# CoRe: A robustly benchmarked R package for the identification of core fitness genes in genome-wide pooled CRISPR knock-out screens

Alessandro Vinceti[1], Emre Karakoc[2], Clare Pacini[2], Umberto Perron[1], Riccardo Roberto De Lucia[1], Mathew J. Garnett[2], Francesco Iorio[1,2,†]

[1] Human Technopole, Milano, Italy
[2] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK
[†]Correspondence: francesco.iorio@fht.org

## Abstract

**CRISPR-Cas9 recessive genome-wide pooled screens have allowed systematic explorations of weaknesses and vulnerabilities existing in cancer cells, across different tissue lineages at unprecedented accuracy and scale. The identification of novel genes essential for selective cancer cell survival is currently one of the main applications of this technology. Towards this aim, distinguishing genes that are constitutively essential (invariantly across tissues and genomic contexts, i.e. core-fitness genes) from those whose essentiality is associated with molecular features peculiar to certain cancers is of paramount importance for identifying new oncology therapeutic targets. This is crucial to assess the risk of a candidate target's suppression impacting critical cellular processes that are unspecific to cancer. On the other hand, identifying new human core-fitness genes might also elucidate new mechanisms involved in tissue-specific genetic diseases.**

**We present CoRe: an open-source R package implementing established and novel methods for the identification of core-fitness genes based on joint analyses of data from multiple CRISPR-Cas9 screens. In addition, we present results from a fully reproducible benchmarking pipeline demonstrating that CoRe outperforms other state-of-the-art methods, and it yields more reliable sets of core-fitness and common-essential genes with respect to existing reference sets and methods.**

# Background

The ability to perturb individual genes at scale in human cells holds the key to elucidating their function and it is a gateway to the identification of new therapeutic targets across human diseases, including cancer. In this context the CRISPR-Cas9 genome editing system is widely considered the state-of-the-art tool [1–3].

Several genome-scale CRISPR-Cas9 single guide RNA (sgRNA) libraries have been designed and are available to date for genetic perturbation screens in human cells, showing significantly improved precision and scale with respect to previous technologies [4–8]. Some of these libraries have been employed in large-scale in vitro screens assessing each gene's potential in reducing cellular viability/fitness upons inactivation, across hundreds of immortalised human cancer cell lines [7,9–12]. This robust approach has led to comprehensive identifications of cellular fitness genes, providing a detailed view of genetic dependencies and weakness spots existing in cancer cells.

A major goal of the aforementioned studies, and similar future efforts, is to classify and distinguish genetic dependencies involved in normal essential biological processes from disease- and genomic-context specific vulnerabilities. Identifying context specific essential genes, and distinguishing them from constitutively essential genes shared across all tissues and cells, i.e. *core-fitness genes*, is crucial for elucidating the mechanisms involved in tissue-specific diseases. Moving forward and focusing on very well-defined genomic contexts in tumors, this allows identifying cancer synthetic lethalities that could be exploited therapeutically [13]. In fact, genes essential in cells with a tumour specific molecular feature should make ideal therapeutic targets with high effectiveness and selectivity, thus minimal side effects.

Gene dependency profiles, generated via pooled CRISPR-Cas9 screening across large panels of human cancer cell lines, are becoming increasingly available [14,15]. However, identifying and discriminating core-fitness and context-specific essential genes from this type of functional genetics screens remains a not trivial task.

The recently-described *Daisy Model* (DM) aims to identify core-fitness genes (CFGs) by jointly analysing data from genetic screens of multiple cancer cell lines. In this approach, sets of fitness genes for each screened cancer cell line are conceptually represented by the petals of a daisy [10]. These sets have different extents of overlap, but they generally tend to share a common set of CFGs (the core of the daisy). Based on this idea, genes that are essential in most of the screened cell lines are predicted to be CFGs. This approach has been shown to identify CFGs that are enriched for fundamental cellular processes such as transcription, translation, and replication [10]. Nevertheless, in [10] the minimal number of cell lines (3 out of 5 screened cell lines), in which a gene should be significantly essential in order to be predicted as CFG, is arbitrarily defined with no indications on how to determine this threshold on a numerically grounded basis when applying the DM to larger collections of screens.

To overcome this limitation, in [12] we have introduced the *Adaptive Daisy Model* (ADaM): a generalisation of the DM which is able to determine the minimal threshold on the number of cell lines that are dependant on the putative CFGs in a semi-supervised manner, via a joint analysis of a large number of multiple CRISPR-Cas9 screens. ADaM first identifies multiple sets of tissue specific CFGs, then it iterates the process across these gene sets to identify a set of pan-cancer CFGs.

We have also recently proposed an alternative unsupervised approach within the Broad and Sanger Institutes' Cancer Dependency Map collaboration [16,17], where data from screening hundreds of cell lines are analyzed in a pooled fashion, independently of their tissue of origin. This method builds on the basic intuition that if a gene is universally essential then it should rank among the top essential genes in the vast majority of screened models, including those that are the least dependant on it, or generally showing a moderate to weak loss-of-fitness phenotype upon CRISPR-Cas9 targeting.

Finally, a logistic regression based method for classifying genes into CFGs or context specific essentials has been recently introduced by Sharma and colleagues [18] as part of the CEN-tools suite, using reference sets of essential and non-essential genes for the training phase [19].

Although the number of CRISPR-Cas9 and genome-scale RNAi experiments is increasing rapidly, no robustly benchmarked method to identify sets of CFGs has been devised yet in a unique and easy-to-use software package.

We present *CoRe*: an R package implementing recently proposed as well as novel versions of algorithms for the identification of CFGs from a joint analysis of multiple genome-wide pooled CRISPR-Cas9 knock-out screens. Furthermore, we present results from a comparison of CoRe's output (when applied to the largest integrative cancer dependency dataset generated to date [20]) against widely used [10,19], or more recent [18] sets of CFGs obtained via an alternative approach (also tested on the same recent cancer dependency dataset). We report an increased coverage of prior known human essential genes, new potential core-fitness genes, and lower false positive rates for CoRe's methods with respect to other state-of-the-art core-fitness sets and available methods. Finally CoRe's methods are computationally more efficient than others and the CFGs obtained with CoRe could be used in the future as a template classifier of a single screen's specific essential genes, via supervised classification methods, such as BAGEL [21].

## Results

### *Overview of the CoRe package and implemented methods*

We have developed and extensively benchmarked *CoRe*: an R package able to identify core fitness genes (CFGs) from the joint analysis of multiple genome-wide CRISPR knock-out screens. CoRe implements two methods at two different levels of stringency yielding, respectively, two types of gene sets, here referred for simplicity as (i) CFGs and (ii) common essential genes (CEGs), with the first set reflecting a higher level of stringency/confidence.

The first and more stringent method implemented in CoRe is the *Adaptive Daisy Model* (ADaM) [12]: an adaptive version of the *Daisy Model* (DM) [10] that operates in a cascade of two steps, and it is usable on data coming from large-scale CRISPR-cas9 knock-out screens performed in heterogeneous *in vitro* models, for example immortalized human cancer cell lines from multiple tissue lineages (**Fig. 1A-D**)

The first step of ADaM identifies CFGs on a tissue/cancer-type basis, defining them as those exerting a significant fitness effect upon CRISPR-Cas9 targeting on a minimal number of cell lines, which is adaptively determined. In the second step, ADaM identifies as pan-cancer CFGs those that are called as tissue/cancer-type specific CFGs by the first step for a minimal number of tissues/cancer-types, also adaptively determined (**Fig. 1D**).

The second and less stringent method, implemented in CoRe in four different novel variants, is the *Fitness Percentile* (FiPer), which identifies CEGs via a pooled (pan-cancer) analysis of data from large-scale CRISPR-Cas9 knock-out screens, performed in cell lines from multiple tissues/cancer-types [16] (**Fig. 1EF**). For each screened cell line, this approach considers the gene rank positions resulting from sorting them based on their fitness effect upon inactivation, in decreasing order. FiPer then exploits the intuition that CEGs will always rank among the top fitness genes for the vast majority of cell lines, including those for which the fitness reduction is overall less pronounced.

While ADaM takes as input strictly defined binary scores of gene essentiality and it outputs discrete sets of tissue-specific and pan-cancer CFGs, FiPer takes as input quantitative descriptors of gene essentiality and it outputs a unique set of CEGs, providing also a visual means for quickly assessing the tendency of individual genes to be a CEG.

CoRe is publicly available as an open source R package at https://github.com/DepMap-Analytics/CoRe. An interactive vignette, with demonstrations and examples is available at https://rpubs.com/AleVin1995/CoRe. The package includes built-in visualisation and benchmarking functions and their related data objects. It also contains interface functions for downloading and processing state-of-the-art cancer dependency datasets from *Project Score* [15], as well as updated cancer cell line annotations from the *Cell Models Passports* [22]. Finally, results from benchmarking CoRe against state-of-the-art sets of CFGs and other CFGs identification methods, with corresponding figures, are fully reproducible executing the Jupyter

notebook (also compatible with Google CoLab) available at:

https://github.com/DepMap-Analytics/CoRe/blob/master/notebooks/CoRe_Benchmarking.ipynb.

