Genome Medicine

SOFTWARE Open Access

A global cancer data integrator reveals principles of synthetic lethality, sex disparity and immunotherapy

Christopher Yogodzinski[1,2,3*], Abolfazl Arab[1,2,4], Justin R. Pritchard[5], Hani Goodarzi[1,2,4] and Luke A. Gilbert[1,2,6*]

Abstract

Background: Advances in cancer biology are increasingly dependent on integration of heterogeneous datasets. Large-scale efforts have systematically mapped many aspects of cancer cell biology; however, it remains challenging for individual scientists to effectively integrate and understand this data.

Results: We have developed a new data retrieval and indexing framework that allows us to integrate publicly available data from different sources and to combine publicly available data with new or bespoke datasets. Our approach, which we have named the cancer data integrator (CanDI), is straightforward to implement, is well documented, and is continuously updated which should enable individual users to take full advantage of efforts to map cancer cell biology. We show that CanDI empowered testable hypotheses of new synthetic lethal gene pairs, genes associated with sex disparity, and immunotherapy targets in cancer.

Conclusions: CanDI provides a flexible approach for large-scale data integration in cancer research enabling rapid generation of hypotheses. The CanDI data integrator is available at https://github.com/GilbertLabUCSF/CanDI.

Keywords: Functional genomics, Multiomics, Data integration, Synthetic lethality

Background

Large-scale but often independent efforts have mapped phenotypic characteristics of more than one thousand human cancer cell lines. Despite this, static lists of univariate data generally cannot identify the underlying molecular mechanisms driving a complex phenotype.

We hypothesized that a global cancer data integrator that could incorporate many types of publicly available data including functional genomics, whole genome sequencing, exome sequencing, RNA expression data, protein mass spectrometry, DNA methylation profiling, chromatin immunoprecipitation sequencing (ChIP-seq), assay for transposase-accessible chromatin sequencing (ATAC-seq), and metabolomics data would enable us to link disease features to gene products [1–15]. We set out to build a resource that enables cross platform correlation analysis of multi-omic data as this analysis is in and of itself is a high-resolution phenotype. Multi-omic analysis of functional genomics data with genomic, metabolomic or transcriptomic profiling can link cell state or specific signaling pathways to gene function [2, 3, 13, 15–18]. Lastly, co-essentiality profiling across large panels of cell lines has revealed protein complexes and co-essential modules that can assign function to uncharacterized genes [19].

Problematically, in many cases publicly available data are poorly integrated when considering information on all genes across different types of data and the existing data portals are inflexible. For example, lists of genes cannot be queried against groups of cell lines stratified by mutation status or disease subtype. Furthermore, one

* Correspondence: cyogodzi@unc.edu; luke.gilbert@ucsf.edu
[1]Department of Urology, University of California, San Francisco, CA, USA
Full list of author information is available at the end of the article

cannot integrate new data derived from individual labs or other consortia.

We created the Cancer Data Integrator (CanDI) which is a series of python modules designed to seamlessly integrate genomic, functional genomic, RNA, protein, and metabolomic data into one ecosystem [20]. Our python framework operates like a relational database without the overhead of running MySQL or Postgres and enables individual users to easily query this vast dataset and add new data in flexible ways. This was achieved by unifying the indices of these datasets via index tables that are automatically accessed through CanDI's biologically relevant Python Classes. We highlight the utility of CanDI through four types of analysis to demonstrate how complex queries can reveal previously unknown molecular mechanisms in synthetic lethality, sex disparity, and immunotherapy. These data nominate new small molecule and immunotherapy anti-cancer strategies in KRAS-mutant colon, lung, and pancreatic cancers.

Implementation

CanDI

The CanDI data integrator is available at https://github.com/GilbertLabUCSF/CanDI[20].

CanDI module structure

CanDI is a python library built on top of Pandas specializing in retrieving, formatting, and integrating the publicly available data from The Cancer Dependency Map (DepMap) [12, 20], The Cancer Cell Line Encyclopedia (CCLE) [1], The Pooled In-Vitro CRISPR Knockout Essentiality Screens Database (PICKLES) [21], The Comprehensive Resource of Mammalian Protein Complexes (CORUM) [8] and protein localization data from The Cell Atlas [4], The Map of the Cell [11], and The In Silico Surfaceome [7, 22] (Additional file 1: Fig. S1). The data we present is sourced from the 2021Q2 releases of DepMap and CCLE data and the Avanna 2018Q3 release of PICKLES data [23]. CanDI is designed to retrieve data from the most current release of DepMap and CCLE data. While some data we present is from analysis of Bayes Factors take from PICKLES, CanDI's current version no longer makes uses of these data as a measure of gene essentiality. At its core, CanDI is a software that allows for automated retrieval and formatting of publicly available datasets, as well as computational tools that allow for quick data dataset sub-setting, integration, and cross-referencing. Data retrieval is not tied to specific release of the data of which CanDI classes are built around.

Access to all datasets is controlled via a python class called Data. Upon import, the data class reads the config file established during installation and defines unique paths to each dataset and automatically loads the cell line index table and the gene index table. Installation of CanDI, configuration, and data retrieval is handled by a manager class that is accessed indirectly through installation scripts and the Data class. Interactions with this data are controlled through a parent Entity class and several handlers. The biologically relevant abstraction classes (Gene, CellLine Cancer, Organelle, GeneCluster, CellLineCluster) inherit their methods from Entity. Entity methods are wrappers for hidden data handler classes who perform specific transformations, such as data indexing and high-throughput filtering.

Differential expression

In all cases where it is mentioned, differential expression was evaluated using the DESeq2 R package (Release 3.13) [24]. Significance was considered to be an adjusted p value of less than 0.01.

Differential essentiality

Essentiality scores (CERES gene effect scores [12]) are taken from the DepMap database (2021Q2). To reduce the number of hypotheses posed during this analysis, the mutual information of gene essentiality was calculated using the mutual information metric from the python package SciKitLearn (Version 0.22.0). Genes with mutual information scores greater than one standard deviation above the median were removed from consideration. Differential essentiality was evaluated by performing a Mann-Whitney U test between two groups on every gene that passed the mutual information filter. Significance was considered to be a p value of less than 0.01. Magnitude of differential essentiality of a given gene was shown as the difference in mean CERES scores between two groups of cell lines.

Protein localization confidence

Protein localization data was assembled from The Cell Atlas [4], The Map of the Cell [11], and The In Silico Surfaceome [7, 22]. Confidence annotations were taken from the supplemental data of each paper and put on a number scale from 0 to 4 and summed for a total confidence score for each localization annotation for every gene across all three databases. The analysis shown in Fig. 4 represents a gene list that was further manually curated to remove the genes that are localized to the intracellular space at the cell membrane revealing cell surface protein targets that are highly expressed in non-small cell lung cancer (NSCLC) cancer models over normal lung bronchial epithelial cells [4, 7, 11, 22].

DepMap Creative Commons license

When an individual uses CanDI they are downloading DepMap data and thus are agreeing to a CC Attribution 4.0 license (https://creativecommons.org/licenses/by/4.0/).

