Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens.

Prioritisation of cancer therapeutic targets using CRISPR-Cas9 screens

Fiona M Behan[1,2,†], Francesco Iorio[1,2,3,†], Gabriele Picco[1,†], Emanuel Gonçalves[1], Charlotte M Beaver[1], Giorgia Migliardi[4,5], Rita Santos[6], Yanhua Rao[6], 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[6,2], Carolyn Buser-Doepner[6,2], Andrea Bertotti[4,5], Livio Trusolino[4,5], Euan A. Stronach[6,2], Julio Saez-Rodriguez[2,3,7,8], Kosuke Yusa[1,2,9]*, Mathew J Garnett[1,2,]*

Affiliations:

[1] Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge CB10 1SA, UK.

[2] Open Targets, Wellcome Genome Campus, Cambridge CB10 1SA, UK.

[3] European Molecular Biology Laboratory – European Bioinformatics Institute, Wellcome Genome Campus, Cambridge CB10 1SD, UK.

[4] Candiolo Cancer Institute-FPO, IRCCS, 10060 Candiolo, Turin, Italy.

[5] Department of Oncology, University of Torino, 10060 Candiolo, Turin, Italy.

[6] GlaxoSmithKline Research and Development, Gunnels Wood Road, Stevenage, SG1 2NY, UK and Collegeville, Pennsylvania, 19426-0989, USA.

[7] Faculty of Medicine, Joint Research Centre for Computational Biomedicine, RWTH Aachen University, Aachen 52057, Germany.

[8] Heidelberg University, Heidelberg, Germany

[9] Present address: Stem Cell Genetics, Institute of Frontier Life and Medical Sciences, Kyoto University, Kyoto 606-8507 Japan

* Correspondence to: k.yusa@infront.kyoto-u.ac.jp and mathew.garnett@sanger.ac.uk

† Equally contributing authors

Summary:

Functional genomics approaches can overcome limitations that hamper oncology drug development such as lack of robust target identification and poor clinical efficacy. Here we performed genome-scale CRISPR-Cas9 screens in 324 human cancer cell lines from 30 cancer types and developed a data-driven framework to prioritise cancer therapeutic candidates. We integrated gene knockout cell fitness effects with genomic biomarkers and target tractability for drug development to systematically prioritise new targets in defined tissues and genotypes. We verified one of our most promising dependencies, Werner syndrome ATP-dependent helicase, as a target in tumours from multiple cancer types with microsatellite instability. Our analysis provides a comprehensive resource of cancer dependencies, generates a framework to prioritise oncology targets, and nominates specific new targets. The principles described in this study can inform the initial stages of drug development by contributing a new, diverse and more effective portfolio of oncology targets.

Main Text:

The molecular features of a patient's tumour impact clinical responses and can be used to guide therapy, leading to more effective treatments and reduced toxicity(Garraway 2013). Most patients however do not benefit from such targeted therapies in part due to limited knowledge of candidate targets(Zehir et al. 2017). Lack of efficacy is a leading cause of the 90% attrition rate in oncology drug development, and fewer molecular entities to new targets are being developed(Hay et al. 2014). Unbiased strategies that effectively identify and prioritise oncology targets could expand the range of targets, improve success rates, and accelerate development of new therapies.

CRISPR-Cas9 screens utilising libraries of single guide RNAs (sgRNAs) have been used to study gene function and their role in cellular fitness(Shalem et al. 2014; Koike-Yusa et al. 2014; Meyers et al. 2017). CRISPR-Cas9-based genome editing provides high specificity and produces penetrant phenotypes since null alleles can be generated. Here, we present genome-scale CRISPR-Cas9 fitness screens in 324 cancer cell lines and an integrative analysis to prioritise candidate cancer therapeutic targets (Fig. 1a), illustrated by the identification of Werner syndrome ATP-dependent helicase (WRN) as a target for tumours with microsatellite instability (MSI).

Project Score: Genome-scale CRISPR-Cas9 screens in cancer cell lines

To comprehensively catalogue genes required for cancer cell fitness (defined as genes required for cell growth or viability) we performed 941 CRISPR-Cas9 fitness screens in 339 cancer cell lines, targeting 18,009 genes (Extended Data Fig. 1a, b**,** and Supplementary Table 1). Following stringent quality control (Extended Data Fig. 1c-h), the final analysis set included 324 cell lines from 30 different cancer-types, across 19 different tissues **(**Extended

Data Fig. 1i). These cell lines are part of the Cell Model Passports collection of highly genomically-annotated cell lines(van der Meer et al. 2018), broadly represent the molecular features of patient tumours(Iorio et al. 2016), and include common cancers (lung, colon, breast) and cancers of particular unmet clinical need (lung and pancreas)(Bray et al. 2018). Analysis of screen data for these 324 cell lines demonstrated high sensitivity, specificity and precision in classifying essential and non-essential genes(Hart et al. 2015) (Extended Data Fig. 1g, h, j), and results were not biased by experimental factors (Extended Data Fig. 2a-e).