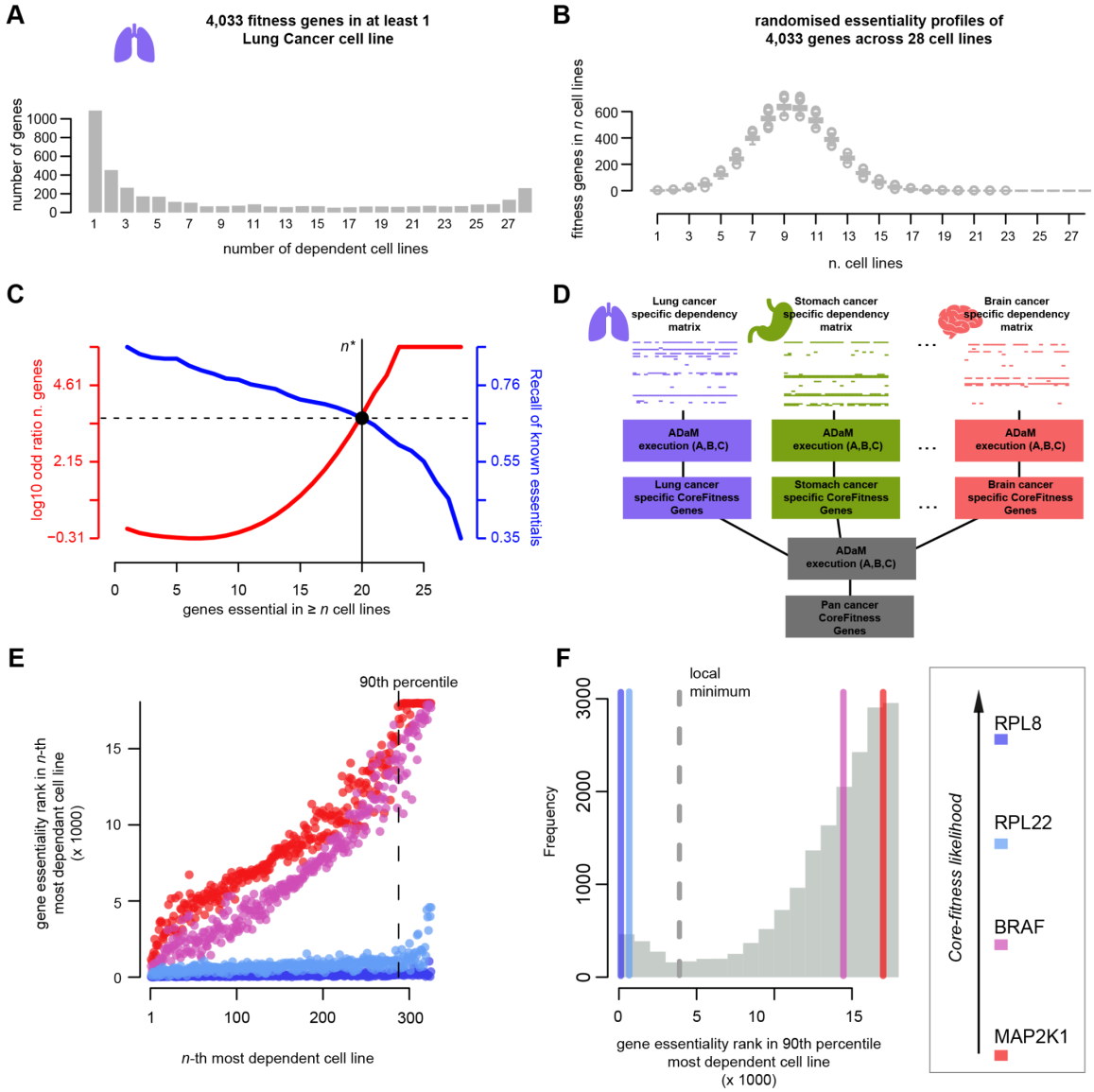

**Fig. 1 - Overview of the methods implemented in CoRe. A.** Typical bimodal distribution of number of fitness genes across fixed numbers of cell lines from a given tissue lineage. **B.** Number of fitness genes in a fixed number of cell lines across 1,000 randomised versions of a binary cancer dependency matrix. **C**. Optimisation criteria implemented in ADaM. The blue curve indicates the recall of a reference set of prior known essential genes computed across sets of genes that are essential in at least *n* cell lines. The red curve indicates, for each *n* in the x-axis, the deviance from expectations of the number of genes that are essential in at least *n* cell lines (derived from the simulations in B). The *n** corresponds to the trade-off between these two quantities and it estimates the minimal number of cell lines in which a gene should be essential in order to be predicted as a core-fitness essential gene. **D**. Schematic of ADaM execution to identify pan-cancer core-fitness essential genes. The first iteration adaptively determines sets of core-fitness genes across tissues/cancer-types. The second iteration computes pan-cancer core-fitness as those predicted as tissues/cancer-type specific core-fitness genes for at least *t** tissues/cancer-types. Where *t** is determined as *n** in C. **E.** CoRe.VisCFness visualisation of fitness percentile (FiPer) curves for four example genes (RPL8, RPL22, BRAF and MAP2K1), showing their common-essentiality likelihood. Each point indicates a screened cell line, with coordinates corresponding to the rank of that cell line based on its dependency on the gene under consideration and the rank of the gene under consideration based on its fitness effect in that cell line (when considering all screened genes), respectively for x- and y-axis. **F.** Distribution of all genes' fitness-rank-positions in their 90th-percentile most dependent cell line. The density of these scores is estimated using a Gaussian kernel and the central point of minimum density is identified. Genes whose score falls below the local minimum are classified as common essential.

## *The Adaptive Daisy Model*

The Adaptive Daisy Model (ADaM) [12] is implemented in the function CoRe.ADaM, which takes as input (i) a binary dependency matrix, where rows correspond to genes and columns to samples (screens or cell-lines), with a 1 in position $[i, j]$ indicating that the inactivation of the *i*-th gene through CRISPR-Cas9 targeting exerts a significant loss of fitness in the *j*-th sample, i.e. that the *j*-th cell line is dependent on the *i*-th gene; (ii) a reference set of prior known CFGs. Binary dependency matrices encompassing data for hundreds of cancer cell lines can be downloaded from Project Score [15] and used with this function by calling CoRe.download_BinaryDepMatrix.

In order to identify CFGs using data from screening *N* cell lines, the Daisy Model introduced in [10] computes a fuzzy intersection of genes that are essential, i.e. fitness genes, in at least *n** cell lines, where this number is defined a priori (typically a number corresponding to a large majority of cell lines). ADaM generalizes this approach by (i) exploiting the bimodality of the distributions of the number of genes essential in a given number of cell lines (**Fig. 1A**), and (ii) adaptively determining an

optimal discriminative threshold of minimal number of cell lines $n^*$ that should be dependent on a given gene in order for calling that gene a CFG.

Briefly, for a binary matrix encompassing gene dependency profiles of $n$ cell lines across thousands of screened genes, ADaM computes fuzzy intersections of genes $I_n$, for each $n = 1, \ldots, N$. These fuzzy intersections include genes with at least $n$ dependent cell lines according to the input matrix. For each tested $n$, ADaM computes the true positive rate $TPR(n)$ yielded by each $I_n$ using the reference CFGs provided in input as positive controls. In parallel, ADaM also computes the number of genes that are expected to be essential in at least $n$ cell lines by chance, by randomly perturbing the input matrix a large number of times (shuffling the entries of each column) (**Fig. 1B**). Finally, ADaM determines the optimal $n^*$ as the largest value providing the trade-off between $TPR(n)$ (inversely proportional to $n$) and the deviance of the number of genes with $n$ dependent cell lines (directly proportional to $n$) from its expectation (**Fig. 1C**). The genes in the corresponding fuzzy intersection $I_{n^*}$ are predicted to be CFGs for the cell lines in the input dependency matrix.

As the distribution of genes that are CFGs in a specific number of tissue-lineage/cancer-types is also bimodal [12], this procedure can be executed in a two step approach on large datasets of cancer dependency profiles, accounting for hundreds of cancer cell lines from multiple tissues, to predict pan-cancer CFGs. In the first step ADaM predicts tissue-lineage/cancer-type specific CFGs, then it iterates by adaptively determining the minimum number $t^*$ of tissue-lineages/cancer-types for which a gene should have been predicted as a specific CFG in order to be now predicted as a pan-cancer CFG. $t^*$ is determined by applying the same algorithm and criteria used to determine the $n^*$ across the tissue-lineages/cancer-types specific executions of ADaM (**Fig. 1D**). Particularly, this last operation is performed on a binary membership matrix with genes on the rows, tissue-lineages/cancer-types on the column and a 1 in position $[i, j]$ indicating that the $i$-th gene is a CF for $j$-th tissue-lineage/cancer-type.

All the functions called by CoRe.ADaM are exported and fully documented in the CoRe package. In addition, CoRe is equipped with the CoRe.PanCancer_ADaM wrapper function, implementing the two-step procedure to identify pan-cancer CF genes, and the CoRe.CS_ADaM

function executing ADaM on a user-defined tissue-lineage/cancer-type, which can be used on dependency matrices from Project Score [15] and cell line annotations from the Cell Model Passports [22].

### *The Fitness Percentile Method*

Differently from the ADaM method, the Fitness Percentile (FiPer) method works in an unsupervised manner. It identifies a set of common essential genes (CEGs) by executing a single pooled analysis of data from multiple CRISPR-Cas9 screens. In addition, it takes as input a dependency matrix with quantitative fitness effect indicators of screened genes across cell lines.

We have designed and implemented in CoRe four novel variants of this method, all sharing the same initial step, which is executed for each individual gene in the input dependency matrix in turn. In this step (i) all cell lines are sorted according to their dependency on the gene under consideration in decreasing order; (ii) the rank position of the gene under consideration resulting from sorting all screened genes according to their fitness effect is determined, for each screened cell line; (iii) a curve of the rank positions computed in (ii) is assembled considering the cell lines ordered as in (i): the fitness rank versus dependency percentile curve (FiPer curve, **Fig. 1E**).

It is reasonable to assume that genes involved in fundamental cellular processes (likely to be CEGs, such as RPL8 and RPL22 in **Fig. 1E**) will generally tend to rank amongst the most significant fitness genes for all the screened cell lines, including those that are the least dependent on them. This tendency can be extrapolated from the FiPer curves (thus measured in data coming from multiple CRISPR-Cas9 screens) and used to estimate the likelihood of a gene to be a CEG.

The CoRe.FiPer function implements four different methods to assess this tendency assigning a FiPer score to each gene differently. This is followed by a procedure that finally partitions all screened genes into two groups, with the first one containing the predicted CEGs.

The first method, the *Fixed percentile* (**Fig. 1EF**), considers as the FiPer score of a gene its fitness rank position in the cell line falling at the highest boundary of a very large dependency percentile of

cell lines (90th by default). The *Average* method considers the average gene rank position in all the cell lines falling over a very large dependency percentile (90th by default). The *Slope* method fits a linear model onto each gene's FiPer curve, then considers the slope of such a model as the gene FiPer score. In the final *AUC* method, the FiPer score of a gene is computed as the area under its FiPer curve.

Finally, a density function is determined with a kernel estimator and fitted onto the gene FiPer scores' observed distribution (which is typically bimodal) and the score corresponding to the point of central local minimal density is used as a discriminative threshold to predict CE genes, which will be those with a FiPer score less than or equal to it (**Fig. 1F**).

CoRe includes also the CoRe.VisCFness function which visualises the tendency of a given gene to be a CEG within a dependency dataset provided in input, and compares this tendency against that of a positive (RPL8 by default) and a negative (MAP2K1 by default) control, and producing the plots shown in **Fig. 1E**.

### *Comparison with existing methods and state-of-the-art sets of core-fitness genes*

We compared the sets of CFGs and CEGs predicted by CoRe when applied to the largest integrative dataset of cancer dependency assembled to date [20] with state-of-the-art sets of core-fitness genes derived from recent functional genetic screening datasets [10,12,18,19], as well as with the output of a logistic-regression based method, part of the recent CEN-tools software proposed in [18], applied to the same dataset [20].

Collectively, we considered 5 state of the art sets of CFGs, illustrated in Table 1.

| Set name | Set Type | Description and Source | Dataset of origin and method |
|---|---|---|---|
| *Hart2014* | State-of-the-art reference set of core-fitness essential genes | A set of 360 genes presented in [23] and used as a classification template by BAGEL: a supervised computational framework for quantifying gene essentiality significance in pooled library screens [10,21]. | Large collection of shRNA gene dependency profiles analysed with a linear algebra approach. |
| *Hart2017* | State-of-the-art reference set of core-fitness essential genes | A set of 684 genes introduced in [19]. | BAGEL reanalysis of 17 genome-scale knockout screens in human cell lines performed with different libraries. |
| *Behan2019* | State-of-the-art reference set of core-fitness essential genes | A set of 553 genes presented in [12]. | ADaM analysis of a large collection of gene dependency profiles from CRISPR-screens of 325 human cancer cell lines from different tissue-lineages/cancer-types (now part of the Project Score database [15]), using a manually curated version of the Hart2014 set (the *curated Hart2014* CFGs), as training. This was obtained by excluding from the Hart2014 set 34 genes, such as for example KRAS and CHD4, predicted to be cancer drivers by the intOGen pipeline [24,25] |
| *Sharma2020* | State-of-the-art reference set of core-fitness essential genest | A set of 519 genes presented in [18]. | Logistic regression approach (part of the CEN-tools software), which uses the BAGEL essential/never-essential genes as training, respectively the Hart2017 set and a set of 927 never-essential genes [10,21]. This approach was individually applied to the dependency profiles from Project Score [15] and from the Broad DepMap portal [26] (Release 19Q2). The final predicted set was composed of genes predicted as CFGs in the two analyses, excluding those in the training set. For the comparison with the unsupervised methods, this set was joined with the Hart2017 set (used in its training phase), rising up to 1,182 genes. |

**Table 1 - State of the art sets of core-fitness essential genes considered to benchmark CoRe.**

Furthermore, we considered new sets of genes (**Table 2**) yielded by executing the CEN-tools logistic regression method [18] and the CoRe methods (ADaM and all the variants of FiPer, as detailed in the Methods) on the largest integrative dataset of cancer dependency assembled to date [20]. This dataset is composed of dependency matrices accounting for 17,486 genes and 855 cell lines from 30 different tissue-lineages and 43 cancer types (the DepMap dataset, **Fig. 2AB)**.

