nature

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41586-019-1103-9

In the format provided by the authors and unedited.

Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens

Fiona M. Behan[1,2,12], Francesco Iorio[1,2,3,12], Gabriele Picco[1,12], Emanuel Gonçalves[1], Charlotte M. Beaver[1], Giorgia Migliardi[4,5], Rita Santos[6], Yanhua Rao[7], Francesco Sassi[4], Marika Pinnelli[4,5], Rizwan Ansari[1], Sarah Harper[1], David Adam Jackson[1], Rebecca McRae[1], Rachel Pooley[1], Piers Wilkinson[1], Dieudonne van der Meer[1], David Dow[2,6], Carolyn Buser-Doepner[2,7], Andrea Bertotti[4,5], Livio Trusolino[4,5], Euan A. Stronach[2,6], Julio Saez-Rodriguez[2,3,8,9,10], Kosuke Yusa[1,2,11,13]* & Mathew J. Garnett[1,2,13]*

[1]Wellcome Sanger Institute, Cambridge, UK. [2]Open Targets, Cambridge, UK. [3]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, UK. [4]Candiolo Cancer Institute-FPO, IRCCS, Turin, Italy. [5]Department of Oncology, University of Torino, Turin, Italy. [6]GlaxoSmithKline Research and Development, Stevenage, UK. [7]GlaxoSmithKline Research and Development, Collegeville, PA, USA. [8]Faculty of Medicine, Joint Research Centre for Computational Biomedicine, RWTH Aachen University, Aachen, Germany. [9]Institute for Computational Biomedicine, Heidelberg University, Faculty of Medicine, Bioquant, Heidelberg, Germany. [10]Heidelberg University Hospital, Heidelberg, Germany. [11]Present address: Stem Cell Genetics, Institute for Frontier Life and Medical Sciences, Kyoto University, Kyoto, Japan. [12]These authors contributed equally: Fiona M. Behan, Francesco Iorio, Gabriele Picco. [13]These authors jointly supervised this work: Kosuke Yusa, Mathew J. Garnett. *e-mail: k.yusa@infront.kyoto-u.ac.jp; mathew.garnett@sanger.ac.uk

Index

Supplementary Information - CRISPR screen data analyses: screening performance assessment

This analysis was restricted only to genes belonging to predefined sets of essential/non-essential genes[1], respectively *E* and *N*, and hereby defined as set *G*. For each cell line we assembled a *predictor* vector *P* containing the depletion fold-changes (FCs) for all the genes in *G*, and a boolean *response* vector with one entry per gene in *G*, which was equal to *TRUE* if the corresponding gene belonged to *E*. Specificity, Precision and Sensitivity curves were computed matching these two vectors, by making use of the *roc* function of the pROC R package[2]. The resulting ROC and Precision/Recall curves are shown in Extended Data Fig. 1g.

In addition, we computed for prior known essential (*E*) and ribosomal protein (*R*) genes a Glass's Δ scores, quantifying depletion signal magnitudes and discriminative distance of their distribution from that of non-essential genes (N), defined as:

$$| \mu(\mathrm{FC}(x \in X)) - \mu(\mathrm{FC}(n \in N)) | / \sigma(\mathrm{FC}(x \in X)),$$

Where $X \in \{E, R\}$, and μ, σ indicate mean and standard deviation, respectively. Results across all screened cell lines are shown in Supplementary Extended Data Fig. 1h.

Supplementary Information - CRISPR screen data analyses: Calling CRISPR-Cas9 gene knockout fitness effects

We ranked all the screened genes in a given cell line based on their BAGEL Bayesian factor[1] (computed as detailed in the Methods) in decreasing order, from most to least depleted. Then for each rank position r, a corresponding set of genes G_r was defined by pooling together all the genes whose rank position was $\leq r$. Subsequently a false discovery rate FDR(r) was computed as $100 \times |G_r \cap N| / |G_r \cap (E \cup N)|$, where E and N are the reference sets of *a priori* known essential and non-essential genes described in the previous section, respectively. Finally, the BF of the gene in the rank position $r^* = \max_r \{\mathrm{FDR}(r) < 5\%\}$, was defined as the significance threshold for the cell line under consideration and all the genes with a BF above this threshold were deemed as significantly essential (at a 5% FDR).