Synthetic lethality of Fanconi anemia genes in ovarian and breast cancer models

Using CanDI, the essentiality scores of 50 top hits identified by a CRISPR screen in Hela cells that confer sensitivity to PARP inhibition [25] were visualized across all ovarian cancer cell models in DepMap (2021Q2) (Fig. 1). FANCA and FANCE showed selective essentiality in the BRCA1 mutant ovarian cancer cell lines. Following this observation, CanDI was used to gather the gene essentiality for all FANC genes in the Fanconi anemia pathway. CanDI was then used to visualize these data across all ovarian and breast cancer cell lines, sorting by BRCA1 mutation status.

Synthetic lethality in KRAS- and EGFR-mutant cell lines

CanDI was leveraged to bin NSCLC cell lines present in both CCLE (Release: 2021Q2) and DepMap (Release 2021Q2) into 8 groups. KRAS-mutant and KRAS-wildtype cell lines with and without EGFR mutants removed as well as EGFR-mutant and EGFR-wildtype cell lines with and without KRAS mutants removed. We present genes that are synthetic lethal with KRAS and EGFR mutations by plotting the mean gene essentiality for all genes in the genome of mutant cell lines against wild type cell lines (Fig. 2, Additional file 1: Fig. S2, S3). Synthetic lethality can be interpreted as a gene's shift off of the $y = x$ line of these figures.

Pan-Cancer synthetic lethality analysis

A set of core oncogenes and tumor suppressor driver mutations was chosen for analysis [26]. To test the effect of these gene's mutations on gene essentiality, CanDI was leveraged to split mutations into two groups: a nonsense mutation group containing genes annotated as tumor suppressors ($N = 153$) and a missense mutation group containing genes annotated as oncogenes with specific driver protein changes ($N = 53$). CanDI was then used to collect a core set of genes with highly variable essentiality. To do this, the Bayes factors from the PICKLES database (Avana 2018Q4) were converted to binary numeric variables. Bayes factors over 5 were assigned a 1 = essential and Bayes factors under 5 were

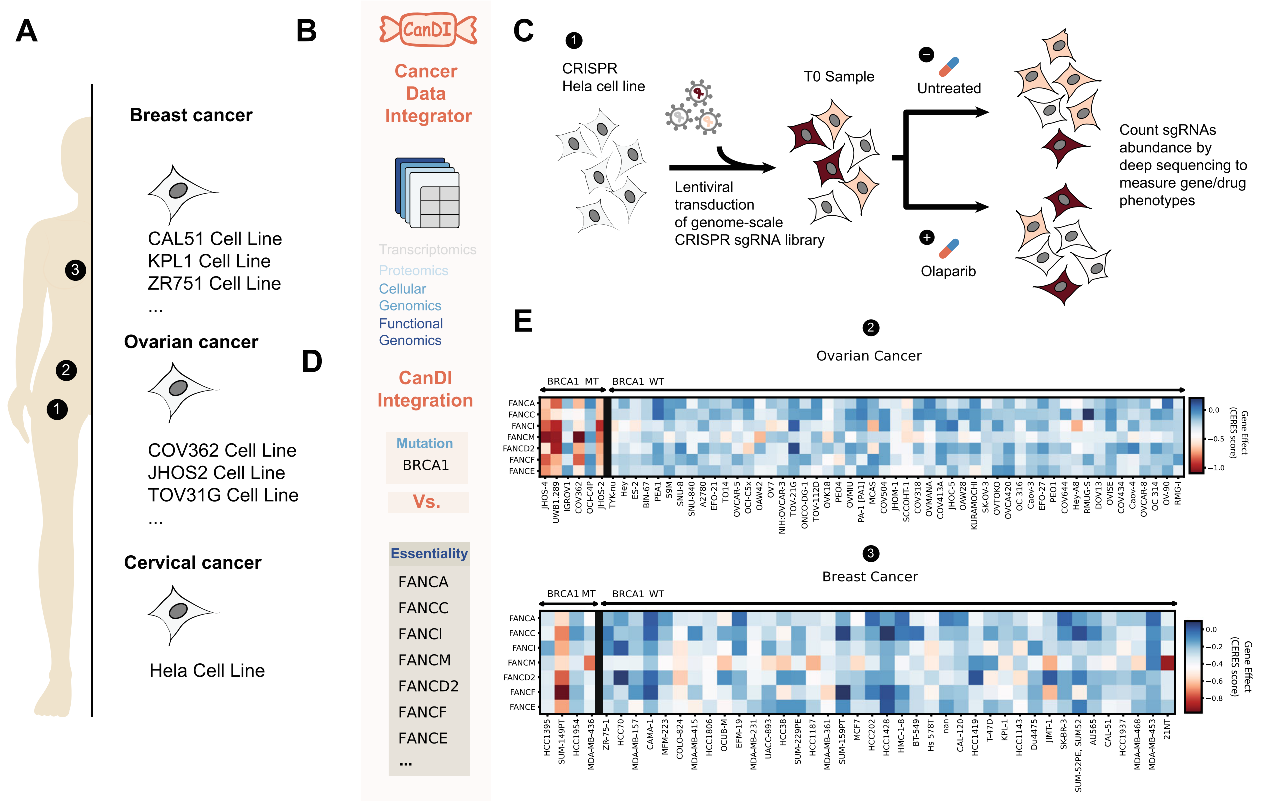

Fig. 1 CanDI integrates multiomics data enabling discovery. **A** A schematic showing human cell models integrated by CanDI. **B** A schematic illustrating types of data integrated by CanDI. **C** A cartoon of a genome-scale CRISPR screen to identify genes that modulate response to PARP inhibition by Olaparib. **D** A schematic depicting data feature inputs parsed by CanDI. **E** Essentiality of Fanconi anemia genes in ovarian and breast cancer cell lines separated by BRCA mutation status. Gene essentiality (CERES gene effect score) is displayed by a heat map. $N = 6$ BRCA1-mutant ovarian cancer, $N = 51$ BRCA1-wildtype ovarian cancer, $N = 4$ BRCA1-mutant breast cancer, $N = 39$ BRCA1-wildtype breast cancer

