Bioinformatics, 2026, **42**(6), btag337
https://doi.org/10.1093/bioinformatics/btag337
Published: 27 May 2026
Original Paper | Data and text mining

OXFORD

Using cancer profiles to identify synthetic lethal therapeutic targets and predictive biomarkers in cancer gene dependency data

Laurence H. Pearl[1,2,*,] and Frances M.G. Pearl[3,*,]

[1]Genome Damage and Stability Centre, School of Life Sciences, University of Sussex, Falmer, Brighton, BN1 9RQ, United Kingdom
[2]Division of Structural Biology, Institute of Cancer Research, Chester Beatty Laboratories, London, SW3 6JB, United Kingdom
[3]Bioinformatics Lab, School of Life Sciences, University of Sussex, Falmer, Brighton, BN1 9RQ, United Kingdom
*Corresponding authors. Laurence H. Pearl, Genome Damage and Stability Centre, School of Life Sciences, University of Sussex, Falmer, Brighton, BN1 9RQ, United Kingdom; Division of Structural Biology, Institute of Cancer Research, Chester Beatty Laboratories, 237 Fulham Road, London, SW3 6JB, United Kingdom. E-mail: laurence.pearl@sussex.ac.uk; Frances M. G. Pearl, Bioinformatics Lab, School of Life Sciences, University of Sussex, Falmer, Brighton, BN1 9RQ, United Kingdom, Email: f.pearl@sussex.ac.uk
Associate Editor: Daisuke Kihara

Abstract

Motivation: **Large scale loss-of-function screens utilising CRISPR or siRNA can provide profound insights into the importance of individual genes for the survival of a cancer cell and can drive the identification of therapeutic targets and biomarkers, and the development of targeted drugs. However, the analysis of these data and the substantial bodies of metadata that relate to them, is technically challenging and typically requires substantial expertise in data science and computer coding.**

Results: **To facilitate the analysis of cancer gene dependency data by cancer biologists and clinical scientists, we have developed DepMine–a computational toolkit providing a powerful system for framing complex queries relating cancer gene dependency to the underlying genetic changes that occur in cancer cells. DepMine identifies synthetic lethal relationships between putative target genes and complex 'cancer profiles' built from user-specified combinations of mutations, copy-number variation, and expression levels, and can refine these to optimal biomarker definitions for target dependency.**

Availability: **The Python implementation of *DepMine* and associated data files can be obtained at https://github.com/UOSbioinformaticslab/depmine and is free to academics and Not-For-Profit organisations. The *DepMine* release referenced in this paper is archived as DOI: 10.5281/zenodo.19570601**

1 Introduction

It is estimated that between 1,000 to 4,000 genes are essential for the survival of a human cell, depending on the method used to define essentiality (Rancati *et al.* 2018, Larrimore and Rancati, 2019, Funk *et al.* 2022). In cancers, as tumours develop and acquire genetic and epigenetic alterations, the tumour cell rewires its regulatory pathways and different sets of genes become essential. Examples of acquired essential genes have been identified by large-scale loss-of-function (LoF) genetic screens (e.g. DRIVE (McDonald *et al.* 2017) and DepMap (Tsherniak *et al.* 2017)).

The DepMap dataset, in which the mutational status and expression profiles of >19,000 genes in >1000 different cell lines has been measured, is the largest amalgamation of data for this. Although not universally true, the dependency of a tumour cell on a gene as determined by CRISPR screens, generally mirrors the susceptibility of that cell to small molecule drugs targeted against the product of that gene (Goncalves *et al.* 2020). Consequently, systematic analysis of the data from such screens is likely to be of great value in development of novel precision therapies and biomarkers.

As cancers evolve, genetic changes are positively selected that directly impact the function of specific genes, with important consequences for the biology of the mutated tumour cell. Typically, these consist of loss-of-function (LoF) mutations in tumour suppressors or gain-of-function (GoF) mutations in proto-oncogenes. LoF of tumour suppressors can be targeted by synthetic sickness/lethality (SSL), a negative genetic interaction whereby a cell can survive the loss of protein product from the tumour suppressor or from its synthetic lethal partner gene, but not from both (Lord and Ashworth, 2017, O'Neil *et al.* 2017). The best-known example is treatment of BRCA1/2 deficient tumours in breast, ovarian and other tumours with PARP inhibitors

Received: 1 December 2025. **Revised:** 14 April 2026. **Accepted:** 11 May 2026

(Mateo *et al.* 2019). Generally, SSL relates pairs of genes, with one gene being the therapeutic target in cells where the other gene is defective. However, SSL may also relate a target gene to a phenotype or complex genetic state as in the identification of Werner syndrome ATP-dependent helicase (WRN) as a synthetic lethal (SSL) target in tumours displaying microsatellite instability (MSI) (Chan *et al.* 2019, Picco *et al.* 2021); a Werner inhibitor (NCT05838768) is now in Phase 1 clinical development for MSI solid tumours (Ferretti *et al.* 2024).

Oncogene addiction occurs when a tumour cell becomes dependent on the presence of an activated proto-oncogene with a gain-of-function (GoF) mutation (Weinstein and Joe 2008). Oncogenes can often be targeted directly, but the therapeutic response to even highly specific and potent drugs is often very mixed across patient populations. Synthetic dosage lethality (SDL), where activation of a gene (e.g. the oncogene) is synthetically lethal with LoF of another gene, can be utilised to identify biomarkers where oncogene-directed therapy is most likely to be effective (Megchelenbrink *et al.* 2015).

Given the cost and technical challenge posed by mounting 'wet' screens to find SSL relationships, there has been substantial interest in developing analysis pipelines that exploit large cancer cell-line or patient genetic and clinical data collections such as TCGA (Liu *et al.* 2018) and COSMIC (Tate *et al.* 2019)) to identify synthetic lethality relationships, and/or targetable oncogenes and response-predictive biomarkers computationally–for example (Ryan *et al.* 2014, Shao *et al.* 2016, Benstead-Hume *et al.* 2019, Das *et al.* 2019, Lee *et al.* 2021, Srivatsa *et al.* 2022, Wang *et al.* 2022, Markowska *et al.* 2023, Pacini *et al.* 2024, Wooller *et al.* 2024). To date these methods generally focus on identifying synthetically lethal gene pairs.

Here we describe *DepMine*, a computational toolkit for the analysis of cancer gene dependency data to identify synthetically lethal relationships and the genetic backgrounds in which they might best be exploited. *DepMine* goes substantially beyond pair-wise SSL relationships by allowing specification of complex 'cancer profiles' which can be used to probe cancer gene dependency data for synthetic lethality with a target gene of interest within cell lines that match the profile. As well as individual gene descriptors, (i.e. GoF or LoF mutations), individual gene or regional copy number variation (CNV), gene expression levels, and cellular properties (e.g. tissue type, mutational signatures, age, sex), can all be combined in the profile definition using a simple set-theory based algebraic formalism. Profiles can be automatically refined to optimise the strength and breadth of the predicted SSL effect with a target gene, thereby defining the genetic backgrounds in which inhibitors or degraders are most likely to be effective. *DepMine* profiles can then be used as biomarkers to select cancer cell lines or patient samples, in which precision therapies of known drug targets, or specific inhibitors of novel drug targets, are most likely to be of therapeutic benefit.

2 Methods

2.1 Data sources

DepMine utilises publicly available datasets centred on DepMap 2023Q2 that can be independently downloaded by the user. The source data files are given in the Supplementary Data, available as supplementary material at *Bioinformatics* online. Expression thresholds are calculated with an ancillary program *expr_scorer.py*, and mutational signatures for the cell lines documented in DepMap are calculated using *SigProfilerAssignment* https://osf.io/mz79v/wiki/home/ (Díaz-Gay *et al.* 2023) with context_type = 96 and genome_build=GRCh38. The input .vcf files required by *SigProfilerAssignment* are generated from the DepMap mutation data by an ancillary program, *MakeVCF.py*.

2.2 Cancer profiles

At the heart of *DepMine* is a set-theory based system in which membership of a set by a cell line is determined by whether or not that cell line matches a user-defined 'cancer profile'. *DepMine* cancer profiles are logical statements constructed from primitive functions that determine the truth or otherwise of postulates about the tissue origin of the cell line, sex and/or age of the patient from which it was derived, and the genetic status or expression level of one or more genes within that cell line. A function returning the Universal Set–i.e. all the available cell lines—is also provided. Primitive functions can be combined by the set operations: union |, intersection **&** and exclusion ^, and expressions can be fully bracketed to achieve cancer profiles of considerable complexity (see **RESULTS**). The definitions of all DepMine cancer profiles referenced in this paper are provided in Table 1 in the Supplementary Data, available as supplementary material at *Bioinformatics* online, and will provide the reader with sufficient examples to define their own.

Operationally, postulate functions are represented by a reserved single character prefix operator followed by its operands. These are summarised in Table 1.

The ***prof*** module provides utilities for showing, renaming and deleting profiles, flattening profiles to primitive functions, and for reading, writing and editing profile collections. Additional modules permit analysis of the tissue distribution of defined profiles—***type***—and the frequency of Cosmic mutational signatures (Díaz-Gay *et al.* 2023, Sondka *et al.* 2024) within that cell line selection—***sig***.

Although developed on the DepMap dataset, the *DepMine* cancer profile system is not restricted to the ≈1100 DepMap data for which dependency data is available, and is readily applicable to any public (e.g. TCGA https://www.cancer.gov/tcga, NCI Genomic Data Commons https://gdc.cancer.gov, Hartwig Medical Database https://www.hartwigmedicalfoundation.nl) or private tumour collection where DNA sequence, copy-number and expression level data are available.

2.3 Defining expression thresholds

Unlike copy number variation (CNV), where the baseline 'dosage' of almost all genes is the same and common thresholds for deletion and amplification can be defined, expression levels vary widely between different genes, so that no simple global definition of relative over- or under-expression can be applied based on absolute expression levels. Instead, *DepMine* assigns over- or under-expression of any gene in a particular cell line by reference to the observed range of expression values of that gene across all available cell lines.

Table 1 DepMine postulate function definitions.

Operator	Operand	Resultant set	Example
+	gene	cells where gene is amplified [1]	**+myc**
+	cytogenetic locus	cells where >=1 gene in locus is amplified [1]	**+17q12.4**
−	gene	cells where gene is deleted [1]	**−tp53**
−	cytogenetic locus	cells where >=1 gene in locus is deleted [1]	**−9p21**
=	gene	cells where gene is diploid [1]	**=snca**
~	gene	Cells where gene is involved in a gene fusion	**~alk**
>	gene	cells where gene is over-expressed [2]	**>erbb2**
<	gene	cells where gene is under-expressed [2]	**<tert**
!	gene	cells where gene has a reading-frame-disruptive loss-of-function mutation	**! brca2**
/	gene: mut	cells where gene has defined missense mutation '?' matches any change	**/braf: v600e** **/kras: g12?**
/	gene: mut, mut[, mut]	cells where gene has one of several defined missense mutations	**/egfr: l834r, t766m**
/	gene+	Cells where gene has an annotated gain-of-function mutation	**/kras+**
/	gene−	Cells where gene has an annotated loss-of-function mutation	**/atm−**
$	gene	cells where gene has no non-synonymous mutations	**$rb1**
{	Cosmic signature[3]	Cells displaying a defined Cosmic mutational signature	**{sbs7a**
{	Cosmic signature aetiology	Cells displaying one or more Cosmic mutational signature attributed to a defined aetiology	**{tobacco**
#		all cells (Universal Set)	
%	tissue	cells derived from tissue	**%breast**
*	sex (Male, Female, Unknown)	cells with sex of patient	***fem**
;	age (Adult, Paediatric, Unknown)	cells with age of patient	**; p**
@	Filename	externally sourced CSV file of cell ACH codes	**@CHAN-MSI**

[1] Definitions of amplification, deletion and diploid status are based on (Mina *et al.* 2020) , adapted to the $\log_2$(CN ratio + 1) formalism used by DepMap/CCLE.