Supplementary Information - High-level data analyses: Pan-cancer and cancer-type core fitness genes using the Adaptive Daisy Model (ADaM)

To identify genes that are consistently depleted across multiple cell lines (hence considered as a proxy set of core fitness (CF) genes) the Daisy Model[3] computes a *fuzzy* intersection I_{m*} (the core of the daisy) composed of genes that are significantly depleted (fitness genes) in at least $m*$ cell lines, where $m*$ is aprioristically defined. The genes not belonging to I_{m*} (thus falling on the petals of the daisy) are deemed to be context-specific essential genes. In [3] the authors screened 7 cell lines from different tissues and defined genes significantly essential in at least $m* = 3$ cell lines as CF genes. Generalising this approach, ADaM (i) applies the Daisy Model but it adaptively determines $m*$ in a data-driven way, predicting sets of CF genes (one set per cancer-type in input), (ii) applies the Daisy Model to the obtained sets of cancer-type CF genes by adaptively defining the number of cancer-types k for which a gene should have been predicted as CF in order to be considered as a pan-cancer CF gene. The whole process is illustrated in Extended Data Fig. 3, for an example cancer-type (ovary) and for determining pan-cancer CF genes.

Briefly, for a given cancer-type T for which M cell lines have been screened, ADaM computes fuzzy intersections of genes I_m, for each $m = 1, \ldots, M$, including genes that are significantly depleted in at least m cell lines. Subsequently for each $m = 1, \ldots, M$, a true positive rate (TPR(m)) is computed by considering as true positives the genes included in a priori known essential gene set E:

$$\mathrm{TPR}(m) = |E \cap I_m| / |E \cap G|,$$

where G is the set of all screened genes.

At the same, time the deviance of $|I_m|$ from its expectation π_m is computed as follows:

$$\mathrm{D}(m) = \log_{10}(|I_m| / \pi_m).$$

To estimate π_m, 1,000 randomised versions of the binary depletion scores (computed as detailed in the previous section) of all the genes across the cell lines from T, are generated while preserving the total number of depleted genes per cell lines, to account for their overall

vulnerability. Then π_m is defined as the average value of the $|I_m|_i$ (for i = 1, …, 1000) computed across the randomised versions of the binary depletion scores.

Finally, m^* is defined as the maximal value of m providing the best trade-off between TPR(m) (inversely proportional to m) and D(m) (proportional to m) (example in Extended Data Fig. 3d).

We applied ADaM to the depletion scores observed in cell lines from each of the cancer-types with at least 8 screened cell lines in turn, defining an ADaM threshold m^* for each of them, and considering the corresponding I_{m*} sets as proxies of cancer-type CF essential genes and those not belonging to I_{m*} as proxies of context-specific genes.

Supplementary Information - High-level data analyses: Comparison between the ADaM pan-cancer CF genes and other reference sets of essential genes.

To estimate false positive rates for the ADaM pan-cancer core fitness genes, as well as for the reference sets of core fitness essential genes in refs.[1] and [4] we assembled sets of negative controls from an independent publication[5]. In this independent study, the essentiality signal of each gene is analysed across a large dataset obtained from an RNAi screen across hundreds of cancer cell lines, and its tendency to distribute according to a skewed Student-t distribution (indicative of that gene being strongly essential in a minority of cell lines) is estimated. In our comparison we considered sets of putative strongly context-specific essential genes (thus false positives) at different level of likelihood of a skewed Student-t distribution.

Supplementary Information - Statistical analyses: Analysis of Variance to identify genomic correlates with gene fitness

We focused our analysis on Cancer Driver Events[6] (CDEs) because these represent a causal link with carcinogenesis and thus increase interpretability of identified associations and facilitate development of genetic biomarkers. Each analysis included only genes that were significantly depleted in at least 2 cell lines (from the considered cancer-type, or across the whole panel of cell lines for the pan-cancer analysis), but excluding genes in the BAGEL[1]

curated set of essential genes (*E*) set, the set of ribosomal protein genes and the other MSigDB[7] genes sets (detailed in the Methods). Genes predicted to be pan-cancer or cancer-type core fitness (CF) genes by ADaM were also excluded from respective ANOVA. CRISPRcleanR[8] corrected gene-level FC were quantile normalised on an individual cell line basis and used as indicators of gene essentiality.