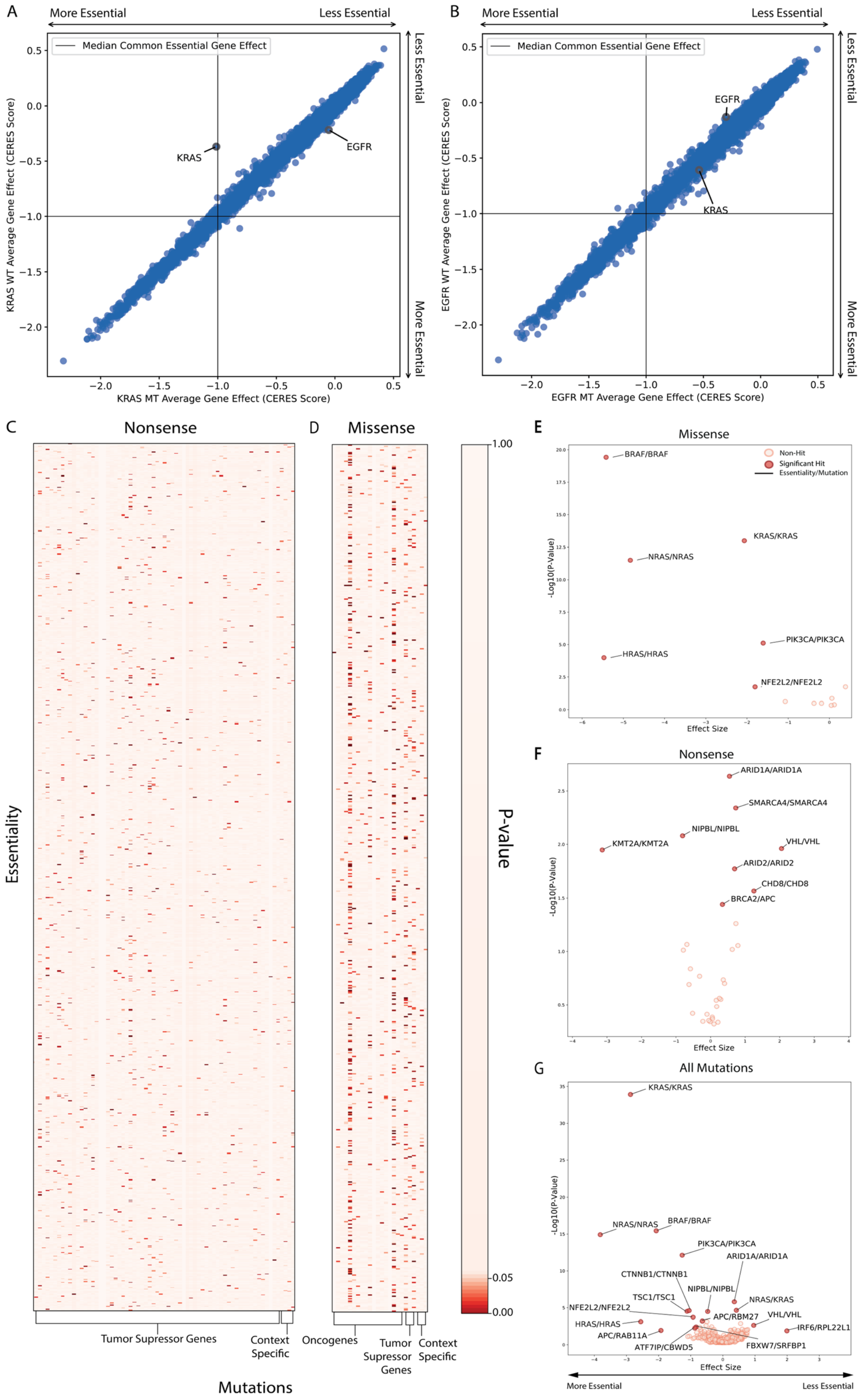

Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 CanDI enables a global analysis of conditional essentiality in cancer. **A** Average gene essentiality for *KRAS* and *EGFR* in groups of NSCLC cell lines stratified by KRAS mutation status. $N = 61$ for KRAS-wildtype shown in blue. $N = 33$ for KRAS-mutant shown in blue. Gene essentiality is the average gene effect (CERES score) across all cell lines for the given group. **B** Average gene essentiality for *KRAS* and *EGFR* in groups of NSCLC cell lines stratified by EGFR mutation status. $N = 77$ for EGFR-wildtype shown in blue, $N = 17$ for EGFR-mutant shown in blue. Gene essentiality is the average gene effect (CERES score) of all cell lines within a given group. **C** *P* values from chi-square tests of gene essentiality and nonsense mutations. **D** *P* values from chi-square tests of gene essentiality and missense mutations. **E** A scatter plot showing effect size of the change in gene essentiality with select missense mutations and the − Log10(*P* value) of each essentiality/mutation pair. **F** A scatter plot showing effect size of the change in gene essentiality with select nonsense mutations and the − Log10(*P* value) of each essentiality/mutation pair. **G** A scatter plot showing effect size of the change in gene essentiality with all mutations and the − Log10(*P* value) of each essentiality/mutation pair

assigned a 0 = non-essential. Genes were then sorted buy their variance across cell lines and genes between the 85th and 95th percentile were used for this analysis ($N = 2340$). To determine a short list of genes with which to follow up on chi-square tests were applied to the 95,940 gene pairs in the missense group and the 603,720 gene pairs in the tumor suppressor group. Three new groups were formed for further analysis: the first consisted of the significant gene/mutation pairs from the oncogenic group, the second consisted of the significant gene/mutation pairs from the tumor suppressor group, and the third was a combination of the significant pairs from both groups with no discrimination on the type of mutations considered.

These groups were further analyzed for differential essentiality via the Mann-Whitney method described above and Cohen's *D* effect size was calculated to measure the extent of the phenotype.

Differential expression and essentiality of male and female KRAS-driven cancers

We used CanDI to gather all cell lines that are present in both DepMap (2021Q2) and CCLE (Release 2021Q2). CanDI was then leveraged to put these cell lines into the following tissue groups: KRAS-mutant colon/colorectal (CRC), pancreatic ductal adenocarcinoma (PDAC), and NSCLC. Each tissue group was then split into male and female subgroups. We chose to analyze KRAS-mutant colon/colorectal, PDAC, and NSCLC cell lines as there are a relatively large number of KRAS-mutant male and female cell lines representative of these types of cancer present in DepMap giving us increased statistical power for subsequent analysis. Differential expression was analyzed by applying the methods described above to raw RNA-seq counts data from CCLE (Release: 2021Q2). Genes with adjusted *p* values less than 0.01 were considered significantly differentially expressed. Differential essentiality was analyzed using the methods described above on the previously described sex subgroups for each tissue type. Genes with *p* values less than 0.05 were considered significantly differentially essential between male and female cell models. For each tissue type, the distributions of the top 7 significantly differentially essential genes were highlighted in comparison with the bottom 3 as a negative control (Fig. 3).

Differential expression of benign and malignant cancer cell lines

We downloaded human bronchial epithelial (HBE) RNA-seq data from Gillen et al. via the European Nucleotide Archive to use as a benign lung tissue model [27]. This data set contains gene expression data for primary HBE cells cultured from three different donors and also normal human bronchial epithelial (NHBE) cells (a mixture of HBE and human tracheal epithelial cells purchased from Lonza Bioscience (CC-2541)). We then used CanDI to put NSCLC models into three different groups: KRAS-mutant, EGFR-mutant, and all cell lines (Fig. 4). For our benign model, raw counts were quantified via kallisto [28]. Raw counts for our malignant cell lines were queried via CanDI. DESeq2 was then applied to evaluate the differential expression between our normal lung tissue model and our three malignant lung tissue groups. The results from DESeq2 were then filtered by significance (adjusted *p* value < 0.01). To filter based on potential immunotherapy targets, we removed all genes not annotated as being localized to the plasma membrane, and genes with localization confidence scores lower than 6. Genes that were obviously misannotated as surface proteins were also manually removed.