[2] Over- and under-expression are defined per gene, based on prior analysis of DepMap expression level values across the full population of cell lines for which data is available (see below).

[3] As defined in (Díaz-Gay *et al.* 2023).

Analysis of expression ranges across multiple genes reveals a considerable disparity of ranges both in terms of the absolute expression levels they span, and the distribution of expression values across those ranges. We found that 99.6% of the 19,193 genes with DepMap expression data drop into one of three expression range types: a) those in which the gene is expressed in most cells with the expression values (expressed as $\log_2$(transcripts-per-million + 1)) approximating to a normal distribution (∼45%); b) a similar number where the majority of cells do not express the gene and where the values for those that do form an extreme positive skewed distribution (∼43%); and c) a smaller set (∼12%) which are bimodal, being effectively superpositions of the other two distributions (Fig. 1a).

DepMine provides useful working definitions for over- and under-expression by assigning bespoke minimum and maximum threshold values for each gene according to the distribution class it belongs to. For bimodal distributions high-expression and low-expression in a given cell line are determined by whether the expression value is respectively greater-than-or-equal to (>=), or less-than-or-equal to (<=) the expression value at the minimum between the two sub-populations, while for skew distributions, high-expression and low-expression are defined as values >= or <= the overall mean of the distribution. For those genes that display an approximately normal distribution, we define the high- and low-expression thresholds as >= $\mu + 2\sigma$ and <= μ-2σ, respectively.

Selection of cell lines by expression level uses the primitive postulate functions **>gene** and **<gene** which test against the pre-calculated threshold values for that gene, which in turn depends on which profile type the gene matches (Fig. 1a). Where a finer degree of control is required, an on-the-fly threshold can be applied by specifying it between the prefix operator and the gene name, e.g. **>3CCND3** will select the very few cell lines in which the CCND3 (cyclin D3) gene is expressed at

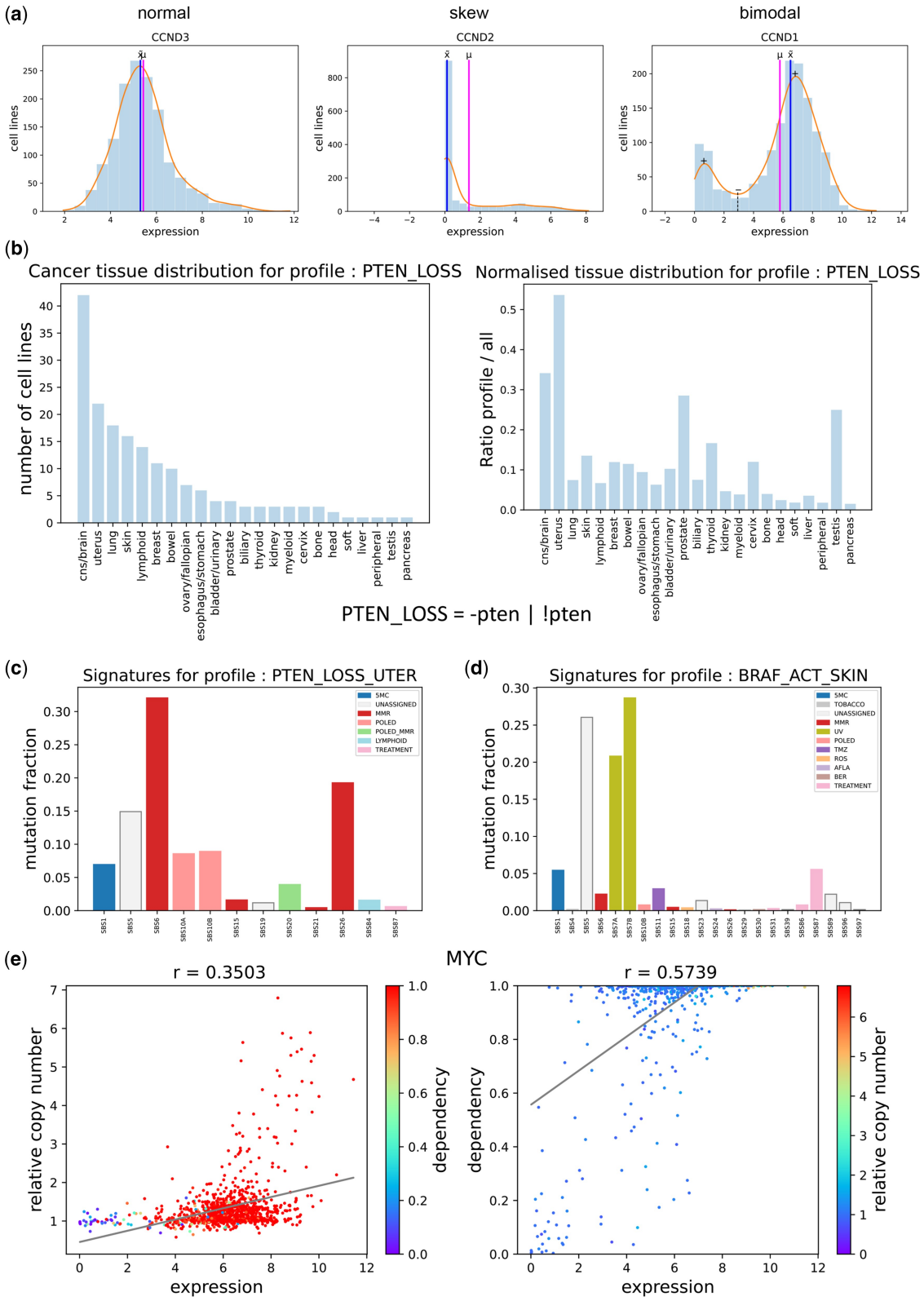

Figure 1 DepMine Tools. a) The three predominant distributions of expression levels - log2(transcripts-per-million+1) – in the DepMap data, exemplified by the three paralogues of cyclin D. CCND3 (left) approximates to a normal distribution where the mean (μ) $\approx$ the median ($\tilde{x}$). Numbered lines indicate 1,2 and 3 standard deviations (either side of the mean; CCND2 (middle) displays a skew distribution where $\mu \neq \tilde{x}$ and the modal value is the lowest bin ($\approx$ no expression); and CCND1 (right) displays a bimodal distribution in which two maxima (+) and a minimum (–) occur. The orange curve shows the probability density function of the three different distributions calculated using kernel density estimation (KDE). Plots were

extremely high levels – $>= \mu + 3\sigma$ relative to the rest of the cells in DepMap. More complex expression-based selections can be made using set algebra. e.g.

CCND3_OUTLIERS $=$ $<$**1ccnd3** $|>$**1ccnd3**

selects those cells where expression of CCND3 is either $>$ = 1σ above or $<$ = 1σ below the mean expression of CCND1 across the DepMap collection, while:

CCND3_CENTRAL=#^CCND3_OUTLIERS

uses exclusion from the universal set (generated by **#^)** to define the effective complement of **CCND3_CENTRAL**, i.e. those cells in which CCND3 is expressed within ± 1σ of the mean.

Graphical interrogation of the expression profile of any specified gene, as in Fig. 1a, is available via the ***expr*** command in the interactive *DepMine* program

2.4 Identifying synthetic sickness/ lethality relationships

The main goal of *DepMine* is the identification of target genes on which specific tumour cell lines depend for their survival, and of the genetic (and epigenetic) factors in such cell lines that confer dependency on those target genes. Where cellular dependence on one potential target gene (GeneA) depends on the genetic or epigenetic status of a second gene (GeneB), these are deemed to display a Synthetic Sickness/Lethality (SSL) relationship. Identification of simple pairwise SSL relationships for a given GeneA is provided by the ***find*** module of *DepMine*.

To find GeneA-GeneB pairs that show SSL we extract the DepMap gene dependency scores for GeneA across all cell lines, and partition these into one of two classes according to whether GeneB in a given cell line harboured a disruptive (e.g. truncating) mutation and/or a homozygous deletion (CNV $<$ $\log_2(0.87/2 + 1)$) and/or showed under-expression (see above for definition) (test class), or whether it harboured no non-synonymous mutations, was diploid ($\log_2(2.64/2 + 1) >$ CNV $> \log_2(1.32/2 + 1)$) and was not under-expressed (control class). Dependency values from cells not matching either of these criteria are omitted from the calculation. These two dependency distributions are then further subdivided according to whether the dependency value is $>=$ or $<$ a threshold dependency (default 0.65, but can be user-specified), and the cell line counts used to populate a 2 x 2 contingency table, which is analysed using a X^2 test (Python **scipy.stats.chi2_contingency**. *P*value $<$ 0.05 are taken as indicating a likely SSL. SSL pairs are presented graphically ranked on *P*value along with the fraction of cell lines in which the GeneB was disrupted/deleted/downregulated, and the magnitude of the effect calculated using Cohen's d value (Cohen 1988). As the ***find*** command for a single GeneA can involve more than 20 million statistical computations, it is distributed across all available CPUs using the Python multiprocessing **Pool.starmap** method, so that its execution time depends on the number of CPUs available. For example, where synthetic lethal partners with disruptive mutations were sought for the ALDOC (aldolase C) gene, the ***find*** command took 386 seconds on a 2020 iMac with a 3.8 GHz 8-Core Intel Core i7 running OSX 12.6. Results of ***find*** searches are displayed graphically as the kernel density estimates (KDE) of the test and control sets of cells, with distributions means indicated.

Using a highly parallel version of *DepMine* using the 2022Q2 release of DepMap, and running on a multi-node HPC cluster, we performed a complete all-by-all equivalent of the ***find*** command looking for SSL partners with Loss-of-Function (LoF—disruptive mutation *or* CNV deletion) for all of the 16,825 genes for which dependency data was available, and which were not genes with high average dependency and therefore unlikely to have utility as therapeutic targets. The results of the all-by-all analysis is discussed below.

DepMine extends the definition of SSL relationships through the use of the powerful Cancer Profile system described above. In this mode the ***switch*** module of *DepMine* applies the same statistical analysis as the pairwise ***find*** search, but instead looks for GeneAs whose dependency are significantly different in a population of cell lines that match a defined cancer profile as compared with a control set of cell lines. Control sets are generated by selecting cells that have no non-synonymous mutations and are diploid for all the genes specified in the test profile, and have the opposite expression pattern to any genes whose expression is specified in the test profile.