For each of the genes included in the analysis, we assembled an essentiality vector consisting of *n* depletion FCs (described above), one entry per cell line (considering the whole panel for pan-cancer analysis, and only cell line from the analysed cancer-type otherwise). The model was linear (no interaction terms) with dependant variables represented by the described vector and factors including tissue (for the pan-cancer analysis only), microsatellite instability status (for the pan-cancer analysis and for those of cancer-types with at least 2 positive samples for this feature) and the status of a cancer driver event (CDE, as defined in[6]), one model for each CDE. Differently from the analyses described in[6] here we used a common set of CDEs (the pan-cancer ones) across the different analyses. Only CDEs occurring in at least 3 cell lines were considered and CDEs with identical patterns of positive occurrence were merged together.

Supplementary Information - Target priority score

Each gene *t* was assigned a target priority score $P(t)$ defined as:

$$P(t) = 0.3\, L_1(t) + 0.7\, L_2(t)$$

To formally define the two individual terms $L_1(t)$ and $L_2(t)$, we introduce the boolean function $f(p)$ which evaluates the status of the property p related to a given gene and possibly a given cell line. This functions assumes a value equal to 1 when p is true and equal to 0 otherwise. Additionally, let us consider the following properties:

$p_1(t) = \{t \text{ has at least a class A marker}\}$,

$p_2(t) = \{t \text{ has at least a class B marker}\}$,

$p_3(t) = \{t \text{ has at least a class C marker}\}$,

$p_4(t) = \{t \text{ has at least a } \textit{weaker} \text{ marker}\}$,

To determine the marker classes, the considered ANOVA associations were those specific to the cancer type under consideration, and the classes were defined as specified in the Methods.

$L_1(t)$ was then defined as follows:

$$L_1(t) = g(t)\,\{0.8\sum_{i=1}^{4} f(p_i(t)) + 0.2\,f(\,t \text{ is somatically mutated in at least 2\% of } T \text{ matched primary tumours})\},$$

where somatic mutations from large cohorts of primary tumours were derived from[6] and $g(t)$ is a filter function defined as follows:

$$g(t) = \{f(t \text{ is targeted by more than 1 sgRNA in the employed library}) \times f(t \text{ does not belong to any reference set of predefined essential genes}) \times f(t \text{ is not predicted as pan-cancer CF gene by ADaM}) \times f(t \text{ is not predicted as } T\text{-specific CF gene by ADaM})\},$$

where the reference set of predefined essential genes are described in the previous section and the last factor is omitted for pan-cancer priority scores.

Finally, $L_2(t)$ was defined as follows:

$$L_2(t) = \sum_c h(t,c) / \sum_c f(t \text{ is significantly essential in } c),$$

where c are the screened cell lines from T (or in the whole panel for pan-cancer scores), essentiality is meant at a BAGEL 5% FDR (as detailed in the previous section *Supplementary Information - CRISPR screen data analyses: Identification of fitness genes*). To defined a $h(t,c)$ let us introduce the following properties:

$$q_1(t,c) = \{\text{scaled BF of } t \text{ in } c > 1\},$$

$$q_2(t,c) = \{\text{scaled BF of } t \text{ in } c > 2\},$$

$q_3(t,c)$ = {scaled BF of t in $c > 3$ },

$q_4(t,c)$ = {MAGeCK depletion FDR of t in $c < 10\%$ },

$q_5(t,c)$ = {MAGeCK depletion FDR of t in $c < 5\%$ },

$q_6(t,c)$ = {t is highly expressed in c at the basal level},

$q_7(t,c)$ = {t is somatically mutated in c },

$q_8(t,c)$ = {t belongs to a biological pathway that is statistically enriched among the genes significantly depleted in c },

where the scaled BF and the MAGeCK depletion FDRs are indicators of gene essentiality and are computed in the previous section, t is highly expressed if it falls over the 95% quantile of basal expression in c according to the FPKM values derived from[9] as described in the following section; somatic mutations for all the cell lines are derived from[6], and pathway enrichments in the set of genes significantly depleted in c are computed with a hypergeometric test using pathways gene sets from Pathway Commons[10] post-processed to reduce redundancies across different sets as detailed in[11].

$h(t,c)$ is then defined as follows:

$$h(t,c) = l(t,c)\ \{0.125 \sum_{i=1}^{8} f(q_i(t,c))\ \},$$

where $h(t,c)$ is a filter function defined as follows:

$h(t,c)$ = {$f(t$ is significantly essential in $c)$ ×

$f(t$ is express in $c)$ ×

$f(t$ is not homozygously copy number deleted in $c)$},

where, as before, significant essentiality is meant at a BAGEL 5% FDR; a gene is expressed in a cell line if its FPKM is ≥ 0.05 in that cell line; the gene level copy number status across screened cell lines has been downloaded from (www.cancerrxgene.org[12]).

