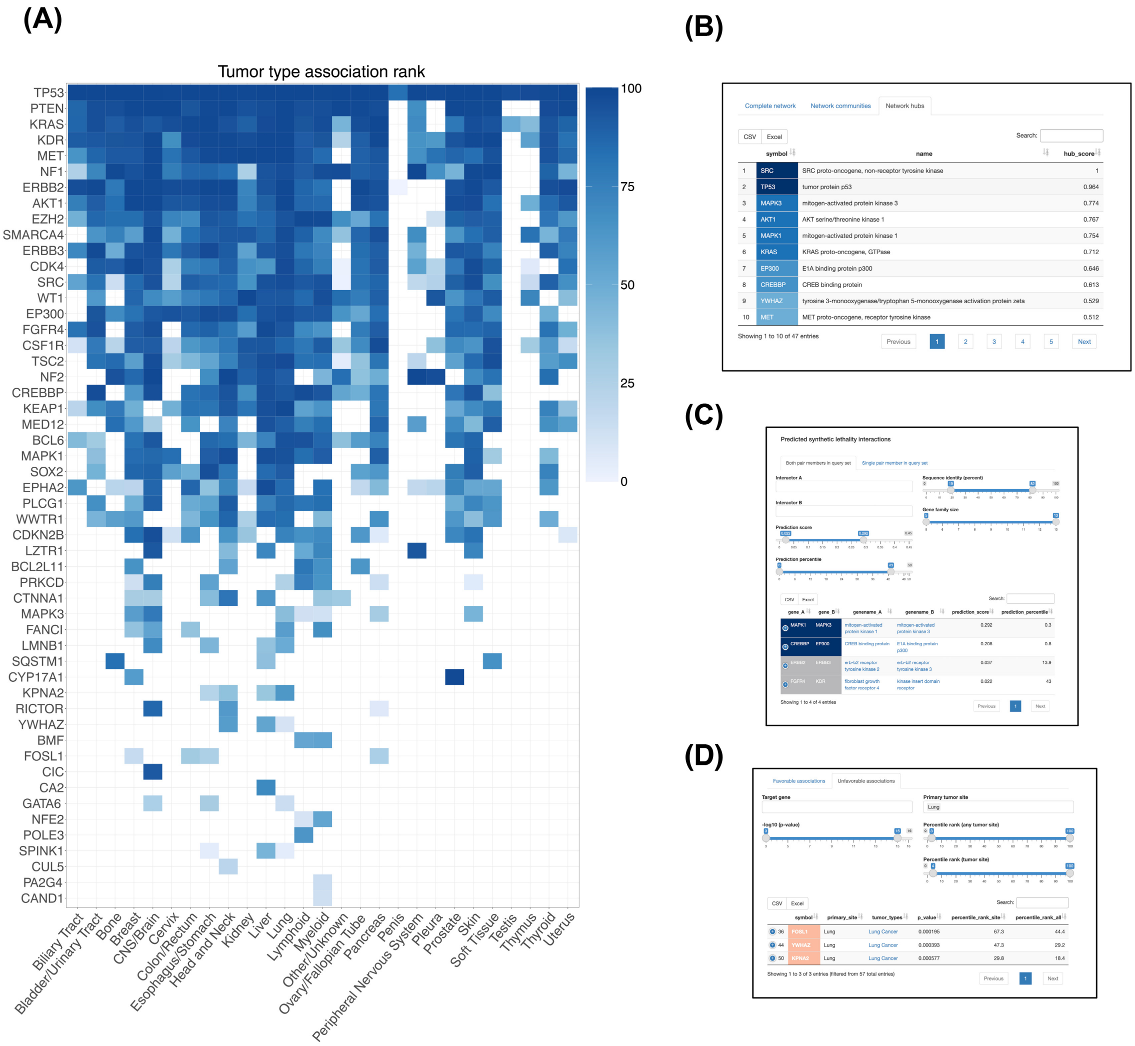
To showcase the usability of oncoEnrichR in the biological interpretation of gene lists, we explored the output of the tool using two candidate hit lists coming from CRISPR and protein proximity screens, respectively (complete output reports are available at https://doi.org/10.5281/zenodo.8051181).