2.5 Mining for secondary effects

While analysis of dependency switching using cancer profiles allows identification of target genes with a strong and significant SSL relationship to a complex cancer profile, the compilation of the cancer profiles themselves relies on prior knowledge of the genes whose mutation and/or copy number variation and/or under/over-expression defines the state or phenotype of interest. Thus, in almost all significant and strong SSL relationships identified, as well as a population that matches the test profile and has a high dependency, there will be cell lines that match the test profile, but nonetheless display low dependency

generated using the expr function of DepMine. b) Tissue distribution of 179 cell lines matching the cancer profile : PTEN_LOSS shown (left) as absolute cell line counts, and (right) as fraction of tissue type normalised by number of that type in the DepMap collection. The distribution of tissue types in DepMap is very uneven, so although it appears that PTENLoF is predominantly in brain tumours, once corrected for the unequal tissue distribution uterine tumours are the most likely to have this genetic background. The figures are generated by the type command in the interactive DepMine program. c) Mutational signature frequencies for cells matching the profile PTEN_LOSS_UTER showing strong signals related to mismatch repair deficiency (SBS6, SBS15, SBS26) and/or loss of proof-reading activity in replicative DNA polymerases (SBS10A, SBS10B, SBS20). d) Mutational signature frequencies for cells matching the profile BRAF_ACT_SKIN, showing a strong signal for mutations resulting from UV-driven DNA damage (SBS7A, SBS7B), as well as SBS5, whose aetiology is unknown. c) and d) are generated by the sig command in DepMine. e) The MYC gene shows a degree of copy number amplification in tumours that is moderately correlated (r=0.35) with expression level (left). High expression of MYC is strongly correlated (0.57) with increased cell dependency (right) consistent with its status as an oncogene (see below).

on the SSL target. In many ways this mirrors the common observation in cancer clinical trials that there are often a group of patients selected on the same criteria as those who respond to the new therapy, but nonetheless fail to respond. In both cases this is likely to be due to the inadequacy of the profile used to select the cell lines or patients, and begs the question as to how the initial profile, which proves so useful in initially identifying the SSL target or treated patient cohort, can be refined to give a strong predictive 'biomarker' for those tumour cell lines (and patients) which are mostly likely to respond to a therapy directed at the identified SSL target. To address this, *DepMine* incorporates a 'mining' feature that analyses cell line populations that match the profile, to find secondary features that improve the size of the SSL effect and increase the proportion of the cell lines in the matching distribution showing high dependency on the SSL target.

Secondary mining is implemented in the *DepMine* ***mine*** module, and allows for analysis of the profile-matching-population on the basis of all the factors that can be used to define cancer profiles: sex, age, tissue type, LoF mutations, GoF mutations, gene fusion, mutational signatures, deletion, amplification, downregulation of expression, upregulation of expression. The ***mine*** algorithm looks for distributions of profile-matching cell lines that additionally match the different secondary factors being tested and determines whether those cells are statistically over-represented in the part of the overall cell line population with low dependency on the SSL target, or are over-represented in the population with high dependency on the SSL target. In the former case, refinement of the initial profile to exclude cells matching the additional factor will be expected to improve the size of the SSL effect, while for the latter case, specific inclusion of matching cells will be expected to achieve this.

Inevitably there is a complex interplay and interdependence between the influence different secondary factors have on the effect size, and the effects themselves are not simply additive. So, different combinations of secondary effects need to be explored to find the combination that maximises the predicted effect size, and this may be open to further improvement by a further round (or rounds) of secondary mining. However, a balance needs to be struck between the effect size a cancer profile can achieve and the number of cell lines it ultimately matches. Taken to absurdity, one can envisage an ultimately refined profile that only matches a single cell line, albeit one that has maximal (1.0) dependence on the target gene.

To help find solutions that balance effect size and cell line coverage, *DepMine* provides a Pareto optimisation function (Hu *et al.* 2013) that evaluates a 'forest' of all possible decision tree combinations of factors that individually show a strong effect when added to the 'root' profile, and identifies those that give an optimal combination of effect size and the number of cell lines they match.

2.6 Cell line selection with cancer profiles

DepMine cancer profiles provide a powerful tool for selecting cell lines on the basis of complex combinations of tissue, age and sex origin of the tumour and the presence of defined mutational signatures, as well as mutational status of genes and genomic loci, and relative expression levels of individual genes. The basic 'currency' of SSL analysis—loss of a functional gene product (e.g. PTEN) through deletion or frame-disrupting mutation—can be simply represented and used to analyse the tissue distribution of the loss of that important tumour suppressor (Fig. 1b). The profile can be further defined to reference only those tumours of uterine origin in which PTEN is lost, and the distribution of mutational signatures in that subset analysed (Fig. 1c). Activation of an oncogene such as BRAF through gain of function (GoF) mutation is also effectively represented as a *DepMine* profile and combined with a tissue origin, reveals the distinctive UV signature of BRAF-driven melanomas (Fig. 1d).

Very specific GoF profiles can be developed to encapsulate activation of pathways that can arise due to a variety of genetic events e.g.

$$\begin{aligned}\textbf{MAPK\_ACTIVATED} = {} & /\textbf{hras} : \textbf{g12}?, \textbf{g13}?, \textbf{q61}? \mid /\textbf{kras} \\ & : \textbf{g12}?, \textbf{g13}?, \textbf{q61}? \mid /\textbf{nras} \\ & : \textbf{g12}?, \textbf{g13}?, \textbf{q61}? \mid /\textbf{raf1} \\ & : \textbf{s257l}, \textbf{s259f} \mid /\textbf{braf}+\end{aligned}$$

defines the set of cells in which any of the RAS family genes have one of the well-characterised oncogenic mutations, or CRAF has either of the missense mutations that preventing its downregulation by 14–3-3 proteins, or has any annotated gain-of-function mutation in BRAF, any or all of which result in an increase in signalling through the canonical RAF-MEK-ERK MAP-kinase cascade.

Selection based on copy number variation e.g. the MYC amplification seen in many tumours, can also be effectively represented, optionally combined with selection based on expression level—e.g. **MYC_UP = +myc | >myc —**although these may often be correlated and redundant. *DepMine* provides a tool in the ***expr*** module to visualise any correlations between copy number and expression, and relates these to the pattern of tumour cell dependency on that gene (Fig. 1e).

3 Results

3.1 Identifying All-By-All SSL gene pairs

Although the power of DepMine primarily resides in its ability to use complex cancer profiles to find SSL relationships, we tested the effectiveness of our statistical model using an all-by-all pairwise equivalent of the ***find*** command looking for SSL partners with severe LoF (frame disruptive mutation) *or* CNV deletion for all of the 16,825 genes for which dependency data was available, and which were not genes with high average dependency and therefore unlikely to have utility as therapeutic targets. We found 34,256 GeneA: GeneB pairs where the $|\text{dep(GeneA)}| < 0.65$, and dep(GeneA) in cells with LoF/CNV-del of GeneB was significantly $>$ than GeneB-WT cells ($P < 0.05$). A full-listing of the SSL pairs identified is provided in Supplementary Database 1, available as supplementary material at *Bioinformatics* online.

We identified strong and highly significant signals for several paralogous gene pairs that have been previously noted to display SSL in experimental screens (Ryan *et al.* 2023), including ARID1A/ARID1B, SMARCA2/SMARCA4 and EP300/CREBBP (Fig. 2a-c). These

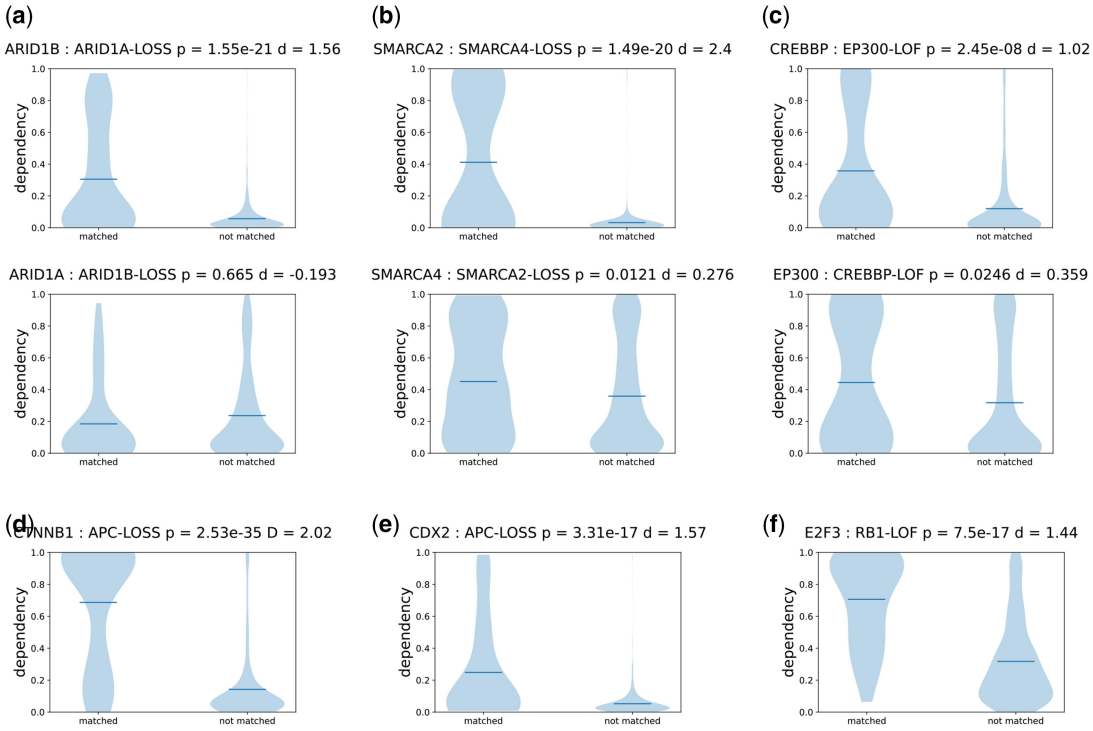

Figure 2 Pairwise SSL Analysis. a) Kernal Density Estimate (KDE) plots for distribution of dependency on (top) ARID1B in cells with ARID1A-LOF (matched) compared to cells with ARID1A-WT (not matched) – the dark line indicates the mean values of those distributions. The difference is highly significant with a very strong effect ($d > 0.8$); (bottom) as (top) but dependency on ARID1A in cells where ARID1B is either LoF or wild-type – unlike (top) the distributions show no significant difference. b) as a) but for SMARCA2 and SMARCA4. c) as a) but for CREBBP and EP300. d) as a) but for CTNNB1 dependency in cells with APC-LoF compared to cells with APC-WT. e) as d) but for CDX2 dependency. f) as a) but for dependency on E2F3 in cells with RB1-LOF compared to cells with RB1-WT.

SSL relationships are not generally commutative, so that while loss of ARID1A, SMARCA4 or EP300 in tumour cells enhances dependency on ARID1B, SMARCA2 or CREBBP respectively, loss of ARID1B. SMARCA2 or CREBBP had little effect on the requirement for ARID1A, SMARCA4 or EP300 respectively for tumour cell viability. Whether this reflects underlying biology or results from few observed mutations for the second gene, is unclear. It is worthy of note, that even in the orientations that display the effect, these paralogue SSL relationships are not absolute and there remain many cell lines lacking the sensitising paralogue that aren't dependent on the compensating paralogue. Indeed, searching DepMap reveals more than 20 viable cell lines in which functional ARID1A and functional ARID1B are both lost. Clearly there are secondary factors that define the settings in which these paralogue SSLs operate, but these are not revealed by simple pairwise LoF analysis, experimental or computational–hence our motivation in DepMine to facilitate the representation of complex and multigenic SSL relationships.

Significant non-paralogous SSL relationships were also identified for important tumour suppressors, several of which have already been observed experimentally. These include a very strong SSL between CTNNB1 (encoding β-catenin) and deletion/LoF of the adenomatous polyposis cancer associated APC, which is required for regulating β-catenin levels in the Wnt-signalling pathway (Fig. 2d). APC LoF occurs in ~30% of sporadic colorectal cancers and correlates with poor prognosis (Li *et al.* 2023). Other genes showing medium/strong SSL with APC include CDH1, JUP, TCF7L2 and CDX2 (Fig. 2e).