For the training phase of CEN-tools, we used the curated Hart2014 CFGs [12], which we also used as reference set of positives while running ADaM, and the BAGEL never-essential genes [10], also curated as described in [12] (the curated BAGEL non-essential set). In order to provide a fair benchmark with respect to sets outputted by the unsupervised methods, we also joined the Sharma2020 set and the CEN-tools set with the reference CFGs used in their respective training phases, i.e. the Hart2017 set and the curated Hart2014 set. All the compared sets of CFGs and CEGs,

the curated Hart2014 essential and curated BAGEL non-essential genes are included in **Supplementary Table 1.**

| Set name | Set Type | Number of genes | Dataset of origin |
|---|---|---|---|
| CEN-tools | Novel analysis | 756 [For the comparison with the unsupervised methods, this set was joined with the curated Hart2014 set (used in its training phase), rising up to 1,082 genes] | DepMap dataset [20] |
| CoRe ADaM | Novel analysis | 1,075 | DepMap dataset [20] |
| CoRe FiPer average | Novel analysis | 1,424 | DepMap dataset [20] |
| CoRe FiPer slope | Novel analysis | 1,704 | DepMap dataset [20] |
| CoRe FiPer AUC | Novel analysis | 1,987 | DepMap dataset [20] |
| CoRe FiPer Fixed | Novel analysis | 1,947 | DepMap dataset [20] |
| CoRe FiPer consensual | Novel analysis | 1,673 | DepMap dataset [20] |

**Table 2 - Sets of core-fitness and common-essential genes obtained by novel analyses of the DepMap dataset and considered to benchmark CoRe.**

Amongst the predicted CFG sets derived from old and new executions of supervised methods, ADaM yielded the largest number of CFGs (460) not included in any of the training sets (curated Hart2014, Hart2017 and never-essentials), when applied to the DepMap dataset (**Fig. 2A**). The Sharma2020 set ranked second (with 441), followed by the novel execution of the CEN-tools (379)(**Fig. 2A**). As expected, all these sets, included more novel CFGs than Behan2019 (157), likely due to its derivation from a sensibly smaller cancer dependency dataset (325 cell lines against 855 for ADaM and CEN-tools, and 325 + 489 for Sharma2020, **Fig. 2A**).

The 4 variants of the CoRe FiPer method yielded much larger and highly concordant sets of predicted CEGs (median = 1,825.5, min = 1,424 for FiPer average, max = 1,987 for FiPer AUC, **Fig. 2B**), as well as novel hits (median = 1,115, min = 743 for FiPer average, max = 1,262 for FiPer AUC, **Fig. 2B**). The set of CEGs predicted by FiPer average was included in those predicted by all the other FiPer variants. For this reason we decided to assemble a 5th FiPer set by intersecting the output of FiPer Slope, AUC and Fixed: the FiPer consensual set. This yielded 1,673 genes, of which 975 were novel hits (**Fig. 2A**).

All the sets of CFGs/CEGs outputted by the CoRe methods covered most of the state-of-the-art sets of CFGs (ADaM median Recall across prior known sets: 77.24%, FiPer median Recall across prior known sets, averaged across variants: 89.31%, **Fig. 2C**).

While comparing overall CFG/CEG sets similarities, we observed three major clusters composed respectively by (i) the sets outputted by the FiPer variants, then (ii) Sharma2020, CEN-tools (both joined with respective training sets) and ADaM sets, and (iii) Hart2014, Hart2017 and Behan2019 sets (**Supplementary Fig. 1AB).** Taken together, these results suggest that the ADaM, CEN-tools and Sharma2020 sets might include similar numbers of novel CFGs, thus potentially extending in a similar way the other state-of-the-art CFG sets.

To investigate and compare true/false positives rates of the putative novel CFG/CEGs, we assembled, respectively, (i) a set of prior known CFGs (not included into any of the training sets) to be used as positive controls, and (ii) considered genes not expressed in human cancer cell lines or whose essentiality is statistically associated with a molecular feature (thus very likely to be linked to specific molecular contexts) as negative controls.

To assemble the set of positive controls, we collected signatures of genes involved in fundamental biological processes and universally essential genes: such as genes coding for ribosomal proteins, RNA polymerases, histones, or genes involved in DNA replications, etc. curated in [27] and [20] from MsigDB [28]. As negative controls we assembled a set of genes never expressed (fragments per kilobase of transcript per million mapped reads (FPKM) < 0.1) in more than 1,000 human cancer cell lines (from the Cell Model Passports [22]), or whose fitness signal across hundreds of cell lines has a t-skewed normal distribution (according to the normLRT score introduced and applied to an independent shRNA-based cancer dependency dataset in [29]) and it is statistically associated with a genomic marker [20]. Excluding genes included in at least one of the training sets yielded a final set of 408 positive controls and 7,767 negative controls (**Supplementary Table 2**). Of these, 265 positive controls and 555 negative controls were included in the DepMap dataset.

Of the CGFs outputted by the supervised methods, ADaM had the best TPR, covering 29% of the positive controls included in the DepMap dataset. Sharma2020 ranked second (23.4%) followed by CEN-tools (23%) and Behan2019 (15%) (**Fig. 2D**). The median TPR for the FiPer variants was 47%,

with FiPer AUC ranking first (54%) and FiPer Average last (42%). In terms of FPRs, Behan2019 performed the best, covering only 1.2% of the negative controls included in the DepMap dataset. ADaM ranked second (1.5%), followed by CEN-tools (1.7%) and Sharma2020 (2.3%). The median relative FPR for the FiPer variants was equal to 4% with FiPer average performing best (2.5%) and FiPer fixed worst (7%).

To account for differences in set sizes, which impact the observed TPRs/FPRs, we sought to compare the observed FPRs with those expected when using a baseline daisy model (DM) predictor of CFGs on the DepMap dataset, considering the thresholds $n^*$ providing the observed TPRs of independent positive controls (**Fig. 2E** and **Supplementary Fig. 2A-D**).

When considering the supervised methods, CoRe outperformed both CEN-tools and Sharma2020, yielding better ratios of FPRs divided by those obtained at the observed TPRs by the DM (1.1 and 1.2 respectively for Behan2019 and ADaM, against 1.4 for CEN-tools and 1.8 for Sharma2020 (**Fig. 2F**)). Much better performances were obtained by the FiPer variants (median FPR / baseline ratio = 0.72) with FiPer AUC performing the best (0.64) and FiPer average the worst (0.83).

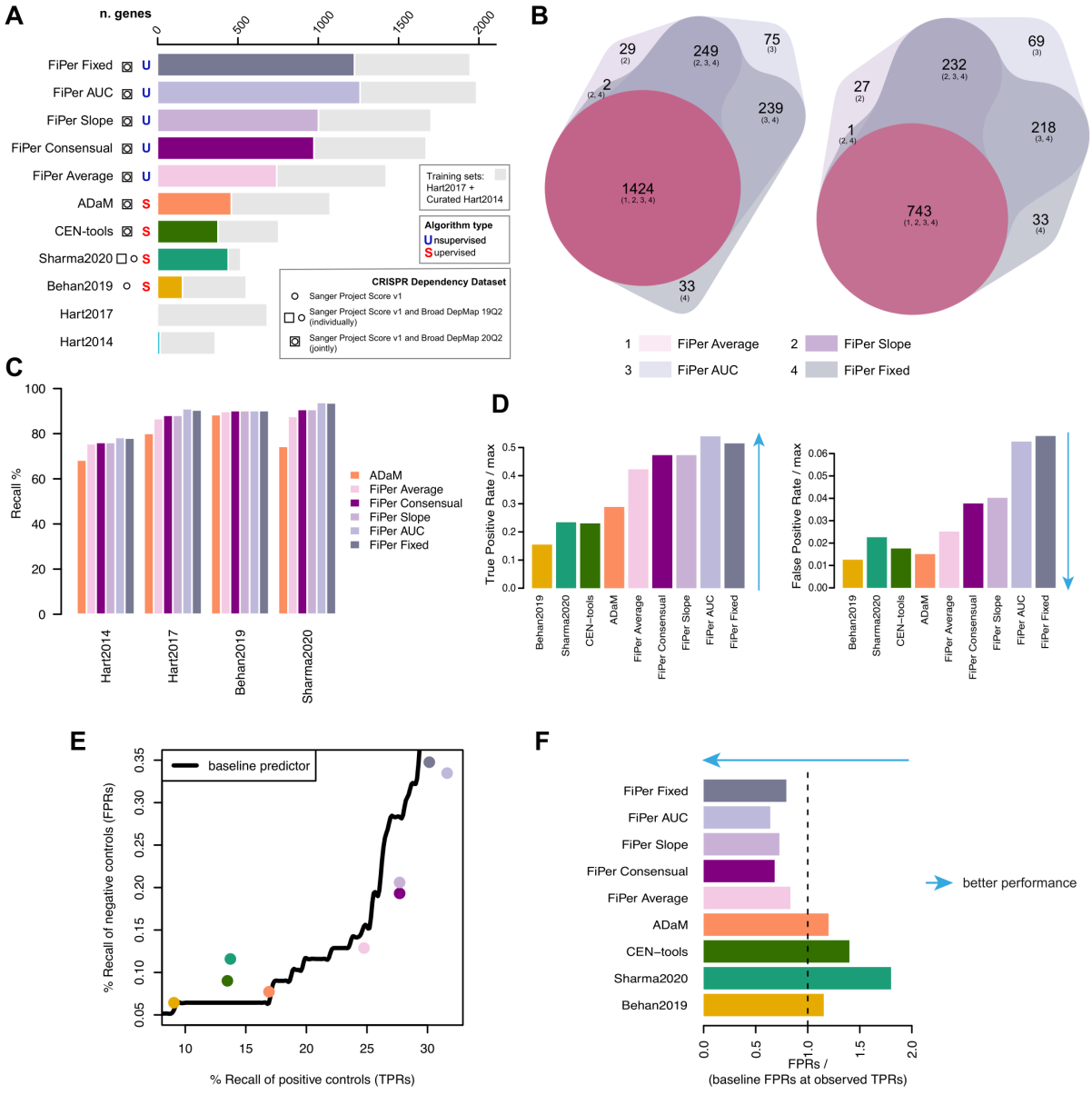

**Fig. 2 - Core fitness and common essential genes predicted by CoRe in comparison with state-of-the-art sets and those predicted by other methods. A.** For each method predicting core fitness essential genes (CFGs),common essential genes (CEGs), or state-of-the-art (SOA) sets of CFGs, the overall length of the bar indicates the total number of genes, whereas the length of the coloured bar indicates the total number of predicted genes not included in any of the training sets. Squares/circles indicate the dataset analysed by each method or used to derive the considered SOA set, and letters indicate the nature of the method, i.e. (S)upervised or (U)supervised. **B**. Comparison of common essential gene sets predicted by the four variants of the FiPer method (left) and considering novel hits only, i.e. excluding any gene belonging to any of the training sets (right). **C.** Recall of SOA sets of CFGs genes across CoRe methods' predictions. **D.** True and False positive rates (TPRs/FPRs) of independent true/negative controls across SOA sets of CFGs, CoRe and other methods, relative to the maximal TPRs/FPRs attainable by the basal daisy model (DM) predictor of CFGs. **E.** Performance assessment accounting for set sizes. Each point corresponds to a different method or SOA set, with coordinates indicating their TPR/FPR, respectively along x- and y-axis. Black curve indicates the FPRs obtained by a baseline DM predictor at given TPRs. **F.** FPRs of all tested methods and SOA sets of CFGs relative to baseline performances. The length of each bar indicates the ratio between the FPR of the method or set under consideration and that of the baseline DM classifier at a TPR equal to that observed for the method or set under consideration.