Defining core and context-specific fitness genes

Genes required for cell fitness in specific molecular or histological contexts are likely to encode favourable drug targets because of reduced likelihood of inducing toxic effects in healthy tissues(Narasimhan et al. 2016). Conversely, fitness genes common to the majority of cell lines tested or common within a cancer type (referred to as pan-cancer or cancer-type core fitness (CF) genes, respectively), may be involved in cell essential processes and have greater toxicity. It is therefore important to distinguish context-specific fitness genes from CF genes.

We identified a median of 1,459 fitness genes in each cell line (Extended Data Fig. 2f-n, and Supplementary Table 2). In total, 41% (n = 7,470) of all targeted genes induced a fitness effect in one or more cell lines (Fig. 1b). The majority (83%) of fitness genes induced a dependency in less than 50% of cell lines. To distinguish context-specific and CF genes, rather than selecting an arbitrary threshold, we developed a statistical method, ADaM (adaptive daisy model, Extended Data Fig. 3a-d), which adaptively determined the minimum number of dependent cell lines required for a gene to be classified as a CF gene (Fig. 1c). Genes which were defined as CF in at least 12 of 13 cancer types (also adaptively

determined) were classified as a pan-cancer CF genes (Extended Data Fig. 3e-g). This yielded a median of 866 cancer-type and 553 pan-cancer CF genes (Fig. 1c and Supplementary Table 3).

Of the ADaM pan-cancer CF genes, 399 were previously defined as essential genes(Hart et al. 2015; Hart et al. 2017) and 125 are genes involved in essential cellular processes(Tzelepis et al. 2016; Wang et al. 2014) (Extended Data Fig. 4a). The remaining 132 (24%) genes were newly identified and are also significantly enriched in cellular housekeeping genes and pathways (Extended Data Fig. 4b, c and Supplementary Table 4). Compared to reference CF sets(Hart et al. 2015; Hart et al. 2017), our pan-cancer CF gene set showed greater recall of genes involved in essential processes (median = 67%, versus 28% and 51% respectively, Extended Data Fig. 4d), with similar false discovery rates for putative context-specific fitness genes (taken from(McDonald et al. 2017), Extended Data Fig. 4e). Blood cancer cell lines had the most distinctive profile of CF genes (31 exclusive CF genes; Extended Data Fig. 4f). Cancer-type CF genes are generally highly expressed in matched healthy tissues (Extended Data Fig. 4g), consistent with their predicted role in fundamental cellular processes and suggesting potential toxicity as targets. Interestingly, five genes were CF in a single cancer-type and were lowly or not expressed at the basal level in the matched normal tissues (Extended Data Fig. 4g)**,** suggesting they could be cancer cell specific dependencies in these tissues.

Overall, using a statistical approach, we have refined and expanded current knowledge of human CF genes and identified genes that have a high likelihood of toxicity, thereby representing less favourable therapeutic targets. Furthermore, owning to the large scale of our dataset, we could now define context-specific fitness genes (median n = 2,813 genes per

cancer type), many of which had a loss-of-fitness effect similar to or stronger than CF genes (Fig. 1c).

A quantitative framework for target prioritization

To nominate promising therapeutic targets from our list of context-specific fitness genes, we developed a computational framework that integrated multiple lines of evidence to assign each gene a target priority score (0 - 100), and generated ranked lists of candidates for an individual cancer type or pan-cancer (Fig. 1a and Extended Data Fig. 5a). To exclude genes likely to be poor targets due to potential toxicity, CF genes were scored as zero, as were potential false-positives such as not expressed or homozygously deleted genes. Seventy percent of the priority score was derived from CRISPR-Cas9 experimental evidence and averaged across dependent cell lines based on (i) fitness effect size, (ii) significance of fitness deficiency, (iii) target gene expression, (iv) target mutational status, and (v) evidence for other fitness genes in the same pathway. The remaining 30% was based on (i) evidence of a genetic biomarker associated with a target dependency and (ii) the frequency at which the target is somatically altered in patient tumours(Iorio et al. 2016). For the biomarker analysis, we performed an analysis of variance (ANOVA; Fig. 2a, Extended Data Fig. 5b and Supplementary Data 1) to test associations between fitness genes and the presence of 381 cancer driver events (105 SNVs and 276 CNVs)(Iorio et al. 2016) or MSI, in each cancer type with sufficient sample size (n = 10 cell lines) and pan-cancer. Lastly, we derived a priority score threshold (55 and 41 for pan-cancer and cancer-type specific, respectively) based on scores calculated for targets with approved or pre-clinical cancer compounds (Extended Data Fig. 5c and Supplementary Table 5).

In total, we identified 628 unique priority targets, including 92 pan-cancer and 617 cancer-type specific targets (Figs. 2b, 3a, and Supplementary Tables 6 and 7). The number of priority targets varied ~3-fold across cancer types with a median of 88 targets. The majority of cancer-type priority targets (n = 457, 74%) were detected in only one (56%) or two (18%) cancer-types, underscoring their context specificity. Most priority pan-cancer targets (88%) were also identified in the cancer-type analyses (Extended Data Fig. 5d). Eleven genes identified uniquely as pan-cancer priority targets typically included dependencies in a small subset of cell lines across multiple cancer types (e.g. *CREBBP* and *JUP*), or in a cancer-type where limited numbers of cell lines were available, thereby being excluded from the cancer-type specific analysis (e.g. *SOX10* in melanoma; Extended Data Fig. 5e).