SSLs were identified for LoF of RB1, a common occurrence in cancers with a poor prognosis. The strongest effect was seen with the transcription factor E2F3, whose inhibitory interaction with the RB1 protein regulates entry to S-phase (Chen *et al.* 2009), suggesting that RB1-defective tumours specifically develop a level of 'oncogene addiction' to E2F3, but not to other RB1-interacting E2F paralogues which display no significant SSL (Fig. 2f). The E3-ubiquitin ligase subunit SKP2 also shows a strong SSL with RB1-loss, which has previously been noted experimentally (Gupta *et al.* 2022).

3.2 Unravelling passenger effects in regional deletions

Within the all-by-all pairwise analysis using either frame-disrupting mutations or CNV deletion of GeneBs, a number of GeneAs emerged as promiscuous partners, giving significant signals for many GeneBs. Promiscuous SSL partners included VPS4A, RAB6, HBS1L, KRAS, ITGB1, HSPA13 and WRN. With the exception of WRN (but see later) the multiple GeneBs whose LoF/deletion switched the dependency of these promiscuous GeneAs, showed strong clustering into a small number of chromosomal loci, commonly involving one or both of 9p21 and 18q21 (Fig. 3a).

The power of *DepMine*'s cancer profiles to unravel complex synthetic lethal relationships is substantial. For example: strong SSL relationships were identified in the all-by-all pairwise analysis (see above and Supplementary Database 1, available as

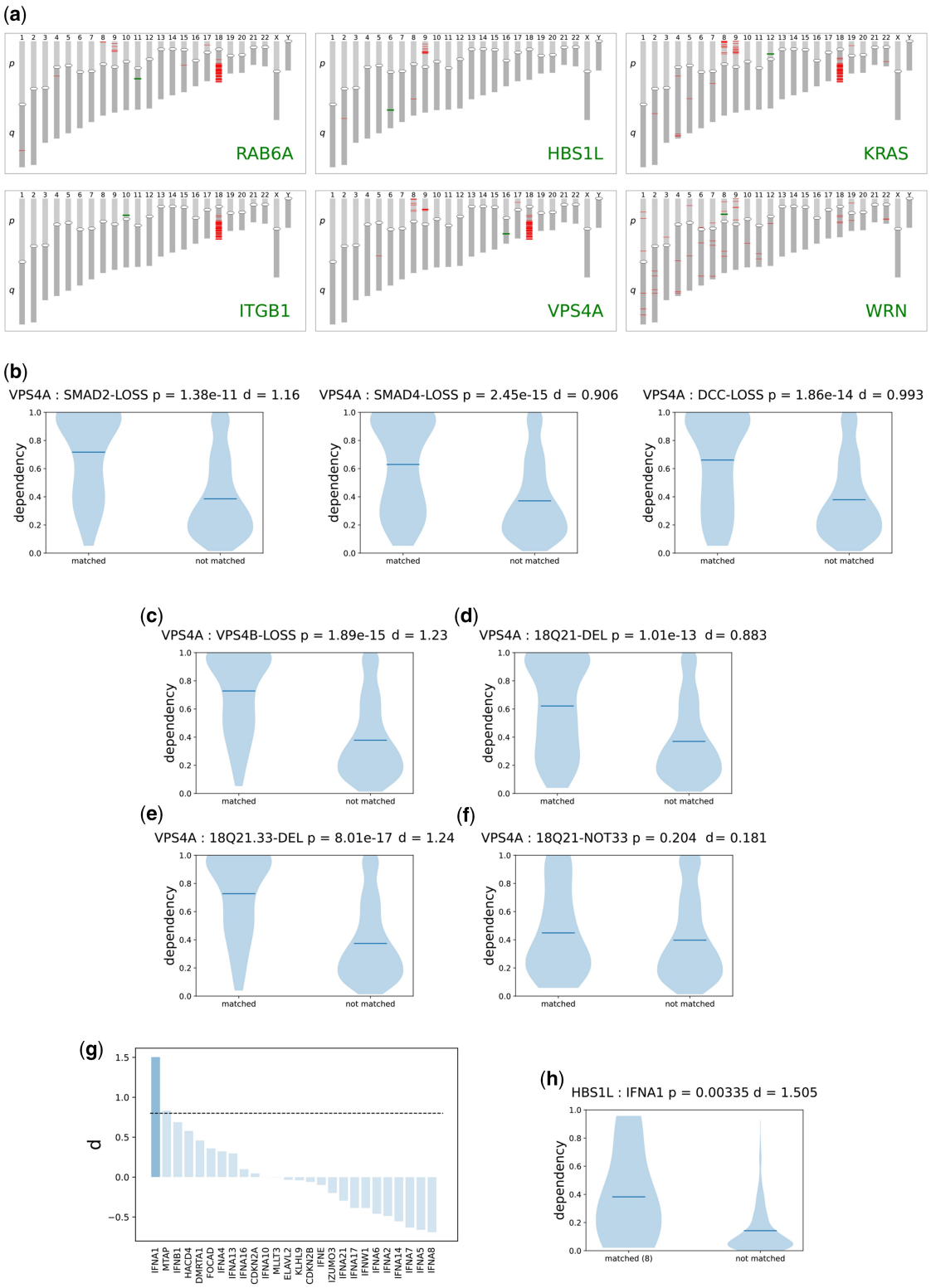

Figure 3 Chromosomal Passenger SSL Analysis. a) Chromosomal locations (red lines) of the genes whose LoF/deletion showed a statistically significant SSL effect with one of the promiscuous SSL GeneA partners shown. The location of the GeneA itself is indicated by the thick green line. The positions of centromeres are indicated by white ovals. This plot is generated by the chro module of DepMine. b) KDE plots for VPS4A dependency in cells with LoF of (left) SMAD2, (centre) SMAD4, or (right) DCC compared to cells in which these genes are diploid and harbour no non-synonymous mutations. All three genes show strong, statistically significant SSL effects. c) as b) but for VPS4A dependency in cells with LoF of VPS4B. d) as b) but for deletion of any gene(s) in cytogenetic band 18q21. e) as d) but for deletion of any gene(s)in sub-band 18q21.33. f) as d) but for deletion of any

supplementary material at *Bioinformatics* online) between VPS4A and LoF/deletion of a number of other genes, including important tumour suppressors such as SMAD2, SMAD4 and DCC which are disrupted in a range of cancers (Fig. 3b). While no obvious function or role links the VPS4A gene product, a AAA +-ATPase involved in vacuolar sorting, with the various genes with which it displays an SSL effect, those genes share a common chromosomal location, with the majority mapping to cytogenetic band 18q21 on the long arm of chromosome 18 (Fig. 3a), a genetic locus often deleted in gastrointestinal and pancreatic cancers, and associated with poor prognosis and liver metastases (Tanaka *et al.* 2006).

At first sight this appears to be a promising 'hit' as VPS4A is a druggable target which could be used to treat cancers where SMAD2/SMAD4 signalling has been lost. However, one of the other genes in the 18q21 region is VPS4B, a VPS4A paralogue, with which it has a well-documented SSL relationship (Neggers *et al.* 2021) readily detected by *DepMine* (Fig. 3c). It therefore becomes difficult to determine whether the effects seen for SMAD2, SMAD4, DCC and other genes in 18q21 are direct effects due to a genuine functional synthetic lethality with VPS4A, or whether it is because of their proximity to the VPS4B gene which is often lost with them as a 'passenger' in cancer-associated deletions of 18q21, and so indirectly reflect the paralogue SSL between VPS4A and VPS4B. Certainly, VPS4A generates a statistically significant and reasonably strong SSL effect when tested against cells with a cancer profile defined by deletion of any gene within 18q21 (**18Q21_DEL**), compared to cell lines without any such deletions (Fig. 3d). So, a paralogue passenger proximity effect could be operating even though SMAD2, SMAD4 and DCC map to sub-bands 18q21.1 and 18q21.2 at the centromere-proximal end of 18q21, while VPS4B maps to sub-band 18q21.33 at the centromere-distal end of 18q21. This type of conundrum in *DepMine* can be resolved by defining a complex cancer profile **18Q21.33_DEL** that selects cell lines with any genes deleted in 18q21.33, and profile **18Q21_NOT33** that selects cell lines with deletion of genes anywhere in 18q21 except 18q21.33. Consistent with the apparent VPS4A SSL with genes other than VPS4B being primarily due to their collateral loss along with VPS4B in regional deletions, VPS4A shows a significant and strong SSL effect with **18Q21.33_DEL** (Fig. 3e) comparable to the that with LoF of VPS4B (Fig. 3c), whereas the SSL effect is completely lost when sub-band 18q21.33 deletions are excluded (Fig. 3f), even though loss of the SMAD2, SMAD4 and DCC genes would still be present in the matched cell line set.

We can therefore conclude that there is unlikely to be any direct functionally related SSL between VPS4A and these important tumour suppressors, but that their chromosomal proximity to the VPS4A paralogue VPS4B nonetheless opens up a potential therapeutic vulnerability in the many tumours in which SMAD2, SMAD4 and DCC loss has been positively selected, so long as the deletion that eliminates these tumour suppressors encompasses 18q21.33 and removes VPS4B as a passenger.

HBS1L—which encodes a GTPase involved in release of stalled ribosomes (Shao *et al.* 2016) – shows significant SSL relationships with a cluster of genes mapping to chromosome band 9p21.3—a locus that is lost in a range of tumours (Spiliopoulou *et al.* 2022, Jiang *et al.* 2023) and increasingly associated with resistance to immune checkpoint therapy (Han *et al.* 2021). This locus contains CDKN2A—which encodes the well characterised tumour suppressors $p16^{INK4A}$ and $p14^{ARF}$, MTAP—encoding methyladenosine phosphorylase whose loss is associated with poor survival in glioblastoma and other tumours (Patro *et al.* 2022), and a cluster of 13 homologous genes encoding α-interferon isoforms. As HBS1L has no essential paralogues within this region, it is not obvious which gene or genes lost in the 9p21.3 deletion are responsible for switching HBS1L dependence. To address this, the ***locu*** module of *DepMine* performs a scan of all GeneBs in the locus and calculate the switch effect size on the target gene—HBS1L in this case–based only on mutational LoF (nonsense or frameshift), but not copy number variation (Fig. 3g). For HBS1L, this scan identifies the IFNA1 gene, encoding interferon-α1, as having a very strong and significant direct SSL relationship with HBS1L (Fig. 3h), suggesting that it is its loss in the 9p21.3 deletion that is most important in sensitising tumours with this deletion to HBS1L loss, and not the more 'usual suspects' CDKN2A or MTAP.

Interestingly, experimental dissection of the 9p21.3 locus has shown that specific loss of the interferon cluster in substantial 9p21.3 deletions in a mouse pancreatic cancer model, plays an important role in facilitating immune evasion and immunotherapy resistance (Barriga *et al.* 2022). The SSL relationship we observe here suggests that HBS1L may be a useful target for that context. In similar analyses (see Supplementary Data, available as supplementary material at *Bioinformatics* online) KRAS, RAB6A and ITGB1 all show significant SSL with mutational LoF of SMAD4 in the 18q21 locus, but only KRAS shows a 'strong' effect.