Optimal sets of CFGs/CEGs are expected to be essential in a vast majority of cancer cell lines: they have an average large negative impact on cellular fitness upon inactivation and are constitutively expressed in non-diseased tissues. To evaluate these properties across the output of compared methods and SOA sets, we first measured the median number of cell lines dependent on the predicted sets of CFGs/CEGs (**Fig. 3A**). This was generally high for all the supervised methods, with the Behan2019 CFGs being essential (scaled fitness score < -0.5, Methods) in a median percentage of 99.8% cell lines of the DepMap dataset, followed by CEN-tools (98.9%), ADaM (98.1%) and Sharma2020 (96.8%). As expected, the CEGs yielded by the FiPer variants, were generally essential in smaller but still large percentages of cell lines (grand median = 82.3%, min = 70.2% for FiPer AUC - max = 92% for FiPer average). Nevertheless, when looking at the $n^*$ thresholds required by the baseline DM to attain the observed TPRs across predicted CFGs/CEGs (**Fig. 3B**), among the supervised methods the ADaM set showed again the best ratio between median number of dependent cell lines versus baseline (1.14, 98.1% against 86%), followed by CEN-tools (1.06, 98.9% against 93%), Sharma2020 (1.05, 96.8% against 92%) and Behan2019 (1.01, 99.8% against 98.6%) (**Fig. 3C**). The FiPer variants CEGs showed a median ratio between number of dependent cell lines versus DM thresholds at same TPR

that was generally strikingly large across methods (median = 2.62, max 4.26 for FiPer AUC - min 1.95 for FiPer average).

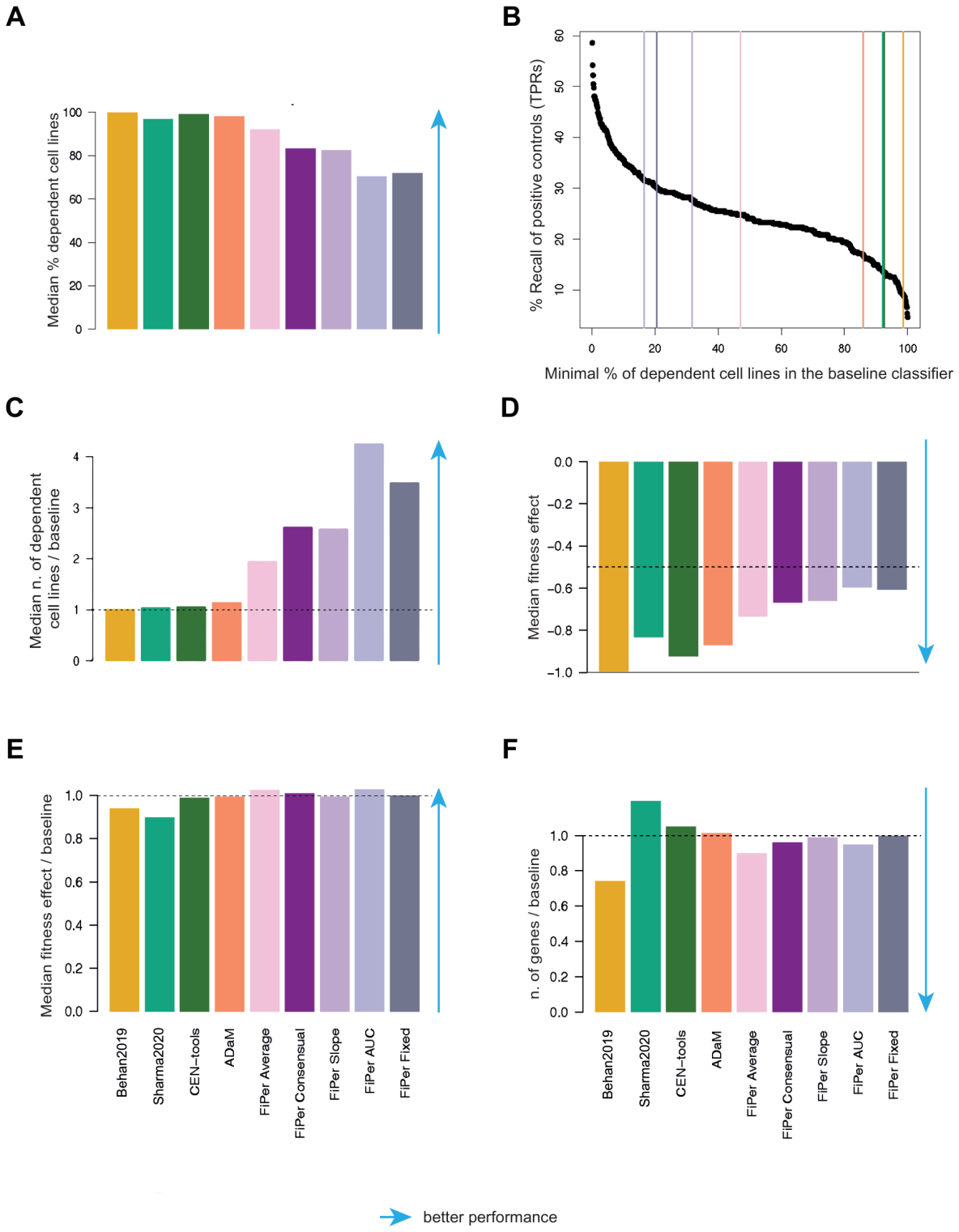

**Fig. 3 - Fitness effects of CFG sets across cell lines. A.** Median percentage of cell lines in which the genes in the predicted sets or core-fitness gene (CFG) or common essential gene (CEG) sets are significantly essential. **B.** Threshold of minimal number of dependent cell lines *n* required by the baseline daisy model predictor (DM) to attain the true positive rates (TPRs) observed across tested methods. **C.** Ratios between median numbers of dependent cell lines for predicted sets divided by the threshold *n* of the DM to attain their TPRs. **D.** Median fitness effect exerted by the genes in the predicted CFG/CEG sets. **E.** Ratio between the median fitness effect in D and the median fitness effect exerted by the DM at the observed TPRs. **F.** Ratio between the number of genes in the predicted sets and those predicted by the DM at the observed TPRs.

The proximity to 1 of all the ratios for the supervised methods indicate that, generally, they all implicitly discover the DM's optimal *n**. ADaM goes further and selects a set of genes providing a TPR that would require a much lax minimal number of dependent cell lines to be achieved by the DM, thus resulting in an increased FPR. Furthermore, in these circumstances, the unsupervised methods massively outperform the supervised ones, showing the effectiveness of the FiPer criteria used to pick CEGs.

Next, we measure the median scaled fitness effect of the predicted CFGs/CEGs across cell lines, and we find it comfortably below -0.8 -- i.e. 80% of the median effect for curated Hart2014 (Methods) -- for all the supervised methods (strongest effect = -0.99 for Behan2019, weakest for Sharma2020 = -0.83) and below -0.5, i.e. half the fitness effect of the curated Hart2014, for the FiPer variants (strongest for FiPer average = -0.73, weakest for FiPer AUC= -0.59) (**Fig. 3D**). Nevertheless, when comparing these values with their equivalent for the CFGs predicted by the baseline DM at the observed TPRs (excluding genes belonging to the training sets), ADaM was again the best performing supervised method (ratio between median fitness effect and baseline = 0.99), followed by CEN-tools (0.98), Behan2019 (0.93), and Sharma2020 (0.89). The median ratio for the FiPer variants was equal to 1.01 with FiPer AUC performing best (1.02)(**Fig. 3E**).

Finally, we found that all the compared methods predicted sets of CFGs/CEGs that were constitutively expressed in normal tissues at similar median levels (**Supplementary Fig. 3**). In addition, the CFG sets' cardinality was systematically comparable or lower than that of CFG sets

outputted by the baseline DM at the observed TPRs , with the exception of Sharma2020 and CEN-tools (**Fig. 3F**). Thus, these two sets were confirmed to be suboptimal and predicting larger numbers of CFGs with respect to the baseline DM but with worse FPRs at the observed TPRs (**Fig. 2EF**).

All these results were confirmed when the benchmark analyses were extended to the Hart2014 and Hart2017 sets, adding to CEN-tools and Sharma2020 their corresponding positive training sets and not excluding training set genes from positive/negative controls (thus considering 905 positive and 8,040 negative controls - of which respective 466 and 695 are in the DepMap dataset) (**Supplementary Fig. 4A-C**).

When considering all state-of-the-art sets of CFGs and supervised methods, we again established that ADaM provides the best TPRs and FPRs (both absolute and relative to baseline, **Fig. 4A-D**).

The Hart2014 set showed the best FPRs versus baseline ratio, although this had to be extrapolated. In fact, this set had a TPR (21.7%) that was lower than that of the baseline DM classifier at the most stringent $n^*$ threshold (TPR = 24%, for 343 CFGs that are significantly essential in 100% of the screened cell lines) (**Fig. 4C**), and strikingly did not include 66 positive controls that are significantly essential in all the cell lines of the DepMap dataset (**Fig. 4E**). These 66 genes were all covered by all the methods executed on the DepMap dataset and only partially recalled by the Hart2017 (73%), the Behan2019 (82%) and the Sharma2020 (94%) sets.

Taken together, these results strongly indicate that the CFGs derived from the DepMap dataset reliably extend state-of-the-art CFG sets and that, among those derived with supervised methods, the ADaM set is the most robust one. This was also confirmed in terms of number of cell lines dependent on the predicted CFGs (**Fig. 5AB**) and their median fitness effect (**Fig. 5CD**), relative to baseline performances.