Of the 628 priority targets, 120 (19%) were associated with at least one biomarker with high significance and large effect size (defined as Class A targets), and thus would be of particular interest for drug development (Figs. 2a, c). For example, *PIK3CA* is a Class A target in breast, esophagus, colorectal and ovarian carcinoma; PI3K inhibitors are in clinical development for *PIK3CA*-mutated cancers(Massacesi et al. 2016). Using less stringent thresholds expanded the targets with at least one biomarker association, thus defining Class B (n = 61, 10%) and C (n = 117, 19%) targets, some of which were identified in multiple cancer types (Supplementary Table 8). Taken together, these results highlight the potential of a data-driven quantitative framework in prioritising targets by aggregating multi-cell line CRISPR-Cas9 screening data with associated genomic features.

Tractability assessment of priority targets

Based on current drug development strategies, targets vary in their suitability for pharmaceutical intervention and this informs target selection. Using a target tractability assessment for small molecule and antibody development, we assigned each gene to one of

10 tractability buckets (1 being highest tractability)(Brown et al. 2018). We cross-referenced the 628 priority targets with their tractability and categorised them into three tractability groups (Figs. 2b, 3a, and Supplementary Table 9).

Tractability Group 1 (buckets 1-3) comprises targets of approved anti-cancer drugs or compounds in clinical/pre-clinical development, and included 40 unique priority targets including, for instance, *ERBB2*, *ERBB3*, *CDK4*, *AKT1*, *ESR1*, *TYMS* and *PIK3CB* in breast carcinoma, and *PIK3CA*, *IGF1R*, *MTOR*, and *ATR* in colorectal carcinoma (Figs. 3a, 4 and Extended Data Fig. 6). Of these, 20 priority targets have at least one drug developed for the cancer-type in which the target was identified as priority, whereas the remaining 20 have drugs that have been used or developed for treatment of other cancer types, representing drug repurposing opportunities. A third of the Group 1 priority targets have a Class A biomarker, indicating highly desirable targets (Supplementary Tables 8 and 9). An example is *CSNK2A1*, which is a highly significant fitness gene in colorectal cancer cell lines with amplification of a chromosomal segment containing *FLT3* and *WASF3* (p = 6.65 x 10^{-6}, GΔs > 2.9, Fig. 3b) and targeted by silmasertib. Other Group 1 priority targets with markers are: *ERBB2* or *ERBB3* dependency in the presence of *ERBB2* amplification; *CDK2* dependency in *ASXL*-amplified esophageal cancer cell lines; *PIK3CA* dependency in the presence of *PIK3CA* mutations; and *PIK3CB* dependency in breast cancers cell lines with *PTEN* mutations (Fig. 3b and Supplementary Data 1).

Tractability Group 2 (buckets 4-7) contained 277 priority targets without drugs in clinical development but with evidence supporting target tractability (Figs. 3a, 4, Extended Data Fig. 6 and Supplementary Table 9). Of these, 18% have a Class A biomarker including: *KRAS* dependency in *KRAS* mutant cell lines; *USP7* in *APC* wild-type colorectal cell lines; *KMT2D* in breast cancer cell lines with amplification of a chromosomal segment containing *PPM1D*

and *CLTC*; and *TRIAP1* in *MYC*-amplified bone and gastric cancer cell lines (Fig. 3b and Supplementary Data 1). Of note, we observed a Class A biomarker dependency on *WRN* in MSI colorectal and ovarian cell lines, and pan-cancer (Fig. 3b). Of the Group 2 priority targets that were not associated with a biomarker, *GPX4* is a target in multiple cancer types (Fig. 4, Extended Data Fig. 6 and Supplementary Table 9). Sensitivity to GPX4 inhibition has been associated with epithelial-mesenchymal transition (EMT)(Viswanathan et al. 2017) and we observed differential EMT marker expression in GPX4-dependent cell lines (Extended Data Fig. 7a and Supplementary Data 2). This example is indicative of future refinements of our target prioritisation scheme to capture priority targets associated with an expanded set of molecular features including gene expression, chromatin and differentiation states.

Lastly, Group 3 (buckets 8 - 10) included 311 priority targets with no support or a lack of information to inform tractability (Figs. 3a, 4, and Extended Data Fig. 6), and is significantly enriched for transcription factors (Extended Data Fig. 7b and Supplementary Data 3). Examples of Group 3 priority targets with Class A biomarkers include *FOXA1* and *GATA3* in breast, *MYB* in hematological and lymphoid, *STX5* in ovarian, and *PFDN5* in neuroblastoma cell lines (Fig. 3b).