3.3 Analysis with cancer profiles

The use of cancer profiles in *DepMine* extends the use of DepMap data for identifying SSL relationships beyond gene pairs, allowing the identification of genes whose dependency is switched on the basis of sets of factors that define complex states. A good example of this is the phenotype of microsatellite instability (MSI) which manifests as expansion/contraction of short repeating DNA mononucleotide and dinucleotide sequences (Ellegren, 2004) and is particularly associated with highly mutated colorectal, gastric and endometrial tumours (Sato *et al.* 2023, Riedinger *et al.* 2024) as well as some haematological malignancies (Maletzki *et al.* 2013). MSI results from LoF or transcriptional downregulation of the genes encoding components of the DNA mismatch repair system (MMR) (Li, 2008) and/or from loss or defects in proof-reading replicative DNA polymerases

gene(s) in band 18q21 apart from sub-band 18q21.33 g) Waterfall plot of d values for switching of HBS1L dependent on only LoF (missense or frameshift) mutations, but not deletion, of genes in the 9p21.3 cytogenetic locus for which mutations are recorded in DepMap. Dark bars indicate statistical significance ($p < 0.05$) Only LoF of IFNA1 produces a significant effect > 0.8 – the usual threshold for a strong signal. The plot was generated by the locu module in DepMine h) KDE plot of HBS1L dependency in cell line with IFNA1 LoF but not deletion, compared with cell lines in which IFNA1 has no mutation or deletion.

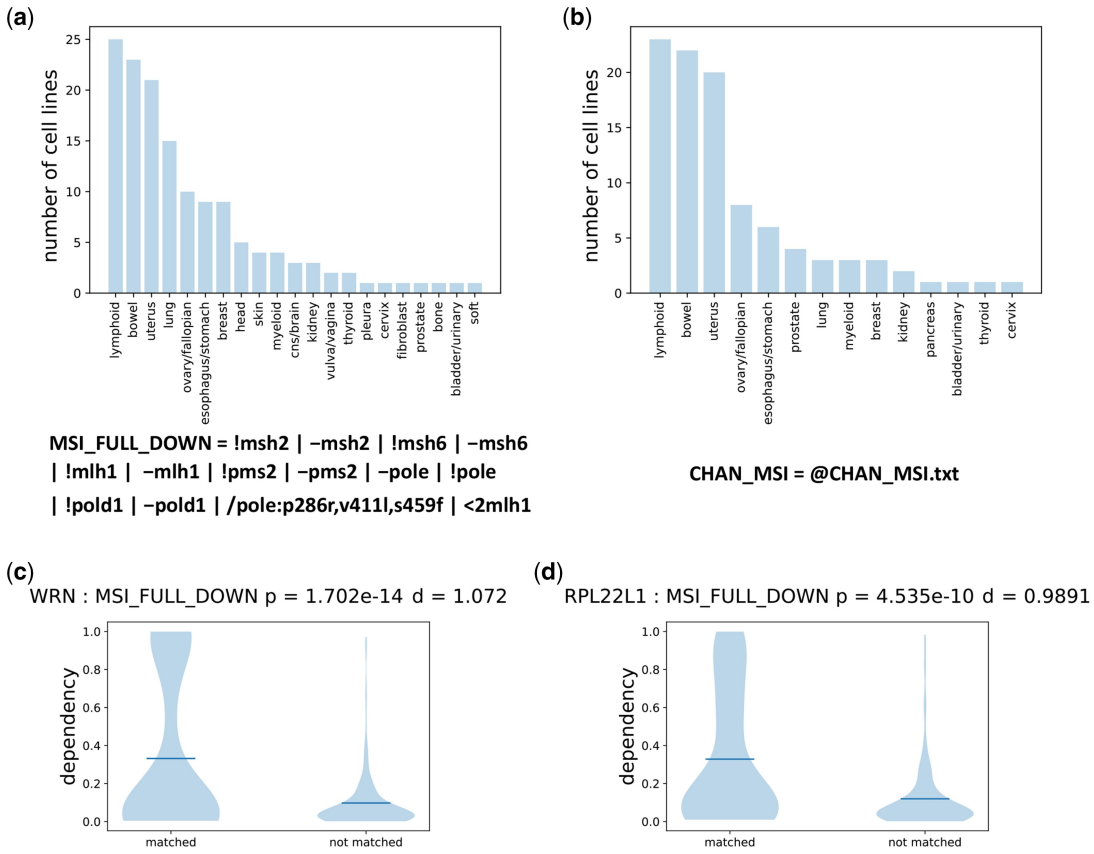

Figure 4 SSL Using Complex Cancer Profiles. a) Tissue distribution of 142 cell lines matching the cancer profile : MSI_FULL_DOWN. b) Tissue distribution of 99 cell lines in the cancer profile CHAN_MSI designated MSI-High in (Chan, et al., 2019) c) KDE plot for WRN dependency in cells matching the MSI_FULL_DOWN profile compared with a control set of cells d) as c) but for RPL22L1.

(Haradhvala *et al.* 2018). Based on this we constructed a *DepMine* cancer profile for microsatellite instability, **MSI_FULL_DOWN**, which matches 142 cell lines in DepMap across a range of tissue types (Fig. 4a), but dominated by tumours of lymphoid, bowel and uterine origin.

To 'sense check' our MSI profile, we defined a second profile (**CHAN_MSI** = @**CHAN_MSI.txt)** by uploading a curated list of cell lines classified as MSI-High based on experimental analysis (Chan *et al.* 2019). **CHAN_MSI** matches 99 cell lines in DepMap also dominated by tumours of lymphoid, bowel and uterine origin (Fig. 4b), but with a lower representation of lung samples than our computationally defined **MSI_FULL_DOWN** profile. The two profiles have 71 cell lines in common (determined by **CHAN_MSI & MSI_FULL_DOWN**).

There has been considerable interest in identifying druggable targets specific to MSI cells, and a number of experimental CRISPR screens have been performed for this purpose (Chan *et al.* 2019, Picco *et al.* 2021). To determine whether this question was amenable to our computational approaches we utilised the ***switch*** module in *DepMine* which applies a similar analysis to the pairwise ***find*** search, but looks for GeneAs whose dependency is significantly different in a population of cell lines that match a defined cancer profile–e.g. **MSI_FULL_DOWN** as compared with a control set of cell lines. Control sets are generated by selecting cells that have no non-synonymous mutations and are diploid for all the genes specified in the test profile, and have the opposite expression pattern to any genes whose expression is specified in the test profile. Dependency switching analysis using the **MSI_FULL_DOWN** profile found 516 genes displaying SSL with $P < 0.05$ ($-\log_{10}(p) > 2$) and took approximately 9 minutes on a 2020 iMac with a 3.8 GHz 8-Core Intel Core i7 running OSX 12.6.

Although many genes displayed statistically significant differences between the dependency distributions for the profile matching and control sets of cell lines, only the highest ranking (based on *P*value) two genes also showed strong effects (Cohen d > 0.8). The first of these was WRN—the gene encoding the RECQ-family DNA helicase whose germline mutation underlies the cancer-prone progeria Werner's syndrome (Fig. 4c) (Monnat, 2010). WRN is the prime target identified in the several laborious experimental screens that have been conducted (Chan *et al.* 2019, Lieb *et al.* 2019, van Wietmarschen *et al.* 2020, Picco *et al.* 2021), so its unambiguous *ab initio* identification here as a strong SSL partner for MSI by a completely hypothesis-free computational analysis in under 10 minutes on a standard desktop computer, completely vindicates our approach. The second gene showing significance and a strong effect was RPL22L1 (Fig. 4d) which encodes a ribosomal protein that has previously been implicated in progression and treatment resistance in a range of different cancer types (Rao *et al.* 2019, Chen *et al.* 2023, Yi *et al.* 2023), although the mechanism is not understood. WRN and RPL22L1 were also the top two

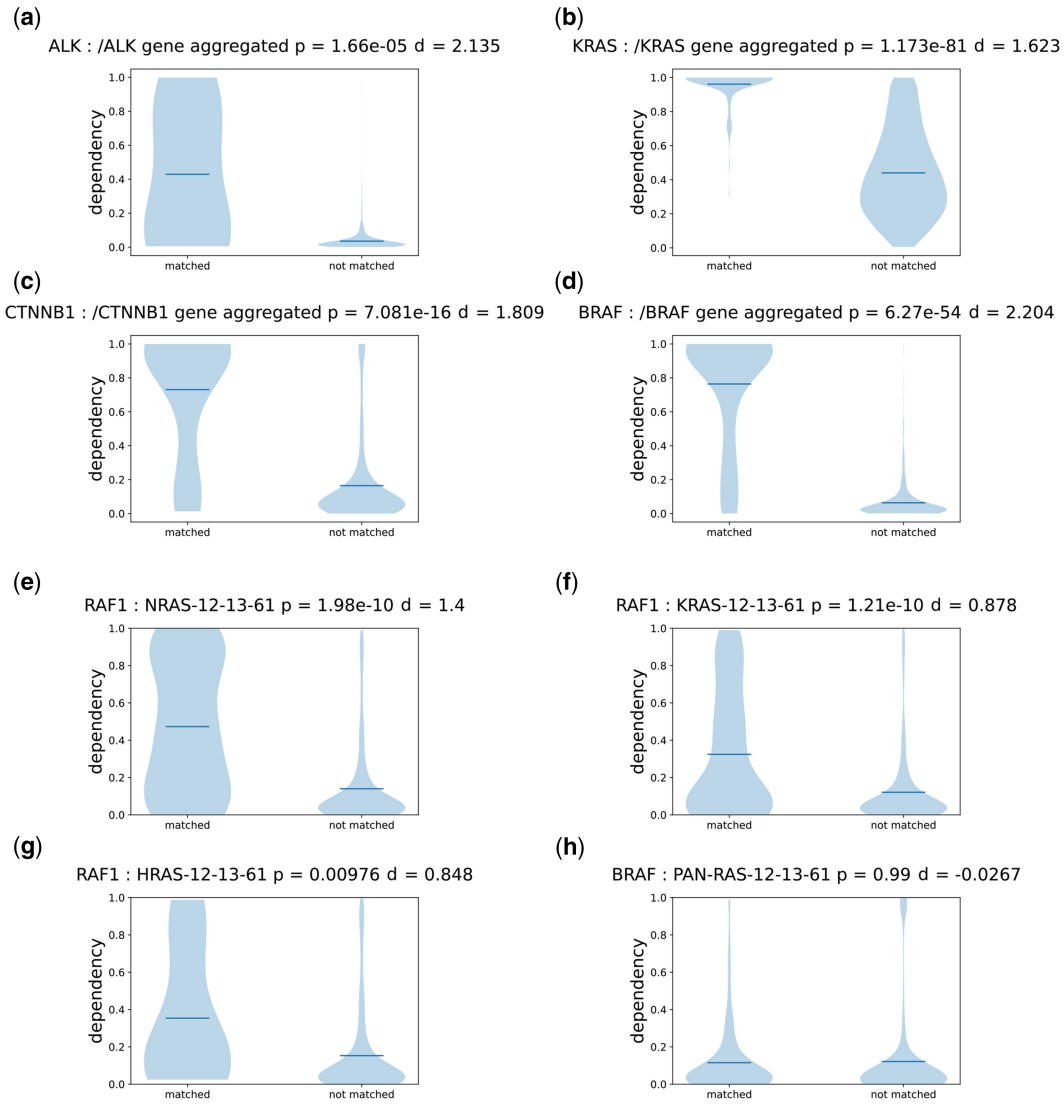

Figure 5 SSL with Gain-of-Function. a) KDE plot of strong self-dependencies found in the GOF analysis demonstrating cellular 'addiction' to ALK. Definitions of the gene-aggregated GoF profiles used are listed in Supplementary Table 1. b) as a) but for KRAS. c) as a) but for CTNNB1 d) as a) but for BRAF. e) KDE plot of CRAF (RAF1) dependency in cell lines harbouring activating mutations at residues 12,13 or 61 in NRAS. f) as e) but for KRAS. g) as e) but for HRAS. h) KDE plot of BRAF dependency in cell lines harbouring activating mutations in any RAS paralogue. Unlike CRAF (RAF1), BRAF shows no SSL with mutations in any of the RAS genes.

hits when the ***switch*** analysis was repeated with the **CHAN_MSI** profile (results not shown).