3.2.1 | CRISPR screen: resistance to EGFR inhibition

We initially interrogated a list of n = 57 hits originating from a CRISPR/Cas9 screen looking for novel drug resistance genes in non-small cell lung cancer.[34] Of note, the set of 57 hits came here out of a domain-specific recommendation system, subsequently assessed by five independent domain experts with regard to their relevance as potential drivers of resistance to EGFR inhibition. The aim of the oncoEnrichR analysis in this context was thus to showcase how the tool can validate the relevance of known resistance markers, and add supporting evidence and perspectives on previously unknown markers (see also https://www.nature.com/articles/s41467-022-29292-7/figures/3, which illustrates known [panel A] and unknown [panel B and C] resistance markers).

The report produced with oncoEnrichR confirms the strong cancer relevance of well-established resistance markers, including multiple known tumor suppressor genes (eg, TP53, PTEN, NF1/2, SMARCA4) and proto-oncogenes (eg, KRAS, ERBB2, MET, MAPK1, CDK4), as can be seen in the *Cancer associations* section. Three of the previously unknown resistance markers (CIC, EZH2 and CREBBP) are indicated to carry both oncogenic and tumor suppressive roles, a matter which can be further explored in the linked supporting literature from CancerMine, and also through entries shown in the *Cancer Hallmarks* evidence section. The tumor type-specific association overview (Figure 3A) provides further an opportunity to find candidates for which the current available evidence of association to lung cancer is weak (eg, LZTR1 and FOSL1), and also other candidates which are strongly associated to selected tumor types (eg, CYP1A1 in prostate cancer, SQSTM1 in sarcoma).

The *Drug associations* section highlights the availability of a large collection of anticancer drugs, which confirms how several of the unknown resistance markers can be targeted by various inhibitors (eg, Tazemetostat [EZH2], Infigratinib/Erdafitinib [FGFR4]). An important consideration with respect to potential drug targets is their level of interactivity with other proteins, and through its interactome hub score calculation oncoEnrichR highlights the centrality of the SRC protein (Figure 3B). Furthermore, oncoEnrichR allows the interrogation of predicted synthetic lethality of multiple candidate hits, here



Complete network | Network communities | Network hubs

CSV | Excel   Search:

	symbol	name	hub_score
1	SRC	SRC proto-oncogene, non-receptor tyrosine kinase	1
2	TP53	tumor protein p53	0.964
3	MAPK3	mitogen-activated protein kinase 3	0.774
4	AKT1	AKT serine/threonine kinase 1	0.767
5	MAPK1	mitogen-activated protein kinase 1	0.754
6	KRAS	KRAS proto-oncogene, GTPase	0.712
7	EP300	E1A binding protein p300	0.646
8	CREBBP	CREB binding protein	0.613
9	YWHAZ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein zeta	0.529
10	MET	MET proto-oncogene, receptor tyrosine kinase	0.512

Showing 1 to 10 of 47 entries   Previous 1 2 3 4 5 Next

Predicted synthetic lethality interactions

Both pair members in query set | Single pair member in query set

Interactor A

Interactor B

Sequence identity (percent): 19 – 82 (0–100)

Gene family size: 5 – 13

Prediction score: 0.022 – 0.292 (0–0.45)

Prediction percentile: 0 – 43 (0–50)

CSV | Excel   Search:

gene_A	gene_B	genename_A	genename_B	prediction_score	prediction_percentile
MAPK1	MAPK3	mitogen-activated protein kinase 1	mitogen-activated protein kinase 3	0.292	0.3
CREBBP	EP300	CREB binding protein	E1A binding protein p300	0.208	0.8
ERBB2	ERBB3	erb-b2 receptor tyrosine kinase 2	erb-b2 receptor tyrosine kinase 3	0.037	13.9
FGFR4	KDR	fibroblast growth factor receptor 4	kinase insert domain receptor	0.022	43

Showing 1 to 4 of 4 entries   Previous 1 Next

Favorable associations | Unfavorable associations

Target gene

Primary tumor site: Lung

-log10 (p-value): 3 – 15 (3–16)

Percentile rank (any tumor site): 3 – 100 (0–100)