Results

CanDI is a global cancer data integrator.

We set out to integrate multiple types of data by creating programmatic and biologically relevant abstractions that allow for flexible cross-referencing across all datasets [20]. Data from the Cancer Cell Line Encyclopedia (CCLE) for RNA expression, DNA mutation, DNA copy number, and chromosome fusions across more than 1000 cancer cells lines was integrated into our database with the functional genomics data from the Cancer Dependency Map (DepMap) (Fig. 1a, b and Additional file 1: Fig. S1) [1, 12, 21]. We also integrated protein-protein interaction data from the CORUM database along with three additional distinct protein localization databases [4, 7, 11, 22]. CanDI by default will access the most

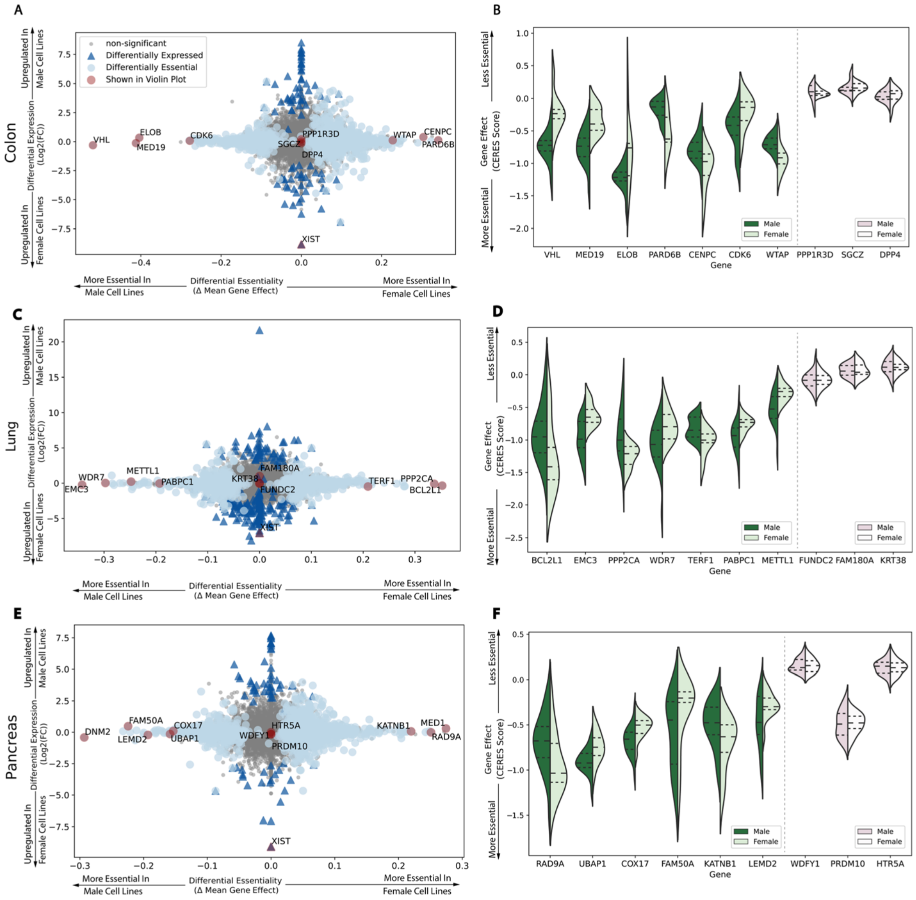

Fig. 3 CanDI reveals female and male context-specific essential genes. **A** Differential gene expression and differential gene essentiality in male and female CRC cell lines. For differential expression analysis: *N* = 24 male cell lines and *N* = 14 female cell lines. For differential essentiality analysis: and *N* = 18 male cell lines and *N* = 9 female cell lines. **B** The distribution of gene effect CERES scores in male and female CRC cell lines. The top seven and bottom three differentially essential genes are shown in violin plots split by the sex of the cell lines. **C** Differential gene expression and differential gene essentiality in male and female NSCLC cell lines. For differential expression analysis: *N* = 30 male cell lines and *N* = 13 female cell lines. For differential essentiality analysis and *N* = 23 male cell lines and *N* = 9 female cell lines. **D** The distribution of gene effect CERES scores in male and female NSCLC cell lines. The top seven and bottom three differentially essential genes are shown in violin plots split by the sex of the cell lines. **E** Differential gene expression and differential gene essentiality in male and female PDAC cancer cell lines. For differential expression analysis: *N* = 26 male cell lines and *N* = 19 female cell lines. For differential essentiality analysis: *N* = 24 male cell lines and *N* = 12 female cell lines. **F** The distribution of gene essentiality CERES scores in male and female PDAC cell lines. The top seven and bottom three differentially essential genes are shown in violin plots split by the sex of the cell lines

recent release of data from DepMap, although users can also specify both the release and data type that is accessed [20]. The key advantage to this approach is that CanDI enables one to easily input user-defined queries with multi-tiered conditional logic into this large integrated dataset to analyze gene function, gene expression, protein localization, and protein-protein interactions [20].

CanDI identifies genes that are conditionally essential in BRCA-mutant ovarian cancer

The concept that loss-of-function tumor suppressor gene mutations can render cancer cells critically reliant on the function of a second gene is known as synthetic lethality. Despite the promise of synthetic lethality, it has been challenging to predict or identify genes that are synthetically lethal with commonly mutated tumor suppressor genes. While there are many underlying reasons for this challenge, we reasoned that data integration through CanDI could identify synthetic lethal interactions missed by others.

A paradigmatic example of synthetic lethality emerged from the study of DNA damage repair (DDR) [29]. Somatic mutations in the DNA double-strand break (DSB) repair genes, *BRCA1*, create an increased dependence on DNA single-strand break (SSB) repair. This dependence can be exploited through small molecule inhibition of PARP1-mediated SSB repair. Inhibition of PARP