3.4 Gain of function dependency switching

While loss of function due to disruption of the reading frame is straightforward to predict based on the DNA sequence of a gene, it is difficult to predict the consequences of missense mutations. While the majority of these will have no significant effect, a minority will cause loss (LoF) or gain of function (GoF) which may have major consequences.

DepMine incorporates a dedicated ***GOF*** module that applies the ***switch*** analysis described above, to a curated list of GoF mutations. This can come from DepMap itself, which annotates 197 GoF mutations across 56 different genes (2023Q2 release), or from an extensive GoF dataset we have hand-compiled over several years from literature reports, which contains 1153 mutations across 67 different proteins (based on (Baeissa *et al.* 2017)).

The **GOF** module can operate at the level of the individual mutation, so that independent ***switch*** analyses are run, for example, for KRAS: G12V (**/kras: g12v**), KRAS: G12D (**/kras: g12d**) and KRAS: G12C (**/kras: g12c**). Alternatively, mutations can be aggregated at the residue level so that a combined ***switch*** analysis is performed for all KRAS: G12 mutations (**/kras: g12?**), or at the gene level so that a combined ***switch*** analysis is performed for all KRAS GoF mutations in the curated dataset (profile/**kras: a146p, a146t, a146v, a59t, g12a, g12c, g12d, g12r, g12s, g12v, g13c, g13d, k117n, q61h, q61k, q61l, q61r**). While hits obtained with aggregated GoF profiles may need further analysis to tease out specific contributions of individual mutations, they are an effective way of increasing the number of

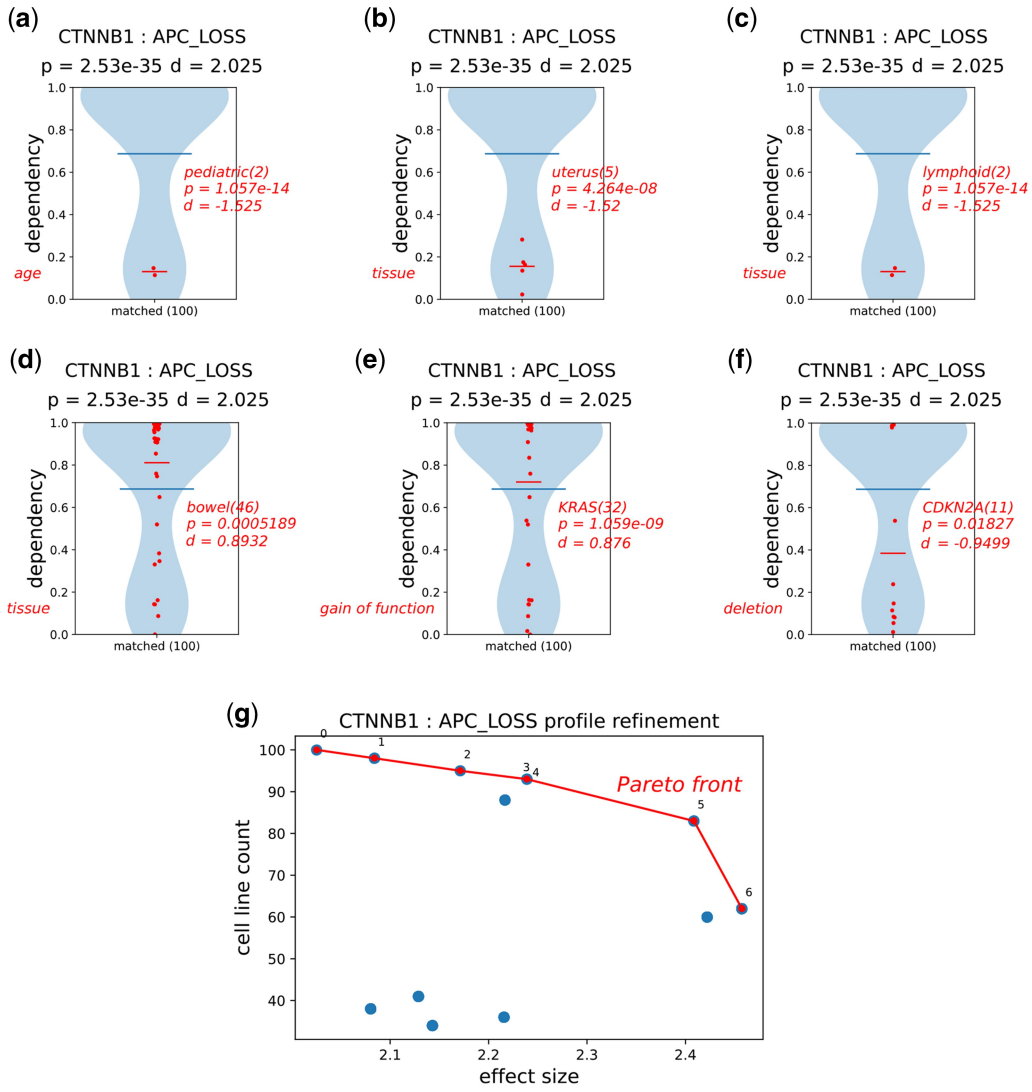

Figure 6 Automated Identification of Secondary Factors. a) KDE plot for dependency on CTNNB1 in cells matching the APC_LOSS profile, overlayed with distributions of cell lines of paediatric origin displaying significant enrichment in the low dependency sub-population (red). b) as a) but for cell lines derived from uterus, displaying significant enrichment in the low dependency sub-population (red). c) as a) but for cell lines derived from lymphoid tissues, displaying significant enrichment in the low dependency sub-population (red). d) as a) but for cell lines derived from bowel, displaying significant enrichment in the high dependency sub-population (red). e) as a) but for cell lines KRAS GoF mutations, displaying significant enrichment in the high dependency sub-population (red). f) as a) but for cell lines with deletion of the CDKN2A gene, displaying significant enrichment in the low dependency sub-population (red). g) Plot of cell line count versus effect size for dependency on CTNNB1 gene by cells selected by the APC_LOSS profile with addition of different combinations of secondary factors. The red line and numbering indicate Pareto optimal solutions.

observations so that sufficient (>5) are obtained for the significance calculation. A full ***switch*** screen of our curated GoF dataset using gene-level aggregation identified 2082 significant hits with strong SSL to a GoF mutated gene, and took under 2 hours. The full hit-list as well as the aggregated profile definitions used is provided in Supplementary Database 2, available as supplementary material at *Bioinformatics* online.

Some of the strongest signals in the ***GOF*** analysis occur between GoF mutated genes and themselves (Fig. 5a**, b, c, d**). Most of the genes that display this effect are well known highly penetrant oncogenes e.g. KRAS, BRAF, NRAS, PIK3CA, CTNNB1, NOTCH1, ALK—and the self-dependence reflects the phenomenon of oncogene addiction (Weinstein and Joe, 2008) that makes the products of these genes compelling targets for drug discovery (Torti and Trusolino, 2011). Many of the genes identified by non-self signals in the ***GOF*** analysis are transcription factors mediating the downstream effects of the activated oncogene with which they consequently display an SSL relationship, but these tend to be very difficult targets for drug discovery.

One notable exception is CRAF (RAF1), a druggable target showing a significantly SSL effect with activated NRAS (Fig. 5e), and weaker but nonetheless significant SSL effects with KRAS and HRAS (Fig. 5f, g). Given the challenge in directly targeting NRAS, CRAF inhibitors (of which several are available) would be expected to be useful in NRAS activated tumours (Bekele *et al.* 2021). Despite the significant degree of functional redundancy

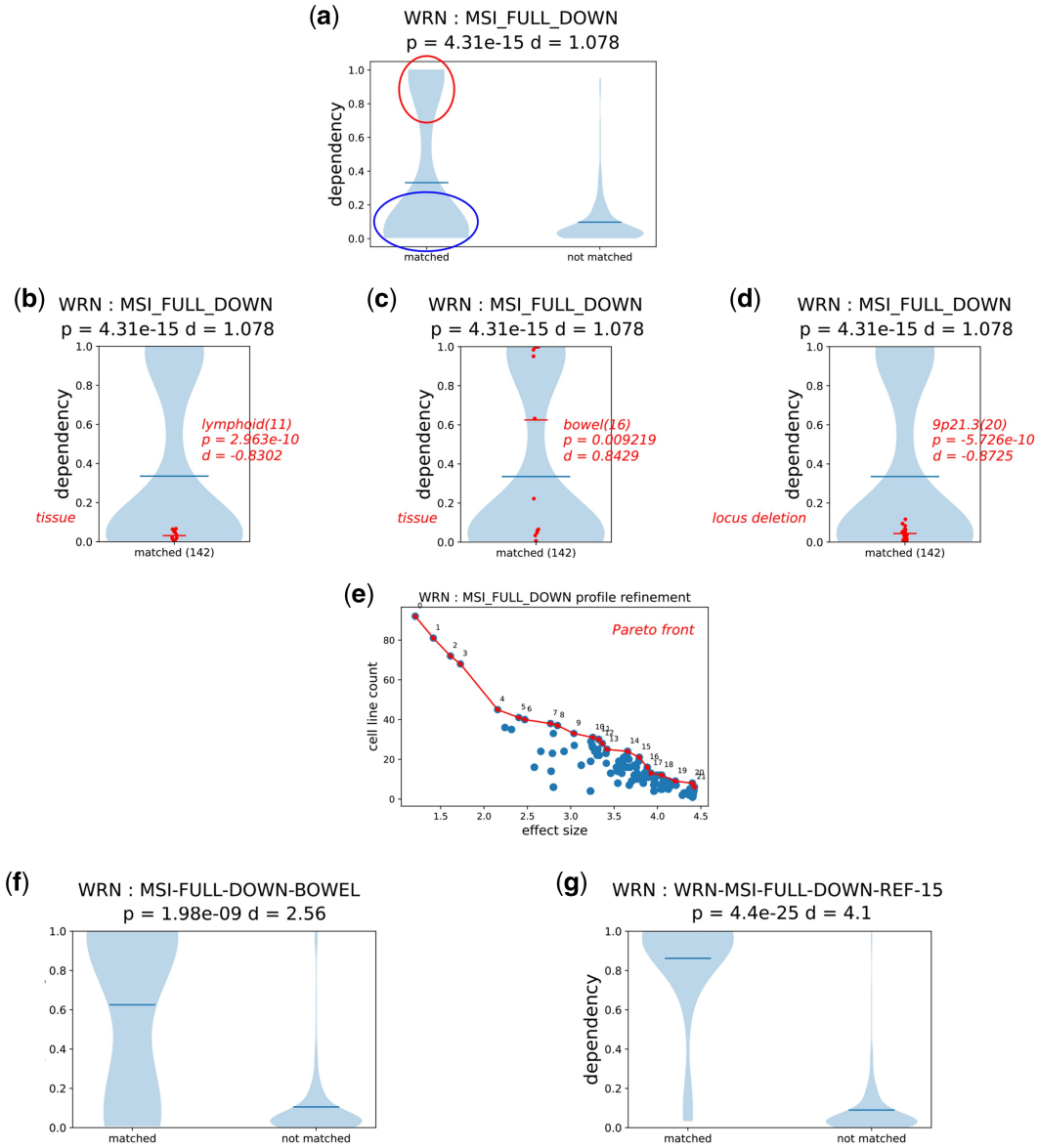

Figure 7 Optimising Predictive Biomarkers. a) Within the distribution of cell lines matching the MSI_FULL_DOWN profile, there are distinct sub-populations which show high dependency on WRN (red oval) and those that show low dependency (blue oval). b) KDE plot for dependency on WRN in cells matching the MSI_FULL_DOWN profile, overlayed with distributions of cell lines displaying significant enrichment in the low dependency sub-population (red) of cells derived lymphoid tissue. c) as b) but for cell lines displaying significant enrichment in the high dependency sub-population (red) of cells derived from bowel d) as b) but for cell lines displaying significant enrichment in the high dependency sub-population (red) of cells with deletion of cytogenetic locus 9p21.3. e) Plot of cell line count versus relative effect size for dependency on WRN in cells selected by the MSI_FULL_DOWN profile with addition of different combinations of secondary factors. The red line and numbering indicate Pareto optimal solutions. f) KDE plot of WRN dependency in cell lines matching the MSI_FULL_DOWN profile, with the additional specification of a bowel tissue origin, compared with the control cell population. g) as f) but with cells selected by profile combination 15 from the automatic optimisation process.

between CRAF and BRAF, the effect is restricted to CRAF, with BRAF showing negligible SSL with any of the RAS genes (Fig. 5h).