In contrast, Tractability Group 1 priority targets were enriched in protein kinases, highlighting a major focus in drug development, compared to Groups 2 and 3, which include a more functionally diverse set of targets (Extended Data Fig. 7b and Supplementary Data 3). Group 2 targets are most likely to be novel and tractable through conventional modalities, and therefore represent good candidates for drug development. Newer therapeutic modalities, such as proteolysis-targeting chimeras (PROTACs)(Sakamoto et al. 2001), may widen the range of proteins amenable to pharmaceutical intervention to include Group 3 targets.

Overall, our framework informed a data driven list of prioritised therapeutic targets serving as strong candidates for cancer drug development.

Werner helicase is a target in cancers with microsatellite instability

To substantiate our target prioritisation strategy, we investigated WRN helicase as a promising target in MSI cancers (Figs. 3 and 4). WRN is one of the five RecQ family DNA helicases, is the only one with both a helicase and an exonuclease domain, and plays diverse roles in DNA repair, replication, transcription and telomere maintenance(Chu and Hickson 2009). The MSI phenotype is caused by impaired DNA mismatch repair (MMR) due to silencing or inactivation of MMR pathway genes. MSI is associated with high mutational load and occurs in more than 20 tumours types, and is frequent in colon, ovarian, endometrial and gastric cancers (3 - 28%)(Cortes-Ciriano et al. 2017).

Dependency on *WRN* was highly associated with MSI in the pan-cancer ANOVA, and colon and ovarian analyses (Fig. 2a, 3b and Supplementary Data 1). Most MSI endometrial and gastric cancer cell lines were dependent on *WRN* but the association with MSI was not significant (for gastric) or not tested due to small sample size (Extended Data Fig. 7c). MSI is rare (<1%) in many other tumour types such as kidney, melanoma and prostate(Cortes-Ciriano et al. 2017), and most (4 of 5 tested) MSI cell lines from these tissues were not dependent on *WRN* (Extended Data Fig. 7c). Other tested RecQ family members (*BLM*, *RECQL* and *RECQL5*) were not associated as fitness genes in MSI cell lines. A focused analysis of non-synonymous mutations, promoter methylation and homozygous deletions of MMR pathway genes confirmed a significant association between *MLH1* promoter hypermethylation (Student's *t*-test FDR = 7.72×10^{-3}), *MLL2* (FDR = 1.43×10^{-4}) and *MSH6* (FDR = 3.85×10^{-2}) mutations and *WRN* dependency (Fig. 5a).

To further validate WRN, we performed a CRISPR-based co-competition assays comparing the relative fitness of *WRN* knockout versus wild-type cells. Knockout of WRN using four individual sgRNAs decreased fitness of *WRN* knock-out compared to wild-type cells exclusively in six MSI cell lines from colon, ovarian, endometrial and gastric cancer (Fig. 5b, Extended Data Fig. 8a, and Supplementary Table 10). In contrast, there was no difference in the fitness of *WRN* knock-out compared to wild-type cells in all microsatellite stable (MSS) cell lines from these four tissues. Similarly, *WRN* was selectively essential for MSI cells in clonogenic assays (Extended Data Fig. 8b, c). Of note, *WRN* knock-out had a potent effect on cell fitness, with an effect size similar to CF genes (Figs. 5a, b). Furthermore, we mined data from systematic RNAi screens and confirmed *WRN* dependency in MSI cancer cell lines (Extended Data Fig. 8d)(McDonald et al. 2017), and confirmed that *WRN* down-regulation by RNAi robustly impaired growth in MSI HCT116 cells (Extended Data Fig. 9a, b), thus providing validation in an independent orthogonal experimental system. Despite the association between MMR-deficiency and WRN dependency, knock-out of *MLH1* in MSS SW620 cells and culturing for 3 months did not induce *WRN* dependency, and conversely re-expression of MLH1 and/or MSH3 to restore MMR function in HCT116 cell lines did not rescue *WRN* dependency (Extended Data Fig. 9).

To determine if the loss-of-fitness effect was selective for WRN and identify a potential drug targeting strategy, we performed functional rescue experiments using wild-type, or hypomorphic versions of mouse *Wrn* (resistant to the *WRN* sgRNAs used) with a mutation in the exonuclease (E78A) or helicase (R599C or T1052G) domain that impairs protein function(Perry et al. 2006; Kamath-Loeb et al. 2004; Ketkar et al. 2017). Expression of wild-type or exonuclease-deficient Wrn rescued knock-out of *WRN* in MSI cells, whereas

expression of helicase-deficient Wrn led to no (R799C) or weak (T1052G) rescue (Fig. 5c, and Extended Data Fig. 8e, f). Thus, the helicase activity of WRN is required and an important domain for therapeutic targeting.

To evaluate *in vivo* sensitivity of MSI cells to WRN depletion, we developed a doxycycline-inducible *WRN* sgRNA system in HCT116 cells (HCT116-WRN)(Extended Data Fig. 9c, d). Following subcutaneous engraftment of HCT116-WRN cells into mice, treatment with doxycycline by oral gavage led to significant growth suppression of established tumours and a reduction in the number of proliferating cells (Fig. 5d-f, and Extended Data Fig. 9e, f). These findings confirm the role of WRN in sustaining *in vivo* growth of MSI colorectal cancer cells.