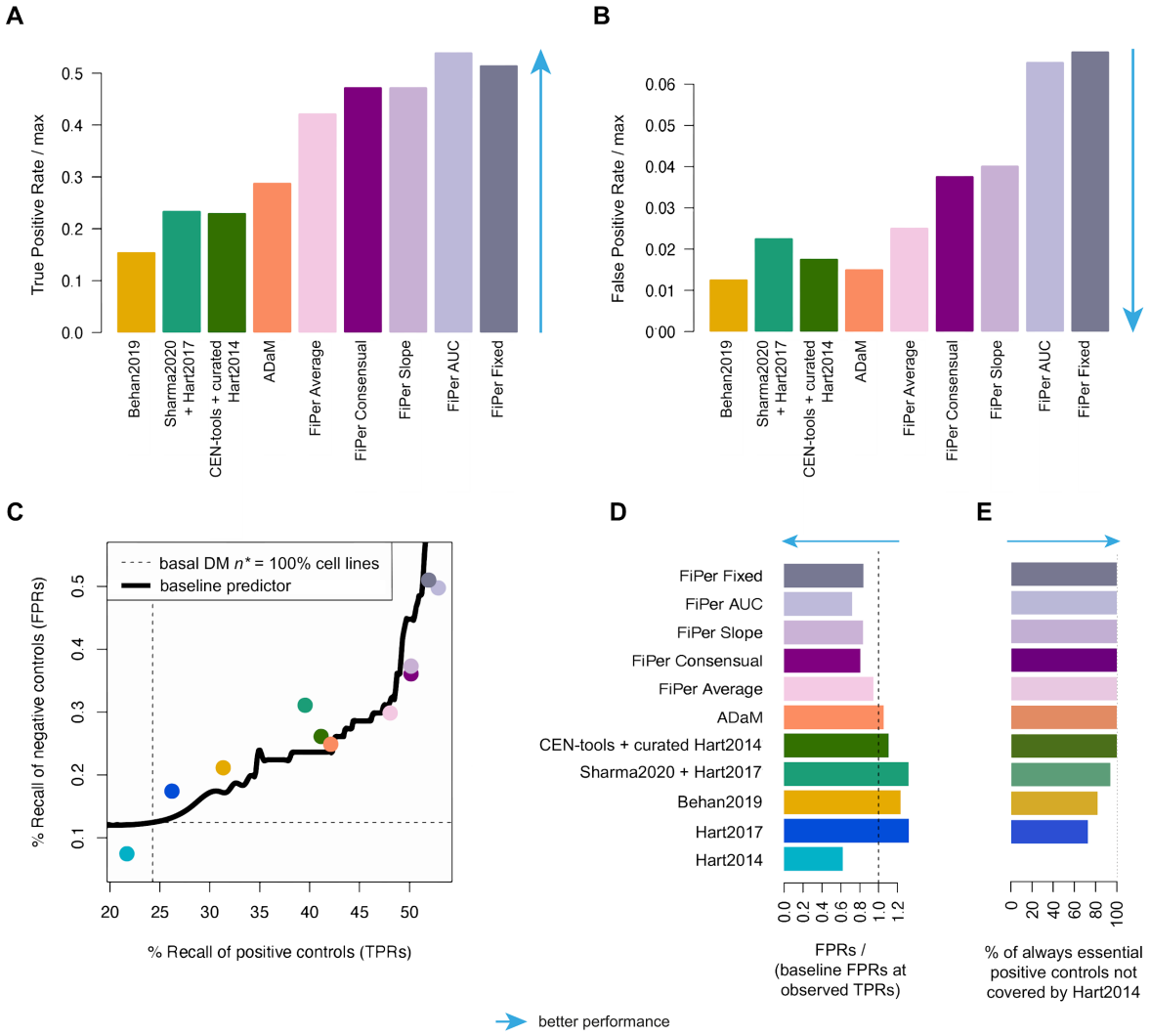


**Fig. 4 - Performances of tested methods when accounting for genes in the training sets. AB.** True and False positive rates (TPRs, FPRs) of independent true and negative controls across state-of-the-art (SOE) sets of core-fitness essential genes (CFGs), and sets outputted by CoRe and other methods, relative to the maximal TPRs/FPRs attainable by a basal daisy model (DM) predictor of CFGs. **C.** Performance assessment accounting for set size. Each point corresponds to a different method or SOA set, with coordinates indicating their TPR and FPR, respectively, along the x- and y-axis. The black curve indicates the FPRs obtained by a baseline DM predictor at given TPRs. **D.** FPRs of all tested methods and SOA sets of CFGs relative to baseline performances. The length of each bar indicates the ratio between the FPR of the set under consideration and that of the baseline DM classifier at a TPR equal to that observed for that set. **E.** Recall of positive control genes that are essential in 100% of the cell lines in the DepMap dataset and are not covered by the Hart2014 set, across all benchmarked sets.

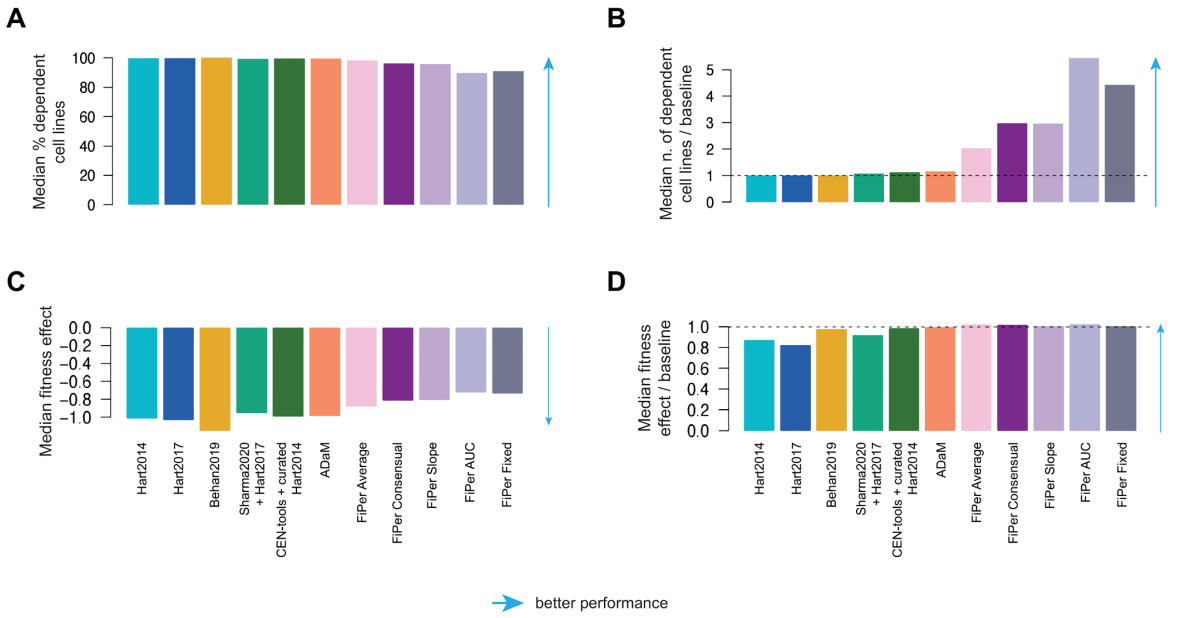


**Fig. 5 - Comparison between CFG/CEG sets' essentiality profile when accounting for genes in the training sets. A.** Median percentage of cell lines in which the genes in the predicted sets or core-fitness gene (CFG) or common essential gene (CEG) sets are significantly essential. **B.** Ratios between median numbers of dependent cell lines for predicted sets divided by the threshold *n* of the baseline daisy model predictor (DM) to attain their TPRs. **C.** Median fitness effect exerted by the genes in the predicted CFG/CEG sets. **D.** Ratio between the median fitness effect in D and the median fitness effect exerted by the DM at the observed TPRs.

## *Methods' performances using an independent cancer dependency dataset*

We sought to compare the CGF and CEG sets outputted by the considered methods in terms of their median fitness effect across multiple screened models when using an independent cancer dependency dataset. To accomplish this, we considered an integrated dependency dataset generated by applying the DEMETER2 model to three large-scale RNAi screening datasets, covering 712 unique cancer cell lines [30], pre-processed as specified in the Methods.

Also, in this case, the two versions of the ADaM CFGs sets outperformed the other supervised methods both in terms of absolute grand median fitness effect (-0.79 and -0.61, respectively, for Behan2019 and ADaM, versus -0.6 and -0.5, respectively for CEN-tools and Sharma2020) and ratio with respect to baseline DM (0.98 and 0.96, respectively for ADaM and Behan2019, versus 0.94 and 0.76, respectively for CEN-tools and Sharma2020) (**Supplementary Fig. 5AB**)

As we previously observed, the FiPer variants' CEGs showed an overall milder grand median fitness effect (median = -0.36) but much better ratios with respect to baseline (median = 0.99).

### *Functional characterisation of predicted sets of Core-fitness-essential and Common-essential genes*

We performed a systematic statistical enrichment analysis of gene families across all sets of CFGs and CEGs considered in our benchmark, to functionally characterise them. This yielded a set of 13 families significantly enriched (FDR < 5%) consistently across all the state-of-the-art sets of CFGs as well as in the CFGs outputted by all tested supervised methods (**Fig. 6A** and **Supplementary Table 3**), thus worthy to be considered as bonafide true positive enrichments in human core fitness essential genes (the core-fitness families). These CFGs encompass most of the true positive controls used in our benchmark (ribosomal protein genes, proteasome, RNA polymerase [28]), as well as other plausible families, such as proteins involved in the initiation phase of eukaryotic translation [31], chaperonins [32], nucleoporins [33,34] and less immediate hits, such as AAA-ATPase [35,36] and WD repeat domain families [37,38].

The coverage of these families was much larger for the more recent CFG sets with respect to the state-of-the art CFGs, with ADaM and Sharma2020 performing best (average Recall across families = 57% and 54%, respectively). The unsupervised methods further extended the coverage of these gene families with average Recalls ranging from 63% (for FiPer average) to 68% (for FiPer AUC), with a median of 65%.

57 gene families were significantly enriched (FDR < 5%) consistently across the CEG sets outputted by the FiPer methods (**Fig. 6B**). These included all the 13 core-fitness families plus 44 additional groups (the common-essential families) such as COP9 signalosome [39,40], mediator complex [41], SNAP complex [42,43] and prefoldin subunits [44] to name a few.

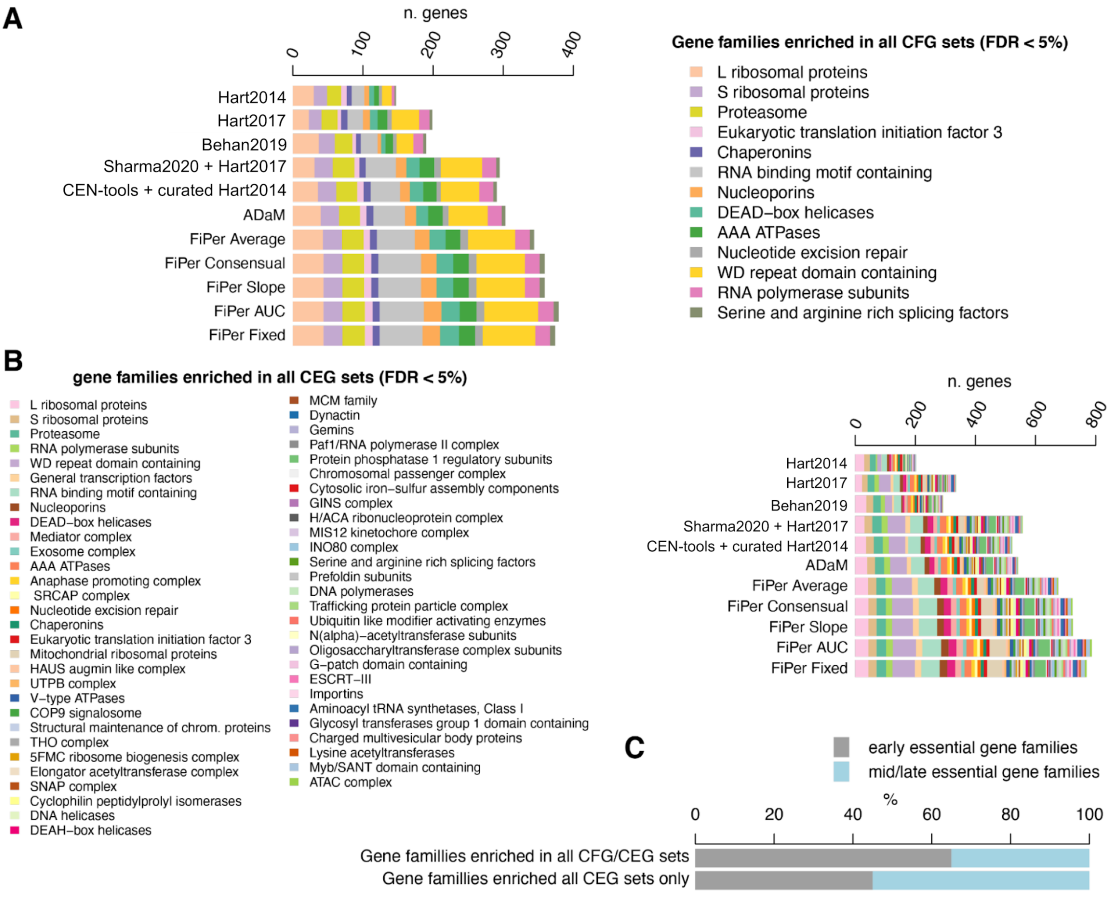


**Fig. 6 - Functional characterisation of predicted core-fitness/common-essential genes. A.** Gene families consistently significantly enriched (FDR < 5%) across all the state-of-the-art set of core-fitness essential genes (CFGs) and those outputted by the supervised methods. **B.** Gene families consistently and significantly enriched (FDR < 5%) across all the common-essential gene (CEG) sets outputted by the CoRe FiPer variants. **C** Percentage of early and mid/late essential gene families that are also always enriched across CFG and CEG sets or in CEG sets only.