3.5 Mining for secondary factors

While strong effects of mutational disruption/deletion of individual genes on the dependency of other genes can be detected by pairwise analyses (Lord *et al.* 2020), these are relatively unusual, and in most circumstances multiple factors play a role in determining whether a given gene is important or not in the survival of a cancer cell line (Ryan *et al.* 2023). Expanding pairwise relationships to a third dimension is feasible but the computation required increases very substantially, and the overwhelming majority of the triple combinations examined will not be informative. The ability of *DepMine* to define cancer profiles allows it to leverage the significant body of prior knowledge available in cancer biology and identify genes whose dependency switches as a result of far more complex situations than the loss of function of a single gene.

For example, the **APC_LOSS** profile that generates a strong and highly significant SSL with CTNNB1 (β-catenin) (Fig. 2d), nonetheless matches a number of cell lines that display a low dependency.

Applying the *DepMine* mining algorithm to this data we find that profile-matching cell lines deriving from paediatric samples or of uterine or lymphoid origin are significantly over-represented in this low-dependency population (Fig. 6a-c), while those originating from bowel are significantly over-represented in the high-dependency population (Fig. 6d). Given the high prevalence of mutations, copy number and expression level changes in cancer cell lines in general, mining for secondary genes is restricted to changes that occur in more than a specified fraction of the cell lines matching the profile—10% empirically provides a useful cut-off. A significant strong effect was observed for GoF mutation of KRAS (Fig. 6e) which occurs in nearly 1/3 of the matching cell lines and is associated with high CTNNB1 dependence, and a significant strong effect was found for deletion of CDKN2A which is associated with low dependency on CTNNB1 (Fig. 6f).

3.6 Automatic generation of complex profiles

Application of the Pareto optimisation function (Hu *et al.* 2013) in *DepMine* analysed all possible combinations of these secondary factors that were individually found to influence the CTNNB1 dependency profile in cells with APC^{LoF}, and identified optimal combinations with respect to effect size and the number of cell lines they match (Fig. 6g). Combination 5 (**CTNNB1_APC_LOSS_REF_5**), which combines APC^{LoF} with tissue specifications that exclude cells of paediatric origin or from uterine or lymphoid tissue, and excludes cells in which CDKN2A (which encodes $p16^{INK4A}$ and $p14^{ARF}$) is deleted, achieves a substantial increase in effect size, while retaining applicability to 83 cell lines and probably constitutes a globally optimal solution. Combination 6 (**CTNNB1_APC_LOSS_REF_6**) which doesn't apply the restriction on CDKN2A but adds a positive requirement for a bowel tissue origin in its place, further improves the effect size, but decreases the cell line cover to only 62. The final choice of cell line selection in which to explore inhibition or targeted depletion of CTNNB1, is up to the investigator and will depend on whether they are more concerned with the breadth of cancers that can be targeted or the size of the effect likely to be elicited.

In a second example we took the observation of a strong SSL relationship between WRN and our microsatellite instability profile **MSI_FULL_DOWN** (Fig. 7a) and mined the profile matching cell line population for secondary factors. The ***mine*** algorithm identified two tissue-origins*: lymphoid*, strongly associated with low WRN dependency, and *bowel*, strongly associated with high WRN dependency, which effectively subdivide the profile-matching cell population (Fig. 7b and c). Indeed, this is essentially the selection criteria for ongoing clinical trials of WRN inhibitors (Moschetta *et al.* 2023) - however it is not identified as an optimal profile in our analysis. Six individual genes were also identified whose mutational disruption in $>$ 20% of the cell lines matching the **MSI_FULL_DOWN** profile showed significant strong effects associated with high WRN dependency: ARID1A a component of SWI/SNF chromatin remodelling complexes mutated in $>$5% of cancers (Mullen *et al.* 2021), KMT2B and KMT2D are lysine methyltransferases frequently mutated in a range of cancer (Rao and Dou 2015). Of lesser impact are RPL22 - a ribosomal protein whose loss has been associated with oncogenesis through multiple mechanisms (Goudarzi and Lindstrom 2016); PTEN—a lipid phosphatase tumour suppressor that negatively regulates signalling through the PI3K/AKT axis (Lee *et al.* 2018); and TTN, which encodes titin–the largest protein in the human genome with $\sim$34,000 amino acids. Although TTN is not directly linked to any tumour suppressor pathway, mutation within the TTN gene has been shown to be an effective indicator for overall mutational burden in a tumour cell line and indicative of MSI in colorectal cancer samples (Oh *et al.* 2020).

Mining also identified two copy number variations: deletion of a region of 9p21.3 containing CDKN2A, CDKN2B and MTAP is strongly associated with low dependency on WRN (Fig. 7d); and amplification of a region of 4p16.1 containing the uncharacterised FAM90A26, DEFB131A encoding a β-defensin isoform, and an extremely unusual multi-gene cluster encoding 20 members of the USPL17 family of deubiquitinating enzymes (Burrows *et al.* 2005), was associated with high WRN dependency.

In contrast to the CTNNB1- APC^{LoF} situation where some of the factors are redundant and several combinations have similar effect size and cell count scores, the combinatorial exploration of the 1024 decision trees built from the 10 secondary factors identified for the **MSI_FULL_DOWN** profile give a complex distribution (Fig. 7e). These are divided into a broad-acting subgroup with small improvements of the original profile, and a large poorly differentiated cluster with substantially higher effect size, but selecting very much smaller cell line populations. The effectiveness of the ***mine*** algorithm can be clearly seen by comparing the switching analysis for a 'hand-made' profile **MSI_FULL_DOWN_BOWEL**, where the clinical trial recruitment for a bowel tissue origin is added onto the base MSI profile (Fig. 7f), with one of the Pareto refined profiles from ***mine*** (solution 15 in Fig. 7e) that additionally specifies avoidance of tumours of lymphoid origin or with 9p21.3 deletion, and has positive requirements for LoF mutations of KMT2D and TTN on top of the basic MSI profile (Fig. 7g). Not only does the refined solution deliver a substantially improved effect size, but it matches 21 DepMap cell lines with dependency data compared to 16 for the tissue focussed profile.

4 Discussion

Cancer gene dependency screens, exemplified by DepMap (Tsherniak *et al.* 2017) are rich sources of information with the potential to accelerate identification of new therapeutic targets, and at least as important, identify selective genetic and epigenetic biomarkers that enable targeted therapies to be used with maximal effect. However, the complexity of such data and its associated metadata, especially in their raw state, limits the depth and breadth to which the average laboratory researcher can explore, and typically this is restricted to browsing data on 'favourite' genes. Analysis that takes on the entirety of the data and metadata requires substantial computational and data-science expertise, and while of considerable value, these studies do not always ask the questions the laboratory researcher or

clinician might wish to ask. DepMap itself provides some powerful query and analysis tools on its website https://depmap.org/portal/, and other groups have developed independent interfaces and packages that provide additional or complementary analysis (Killian and Gatto, 2021, Shimada *et al.* 2021, Wappett *et al.* 2021).

The *DepMine* software presented here seeks to provide a comprehensible but still very powerful toolkit with which highly complex questions can be posed across the entire cell and gene dependency collection, without the need for computational/data science expertise beyond the ability to frame a *cancer profile* as a basic algebraic formula. The breadth of criteria interrogated by the primitive postulate operators allows selection on almost every useful characteristic of a cancer cell line, and these can be elaborated into *cancer profiles* of essentially unlimited complexity using basic set operators and brackets. Defined *cancer profiles* can be assembled into collections, allowing the compilation of a library of cell line definitions that can be used for multiple downstream analyses. Although based on the data architecture of DepMap, the *DepMine cancer profile* selection system is fully adaptable to any cell line collection–public or private–where mutational, copy-number, and expression data has been generated.

Acknowledgements

We are grateful to our colleagues for critical discussions during the development of *DepMine*, and for presenting us with challenging problems to solve, which have helped drive its development.

Author contributions

Laurence H. Pearl (Conceptualization [Equal], Data curation [Lead], Formal analysis [Equal], Investigation [Equal], Methodology [Equal], Project administration [Equal], Software [Lead], Validation [Lead], Visualization [Lead], Writing—original draft [Lead], Writing—review & editing [Equal]), and Frances M. G. Pearl (Conceptualization [Equal], Data curation [Supporting], Formal analysis [Equal], Investigation [Equal], Methodology [Equal], Project administration [Equal], Resources [Equal], Software [Supporting], Supervision [Lead], Validation [Supporting], Visualization [Supporting], Writing—original draft [Supporting], Writing—review & editing [Equal])

Supplementary material

Supplementary material is available at *Bioinformatics* online.

Conflict of interests

None declared.

Funding

This work was supported by The University of Sussex and The Institute of Cancer Research.

Availability

The Python implementation of DepMine and associated data files can be obtained https://github.com/UOSbioinformaticslab/depmine and is free to academics and Not-For-Profit organisations. The DepMine release referenced in this paper is archived as DOI: 10.5281/zenodo.19570601

References

Baeissa H, Benstead-Hume G, Richardson CJ *et al.* Identification and analysis of mutational hotspots in oncogenes and tumour suppressors. *Oncotarget* 2017;**8**:21290–304.

Barriga FM, Tsanov KM, Ho YJ *et al.* MACHETE identifies interferon-encompassing chromosome 9p21.3 deletions as mediators of immune evasion and metastasis. *Nat Cancer* 2022;**3**:1367–85.

Bekele RT, Samant AS, Nassar AH *et al.* RAF1 amplification drives a subset of bladder tumors and confers sensitivity to MAPK-directed therapeutics. *J Clin Invest* 2021;**131**:e150851.

Benstead-Hume G, Chen X, Hopkins SR *et al.* Predicting synthetic lethal interactions using conserved patterns in protein interaction networks. *PLoS Comput Biol* 2019;**15**:e1006888.

Burrows JF, McGrattan MJ, Johnston JA. The DUB/USP17 deubiquitinating enzymes, a multigene family within a tandemly repeated sequence. *Genomics* 2005;**85**:524–9.

Chan EM, Shibue T, McFarland JM *et al.* WRN helicase is a synthetic lethal target in microsatellite unstable cancers. *Nature* 2019;**568**:551–6.

Chen HZ, Tsai SY, Leone G. Emerging roles of E2Fs in cancer: an exit from cell cycle control. *Nat Rev Cancer* 2009;**9**:785–97.

Chen Y, Mu Y, Guan Q *et al.* RPL22L1, a novel candidate oncogene promotes temozolomide resistance by activating STAT3 in glioblastoma. *Cell Death Dis* 2023;**14**:757.