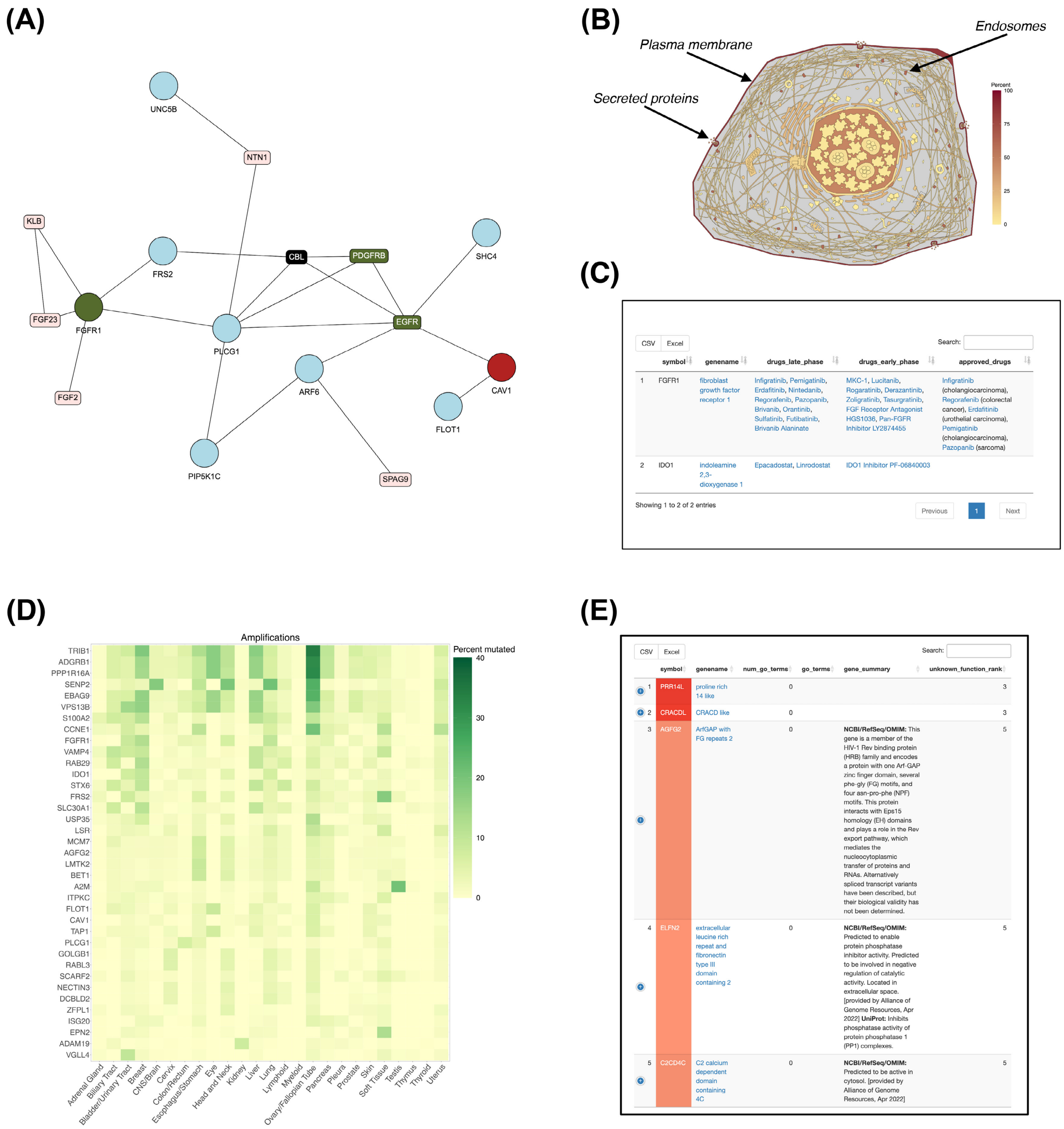
Percentile rank (tumor site): 4 – 100 (0–100)

CSV | Excel   Search:

	symbol	primary_site	tumor_types	p_value	percentile_rank_site	percentile_rank_all
36	FOSL1	Lung	Lung Cancer	0.000195	67.3	44.4
44	YWHAZ	Lung	Lung Cancer	0.000393	47.3	29.2
50	KPNA2	Lung	Lung Cancer	0.000577	29.8	18.4

Showing 1 to 3 of 3 entries (filtered from 57 total entries)   Previous 1 Next

FIGURE 3 CRISPR use case analysis—EGFRi resistance drivers. Specific output views from the oncoEnrichR HTML report, analyzing N = 57 candidate drivers of EGFRi resistance (complete reports available at https://doi.org/10.5281/zenodo.8051181). (A) Tumor-type association rank for all candidate hits found, confirming the strong cancer relevance of multiple known resistance markers. (B) Ranking of candidate hits according to protein-protein network centrality scores, indicating the vast interactome of SRC. (C) Predicted synthetic lethality interactions among members of the candidate set. (D) Prognostic gene expression associations for candidate genes in lung cancer patients.