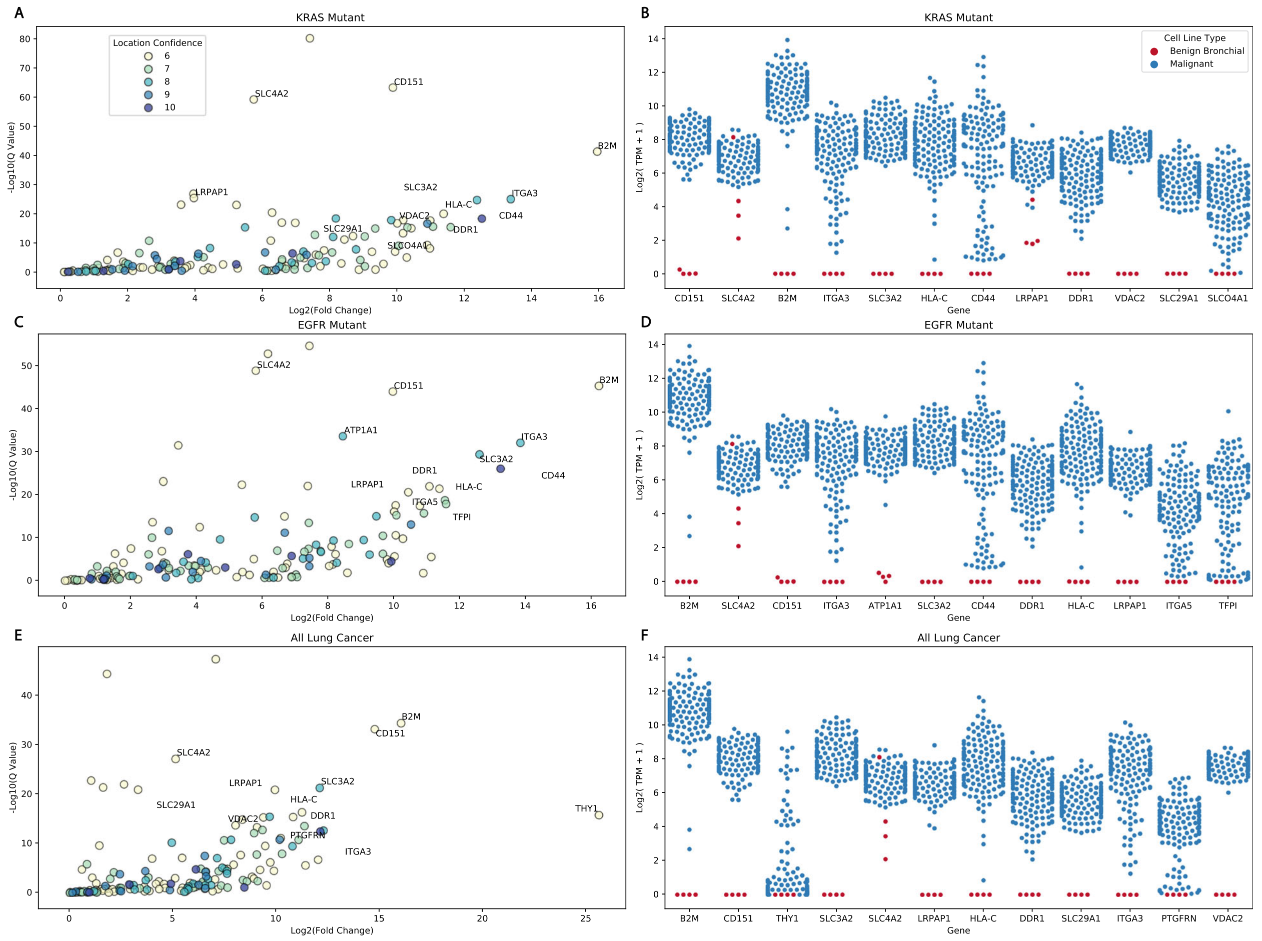

Fig. 4 CanDI enables rapid integration of external datasets to reveal immunotherapy targets. **A** A graph showing genes that are upregulated in KRAS-mutant NSCLC cell lines relative to primary human bronchial epithelial cells. A cell membrane protein localization score is shown for each gene. Higher protein localization scores indicate higher confidence annotations. **B** A scatter plot showing gene expression for genes that encode cell surface proteins in KRAS-mutant NSCLC cell lines and primary human bronchial epithelial cells. *N* = 46 for KRAS-mutant NSCLC cell lines and *N* = 4 for primary human bronchial epithelial cells. **C** A graph showing genes that are upregulated in EGFR-mutant NSCLC cell lines relative to primary human bronchial epithelial cells. A cell membrane protein localization score is shown for each gene. Higher protein localization scores indicate higher confidence annotations. **D** A scatter plot showing gene expression for genes that encode cell surface proteins in EGFR-mutant NSCLC cell lines and primary human bronchial epithelial cells. *N* = 21 for EGFR-mutant NSCLC cell lines and *N* = 4 for primary human bronchial epithelial cells. **E** A graph showing genes that are upregulated in NSCLC cell lines relative to primary human bronchial epithelial cells. A cell membrane protein localization score is shown for each gene. Higher protein localization scores indicate higher confidence annotations. **F** A scatter plot showing gene expression for genes that encode cell surface proteins in NSCLC cell lines and primary human bronchial epithelial cells. *N* = 141 for NSCLC cell lines and *N* = 4 for primary human bronchial epithelial cells

provides significant clinical responses in advanced breast, prostate and ovarian cancer patients with DDR mutations but they ultimately progress [29]. Thus, new synthetic lethal associations with *BRCA1* are a potential path towards therapeutic development PARP refractory patients.

To illustrate the flexibility of CanDI to mine context-specific synthetic sick lethal (SSL) genetic relationships, we hypothesized that the genes that modulate response to a PARP1 inhibitor might be enriched for selectively essential proliferation or survival of BRCA1-mutant cancer cells. To test this hypothesis, we integrated the results of an existing CRISPR screen that identified genes that modulate response to the PARP inhibitor olaparib [25]. We then tested whether any of these genes are differentially essential for cell proliferation or survival in ovarian cancer and in breast cancer cell models that are either BRCA1 proficient or deficient (Fig. 1c, d). This query revealed that the Fanconi Anemia pathway is selectively essential in BRCA1-mutated ovarian cancer models but not in BRCA1-wild type ovarian cancer, BRCA1-mutated breast cancer, or BRCA1-wildtype breast cancer models (Fig. 1e, Additional file 2: Table S1, and Additional file 3: Table S2). To our knowledge, a

SSL phenotype between FANCM and BRCA1 has not been previously reported in human cancer models although our hypothesis is supported by a recent paper characterizing a SSL phenotype between FANCM and BRCA1 in mouse embryonic stem cells [30]. A second recent paper has nominated a role for FANCM and BRCA1 in telomere maintenance [31]. Importantly, FANCM is a helicase/translocase and thus considered to be a druggable target for cancer therapy [32]. Clinical genomics data support this SSL hypothesis, although this remains to be tested in ovarian cancer patient samples [33]. Because the DepMap currently only allows single genes to be queried and does not enable users to easily stratify cell lines by mutation such analysis would normally take a user several days to complete manually. Our approach enabled this analysis to be completed using a desktop computer in less than 2 hours, which includes the visualization of data presented here (Fig. 1e).