Cohen J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: L. Erlbaum Associates.

Das S, Deng X, Camphausen K *et al.* DiscoverSL: an R package for multi-omic data driven prediction of synthetic lethality in cancers. *Bioinformatics* 2019;**35**:701–2.

Díaz-Gay M, Vangara R, Barnes M *et al.* Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment. Bioinformatics 2023;**39.** https://doi.org/10.1093/bioinformatics/btad756

Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* 2004;**5**:435–45.

Ferretti S, Hamon J, de Kanter R *et al.* Discovery of WRN inhibitor HRO761 with synthetic lethality in MSI cancers. *Nature* 2024;**629**:443–9.

Funk L, Su KC, Ly J *et al.* The phenotypic landscape of essential human genes. *Cell* 2022;**185**:4634–53.e22.

Goncalves E, Segura-Cabrera A, Pacini C *et al.* Drug mechanism-of-action discovery through the integration of pharmacological and CRISPR screens. *Mol Syst Biol* 2020;**16**:e9405.

Goudarzi KM, Lindstrom MS. Role of ribosomal protein mutations in tumor development (review). *Int J Oncol* 2016;**48**:1313–24.

Gupta P, Zhao H, Hoang B *et al.* Targeting the untargetable: RB1-deficient tumours are vulnerable to Skp2 ubiquitin ligase inhibition. *Br J Cancer* 2022;**127**:969–75.

Han G, Yang G, Hao D *et al.* 9p21 loss confers a cold tumor immune microenvironment and primary resistance to immune checkpoint therapy. *Nat Commun* 2021;**12**:5606.

Haradhvala NJ, Kim J, Maruvka YE *et al.* Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat Commun* 2018;**9**:1746.

Hu XB, Wang M, Di Paolo E. Calculating complete and exact pareto front for multiobjective optimization: a new deterministic approach for discrete problems. *IEEE Trans Cybern* 2013; **43**:1088–101.

Jiang Y, Wu Y, Zhang L *et al.* Loss of chromosome 9p21 is associated with a poor prognosis in adenosquamous carcinoma of the pancreas. *Precis Clin Med* 2023;**6**:pbad030.

Killian T, Gatto L. Exploiting the DepMap cancer dependency data using the depmap R package. *F1000Res* 2021;**10**:416.

Larrimore KE, Rancati G. The conditional nature of gene essentiality. *Curr Opin Genet Dev* 2019;**58-59**:55–61.

Lee JS, Nair NU, Dinstag G *et al.* Synthetic lethality-mediated precision oncology via the tumor transcriptome. *Cell* 2021; **184**:2487–502.e13.

Lee YR, Chen M, Pandolfi PP. The functions and regulation of the PTEN tumour suppressor: new modes and prospects. *Nat Rev Mol Cell Biol* 2018;**19**:547–62.

Li B, Zhang G, Xu X. APC mutation correlated with poor response of immunotherapy in Colon cancer. *BMC Gastroenterol* 2023; **23**:95.

Li GM. Mechanisms and functions of DNA mismatch repair. *Cell Res* 2008;**18**:85–98.

Lieb S, Blaha-Ostermann S, Kamper E *et al.* Werner syndrome helicase is a selective vulnerability of microsatellite instability-high tumor cells. *Elife* 2019;**8**:e43333.

Liu J, Lichtenberg T, Hoadley KA, Cancer Genome Atlas Research Network *et al.* An integrated TCGA Pan-Cancer clinical data resource to drive High-Quality survival outcome analytics. *Cell* 2018;**173**:400–16.e11.

Lord CJ, Ashworth A. PARP inhibitors: synthetic lethality in the clinic. *Science* 2017;**355**:1152–8.

Lord CJ, Quinn N, Ryan CJ. Integrative analysis of large-scale loss-of-function screens identifies robust cancer-associated genetic interactions. *Elife* 2020;**9**:e58925.

Maletzki C, Stier S, Linnebacher M. Microsatellite instability in hematological malignancies: hypermutation vs. immune control-who is challenging who? *Oncoimmunology* 2013; **2**:e25419.

Markowska M, Budzinska MA, Coenen-Stass A *et al.* Synthetic lethality prediction in DNA damage repair, chromatin remodeling and the cell cycle using multi-omics data from cell lines and patients. *Sci Rep* 2023;**13**:7049.

Mateo J, Lord CJ, Serra V *et al.* A decade of clinical development of PARP inhibitors in perspective. *Ann Oncol* 2019;**30**:1437–47.

McDonald ER, 3rd, de Weck A, Schlabach MR *et al.* Project DRIVE: a compendium of cancer dependencies and synthetic lethal relationships uncovered by Large-Scale, deep RNAi screening. *Cell* 2017;**170**:577–92.e10.

Megchelenbrink W, Katzir R, Lu X *et al.* Synthetic dosage lethality in the human metabolic network is highly predictive of tumor growth and cancer patient survival. *Proc Natl Acad Sci U S A* 2015;**112**:12217–22.

Mina M, Iyer A, Tavernari D *et al.* Discovering functional evolutionary dependencies in human cancers. *Nat Genet* 2020;**52**: 1198–207. https://doi.org/10.1038/s41588-020-0703-5

Monnat RJ. Jr. Human RECQ helicases: roles in DNA metabolism, mutagenesis and cancer biology. *Semin Cancer Biol* 2010; **20**:329–39.

Moschetta M, Doi T, Kuboki Y *et al.* 719TiP a phase I/Ib study of the werner (WRN) helicase inhibitor HRO761 as single agent and in combination with irinotecan or tislelizumab in patients with microsatellite instability-high (MSIhi) or mismatch repair deficient (dMMR) advanced solid tumors. *Ann. Oncol* 2023; **34**:S496.

Mullen J, Kato S, Sicklick JK *et al.* Targeting ARID1A mutations in cancer. *Cancer Treat Rev* 2021;**100**:102287.

Neggers JE, Paolella BR, Asfaw A *et al.* Synthetic lethal interaction between the ESCRT paralog enzymes VPS4A and VPS4B in cancers harboring loss of chromosome 18q or 16q. *Cell Rep* 2021;**36**:109367.

O'Neil NJ, Bailey ML, Hieter P. Synthetic lethality and cancer. *Nat Rev Genet* 2017;**18**:613–23.

Oh JH, Jang SJ, Kim J *et al.* Spontaneous mutations in the single TTN gene represent high tumor mutation burden. *NPJ Genom Med* 2020;**5**:33.

Pacini C, Duncan E, Goncalves E *et al.* A comprehensive clinically informed map of dependencies in cancer cells and framework for target prioritization. *Cancer Cell* 2024;**42**:301–16.e9.

Patro CPK, Biswas N, Pingle SC *et al.* MTAP loss: a possible therapeutic approach for glioblastoma. *J Transl Med* 2022;**20**:620.

Picco G, Cattaneo CM, van Vliet EJ, UK Genom. Cell Model Network *et al.* Werner helicase is a Synthetic-Lethal vulnerability in mismatch Repair-Deficient colorectal cancer refractory to targeted therapies, chemotherapy, and immunotherapy. *Cancer Discov* 2021;**11**:1923–37.

Rancati G, Moffat J, Typas A *et al.* Emerging and evolving concepts in gene essentiality. *Nat Rev Genet* 2018;**19**:34–49.

Rao RC, Dou Y. Hijacked in cancer: the KMT2 (MLL) family of methyltransferases. *Nat Rev Cancer* 2015;**15**:334–46.

Rao S, Peri S, Hoffmann J *et al.* RPL22L1 induction in colorectal cancer is associated with poor prognosis and 5-FU resistance. *PLoS One* 2019;**14**:e0222392.

Riedinger CJ, Esnakula A, Haight PJ *et al.* Characterization of mismatch-repair/microsatellite instability-discordant endometrial cancers. *Cancer* 2024;**130**:385–99.

Ryan CJ, Devakumar LPS, Pettitt SJ *et al.* Complex synthetic lethality in cancer. *Nat Genet* 2023;**55**:2039–48.

Ryan CJ, Lord CJ, Ashworth A. DAISY: picking synthetic lethals from cancer genomes. *Cancer Cell* 2014;**26**:306–8.

Ryan CJ, Mehta I, Kebabci N *et al.* Targeting synthetic lethal paralogs in cancer. *Trends Cancer* 2023;**9**:397–409.

Sato C, Kawakami H, Tanaka R *et al.* Survival impact of microsatellite instability in stage II gastric cancer patients who received S-1 adjuvant monotherapy after curative resection. *Sci Rep* 2023;**13**:10826.

Shao C, Westermann F, Höfer T. 2016. Synlet: an R package for systemically analyzing synthetic lethal RNA interference screen data. *bioRxiv*, https://doi.org/10.1101/059345

Shao S, Murray J, Brown A *et al.* Decoding mammalian ribosome-mRNA states by translational GTPase complexes. *Cell* 2016;**167**:1229–40.e15.

Shimada K, Bachman JA, Muhlich JL *et al.* shinyDepMap, a tool to identify targetable cancer genes and their functional connections from cancer dependency map data. *Elife* 2021; **10**:e67073.

Sondka Z, Dhir NB, Carvalho-Silva D *et al.* COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Res* 2024;**52**:D1210–D1217.

Spiliopoulou P, Yang SYC, Bruce JP *et al.* All is not lost: learning from 9p21 loss in cancer. *Trends Immunol* 2022; **43**:379–90.

Srivatsa S, Montazeri H, Bianco G *et al.* Discovery of synthetic lethal interactions from large-scale pan-cancer perturbation screens. *Nat Commun* 2022;**13**:7748.

Tanaka T, Watanabe T, Kazama Y *et al.* Chromosome 18q deletion and Smad4 protein inactivation correlate with liver metastasis: a study matched for T- and N- classification. *Br J Cancer* 2006;**95**:1562–7.

Tate JG, Bamford S, Jubb HC *et al.* COSMIC: the catalogue of somatic mutations In cancer. *Nucleic Acids Res* 2019; **47**:D941–D947.

Torti D, Trusolino L. Oncogene addiction as a foundational rationale for targeted anti-cancer therapy: promises and perils. *EMBO Mol Med* 2011;**3**:623–36.

Tsherniak A, Vazquez F, Montgomery PG *et al.* Defining a cancer dependency map. *Cell* 2017;**170**:564–76.e16.

van Wietmarschen N, Sridharan S, Nathan WJ *et al.* Repeat expansions confer WRN dependence in microsatellite-unstable cancers. *Nature* 2020;**586**:292–8.

Wang J, Wu M, Huang X *et al.* SynLethDB 2.0: a web-based knowledge graph database on synthetic lethality for novel anticancer drug discovery. *Database (Oxford)* 2022;**2022**:baac044.

Wappett M, Harris A, Lubbock ALR *et al.* SynLeGG: analysis and visualization of multiomics data for discovery of cancer 'achilles heels' and gene function relationships. *Nucleic Acids Res* 2021;**49**:W613–W618.

Weinstein IB, Joe A. Oncogene addiction. *Cancer Res* 2008;**68**: 3077–80; discussion 3080.

Wooller SK, Pearl LH, Pearl FMG. Identifying actionable synthetically lethal cancer gene pairs using mutual exclusivity. *FEBS Lett* 2024;**598**:2028–39.

Yi X, Zhang C, Liu B *et al.* Ribosomal protein L22-like1 promotes prostate cancer progression by activating PI3K/akt/mTOR signalling pathway. *J Cell Mol Med* 2023;**27**:403–11.

Bioinformatics, 2026, 42, 1–17
https://doi.org/10.1093/bioinformatics/btag337
Original Paper