	symbol	genename	drugs_late_phase	drugs_early_phase	approved_drugs
1	FGFR1	fibroblast growth factor receptor 1	Infigratinib, Pemigatinib, Erdafitinib, Nintedanib, Regorafenib, Pazopanib, Brivanib, Orantinib, Sulfatinib, Futibatinib, Brivanib Alaninate	MKC-1, Lucitanib, Rogaratinib, Derazantinib, Zoligratinib, Tasurgratinib, FGF Receptor Antagonist HGS1036, Pan-FGFR Inhibitor LY2874455	Infigratinib (cholangiocarcinoma), Regorafenib (colorectal cancer), Erdafitinib (urothelial carcinoma), Pemigatinib (cholangiocarcinoma), Pazopanib (sarcoma)
2	IDO1	indoleamine 2,3-dioxygenase 1	Epacadostat, Linrodostat	IDO1 Inhibitor PF-06840003	

Showing 1 to 2 of 2 entries

	symbol	genename	num_go_terms	go_terms	gene_summary	unknown_function_rank
1	PRR14L	proline rich 14 like	0			3
2	CRACDL	CRACD like	0			3
3	AGFG2	ArfGAP with FG repeats 2	0		**NCBI/RefSeq/OMIM:** This gene is a member of the HIV-1 Rev binding protein (HRB) family and encodes a protein with one Arf-GAP zinc finger domain, several phe-gly (FG) motifs, and four asn-pro-phe (NPF) motifs. This protein interacts with Eps15 homology (EH) domains and plays a role in the Rev export pathway, which mediates the nucleocytoplasmic transfer of proteins and RNAs. Alternatively spliced transcript variants have been described, but their biological validity has not been determined.	5
4	ELFN2	extracellular leucine rich repeat and fibronectin type III domain containing 2	0		**NCBI/RefSeq/OMIM:** Predicted to enable protein phosphatase inhibitor activity. Predicted to be involved in negative regulation of catalytic activity. Located in extracellular space. [provided by Alliance of Genome Resources, Apr 2022] **UniProt:** Inhibits phosphatase activity of protein phosphatase 1 (PP1) complexes.	5
5	C2CD4C	C2 calcium dependent domain containing 4C	0		**NCBI/RefSeq/OMIM:** Predicted to be active in cytosol. [provided by Alliance of Genome Resources, Apr 2022]	5

FIGURE 4 Proteomics use case analysis—FGFR1 signaling network. Specific output views from the oncoEnrichR HTML report, analyzing N = 74 proteins found through a protein proximity screen using FGFR1 as the bait (complete reports available at https://doi.org/10.5281/zenodo.8051181). (A) Community structure of the protein-protein interaction network (based on STRING interaction data), validating the proximal interaction partners of FGFR1. Query set members as circles, interacting non-query set proteins as squares, cancer gene roles are color-coded (tumor suppressors in red, proto-oncogenes in green, dual roles in black). (B) Subcellular anatogram indicating main localization of hits (secretory pathway, the plasma membrane, the endosomal system and the nucleus). (C) Cancer drugs in early or late phase development targeted against key members of the query set. (D) High-level amplifications of query set members found in tumor samples from TCGA cohorts, showcasing considerable frequency of CCNE1 amplifications in samples originating from the ovary/fallopian tube. (E) List of genes with poorly defined or unknown functions detected in the query set, indicated through lack of gene summary descriptions or curated GO annotation terms.

exemplified through MAPK1/MAPK3 and CREBBP/EP300 interactions (Figure 3C). Finally, exploring potential prognostic associations can provide important supplementary evidence for selection of follow-up candidates. oncoEnrichR indicates that high expression of three candidate hits is significantly associated with poor survival in lung cancer (Figure 3D), interestingly also FOSL1, for which there was weak overall evidence in the *Cancer Associations* section.