Conditional genetic essentiality in KRAS- and EGFR-mutant NSCLC cells

Beyond tumor suppressor genes (TSGs), many common driver oncogenes such as KRAS^{G12D} are currently undruggable, which motivates the search for oncogene-specific conditional genetic dependencies. We reasoned that CanDI enables us to rapidly search functional genomics data for genes that are conditionally essential in lung cancer cells driven by KRAS and EGFR mutations. We stratified non-small cell lung cancer cell (NSCLC) models by EGFR and KRAS mutations and then looked at the average gene essentiality for all genes within each of these 4 subtypes of NSCLC. We chose to analyze KRAS- and EGFR-mutant cell lines as these driver oncogene mutations are well characterized and amongst the most common mutations present in DepMap cell lines giving us increased statistical power for subsequent analysis. We observed that KRAS is conditionally self-essential in KRAS-mutant cell models but that no other genes are conditionally essential in KRAS-mutant, EGFR-mutant, KRAS-wildtype, or EGFR-wildtype cell models (Fig. 2a, b, Additional file 1: Fig. S2,S3 and Additional file 4: Table S3). This finding demonstrates that very few—if any—genes are synthetic lethal with KRAS- or EGFR- in KRAS- and EGFR-mutant lung cancer cell lines. It may be that these experiments are underpowered or it may be that when the genetic dependencies of diverse cell lines representing a disease subtype are averaged across a single variable (e.g., a KRAS mutation) very few common synthetic lethal phenotypes are observed [34]. CanDI provides potential solutions for both of these hypotheses.

CanDI enables a global analysis of conditional essentiality in cancer.

It is thought that data aggregation across vast landscapes of unknown covariates does not necessarily increase the statistical power to identify rare associations [34]. Thus, the global analyses of aggregated cancer data sometimes lie in systematically sub-setting data based on key covariates post aggregation. This has been observed in driver gene identification [35]. Inspired by our analysis of TSG and oncogene conditionally essentiality above, we next used CanDI to identify genes that are conditionally essential in the context of several hundred cancer driver mutations. We first grouped driver mutations (e.g., nonsense or missense) for each driver gene. For this analysis, we selected several thousand genes that are in the 85–90th percentile of essentiality within the DepMap data and therefore conditionally essential, meaning these genes are required for cell growth or survival in a subset of cell lines. Importantly, it is not known why these several thousand genes are conditionally essential. We then tested whether each of these conditionally essential genes has a significant association with individual driver mutations. Our analytic approach does not weight the number of cell models representing each driver mutation nor does this give information on phenotype effect sizes. Our analysis nominates a large number of conditionally dependent genetic relationships with both TSG and oncogenes (Fig. 2c, d and Additional file 5: Table S4). A number of the conditional genetic dependencies identified in our independent variable analysis above are represented by a limited number of cell models and so further investigation is needed to systematically validate these conditional dependencies, but this data further suggests that averaging genetic dependencies across diverse cell lines with un-modeled covariates obscures conditional SSL relationships. For example, this analysis does not weight biological variables such as other mutations, chromosomal amplifications or deletions, aneuploidy, cancer subtype, biological sex, and more. It is currently impossible to account for all variables with the relatively limited number of human cancer cell lines that exist; however, it is predicted that if sufficient biological models representative of each variable were analyzed then a predictable pattern of SSL relationships would emerge.

To further investigate this hypothesis, we analyzed these same conditional genetic relationships with a second analytic approach that weights the number of cell models representing each driver mutation. We observed a limited number of conditional genetic dependencies that largely consists of oncogene self-essential dependencies as previously highlighted for KRAS-mutant cell lines (Fig. 2e-g and Additional file 6: Table S5) [13, 36]. Our analysis of mutant oncogene self-essentiality suggests that mutant oncogenes that drove tumor initiation or progression generally continue to support the proliferation and survival of human cancer cell lines in vitro. Thus, analysis that averages each conditional phenotype across diverse panels of cell lines with unknown covariates masks interesting conditional genetic dependencies.

CanDI reveals female and male context-specific essential genes in colon, lung, and pancreatic cancer

Cancer functional genomics data is often analyzed without consideration for fundamental biological properties such as the sex of the tumor from which each cell line is derived. It is well established that biological sex influences cancer predisposition, cancer progression, and response to therapy [37]. We hypothesized that individual genes may be differentially essential across male and female cell lines. This hypothesis to our knowledge has never been tested in an unbiased large-scale manner. To maximize our statistical power to identify such differences, we chose to test this hypothesis in a disease setting with large number of relatively homogenous cell lines and fewer unknown covariates. Using CanDI, we stratified all KRAS-mutant NSCLC, PDAC, and CRC cell lines by sex and then tested for conditional gene essentiality. This analysis identified a number of genes that are differentially essential in male or female KRAS-mutant NSCLC, PDAC, and CRC models (Fig. 3a-f and Additional file 7: Table S6, Additional file 8: Table S7, Additional file 9: Table S8, Additional file 10: Table S9, Additional file 11: Table S10, Additional file 12: Table S11). The genes that we identify are not common across all three disease types suggesting as one might expect that the biology of the tumor in part also determines gene essentiality. To test whether any association between differentially essential genes could be identified from expression data (e.g., essential genes encoded on the Y chromosome), we first used CanDI to identify genes that are differentially expressed between male and female cell lines within each disease [24]. We then plotted the set of differentially essential genes against the differentially expressed genes in KRAS-mutant NSCLC, PDAC, and CRC models (Fig. 3a, c, e and Additional file 7: Table S6, Additional file 8: Table S7, Additional file 9: Table S8, Additional file 10: Table S9, Additional file 11: Table S10, Additional file 12: Table S11) and found little overlap between these gene lists. Notably, we observed VHL and ELOB, which are thought to form a protein complex, are top hits that are more essential in KRAS-mutant male colon cancer cells [38, 39]. Our analysis demonstrates that stratifying groups of heterogeneous cancer models by three variables, in this case tumor type, KRAS mutation status and sex, reveals differentially essential genes. CanDI enables biologically principled stratification of data in the CCLE and DepMap by any feature associated with a group of cell models. This stratification allows us to identify genes associated with sex, which is not possible with other covariates included.

CanDI enables rapid integration of external datasets to reveal immunotherapy targets.

An emerging challenge in the cancer biology is how to robustly integrate larger "resource" datasets like CCLE with the vast amount of published data from individual laboratories. For example, a big challenge in antibody discovery is identifying specific surface markers on cancer cells. To approach these big questions, we utilized CanDIs ability to rapidly take new datasets, such as raw RNA-seq counts data in a disparate study of interest, then normalize and integrate this data into the CCLE, DepMap, and protein localization databases previously described. Specifically, we rapidly integrated an RNA-seq expression dataset that measured the set of transcribed genes in primary lung bronchial epithelial cells from 4 donors [27]. Classes within CanDI enable rapid application of DESeq2 to assess the differential expression between outside datasets and the CCLE. We used this feature to identify genes that are differentially expressed between primary lung bronchial epithelial cells and KRAS-mutant NSCLC, EGFR-mutant NSCLC, or all NSCLC models in CCLE. We then used CanDI to identify genes that are upregulated in cancer cells over normal lung bronchial epithelial cells with protein products that are localized to the cell membrane. This analysis of KRAS-mutant, EGFR-mutant, and pan-NSCLC generated highly similar lists of differentially expressed surface proteins (Fig. 4a-f and Additional file 13: Table S12). Notably, overexpression of several of these genes, such as *CD151* and *CD44*, has been observed in lung cancer and is associated with poor prognosis [40–42]. These proteins represent potential new immunotherapy targets in KRAS-driven NSCLC.