Finally, we defined a minimum priority score threshold based on scores calculated for targets with approved or pre-clinical cancer compounds as follows (Extended Data Fig. 5c and Supplementary Table 6). For each cancer-type T, let us consider the priority scores of the targets with approved or pre-clinical anti-cancer T-specific compounds D_T (Supplementary Table 6) and those of the targets that are currently non druggable O_T. Let D^* be the union of all the D_T across cancer types and O^* the union of all the O_T. We fitted two kernel-estimated density distributions on D^* and O^*, and we defined as minimal threshold value for priority target the priority score at which the probability mass function of the first distribution was at least twice that of the second one. To define a threshold for pan-cancer priority score we follow the same procedure but consider the scores of the targets with approved or pre-clinical anti-cancer (unspecific to any cancer type) compounds to define D_T.

Supplementary Information: Collective test for genomic markers of gene essentiality shared among tissues

All the targets for which in at least two cancer-types the following condition was satisfied:

$$\sum_{i=1}^{4} f(p_i(t)) \geq 3,$$

with p_i defined as in the previous section, had their differential essentiality re-tested against the status of g in a collective *t-test* considering all the cell lines from the cancer-types in which the condition above held, pooled together, yielding the results shown in Supplementary Table 8b.

Supplementary Information - Target tractability

This was assembled using data from the following resources.

For the small molecule pipeline:

- Uniprot[13], PDBe[14], InterPro[15], Pfam[16] & GO[17]: accessed on 18/11/2016
- Complex portal[18], Biomodels[19] accessed on: 24/06/2017
- ChEMBL[20]: version 22 (October 2016)
- SureChEMBL[21]: SureChEMBL RDF as available in Open PHACTS (version 1.2, March 2015)

- Human genome: NCBI[22], November 2016 (20 912 protein coding genes)

For the antibody pipeline:

- HP [23]: accessed on February 2017
- Uniprot & GO: accessed on 13/07/2017
- Complex portal & Biomodels: accessed on 24/06/2017
- ENSEMBL Compara[24]: accessed on 13/07/2017 (ENSEMBL 89)
- ChEMBL: version 23 (May 2017)
- SureChEMBL: SureChEMBL RDF as available in Open PHACTS (version 1.2, March 2015)
- Human genome: NCBI, November 2016 (20 912 protein coding genes)

References

1. Hart, T. & Moffat, J. BAGEL: a computational framework for identifying essential genes from pooled library screens. *BMC Bioinformatics* **17**, 164 (2016).
2. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
3. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**, 1515–1526 (2015).
4. Hart, T. *et al.* Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3* **7**, 2719–2727 (2017).
5. McDonald, E. R., 3rd *et al.* Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. *Cell* **170**, 577–592.e10 (2017).
6. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740–754 (2016).
7. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545 (2005).
8. Iorio, F. *et al.* Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. *BMC Genomics* **19**, 604 (2018).
9. Garcia-Alonso, L. *et al.* Transcription Factor Activities Enhance Markers of Drug Sensitivity in Cancer. *Cancer Res.* **78**, 769–780 (2018).
10. Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39**, D685–90 (2011).
11. Iorio, F. *et al.* Pathway-based dissection of the genomic heterogeneity of cancer hallmarks' acquisition with SLAPenrich. *Sci. Rep.* **8**, 6713 (2018).
12. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
13. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–12 (2015).
14. Gutmanas, A. *et al.* PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* **42**, D285–91 (2014).
15. Mitchell, A. *et al.* The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* **43**, D213–21 (2015).
16. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–85 (2016).
17. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43**, D1049–56 (2015).
18. Meldal, B. H. M. *et al.* The complex portal--an encyclopaedia of macromolecular complexes. *Nucleic Acids Res.* **43**, D479–84 (2015).
19. Chelliah, V. *et al.* BioModels: ten-year anniversary. *Nucleic Acids Res.* **43**, D542–8 (2015).
20. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2017).
21. Papadatos, G. *et al.* SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Res.* **44**, D1220–8 (2016).
22. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **46**, D8–D13 (2018).
23. Thul, P. J. & Lindskog, C. The human protein atlas: A spatial map of the human

proteome. *Protein Sci.* **27**, 233–244 (2018).
24. Aken, B. L. *et al.* Ensembl 2017. *Nucleic Acids Res.* **45**, D635–D642 (2017).