3.2.2 | Protein proximity screen: FGFR1 interaction network

We have previously performed a proximity labeling BioID experiment using FGFR1 (fibroblast growth factor receptor 1) as the bait, where the potential interaction partners were analyzed by Enrichr and BioGrid at the time.[35] Enrichr was used to identify enriched pathways and molecular functions of the FGFR1 interactome, and BioGrid for comparison of previously reported partners of FGFR1 to hits found in our experiment. As aberrant FGF signaling is implicated in multiple tumor types,[36,37] we wanted to use oncoEnrichR to reassess the dataset and further elucidate FGFR1 networks in cancer.

Proteins associating with FGFR1 upon stimulation with the activating ligand (FGF1) were subjected to oncoEnrichR analysis ($n = 74$ proteins). oncoEnrichR successfully imported gene names from the Maxquant proteomic output file. As the screen was done several years ago, a few names were now obsolete ($n = 6$), but they were all successfully updated by oncoEnrichR with the current official gene names (*Query verification* module).

Several features in oncoEnrichR proved useful for the validation of screen hits. For example, the *protein-protein interaction* module revealed previously identified interactions. In the case of FGFR1, the proximal interaction network included the previously reported interactions of FRS2 and PLCG1 (Figure 4A). Moreover, FGFR1 is known to localize to the secretory pathway, the plasma membrane, the endosomal system and the nucleus. These localizations were clearly reflected in the subcellular anatogram showing that the majority of the hits in the BioID screen localized to these compartments (Figure 4B). As such, the subcellular compartment module may help in validating the results from the screen, but also point to new cell localizations for baits, via their interaction partners. For example, the reported *focal adhesion sites* could imply a new localization and function of FGFR1 with respect to focal adhesions. Finally, gene set enrichment analyses indicated the expected protein networks (eg, secretory pathway, endosomes, clathrin-coated vesicle), as well as some new unexplored functions (eg, focal adhesions).

A useful feature in oncoEnrichR is the annotation of hits as tumor suppressors or oncogenes, including associated links to published literature.[19] oncoEnrichR also informs on druggability and suggests matching inhibitors which can be used to validate and further investigate hits. For instance, oncoEnrichR reports on the cancer associations of IDO1 (indoleamine 2,3-dioxygenase 1), a strong hit in the screen and suggest inhibitors that can be used to target this protein (eg, Epacadostat and Linrodostat), providing an easy way to test its role in biological experiments (Figure 4C). Interestingly, CCNE1, an important oncogene involved in cell cycle progression, could possibly also be involved in FGFR1 signal transduction, as it was found as a strong hit in the screen. Large-scale CRISPR/Cas9 screen data found in the *Gene fitness effects* module can be used to suggest which cancer cell lines are dependent on your genes of interest. For instance, in the case of CCNE1, cell lines derived from ovary cancer depend on this gene, and oncoEnrichR shows a list of cell lines that can be used for further studies. Of note, CCNE1 is amplified in ovary cancer (21% of cases), as shown in the *Tumor aberration frequencies* module (Figure 4D).

The oncoEnrichR also lists proteins where no or little evidence for function has been found, thereby highlighting proteins that could be the starting point for additional research projects. In the FGFR1 dataset, we identified several proteins with a poorly defined function (eg, ELFN2, C2CD4C, PRR14L and CRACDL, Figure 4E). Interestingly, C2CD4C was found to be frequently deleted in several cancers (eg, 6% in cervix cancer), which may suggest a potential role as a tumor suppressor.

4 | DISCUSSION

Building upon existing data integration frameworks and multiple large-scale omics datasets, we have developed a feature-rich gene set interpretation tool that allows researchers to systematically interpret and prioritize long lists of candidate genes for cancer relevance. The frequently time-consuming nature of gene list interpretation can in part be attributed to a scarcity of single tools that comprehensively organize existing knowledge in the context of cancer. In this regard, we believe that oncoEnrichR, through its broad harvest of prior molecular knowledge and its multifaceted and interactive analysis report, fills an important gap. Moreover, the dual availability of the tool (ie, web and command-line), makes it attractive for use both by researchers with a noncomputational background and for bioinformaticians that are developing complete screening analysis pipelines.