Discussion

New approaches are needed to effectively prioritize candidate therapeutic targets for oncology. We performed CRISPR-Cas9 screens in a diverse collection of cancer cells lines and combined this with genomic and tractability data to systematically nominate new cancer targets in an unbiased way. Even a modest improvement in drug development success rates, and an expanded repertoire of targets, through approaches such as ours could bring patient benefit. Our CRISPR-Cas9 screening results are also a rich resource with diverse applications in fundamental and evolutionary biology, genome engineering and disease genetics(Rancati et al. 2018). Results and data mining tools are available through a Project Score database (https://score.depmap.sanger.ac.uk/).

Despite the comprehensive and systematic approach used here, there are limitations to our study. First, we considered cell-intrinsic targets and not those involving the tumour

environment. Second, although CRISPR-Cas9 screens have high efficiency and specificity, we cannot exclude false-positive and false-negative results(Ong et al. 2017). Third, a small molecule inhibitor or therapeutic antibody may not recapitulate the effect of CRISPR-Cas9 genetic deletion, and redundancy between gene family members may mask dependencies that could be targeted with drugs. Finally, *in vitro* cell lines do not recapitulate some aspects of *in vivo* tumour biology. Thus, confirmatory studies are necessary to evaluate candidate targets.

We identified WRN protein as a promising new synthetic-lethal target in MSI tumours. This finding is corroborated by an accompanying study in this issue [Nature to insert reference]. WRN physically interacts with MMR proteins(Saydam et al. 2007), can resolve DNA recombination intermediates(Opresko et al. 2009), and the yeast homologue Sgs1 has a redundant function with MMR proteins to suppress homeologous recombination in regions of nucleotide mismatch(Myung et al. 2001). Together with our finding that rescue of MMR deficiency alone is insufficient to modulate WRN dependency, this suggests a model whereby WRN is required to efficiently resolve genomic structures present in MMR-deficient cells, potentially homeologous recombination structures, and failure to efficiently resolve these underpins the synthetic-lethal dependency. Mutation of *WRN* leads Werner Syndrome, an autosomal recessive disorder characterised by premature ageing and increased risk of cancer(Huang et al. 2006). Thus, loss of WRN is compatible with human development but targeting WRN could result in damage to normal cells so consideration should be given to maximising therapeutic benefit through patient selection and dose scheduling. A possible route for clinical development of WRN antagonists would be as an adjunct therapy to approved immune checkpoint inhibitors in MSI tumours(Le et al. 2015).

In summary, we developed an unbiased and systematic framework that effectively ranks priority targets, exemplified by WRN. Efforts such as ours, and from others(McDonald et al.

2017; Tsherniak et al. 2017; Meyers et al. 2017; Hart et al. 2015; Wang et al. 2017), to build a compendium of fitness genes, and the identification of context specific dependencies as part of a Cancer Dependency Map, could be transformative to improve success rates in oncology drug development.

References:

Acknowledgments:
We thank David Adams, George Vassiliou and Leo Parts for reading the manuscript. F.I. thanks Ethan Julian Iorio for his insightful comments on the visualisations included in this manuscript. We acknowledge members of Garnett lab, Wellcome Sanger Institute Sequencing Core Facility, Core IT Facility and Cytometry Core Facility for support.

Author contributions
M.G., K.Y., and C.B-D. conceived the project. F.B. led CRISPR-Cas9 screening, co-developed Project Score webportal, performed analyses, verified WRN dependency. F.I. led computational analyses and figure preparation, contributed to the Project Score webportal. G.P. performed experiments to verify WRN dependency, carried out analyses, contributed to *in vivo* studies. E.G. contributed to computational analysis and figures. D.vdM. contributed to developing the Project Score webportal. G.M., F.S., M.P., A.B., L.T. performed *in vivo* studies. C.B., R.A., A.D.J., R.McR., R.P., and P.W. performed CRISPR-Cas9 screens. R.M.S. performed tractability analysis. Y.R. performed WRN rescue experiments. C.B., S.H., A.B., L.T., E.S., D.D., and J.S-R. assisted project supervision. F.B., F.I., E.G., G.P., K.Y. and M.G. wrote the manuscript. K.Y. and M.G. directed the project. Funding acquisition: J.S-R., A.B., L.T., M.G. and K.Y. All authors reviewed and approved the manuscript.

Funding:
The work was funded by Open Targets (OTAR015) to M.G., K.Y. and J.S-R. M.G lab is supported by CRUK (C44943/A22536), SU2C (SU2C-AACR-DT1213) and Wellcome Trust (102696 and 206194). Sanger Institute core facilities are funded by Wellcome Trust (206194)
Also supported by AIRC, Associazione Italiana per la Ricerca sul Cancro, Investigator Grants 20697 (A.B.) and 18532 (L.T.) and 5x1000 grant 21091 (A.B. and L.T.); European Research Council Consolidator Grant 724748 – BEAT (A.B.); Fondazione Piemontese per la

Ricerca sul Cancro-ONLUS, 5x1000 Ministero della Salute 2011 and 2014 (L.T.); and Transcan, TACTIC (L.T.).