Discussion

Data integration is a critical requirement in biology research in the era of genomics and functional genomics. Large-scale efforts such as the CCLE have revealed genomic features of more than 1000 cell line models. This data has not to our knowledge previously been integrated with functional genomics data in a manner that individual users can enter batched queries that are stratified by disease subtype or mutation status. This is not just a small improvement in functionality, but rather it is an enabling format that makes possible the types of conditional genomics analyses that drive discovery. Moreover, it fills a fundamental gap in the cancer research community that integrates large-scale projects with investigator-initiated studies

Our data framework enables biologists without specialized expertise in bioinformatics to use the full spectrum of data in the CCLE and DepMap in a higher throughput and precise manner. Using CanDI, we identified genes that are selectively essential in male versus female KRAS-mutant NSCLC, PDAC, and CRC models. To our knowledge, such analysis has never been performed to begin to query the biologic basis of sex disparity in cancer or cancer therapy. We illustrate another feature of our framework by analyzing a list of hit genes nominated by a bespoke CRISPR drug screen for gene essentiality in BRCA1-wild type and BRCA1-mutated breast and ovarian cancer. In a third application, we analyzed the principle of synthetic lethality for 17,427 genes in 33 KRAS-mutant and 17 EGFR-mutant NSCLC models. We then used CanDI to globally identify genes that are conditionally essential in the context of common cancer driver mutations. Finally, we nominated 12 potential new immunotherapy targets in KRAS-mutant, EGFR-mutant, and pan-NSCLC models by using CanDI to identify genes that are differentially expressed in normal bronchial epithelial cells versus NSCLC models that are localized at the plasma membrane.

Conclusions

Our use of CanDI reveals a wealth of new hypotheses that can be rapidly generated from private and publicly available cancer data. By sharing data flows and use cases with a CanDI community, we illustrate the ways in which individual research groups can interact with massive cancer genomics projects without reinventing tools or relying upon DepMap tool releases. We anticipate that CanDI will be widely used in cell biology, immunology, and cancer research.

Availability and requirements

Project Name: CanDI

Project home page: https://github.com/GilbertLabUCSF/CanDI

Operating System: Platform independent

Programming language: Python

License: CC Attribution 4.0 for DepMap data.

Any restrictions for non-academics: No

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-021-00987-8.

Additional file 1: Figures S1-3. **Figure S1.** An Object-oriented schema diagram showing core structure of CanDI software v0.2-alpha. Gene, CellLine, Oraganelle, Cancer, and CellLine classes all inherit from the Entity Class. **Figure S2.** A scatter plot comparing gene essentiality (Average CERES scores) of NSCLC cell lines between KRAS mutant and KRAS wild type Cell lines with EGFR mutant cell lines removed from consideration. **Figure S3.** A scatter plot comparing gene essentiality (Average CERES scores) of NSCLC cell lines between EGFR mutant and EGFR wild type Cell lines with KRAS mutant cell lines removed from consideration.

Additional file 2: Table S1. A table containing raw CERES scores displayed in the ovarian cell line heat map of Fig. 1e.

Additional file 3: Table S2. A table containing raw CERES scores displayed in the breast cell line heat map of Fig. 1e.

Additional file 4: Table S3. A table containing mean CERES scores for each series displayed in Fig. 2a, b.

Additional file 5: Table S4. A table containing the data for all chi [2] tests performed to generate Fig. 2c, d.

Additional file 6: Table S5. A table containing the data for scatter plots shown in Fig. 2e, f, g.

Additional file 7: Table S6. A table containing the data from the differential expression analysis of male and female CRC cell lines shown in Fig. 3a.

Additional file 8: Table S7. A table containing the data from the differential essentiality analysis of male and female CRC cell lines Fig. 3a.

Additional file 9: Table S8. A table containing the data from the differential expression analysis of male and female NSCLC cell lines Fig. 3c.

Additional file 10: Table S9. A table containing the data from the differential essentiality analysis of male and female NSCLC cell lines Fig. 3c.

Additional file 11: Table S10. A table containing the data from the differential expression analysis of male and female PDAC cell lines Fig. 3e.

Additional file 12: Table S11. A table containing the data from the differential essentiality analysis of male and female PDAC cell lines Fig. 3e.

Additional file 13: Table S12. A table containing the differential expression analysis data merged with the location data for all three tissues shown in Fig. 4.

Acknowledgements

We thank everyone in the Gilbert lab for helpful comments and discussion. All DepMap data made available through CanDI was generated by the Broad Institute DepMap for research purposes and not intended for clinical use. This data is distributed under the Creative Commons license (Attribution 4.0 International (CC BY 4.0); https://creativecommons.org/licenses/by/4.0/).

Authors' contributions

C.Y. and A.A. wrote all code. A.A. and C.Y. created figures. C.Y, A.A., J.R.P, H.G., and L.A.G created the project, analyzed the data, and wrote the manuscript. All authors read and approved the final manuscript.

Funding

LAG is supported by K99/R00 CA204602 and DP2 CA239597, a Pew-Stewart Scholars for Cancer Research award as well as the Goldberg-Benioff Endowed Professorship in Prostate Cancer Translational Biology.

Availability of data and materials

All data and code is publicly available at https://github.com/GilbertLabUCSF/CanDI [20] and as described in the "Methods".

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

LAG has filed patents on CRISPR functional genomics and is a co-founder of Chroma Medicine. The remaining authors declare that they have no competing interests.

Author details

[1]Department of Urology, University of California, San Francisco, CA, USA. [2]Helen Diller Family Comprehensive Cancer Center, San Francisco, San Francisco, CA, USA. [3]Present Address: University of North Carolina Chapel Hill School of Medicine, Chapel Hill, NC, USA. [4]Department of Biochemistry and Biophysics, University of California, San Francisco, CA, USA. [5]Department of Biomedical Engineering, Pennsylvania State University, University Park, PA, USA. [6]Department of Cellular & Molecular Pharmacology, University of California, San Francisco, CA, USA.