The output of a gene set analysis performed with oncoEnrichR reflects the current, temporary state of molecular cancer knowledge harvested from publicly available databases, for which update frequencies vary considerably (see Table S2). Considering future maintenance and updates of oncoEnrichR, we have established and made available a suite of data preprocessing and quality control procedures, where the ambition is to release data updates for the tool on a biannual basis (see *Data Availability Statement*). Furthermore, we believe that the modular nature of the tool and output report sections will simplify the potential addition of new functionality that can further enhance the mapping of cancer relevance. With respect to specific developments of new functionality, we are considering the support for query lists coming from screens conducted in model organisms (eg, flies), so that users can interrogate the cancer significance of respective human orthologs. We are also planning for the possibility to create a tumor-type focused analysis report, in which prior knowledge is specifically focused on a user-defined cancer type, and where hits for example can be filtered or highlighted based on tumor-type specific expression levels harvested from large-scale cohorts (eg, TCGA). While oncoEnrichR is currently available both as a GitHub-hosted R package and through a web-based interface in Galaxy, we aim to distribute it even more widely using common package repositories, such as CRAN (https://cran.r-project.org/) or Bioconductor.[38]

Like any gene set interpretation tool, we acknowledge that the user needs to interpret the output carefully with reference to the screen that produced the candidate hits. While the simple, unranked gene list input makes oncoEnrichR easy to use, we recognize that an input handling scheme that incorporates effect size or rank from the originating screen will open the way for more sophisticated approaches of hit prioritization.

Importantly, genome-scale screens in cancer have many different objectives, and while oncoEnrichR is not tailored to a specific type of screen or designed to rank the hits according to a particular scientific question, we believe that the multiple perspectives provided by the tool can offer significant interpretation and prioritization support across a wide range of screening applications.

AUTHOR CONTRIBUTIONS

Sigve Nakken: Conceptualization, Software, Methodology, Data curation, Writing - original draft, Writing - review and editing. **Sveinung Gundersen:** Software, Writing - review and editing. **Fabian L. M. Bernal:** Software, Writing - review and editing. **Dimitris Polychronopoulos:** Investigation, Writing - review and editing. **Eivind Hovig:** Resources, Writing - review and editing. **Jørgen Wesche:** Conceptualization, Methodology, Data curation, Writing - review and editing.

ACKNOWLEDGEMENTS

This work was supported by the Research Council of Norway through its Centers of Excellence funding scheme (grant number: 262652). JW was funded by the Norwegian Cancer Society (project number: 198093). The authors wish to acknowledge the University Center for Information Technology (USIT) at the University of Oslo for infrastructure support, ELIXIR Norway for support with Galaxy development, Carlos Company (AstraZeneca UK) for feedback on the article and Dr. Ultan McDermott (Oncology R&D, AstraZeneca UK) for valuable feedback and input on functionality.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

All source code for oncoEnrichR is publicly available at https://github.com/sigven/oncoEnrichR. User documentation is provided at https://sigven.github.io/oncoEnrichR. The raw molecular datasets that are integrated in the data backend of oncoEnrichR, as well as code for preprocessing procedures, are available at https://doi.org/10.5281/zenodo.8051239. All data sources and handling of these data are described in the Supplementary Materials and Methods, and in Supplementary Table S2. Further information is available from the corresponding author upon request.