When comparing the core-fitness and common-essential families with the gene-essentiality timing characterisation presented in [45], we observed in the former more genes exerting a negative fitness effect at an early time point upon knock-out (early-essential genes), whereas the latter included more families enriched in genes whose effect on fitness can be detected only at a later time point (late-essential genes) (**Fig. 6C**), such as exosome complex [46], dynactin [47] and ubiquitin-like modifier activating enzymes [48,49].

*Evaluation of core-fitness gene sets as template predictors of cell line specific essential genes*

We performed a final analysis evaluating each state-of-the-art set of core-fitness essential genes (CFGs), as well as those outputted by CEN-tools and ADaM when applied to the DepMap dataset, as a template classifier of cell line specific essential genes with BAGEL: a widely used bayesian method to estimate gene essentiality significance in pooled CRISPR-cas9 screens [21].

To this aim, we reprocessed with BAGEL the gene depletion fold-changes in the DepMap dataset producing 7 instances of BAGEL Bayes Factor (BF) matrices, quantifying the likelihood of each gene to be essential in each cell line, using each of the benchmarked set in turn as positive reference set of essential genes. To evaluate the obtained cell line specific BFs we assembled a set of cell line specific positive/negative controls. As positive control, we considered putative oncogenetic dependencies arising from oncogenes (from [25]) found mutated or copy number amplified in a cell line (using data from the Cell Model Passports [22]), whereas wild-type and non-expressed (FPKM < 0.1) oncogenes were considered as negative controls (**Supplementary Table 4**). Then, we assessed the 7 BF matrices, pooling all included values together and considering them as a unique rank based predictor (the larger the BF the higher the likelihood of a gene to be essential) of cell line specific essential genes, by means of receiver operating characteristic (ROC) analyses (Methods). Particularly, for each benchmarked set we computed the area under the BF-rank induced precision-recall curve (AUPRC) (**Fig. 7A** and **Supplementary Figure 6A-G**) and the recall of positive controls at 10% FDR (**Fig. 7B**). All the sets of CFGs outputted by CEN-tools and CoRe applied to the DepMap dataset (**Table 2**) outperformed the state-of-the-art sets of CFGs, showing a better ability in detecting as significant significantly essential mutated oncogenes, when used as a template for BAGEL. Above all, ADaM achieved the highest recall at 10% FDR (Methods).

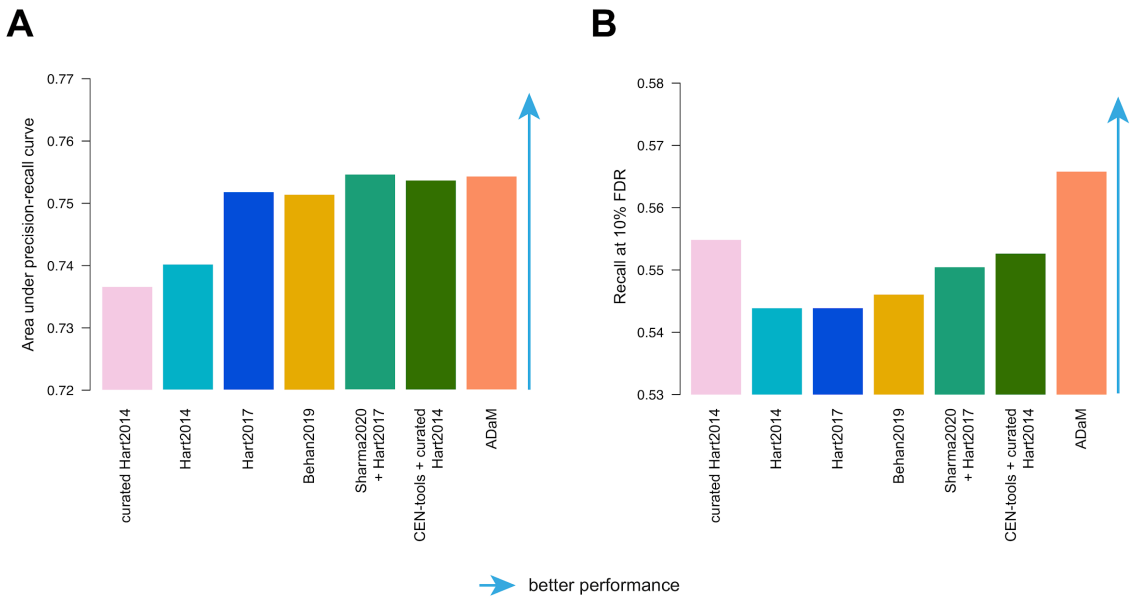


**Fig. 7 - Performances of the benchmarked sets as template classifiers of cell line specific essential genes. A.** Area under precision-recall curve obtained when predicting cell line specific oncogenetic addictions versus not expressed oncogenes with rank based classifiers yielded by gene essentiality Bayesian factors. These are computed by BAGEL using each of the benchmarked sets as positive classification template. **B.** Recall of cell line specific oncogenetic addictions at 10% FDR of not expressed oncogenes yielded by each benchmarked set when used as for A.

## *Computational efficiency*

We measured and compared running times obtained on a typical laptop, across different methods (**Table 3**) applied to the DepMap dataset. The CoRe FiPer methods were between 24 to 95 times faster than ADaM and 32 to 130 times faster than CEN-tools. Across FiPer variants, the slope one was the slowest, probably due to fitting of a linear regression model to a discrete distribution of gene fitness-rank-positions. Nevertheless, FiPer's running time was still significantly lower than ADaM and both outperformed CEN-tools, which was the method with the longest running time.

| **Algorithm** | **Running Time** |
|---|---|
| ADaM | 7 mins 38.23 secs |
| CEN-tools | 10 mins 22.76 secs |
| FiPer (average) | 4.93 secs |
| FiPer (AUC) | 5.77 secs |
| FiPer (fixed) | 4.78 secs |
| FiPer (slope) | 18.97 secs |

**Table 3 - Computational efficiency across methods.** Assessments of running time of the six compared methods when executed on a common laptop.

## Discussion and conclusions

We introduced CoRe: an open source R package implementing both existing and novel methods for the identification of core-fitness essential genes --at two different levels of stringency-- from joint analyses of multiple CRISPR-Cas9 pooled recessive screens. We robustly and extensively benchmarked CoRe against state-of-the-art sets of core-fitness genes and other CFGs discovery methods, using the largest integrative dataset of cancer dependency to date. We observed that the sets of core-fitness essential and common essential genes (CFGs, CEGs) predicted by the CoRe methods are much more comprehensive and robust, in terms of true and false positive rates (TPRs, FPRs) both absolute and relative to a baseline classifier. For the latter, we considered a trivial baseline daisy model [10] outputting as predicted CFGs those genes exerting a negative effect on fitness upon CRISPR-cas9 targeting in at least an optimal minimal number of screened models, which is known *a priori*. We also demonstrated that both CoRe and other methods are able to implicitly detect this optimal DM threshold, with the CoRe methods going much further and accurately predicting sets of genes that are essential in numbers of cell lines that are larger than this threshold. This is much more evident for the less stringent methods implemented in CoRe (i.e. the FiPer variants), thus showing the effectiveness of their underlying algorithm (based on genes' fitness percentile curves), which selectively picks likely true CEGs. Particularly, across these variants the FiPer AUC method performs

the best even when compared to a consensual set of CEGs obtained by intersecting the output of all the other FiPer variants.

Contrary to other methods, the sets of CFG/CEG predicted by CoRe are also smaller than those of that would be required by a baseline DM predictor to attain the same true positive rate, and our benchmark results were all confirmed when extending the analysis to gene sets used in the training phase of at least one of the compared methods, and when considering an independent RNAi based cancer dependency dataset.

Furthermore, we found that the CoRe CFGs/CEGs extend gene families covered by previous state-of-the art sets and methods, with the FiPer methods being able to detect more subtle yet consistent fitness effects and late essential genes. Finally, the CoRe CFGs/CEGs are all constitutively expressed in non-diseased tissue, pointing to the primary role which these genes play inside the cell. Indeed, it has been shown that higher essentiality is correlated with higher expression and association in important biological pathways [50].

Importantly, our final benchmark analysis also suggests that the CFGs yielded by our novel analyses of the DepMap dataset might be better suited than the reference positive control sets currently used [19,23] as positive predictor template when estimating cell line specific essential genes with a supervised classification method, such as BAGEL [21].

The identification of core-fitness genes has important implications in different areas of the life sciences: from drug discovery and cancer therapy to the study of genetic networks. However, different strategies are required according to the type of biological question being investigated. From this perspective, the utility of CoRe is two fold. In fact, when performing functional genetic studies or aiming at identifying novel CFGs, we recommend adopting a more stringent approach, such as ADaM, which can guarantee higher confidence. On the other hand, when the focus is on the identification of new therapeutic targets, thus to seek new promising context-specific essential genes, the opposite is true. As a consequence, applying a less stringent algorithm, such as the FiPer method

(particularly the FiPer AUC) allows a larger number of genes to be classified as common essentials, thus ruling out confounding genes that may skew the outcome of the analysis.

With the increasing availability of comprehensive cancer dependency maps [17], tools such CoRe will be arguably more and more needed in the future and they will contribute translating data and findings from such efforts into novel therapeutic targets candidates.

## Methods

### *DepMap dataset acquisition and pre-processing*

We downloaded the latest version of the integrated Sanger and Broad essentiality matrix processed with CERES from the DepMap portal ('https://www.depmap.org/broad-sanger/integrated_Sanger_Broad_essentiality_matrices_20201201.zip'). Among the 908 cell lines/columns, 51 were found to contain missing values and were thus removed. We then kept only the cell lines with an associated cancer tissue in the Cell Model Passport (annotation file version 20210326, https://cog.sanger.ac.uk/cmp/download/model_list_20210326.csv.gz), totalling to 855. This step is required in order to run ADaM tissue-wise. The dataset was then scaled colon-wise in order to have the median of curated BAGEL essential gene fitness scores equal to -1 and the median of curated BAGEL never-essential equal to 0 across all cell lines.

For the execution of ADaM, we binarised the pre-processed CERES dataset considering as essential all genes having a fitness score less than -0.5 in a given cell line, otherwise they were considered as non-essential.

### *CEN-tools Logistic Regression execution*

We downloaded the CEN-tools package [18] from https://gitlab.ebi.ac.uk/petsalakilab/centools/-/tree/master/CEN-tools. In order to decrease the memory

burden for the GitHub repository of the CoRe package, we removed all the python modules and data objects that were not directly called by the LR.py and clustering.R functions, respectively the python script implementing the logistic regression model and the R script performing the subsequent cluster analysis.

In addition, we added a few lines of code to the LR.py script to make it runnable from the command line and compute data objects on the fly. Particularly, CEN-tools uses a python dictionary in pickle format to specify which genes belong to the true positive set (i.e. curated BAGEL essential) or true negative set (i.e. curated BAGEL never-essential). Both scripts were seeded to guarantee reproducibility. All changes applied to the CEN-tools script are detailed in the Supplementary Materials.