Competing interests:

E.A.S, D.D, C. B-D, R.M.S, Y.R are employees of GSK. This works was funded by Open Targets: a public-private initiative involving academia and industry. K.Y. received research funding from AstraZeneca. M.G receives research funding from AstraZeneca and has performed consultancy for Sanofi. All other authors declare no competing interests.

Data and materials availability.

Supplementary Data 1, 2 and 3 available at:

https://figshare.com/projects/CRISPRtargetID/60146

Cell line gene fitness scores, sgRNA raw counts, processed data and results are available on the Project Score webportal: https://score.depmap.sanger.ac.uk (direct links in the Methods). Software code are available through github (URLs in the Methods).

Figure Legends

Fig 1: Target prioritization framework. a, Strategy to assign target priority scores in multiple cancer-types incorporating CRISPR-Cas9 gene fitness effects, genomic biomarkers and target tractability for drug development. ADaM (adaptive daisy model) distinguishes context specific and core fitness genes. Data sets are available on the Project Score website. **b**, Number of genes exerting a fitness defect in a given number of cell lines. The bars indicate the percentage of genes which induce a dependency in less than (top bar) or at least (bottom bar) 50% of cell lines. **c**, (Bottom) Number of core and context-specific fitness genes predicted by ADaM for 13 cancer-types (median = 866, and 2,813, respectively; n = 311 cell lines). The ADaM threshold is the number of cell lines a gene must be called as a fitness gene to be classified as core fitness. (Top) Comparison of the effect size for ADaM core and context-specific fitness genes (only significant genes are shown, 5% BAGEL FDR).

Fig. 2: Target prioritization and biomarker discovery. a, ANOVA differential dependency biomarker analyses. Each point is an association between the fitness effect of a gene (name in bold) and a molecular feature or MSI (italic). Colours indicate results from 13 cancer-type specific (n cell lines indicated in Supplementary Table 1) or pan-cancer (n = 319) analysis. False discovery rates were calculated using the Storey-Tibshirani method. **b**, Cancer-type specific and pan-cancer priority targets classified based on tractability for drug development as Group 1 - 3 (from strong to weak/absence of evidence). **c**, Priority targets with a genomic biomarker and defined as Class A, B or C, depending on statistical significance and effect size.

Fig. 3: Priority targets and biomarker linked dependencies. a, All priority targets from cancer-type and pan-cancer analyses and their tractability. Priority score thresholds are indicated and selected examples labelled. **b**, Differential fitness analysis (quantile normalised logFCs) for selected priority targets comparing cells with (+) or without (-) a genomic marker (classes A - C as previously defined from ANOVA tests). Each data point is a cell line and colors represent cancer-type. Box and whiskers are interquartile ranges and 95th percentiles, centres indicate medians.

Fig. 4: Cancer-type priority targets. Results for 4 of 13 cancer-type specific analyses. Points are target priority scores with shapes indicating approved/pre-clinical compound to the corresponding target (anti-cancer or specific to the cancer-type considered), and circles the absence of a compound. Symbols indicate the strength of a genomic biomarker. Selected priority targets are labelled.

Fig. 5: WRN is a target in MSI cancer cells. a, Circle plot of cell lines showing from outside: (i) fitness effect of *WRN* knockout and mean effect of core fitness genes (red dashed line); (ii) cancer-type; (iii) *MLH1* methylation status; (iv) *MLL2* and (v) *MSH6* mutation status; and (vi) DNA mutation rate. **b**, *WRN* dependency in a co-competition assay. Control sgRNAs targeting essential (sgEss) and non-essential genes (sgNon). Each point represents mean co-competition score for a cell line (7 MSI and 7 MSS lines in duplicate); four WRN sgRNA guides were used. A score less than 1 denotes selective depletion of sgRNA knockout cells. Mean differences were estimated using a two-sided Welch's t-test. Boxplots represent 1.5 of the interquartile range. **c**, *WRN* rescue using wild-type (WT), exonuclease-deficient or helicase-deficient *Wrn* in MSI SW48 cancer cells. Mean values from 3 independent experiments ± SD are shown (*p-value*s are two-sided *t*-test comparison to WT *Wrn;* n.s = not significant). **d**, Tumor volume of HCT116-WRN (clone a) xenografts treated with doxycycline (doxy; yellow line) compared to vehicle (grey line; *p*-value = 0.006, two-way ANOVA). Values represent mean ± SEM. Numbers of mice in each cohort are shown. **e**, Representative Ki-67 immunohistochemical assessment of HCT116-WRN (clone a) tumors explanted after one week of doxy (40x magnification; scale bar = 50 μm). **f**, Quantification of Ki-67 staining. Results are mean ± SD of 10 fields from three different samples (n = 30) and means were compared using a two-sided Welch's t-test.