ORCID

Sigve Nakken https://orcid.org/0000-0001-8468-2050

TWITTER

Sigve Nakken *@sigven*

REFERENCES

1. Bock C, Datlinger P, Chardon F, et al. High-content CRISPR screening. *Nat Rev Methods Primers*. 2022;2:1-23.
2. Sharifi Tabar M, Francis H, Yeo D, Bailey CG, Rasko JEJ. Mapping oncogenic protein interactions for precision medicine. *Int J Cancer*. 2022;151:7-19.
3. Cieślik M, Chinnaiyan AM. Cancer transcriptome profiling at the juncture of clinical translation. *Nat Rev Genet*. 2018;19:93-109.
4. Zhou Y, Zhou B, Pache L, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun*. 2019;10:1523.
5. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4:44-57.
6. Chen EY, Tan CM, Kou Y, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform*. 2013;14:128.
7. Safran M, Dalah I, Alexander J, et al. GeneCards version 3: the human gene integrator. *Database*. 2010;2010:baq020.
8. von Mering C, Jensen LJ, Snel B, et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*. 2005;33:D433-D437.
9. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol*. 2008;9(Suppl 1):S4.
10. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2012;2:401-404.
11. Tang Z, Kang B, Li C, Chen T, Zhang Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res*. 2019;47:W556-W560.
12. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet*. 2013;45:1113-1120.
13. Ochoa D, Hercules A, Carmona M, et al. Open targets platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Res*. 2021;49:D1302-D1310.
14. Malone J, Holloway E, Adamusiak T, et al. Modeling sample variables with an experimental factor ontology. *Bioinformatics*. 2010;26:1112-1118.
15. Kundra R, Zhang H, Sheridan R, et al. OncoTree: A cancer classification system for precision oncology. *JCO Clin Cancer Inform*. 2021;5:221-230.
16. Afgan E, Baker D, Batut B, et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res*. 2018;46:W537-W544.
17. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer*. 2018;18:696-705.
18. Repana D, Nulsen J, Dressler L, et al. The network of cancer genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol*. 2019;20:1.
19. Lever J, Zhao EY, Grewal J, Jones MR, Jones SJM. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat Methods*. 2019;16:505-507.
20. Martínez-Jiménez F, Muiños F, Sentís I, et al. A compendium of mutational cancer driver genes. *Nat Rev Cancer*. 2020;20:555-572.
21. Dwane L, Behan FM, Gonçalves E, et al. Project score database: a resource for investigating cancer cell dependencies and prioritizing therapeutic targets. *Nucleic Acids Res*. 2020;49:D1365-D1372.
22. de Kegel B, Quinn N, Thompson NA, Adams DJ, Ryan CJ. Comprehensive prediction of robust synthetic lethality between paralog pairs in cancer cell lines. *Cell Syst*. 2021;12:1144-59.e6.
23. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012;16:284-287.
24. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*. 2006;34:D535-D539.
25. Tsitsiridis G, Steinkamp R, Giurgiu M, et al. CORUM: the comprehensive resource of mammalian protein complexes-2022. *Nucleic Acids Res*. 2022;51:D539-D545.

26. Meldal BHM, Bye-A-Jee H, Gajdoš L, et al. Complex portal 2018: extended content and enhanced visualization tools for macromolecular complexes. *Nucleic Acids Res*. 2018;47:D550-D558.
27. Vinayagam A, Hu Y, Kulkarni M, et al. Protein complex-based analysis framework for high-throughput data sets. *Sci Signal*. 2013;6:rs5.
28. Drew K, Wallingford JB, Marcotte EM. hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Mol Syst Biol*. 2021;17:e10016.
29. Jin S, Guerrero-Juarez CF, Zhang L, et al. Inference and analysis of cell-cell communication using CellChat. *Nat Commun*. 2021;12:1088.
30. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res*. 2019;29:1363-1375.
31. Jain A, Tuteja G. TissueEnrich: tissue-specific gene enrichment analysis. *Bioinformatics*. 2019;35:1966-1967.
32. Uhlen M, Oksvold P, Fagerberg L, et al. Towards a knowledge-based human protein atlas. *Nat Biotechnol*. 2010;28:1248-1250.
33. Smith JC, Sheltzer JM. Genome-wide identification and analysis of prognostic features in human cancers. *Cell Rep*. 2022;38:110569.
34. Gogleva A, Polychronopoulos D, Pfeifer M, et al. Knowledge graph-based recommendation framework identifies drivers of resistance in EGFR mutant non-small cell lung cancer. *Nat Commun*. 2022;13:1667.
35. Kostas M, Haugsten EM, Zhen Y, et al. Protein tyrosine phosphatase receptor type G (PTPRG) controls fibroblast growth factor receptor (FGFR) 1 activity and influences sensitivity to FGFR kinase inhibitors. *Mol Cell Proteomics*. 2018;17:850-870.
36. Ahmad I, Iwata T, Leung HY. Mechanisms of FGFR-mediated carcinogenesis. *Biochim Biophys Acta*. 2012;1823:850-860.
37. Tanner Y, Grose RP. Dysregulated FGF signalling in neoplastic disorders. *Semin Cell Dev Biol*. 2016;53:126-135.
38. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5:R80.
39. Dingar D, Kalkat M, Chan P-K, et al. BioID identifies novel c-MYC interacting partners in cultured cells and xenograft tumors. *J Proteome*. 2015;118:95-111.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Nakken S, Gundersen S, Bernal FLM, Polychronopoulos D, Hovig E, Wesche J. Comprehensive interrogation of gene lists from genome-scale cancer screens with oncoEnrichR. *Int J Cancer*. 2023;153(10):1819-1828. doi:10.1002/ijc.34666