Received: 12 February 2021 Accepted: 5 October 2021
Published online: 18 October 2021

References

1. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER III, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569(7757):503–8. https://doi.org/10.1038/s41586-019-1186-3.
2. Li H, Ning S, Ghandi M, Kryukov GV, Gopal S, Deik A, et al. The landscape of cancer cell line metabolism. Nat Med. 2019;25(5):850–60. https://doi.org/10.1038/s41591-019-0404-8.
3. Tsherniak A, et al. Defining a Cancer Dependency Map. Cell. 2017;170:564–576.e16.
4. Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, et al. A subcellular map of the human proteome. Science. 2017;356(6340). https://doi.org/10.1126/science.aal3321.
5. Cancer Cell Line Encyclopedia Consortium & Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. Nature. 2015;528:84–7.
6. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603–7. https://doi.org/10.1038/nature11003.
7. Bausch-Fluck D, Goldmann U, Müller S, van Oostrum M, Müller M, Schubert OT, et al. The in silico human surfaceome. PNAS. 2018;115(46):E10988–97. https://doi.org/10.1073/pnas.1808790115.
8. Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, et al. CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res. 2019;47(D1):D559–63. https://doi.org/10.1093/nar/gky973.
9. Nusinow DP, et al. Quantitative Proteomics of the Cancer Cell Line Encyclopedia. Cell. 2020;180:387–402.e16.
10. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017;45(D1):D362–8. https://doi.org/10.1093/nar/gkw937.
11. Itzhak DN, Tyanova S, Cox J, Borner GH. Global, quantitative and dynamic mapping of protein subcellular localization. Elife. 2016;5. https://doi.org/10.7554/eLife.16950.
12. Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, et al. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat Genet. 2017;49(12):1779–84. https://doi.org/10.1038/ng.3984.
13. Behan FM, Iorio F, Picco G, Gonçalves E, Beaver CM, Migliardi G, et al. Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. Nature. 2019;568(7753):511–6. https://doi.org/10.1038/s41586-019-1103-9.
14. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, et al. Identification and characterization of essential genes in the human genome. Science. 2015;350(6264):1096–101. https://doi.org/10.1126/science.aac7041.
15. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, et al. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. Cell. 2015;163(6):1515–26. https://doi.org/10.1016/j.cell.2015.11.015.
16. Wang T, et al. Gene essentiality profiling reveals gene networks and synthetic lethal interactions with oncogenic Ras. Cell. 2017;168:890–903.e15.
17. Chan EM, Shibue T, McFarland JM, Gaeta B, Ghandi M, Dumont N, et al. WRN helicase is a synthetic lethal target in microsatellite unstable cancers. Nature. 2019;568(7753):551–6. https://doi.org/10.1038/s41586-019-1102-x.
18. Adamson B, et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. Cell. 2016; 167:1867–1882.e21.
19. Wainberg M, Kamber RA, Balsubramani A, Meyers RM, Sinnott-Armstrong N, Hornburg D, et al. A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. Nat Genet. 2021;53(5):638–49. https://doi.org/10.1038/s41588-021-00840-z.
20. Yogodzinski C, Arab A, Pritchard JR, Goodarzi H, Gilbert LA. A global Cancer Data Integrator reveals principles of synthetic lethality, sex disparity and immunotherapy. Github https://github.com/GilbertLabUCSF/CanDI. 2021.
21. Lenoir WF, Lim TL, Hart T. PICKLES: the database of pooled in-vitro CRISPR knockout library essentiality screens. Nucleic Acids Res. 2018;46(D1):D776–80. https://doi.org/10.1093/nar/gkx993.
22. Bausch-Fluck D, Hofmann A, Bock T, Frei AP, Cerciello F, Jacobs A, et al. A mass spectrometric-derived cell surface protein atlas. PLoS One. 2015;10(4): e0121314. https://doi.org/10.1371/journal.pone.0121314.
23. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016;34(2):184–91. https://doi.org/10.1038/nbt.3437.
24. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8.
25. Zimmermann M, Murina O, Reijns MAM, Agathanggelou A, Challis R, Tarnauskaitė Ž, et al. CRISPR screens identify genomic ribonucleotides as a source of PARP-trapping lesions. Nature. 2018;559(7713):285–9. https://doi.org/10.1038/s41586-018-0291-z.
26. Bailey MH, et al. Comprehensive characterization of cancer driver genes and mutations. Cell. 2018;173:371–385.e18.
27. Gillen AE, Yang R, Cotton CU, Perez A, Randell SH, Leir SH, et al. Molecular characterization of gene regulatory networks in primary human tracheal and bronchial epithelial cells. J Cyst Fibros. 2018;17(4):444–53. https://doi.org/10.1016/j.jcf.2018.01.009.
28. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525–7. https://doi.org/10.1038/nbt.3519.
29. O'Connor MJ. Targeting the DNA damage response in cancer. Mol Cell. 2015;60(4):547–60. https://doi.org/10.1016/j.molcel.2015.10.040.
30. Panday A, et al. FANCM regulates repair pathway choice at stalled replication forks. Mol Cell. 2021;81:2428–2444.e6.
31. Pan X, Drosopoulos WC, Sethi L, Madireddy A, Schildkraut CL, Zhang D. FANCM, BRCA1, and BLM cooperatively resolve the replication stress at the ALT telomeres. PNAS. 2017;114(29):E5940–9. https://doi.org/10.1073/pnas.1708065114.
32. Lou K, Gilbert LA, Shokat KM. A bounty of new challenging targets in oncology for chemical discovery. Biochemistry. 2019;58(31):3328–30. https://doi.org/10.1021/acs.biochem.9b00570.
33. Narayan G, Arias-Pulido H, Nandula SV, Basso K, Sugirtharaj DD, Vargas H, et al. Promoter hypermethylation of FANCF: disruption of Fanconi anemia-BRCA pathway in cervical cancer. Cancer Res. 2004;64(9):2994–7. https://doi.org/10.1158/0008-5472.CAN-04-0245.
34. Ideker T, Dutkowski J, Hood L. Boosting signal-to-noise in complex biology: prior knowledge is power. Cell. 2011;144(6):860–3. https://doi.org/10.1016/j.cell.2011.03.007.
35. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol. 2016;34(2):155–63. https://doi.org/10.1038/nbt.3391.
36. Lou K, Steri V, Ge AY, Hwang YC, Yogodzinski CH, Shkedi AR, et al. KRASG12C inhibition produces a driver-limited state revealing collateral dependencies. Sci Signal. 2019;12(583):eaaw9450. https://doi.org/10.1126/scisignal.aaw9450.
37. Cancer Disparities - National Cancer Institute. https://www.cancer.gov/about-cancer/understanding/disparities (2016).
38. Duan DR, et al. Inhibition of transcription elongation by the VHL tumor suppressor protein. Science. 1995;269(5229):1402–6. https://doi.org/10.1126/science.7660122.
39. Kibel A, Iliopoulos O, DeCaprio JA, Kaelin WG. Binding of the von Hippel-Lindau tumor suppressor protein to Elongin B and C. Science. 1995; 269(5229):1444–6. https://doi.org/10.1126/science.7660130.
40. Mj K, et al. Prognostic Significance of CD151 Overexpression in non-small cell lung cancer. Lung Cancer (Amsterdam, Netherlands). 2013;81 https://pubmed.ncbi.nlm.nih.gov/23570797/.

41. Ko YH, Won HS, Jeon EK, Hong SH, Roh SY, Hong YS, et al. Prognostic significance of CD44s expression in resected non-small cell lung cancer. BMC Cancer. 2011;11(1):340. https://doi.org/10.1186/1471-2407-11-340.
42. Penno MB, August JT, Baylin SB, Mabry M, Linnoila RI, Lee VS, et al. Expression of CD44 in human lung tumors. Cancer Res. 1994;54(5):1381–7.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.