For the execution of the logistic regression model implemented by CEN-tools on the new version of the CERES dataset, we used the curated BAGEL essential and curated never-essential genes for the training phase [12]. Based on the logistic regression, CEN-tools computes a matrix of continuous probability distributions for each gene being essential across cell lines and discretizes them according to the number of bins specified by the user. We adopted 20 bins as this was the default parameter used in the original CEN-tools run [18]. Following the pipeline, the matrix was normalised and genes not included in the training sets were then clustered through k-means using the Hartigan-Wong algorithm [51] around four centers. The silhouette method identified four as the optimal number of clusters according to their probability essentiality profiles: core essential, context-specific, rare-context-specific and never-essential. The core essential genes are characterized by the highest value of silhouette width and were then used for the downstream benchmarking.

### *Execution of ADaM*

ADaM takes as input a binarised matrix of fitness essentiality scores. For the identification of tissue CFGs, only the $N$ cell lines that are part of the same cancer tissue/type $T$ are selected. ADaM then implements a fuzzy intersection $I_n$ composed of genes exerting a significant depletion in at least $n$ cells out of $N$. The threshold $n^*$ is obtained in a semi-supervised manner: for each possible fuzzy

intersection $I_n$ from $n = 1$ to $N$, a true positive rate (TPR($n$)) is computed considering a set $E$ of a priori known essential genes as true positives, while $G$ is the whole set of screened genes:

$$\mathrm{TPR}(n) = |E \cap I_n| \,/\, |E \cap G|$$

In addition, ADaM computes the $\log_{10}$ odd ratio OR between the observed $I_n$ and the expected value $\mathrm{E}(I_n)$:

$$\mathrm{OR}(n) = \log_{10}(I_n \,/\, \mathrm{E}(I_n)).$$

$\mathrm{E}(I_n)$ is estimated by shuffling the binary matrix column-wise 1000 times. This way the number of essential genes for every cell line in T is preserved. Then $\mathrm{E}(I_n)$ is defined as the average value of $I_n^i$:

$$\mathrm{E}(I_n) = \frac{1}{1000} \sum_{i=1}^{1000} I_n$$

The threshold $n^*$ corresponds to the minimal number of cell lines $n$ whose $I_n$ provides the trade-off between the two monotonic functions, TPR($n$) being inversely proportional to $n$ and OR($n$) being directly proportional to $n$.

This is implemented by the wrapper function CoRe.CS_ADaM that subsets the dataset by taking only the cell lines included in the cancer tissue/type of interest using the Cell Model Passport [22] annotation file. We used the annotation file version 20210326. The submatrix is then passed to the CoRe.ADaM function that computes the CFGs. For a gene $i$ and a cell $j$, if *[i,j]* equals 1, it means that gene $i$ is essential for cell $j$. We also included in the package a general wrapper function, named CoRe.PanCancer_ADaM that executes ADaM tissue by tissue. Once ADaM identifies tissue CFGs, it builds a new binary matrix with genes on the rows and cancer tissues on the columns. For a gene $i$ and a cancer tissue $j$, if *[i,j]* equals 1, it means that gene $i$ is CFG for tissue $j$. Reiterating CoRe.ADaM on the new matrix results in the computation of the pan-cancer CFGs.

ADaM was executed using the CERES binarised dataset and taking the curated BAGEL essential genes as reference true positives. We set the number of random trials for the generation of the null model to 1000 and ran the algorithm only on those cancer tissues with at least 15 cell lines available as detailed in [12].

### *Execution of FiPer variants*

The fitness percentile method builds upon the assumption that if a gene is constitutively essential then it should rank among the top essentials also in the least dependent cell lines. As opposed to ADaM, this method takes directly as input the pre-processed quantitative CERES matrix containing the fitness essentiality scores for every screened gene in the 855 cell lines.

This fitness percentile is implemented by the CoRe.FiPer function. Initially, the method computes a gene-wise cell line ranking $R_{CL}$, where for each gene *g* it ranks every cell line *cl* according to the fitness essentiality score of *g* in *cl*. It also computes a cell-wise gene ranking $R_G$, where for each cell line it ranks every gene according to the fitness essentiality of *g* in *cl*. Then the package implements four different variants of the fitness percentile method:

- Fixed = a distribution of gene fitness-rank-positions in their most dependent $n^{th}$ (determined by the percentile parameter, default is $90^{th}$) percentile cell line is used in the subsequent step.
- Average = a distribution of average gene fitness-rank-positions across cell lines at or over the $n^{th}$ percentile of most dependent (determined by the percentile parameter, default is $90^{th}$) cell lines is used in the subsequent step.
- Slope = for each gene *g*, a linear model is fit on the sequence of gene fitness-rank-positions across all cell lines sorted according to their dependency on *g*, then a distribution of models' slopes is used in the subsequent step.
- AUC = for each gene *g*, the area under the curve (AUC) resulting from considering the sequence of gene fitness-rank-positions across all cell lines sorted according to their dependency on *g* is used in the subsequent step.

Each FiPer variant outputs a discrete distribution of gene fitness-rank-positions. A gaussian kernel estimator is applied to compute a continuous distribution. The kernel density estimator uses a default bandwidth defined as 0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power. The distribution is bimodal and the rank threshold corresponds to the local minimum. All genes having a fitness-rank-position lower

than the threshold are classified as CEGs. In the benchmarking, we assessed all of the four variants on the full pre-processed CERES matrix in order to identify the pan-cancer CEGs.

### *Benchmark of pan-cancer core-fitness gene sets*

We first devised a baseline predictor in order to assess the sets of pan-cancer core-fitness genes computed by each method. In our case, a baseline predictor was defined by considering core-fitness those genes essential in at least *n* cell lines for all possible *n*. To compute the baseline recall, we pooled together independent sets of a priori known essential genes [20,27]. These genes are involved in housekeeping cellular processes such as translation or DNA replication. In addition, we computed a baseline false positive rate, where we considered as false positive (or selective essentials) those genes that are not-expressed in any cell line or those genes whose dependency is associated with a context-specific biomarker. A gene is unexpressed if its FPKM scores (dataset available at https://cellmodelpassports.sanger.ac.uk/downloads, version: rnaseq_20191101) are constitutively less than 0.1 across all the cell lines. The biomarker/dependency associations were derived from [20].

The metrics derived from the baseline predictor were used to assess each CF set. We considered both novel hits, namely the CF sets stripped out of the BAGEL genes used in the training phase of the two CEN-tools runs (i.e. the *Hart2017* set, the BAGEL non-essential genes [23], curated BAGEL essential and never-essential genes [12]), as well as in their entirety. The recall of each set was normalised by the maximum recall achieved by the baseline predictor and so was done for the false positive rate. The two coordinates associated with each set were used to perform a cubic spline interpolation [52] and evaluate the balance between the normalised recall and false positive ratios according to the size of the set.

Next, we computed the thresholds required by the baseline predictor to attain the recalls observed by all tested methods.

*Characterisation of novel pan-cancer core-fitness sets*

To identify biologically grounded novel pan-cancer CF sets, we considered the gene families found enriched across all the predicted sets. For every of the aforementioned sets, we performed an hypergeometric test for each gene family with at least one gene in the CFG set of interest, following the formula:

$$p_{xf}(k) = Pr(x>k) = \sum_{x+1}^{k} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where *p* is the associated probability value of having more genes than observed *k* for a given family *f* in the CFG set under consideration, *K* is the total number of genes in the CFG set associated to any functional family, *N* is the total number of screened genes in the CERES pre-processed matrix associated to any functional family, *n* is the total number of genes belonging to *f* and found either in the CFG set or the remaining screened genes.

The p-values were then pooled and corrected set-wise using the Benjamini-Hochberg procedure. Gene families with an adjusted p-value < 0.05 in a given CF set were deemed significant. The significantly enriched families in common across the supervised methods (i.e. ADaM and CEN-tools in both instances, including also the Hart2014 and Hart2017 sets) were classified as always enriched and the pooled CFG sets used as ground truth. Particularly, we computed the exclusive CEGs in the FiPer AUC set belonging to the always enriched families that were not found in the ground truth. These CEGs were classified as novel hits.

In addition, we repeated the analysis assembling the significantly enriched families in common across the unsupervised methods (i.e. the four variants of the fitness percentile method plus the FiPer consensual set) and showed that unsupervised methods have higher sensitivity in identifying families derived from late time-point essential gene sets [45].

*Benchmark using an independent cancer dependency dataset*

We downloaded DEMETER v6 04/20 (available at https://ndownloader.figshare.com/files/11489669), cancer dependency data derived from genome-wide RNAi screens [30]. This dataset was scaled

colon-wise in order to have the median of curated BAGEL essential gene fitness scores equal to -1 and the median of curated BAGEL never-essential equal to 0 across all cell lines.

For each CF set, the median DEMETER fitness scores for every gene across cell lines were derived. In addition, we also derived the median DEMETER fitness scores of the genes included in the CF sets predicted by the baseline classifier, at the observed TPRs, and computed the normalised scores across sets.

*Retrieval of oncogene addictions in Bayesian factor templates*

We ran BAGEL v115 on the Sanger release 1 cancer dependency dataset (downloaded from: https://score.depmap.sanger.ac.uk/downloads) processed with CRISPRcleanR [27] (shown in [20] to better preserve context-specific essentialities than CERES). As a positive training gene set we used each of the sets among state-of-the-art sets, or CFG sets derived from the supervised methods, in turn, whereas as a negative training gene set we used the curated BAGEL never-essential genes. This led to seven different templates of Bayesian factor (BF) matrices. Each template was scaled cell-wise by subtracting to each gene the 5% false discovery rate (FDR) threshold computed between the BF scores of the two training distributions, for comparability.

Next, we defined a set of true positives and negatives to assess the ability of the templates in recapitulating oncogene addictions. First, we assembled a binary matrix summarizing the status of pan-cancer Cancer Functional Events (CFEs) across Sanger cell lines, namely somatic mutations, copy number alterations, and hypermethylation. The binary matrix was then subset in order to include only genes unambiguously classified as oncogenes in the catalog of driver genes release 2020.02.01 from the IntOGen database. In addition, cells showing copy number gains in genomic segments containing ERBB2 or EGFR or KRAS or MYC or MYCN (typically copy number amplified oncogenes in different cancer types), were considered as positive events too. Secondly, we assembled an additional binary matrix where we deemed as positive events oncogenes not expressed in a cell line. We considered oncogenes only instead of including all not expressed genes in order to avoid unbalanced control sets, favouring the negative controls. By combining the two binary matrices, we obtained three classes:

- Positive instances constituted by oncogenes mutated and expressed in a cell line.
- Null instances constituted either by wild-type and expressed or mutated and not expressed oncogenes in the cell line.
- Negative instances constituted by wild-type and not expressed oncogenes in the cell line.

The positive and negative instances were used on the BF score of each template to assess the area under precision-recall curve and the recall at fixed percentages of FDRs.

### *Hardware and software details*

All the analyses were performed on a typical laptop with a 2.3 GHz Quad-Core Intel Core i7 processor, with 16 GB 3733 MHz LPDDR4 memory and 8 cores. The operating system was Big Sur v11.2.3 (20D91). The software was executed in the RStudio IDE v1.3.1073 with x86_64-apple-darwin17.0 platform and R programming language v4.0.2, python scripts were executed using python v3.9.1. For all the methods shown in Table 1 below but ADaM, we used the quantitative pre-processed CERES matrix. The matrix consisted of 17,846 genes and 855 cell lines containing gene fitness scores. Instead, ADaM used a binarized version of the matrix as explained in the previous section. The binary matrix consisted of 8,496 genes, considering genes classified as essential in at least one cell line, and 820 cell lines, considering cell lines from a cancer tissue with at least 15 cell lines in total.