Legends of Extended data items - Supplementary Figures

Extended Data Figure 1: Project Score CRISPR-Cas9 screening pipeline, data quality control, and analysis set. a, CRISPR-Cas9 screening pipeline workflow including quality control (QC) steps and go/no-go decisions. **b**, Genomic characterisation of the CRISPR-Cas9 screened cell lines. **c**, Average Pearson's correlation of replicate sgRNA counts (n = 86,875) for individual cell lines. **d**, Data QC threshold (vertical line) based on the distributions of Pearson's correlation values of sgRNA fold-changes between replicates of the same cell line (in green) and all possible pairwise comparisons (in gray), considering only highly informative sgRNAs (n = 838, detailed in the Methods). **e**, Percentage of experiments passing the QC filter based on the threshold defined in d. **f**, Pearson's correlation values as described in d for the cell lines in the final analysis set. **g**, ROC and Precision/Recall curves obtained classifying predefined sets of essential (n = 354) and non-essential (n = 747) genes, based on gene-level depletion logFC rank positions. The median area under the curve across all cell lines are reported. **h**, Glass' delta scores quantifying the depletion effect size for ribosomal protein (n = 61) genes and *a priori* known essential (n = 354) genes for all cell lines. **i**, Cell lines in the final analysis set grouped by tissue (inner ring) and cancer-type (outer ring). **j**, Median

logFC values and inter-quartiles for reference gene sets defined in g and h for 324 cell lines in the analysis set.

Extended Data Figure 2: Assessment of technical confounders in CRISPR-Cas9 screening data and summary of fitness genes. a, Absence of association between screening data quality and the number of replicates (as quantified by a Pearson's correlation with respect to the number of replicates, n = 5 distinct values). Data quality was assessed using the fitness effect (the median logFC) of ribosomal protein (n = 61) genes in each cell line as a reference. **b**, Absence of association between data quality (quantified as in a) and average Pearson's correlation between replicates of individual screened cell lines (n = 324). The p-value refers to a two-sample Student t-test, the score on the right plot is a Pearson's correlation. **c**, Weak correlation and significant association between sgRNA library transduction efficiency in cell lines (averaged for replicates) and data quality. **d**, Weak correlation and significant association between cell line Cas9 activity (averaged for replicates) and data quality. **e**, Absence of association between library coverage and data quality. In c, d, and e, *p-values*, R and sample size n are defined as for b. **f**, Number of fitness genes in each cell line (BAGEL FDR<5%; median = 1,459). **g**, Number of cell lines with fixed intervals of numbers of fitness genes. **h**, Absence of correlation between number of significant fitness genes per cell line and number of replicates, R and samples size n defined as for a. **i**, The effect of sgRNA screening library version on the number of fitness genes identified. A new version of the library (v 1.1) with additional guides for a subset of genes yields moderately larger numbers of fitness genes, however this is equally variable in both groups and confounded by the tissue of origin of the cell lines. P-value is defined as for b. **j**, Reproducible calling of fitness genes in HT-29 across sgRNA libraries. Barplot of the number of fitness genes detected with each library (left). Scatter plots of depletion scores at genome-wide level or considering only highly informative sgRNAs for each library (right). In both cases, Fisher Exact test *p*-value is below machine precision ($< 10^{-16}$). R indicate Pearson's correlation, while C indicates the percentage of genes called as significantly depleted with both libraries over those detected as significantly depleted with one library only. **k**, Pearson's correlation between the number of fitness genes per cell line and Cas9 activity level and library transduction efficiency. **l**, Pearson's correlation between number of fitness genes per cell line and the average Pearson's correlation among replicates of the cell line under consideration. **m**, **n**, Pearson's correlation between number of fitness genes per cell line and the ability to detect a defined set of essential genes. For all panels each data point is a cell line coloured by cancer-type (except panels g and j), and box-and-whisker plots indicate median, interquartile range and 95 percentiles.

Extended Data Figure 3: Computation of ovary specific (as an example) and pan-cancer core fitness genes with the adaptive daisy model (ADaM), and summary of context-specific/core-

fitness genes. a, Number of fitness genes in each ovary cell line. **b**, Number of fitness genes in fixed number *m* of cell lines. **c**, Distributions and cumulative distributions of number of fitness genes observed in *m* cell lines across 1,000 randomised versions of the depletion scores for ovary cell lines. **d**, True positive rates (where positive are *a priori* known essential genes) when considering as predictions the genes that are depleted (fitness genes) in at least *m* cell lines (blue curve), and deviance of the number of these gene from expectations (computed using randomised data shown in c), for all possible *m* values (red curve). The x-coordinate (rounded by excess) of the intersection of these two curves estimates the minimal number of cell lines *m** in which a gene should be significantly depleted in order to be predicted as a core-fitness (CF) gene for a cancer-type. **e**, Number of genes predicted as cancer-type specific CF for a fixed number *k* of cancer-types. **f**, Distributions and cumulative distributions of number of CF genes predicted for a fixed number of *k* tissue types for 1,000 randomised versions of the cancer-type specific CF profiles. **g**, True positive rates (where positive are *a priori* known essential genes) when considering as predictions the genes that are CF for at least *k* cancer-types (blue curve), and deviance of the number of these gene from expectation (computed using randomised data in g; red curve). The x-coordinate estimates the minimal number of cancer-types *k** for which a gene should have been predicted as cancer-type CF in order to be classified as a pan-cancer CF gene. All box and whiskers plots are interquartile ranges and 95th percentiles, with centres indicating medians.