## Availability of data and materials

CoRe is publicly available as an open source R package at https://github.com/DepMap-Analytics/CoRe. An interactive vignette, with demonstrations and examples is available at https://rpubs.com/AleVin1995/CoRe. All CFGs/CEGs resulting from the execution of CoRe are available both as supplementary table (**Supplementary Table 1**) as well as precomputed RData format inside the package. All results from benchmarking CoRe against state-of-the-art sets of CFGs and other methods, and related figures presented in this paper can be

fully reproduced executing a Jupyter notebook available at:

https://github.com/DepMap-Analytics/CoRe/blob/master/notebooks/CoRe_Benchmarking.ipynb

(which can also be executed via browser with the Google CoLab environment). CEN-tools Logistic Regression is publicly available at the following GitLab repository:

https://gitlab.ebi.ac.uk/petsalakilab/centools/-/tree/master/CEN-tools.

The DepMap dataset used for the downstream analyses explained above can be downloaded at:

https://www.depmap.org/broad-sanger/integrated_Sanger_Broad_essentiality_matrices_20201201.zip.

The annotation file as well as the RNAseq data can both be downloaded at the Cell Model Passport website: https://cellmodelpassports.sanger.ac.uk/downloads (respectively model list version 20210326 and rnaseq version 20191101). Finally, the independent dataset DEMETER can be downloaded at:

https://ndownloader.figshare.com/files/11489669.

## Author Contributions

AV conceived the study, designed and performed the benchmark analyses, wrote and documented the CoRe package, assembled the interactive vignette and the jupyter notebook, wrote and revised the manuscript. EK wrote and documented the first version of CoRe, and revised the manuscript. CP contributed to package writing and documentation, and revised the manuscript. UP and RRDL contributed to the design of the benchmark analyses, tested the package, and revised the manuscript. MJG contributed to study supervision. FI conceived the study and the package, contributed to the design of the benchmark analyses, wrote and revised the manuscript, supervised the study.

## Competing interests

MJG and FI receive funding from Open Targets, a public-private initiative involving academia and industry. MJG receives funding from GSK, AstraZeneca, and has performed consultancy for Sanofi. MJG is founder of Mosaic Therapeutics. FI performs consultancy for the joint CRUK—AstraZeneca Functional Genomics Center.

## Acknowledgements

We thank Paula Weidemueller and Evangelia Petsalaki for critically reading and discussing the manuscript.

# References

1. Jinek M, Chylinski K, Fonfara I, Hauer M. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. science.sciencemag.org; 2012; Available from: https://science.sciencemag.org/content/337/6096/816.abstract?casa_token=MTvFKPfjo44AAAAA:6QxK7ciRcVI_15IhbYhisFL2hR5qD2iT7lekPLoLpsM3doW3v_-OmyLg_Q3Vx0yL3h3wNa4zgd9C-rM

2. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, et al. RNA-guided human genome engineering via Cas9. Science. 2013;339:823–6.

3. Cho SW, Kim S, Kim JM, Kim J-S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nat Biotechnol. 2013;31:230–2.

4. Koike-Yusa H, Li Y, Tan E-P, Velasco-Herrera MDC, Yusa K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nat Biotechnol. 2014;32:267–73.

5. Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol. 2014;32:1262–7.

6. Sanjana NE, Shalem O, Zhang F. Improved vectors and genome-wide libraries for CRISPR screening. Nat Methods. 2014;11:783–4.

7. 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;350:1096–101.

8. Gonçalves E, Thomas M, Behan FM, Picco G, Pacini C, Allen F, et al. Minimal genome-wide human CRISPR-Cas9 library. Genome Biol. 2021;22:40.

9. Blomen VA, Májek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, et al. Gene essentiality and synthetic lethality in haploid human cells. Science. 2015;350:1092–6.

10. 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;163:1515–26.

11. Meyers RM, Bryan JG, McFarland JM, Weir BA. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. Nature [Internet]. nature.com; 2017; Available from: https://www.nature.com/articles/ng.3984

12. 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;568:511–6.

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

14. Lenoir WF, Lim TL, Hart T. PICKLES: the database of pooled in-vitro CRISPR knockout library essentiality screens. Nucleic Acids Res. 2018;46:D776–80.

15. Dwane L, Behan FM, Gonçalves E, Lightfoot H, Yang W, van der Meer D, et al. Project Score database: a resource for investigating cancer cell dependencies and prioritizing therapeutic targets. Nucleic Acids Res. 2021;49:D1365–72.

16. Dempster J, Behan FM, Green T, Najgebauer H, Krill-Burger J, Allen F, et al. Agreement between two large pan-cancer genome-scale CRISPR knock-out datasets. Nature Communications. 2019;In Press.

17. Boehm JS, Garnett MJ, Adams DJ, Francies HE, Golub TR, Hahn WC, et al. Cancer research needs a better map. Nature. Springer Science and Business Media LLC; 2021;589:514–6.

18. Sharma S, Dincer C, Weidemüller P, Wright GJ, Petsalaki E. CEN-tools: an integrative platform to identify the contexts of essential genes. Mol Syst Biol. EMBO; 2020;16:e9698.

19. Hart T, Tong AHY, Chan K, Van Leeuwen J, Seetharaman A, Aregger M, et al. Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. G3 . 2017;7:2719–27.

20. Pacini C, Dempster JM, Boyle I, Gonçalves E, Najgebauer H, Karakoc E, et al. Integrated cross-study datasets of genetic dependencies in cancer. Nat Commun. 2021;12:1661.

21. Hart T, Moffat J. BAGEL: a computational framework for identifying essential genes from pooled library screens. BMC Bioinformatics. 2016;17:164.

22. van der Meer D, Barthorpe S, Yang W, Lightfoot H, Hall C, Gilbert J, et al. Cell Model Passports—a hub for clinical,

genetic and functional datasets of preclinical cancer models. Nucleic Acids Res. Narnia; 2019;47:D923–9.

23. Hart T, Brown KR, Sircoulomb F, Rottapel R, Moffat J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. Mol Syst Biol. 2014;10:733.

24. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell. 2016;

25. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020;20:555–72.

26. Broad Institute of Harvard and MIT. Cancer Dependency Map [Internet]. Available from: https://depmap.org/

27. Iorio F, Behan FM, Gonçalves E, Bhosle SG, Chen E, Shepherd R, et al. Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. BMC Genomics. 2018;19:604.

28. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545.

29. McDonald ER 3rd, de Weck A, Schlabach MR, Billy E, Mavrakis KJ, Hoffman GR, 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.

30. McFarland JM, Ho ZV, Kugener G, Dempster JM, Montgomery PG, Bryan JG, et al. Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. Nat Commun. 2018;9:4610.

31. Jaiswal PK, Koul S, Palanisamy N, Koul HK. Eukaryotic Translation Initiation Factor 4 Gamma 1 (EIF4G1): a target for cancer therapeutic intervention? Cancer Cell Int. 2019;19:224.

32. Shemesh N, Jubran J, Dror S, Simonovsky E, Basha O, Argov C, et al. The landscape of molecular chaperones across human tissues reveals a layered architecture of core and variable chaperones. Nat Commun. 2021;12:2180.

33. Khan AU, Qu R, Ouyang J, Dai J. Role of Nucleoporins and Transport Receptors in Cell Differentiation. Front Physiol. 2020;11:239.

34. Raices M, D'Angelo MA. Nuclear pore complex composition: a new regulator of tissue-specific and developmental functions. Nat Rev Mol Cell Biol. 2012;13:687–99.

35. Armenteros-Monterroso E, Zhao L, Gasparoli L, Brooks T, Pearce K, Mansour MR, et al. The AAA+ATPase RUVBL2 is essential for the oncogenic function of c-MYB in acute myeloid leukemia. Leukemia. 2019;33:2817–29.

36. Osaki H, Walf-Vorderwülbecke V, Mangolini M, Zhao L, Horton SJ, Morrone G, et al. The AAA+ ATPase RUVBL2 is a critical mediator of MLL-AF9 oncogenesis. Leukemia. 2013;27:1461–8.

37. O'Bryant D, Wang Z. The essential role of WD repeat domain 77 in prostate tumor initiation induced by Pten loss. Oncogene. 2018;37:4151–63.

38. Schapira M, Tyers M, Torrent M, Arrowsmith CH. WD40 repeat domain proteins: a novel target class? Nat Rev Drug Discov. 2017;16:773–86.

39. Sinha A, Israeli R, Cirigliano A, Gihaz S, Trabelcy B, Braus GH, et al. The COP9 signalosome mediates the Spt23 regulated fatty acid desaturation and ergosterol biosynthesis. FASEB J. 2020;34:4870–89.

40. Gutierrez C, Chemmama IE, Mao H, Yu C, Echeverria I, Block SA, et al. Structural dynamics of the human COP9 signalosome revealed by cross-linking mass spectrometry and integrative modeling. Proc Natl Acad Sci U S A. 2020;117:4088–98.

41. Petrenko N, Jin Y, Wong KH, Struhl K. Evidence that Mediator is essential for Pol II transcription, but is not a required component of the preinitiation complex in vivo. Elife. eLife Sciences Publications, Ltd; 2017;6:e28447.

42. Huang X, Sun S, Wang X, Fan F, Zhou Q, Lu S, et al. Mechanistic insights into the SNARE complex disassembly. Sci Adv. 2019;5:eaau8164.

43. Zhao M, Wu S, Zhou Q, Vivona S, Cipriano DJ, Cheng Y, et al. Mechanistic insights into the recycling machine of the SNARE complex. Nature. 2015;518:61–7.

44. Liang J, Xia L, Oyang L, Lin J, Tan S, Yi P, et al. The functions and mechanisms of prefoldin complex and prefoldin-subunits. Cell Biosci. 2020;10:87.

45. Tzelepis K, Koike-Yusa H, De Braekeleer E, Li Y, Metzakopian E, Dovey OM, et al. A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. Cell Rep. 2016;17:1193–205.

46. Gurunathan S, Kang M-H, Jeyaraj M, Qasim M, Kim J-H. Review of the Isolation, Characterization, Biological Function, and Multifarious Therapeutic Approaches of Exosomes. Cells [Internet]. 2019;8. Available from: http://dx.doi.org/10.3390/cells8040307

47. Lee YD, Kim B, Jung S, Kim H, Kim MK, Kwon J-O, et al. The dynactin subunit DCTN1 controls osteoclastogenesis via the Cdc42/PAK2 pathway [Internet]. Experimental & Molecular Medicine. 2020. p. 514–28. Available from: http://dx.doi.org/10.1038/s12276-020-0406-0

48. Aichem A, Sailer C, Ryu S, Catone N, Stankovic-Valentin N, Schmidtke G, et al. The ubiquitin-like modifier FAT10 interferes with SUMO activation. Nat Commun. 2019;10:4452.

49. Hyer ML, Milhollen MA, Ciavarri J, Fleming P, Traore T, Sappal D, et al. A small-molecule inhibitor of the ubiquitin activating enzyme for cancer treatment. Nat Med. 2018;24:186–93.

50. Chen H, Zhang Z, Jiang S, Li R, Li W, Zhao C, et al. New insights on human essential genes based on integrated analysis and the construction of the HEGIAP web-based platform. Brief Bioinform. 2020;21:1397–410.

51. Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. J R Stat Soc Ser C Appl Stat. [Wiley, Royal Statistical Society]; 1979;28:100–8.

52. Hall CA, Meyer WW. Optimal error bounds for cubic spline interpolation. J Approx Theory. 1976;16:105–22.