Extended Data Figure 4: Characterization of ADaM pan-cancer core fitness genes. a, Membership of 553 pan-cancer core fitness genes in reference essential gene sets (gray bars) and respective recall and enrichment significance p-values from a hypergeometric test considering as background population the whole set of genes targeted in the CRISPR-Cas9 screen (N = 17,995). 132 newly identified core fitness genes fall outside of these reference gene sets. **b**, Pathways and **c**, gene families enriched in n = 132 pan-cancer core fitness genes uniquely identified by ADaM (Benjamini-Hochberg adjusted hypergeometric test *p*-value < 0.05). **d**, Comparison of the ADaM core fitness genes and two previously reported reference sets of essential genes in terms of number of genes, estimated precision and recall (considering as true positive the genes included in reference gene sets corresponding to cellular essential process). **e**, False discovery rates of putative context-specific fitness genes at different thresholds of reliability (n = 7393, 2233, 426 and 82, respectively). **f**, Clustering of cancer-types based on core fitness gene similarity (left) and numbers of core fitness genes exclusive to each cancer-type (right). **g**, Basal expression of cancer-type specific core fitness genes (n indicated in Extended Data Fig. 3e) in matched normal tissues. Five genes were identified as core fitness genes in a single cancer-type and are not expressed at the basal level (< 5% quantile) in matched normal tissue (red points). Cancer-types are coloured as shown in panel f. Box and whiskers

are interquartile ranges and 95th percentiles, with sample sizes indicated in f (barplot on the right), centres indicating median values.

Extended Data Figure 5: Pan-cancer and cancer-type specific priority scores. a, Criteria for the target prioritization scoring system. **b**, ANOVA results from a differential dependency biomarker analyses with all 1,001 significant associations classified as pan-cancer or cancer-type specific (inner circle), loss or gain of fitness marker (middle circle), and whether the marker is a mutation, copy number gain or loss (outer circle). **c**, Distributions of pan-cancer (left) and cancer-type specific (right) non-null target priority scores based on the therapeutic indication of approved/pre-clinical compound. The significance threshold was based on the distribution of scores for targets with approved anti-cancer compounds (specific anti-cancer compounds for the cancer-type specific priority score) versus scores for targets with no available anti-cancer compounds. **d**, Overlap between cancer-type specific priority targets (for at least one cancer-type) and pan-cancer priority targets. **e**, Example priority targets identified only in the pan-cancer context. Each symbol is an individual cell line colored by cancer-type and symbol shapes indicate a significant dependency in the cell line. Values across n = 324 cell lines in the analysis set are reported.

Extended Data Figure 6: Priority therapeutic targets in ten cancer-types and pan-cancer. Each data point is a target with a priority score classified into tractability buckets and groups. The shape represent the indication of the approved/pre-clinical compound to the corresponding target (anti-cancer or specific to the cancer-type considered), and circles indicate the absence of a compound. Symbols within each data point indicate the strength of the genomic marker associated with differential dependency on the target (Class A to C, from strong to weak association).

Extended Data Figure 7: GPX4 fitness selectivity for cells undergoing EMT, functional classification of priority targets, and WRN differential fitness in other cancer-types. a, Differentially expressed genes in cell lines that are dependent on *GPX4* (left)(n = 113, moderated t-statistic FDR estimates). Epithelial-mesenchymal transition is the top differentially enriched cancer hallmark gene signature in *GPX4* dependent cell lines (right). Single sample GSEA p-values were obtained by randomly permuting gene signatures 10,000 times and adjusted for multiple testing using Benjamini-Hochberg FDR correction. **b**, Functional classification of priority targets in each tractability group using the PANTHER database. For clarity, kinases (a subset of transferases) and transcription factors are shown separately. Protein classes are indicated by color. Statistical enrichment was calculated using a systematic hypergeometric test across protein families, following correction for multiple hypothesis testing with Benjamini-Hochberg method. Pie charts indicate the percentage of targets in each Group classified in protein families. **c**, *WRN* dependency in multiple cancer-types. Each data point is a cell line indicating quantile normalised *WRN* depletion logFCs

stratified by MSI status. Box and whiskers are interquartile ranges and 95th percentiles. Statistical significance was calculated from the systematic ANOVA analysis for each cancer-type where the number of cell lines was greater than 10 (n = 14 for gastric carcinoma).

Extended Data Figure 8: Verification of WRN as a target in MSI cancers. a, *WRN* dependency using a co-competition assay in MSS (n = 7) and MSI (n = 7) cell lines from four cancer-types. sgRNAs targeting an essential (sgEss) and non-essential gene (sgNon) were controls. **b**, Selective *WRN* dependency in MSI versus MSS cell lines was confirmed using clonogenic assays in four cancer-types. **c**, A reduction in WRN protein levels with all WRN sgRNAs was confirmed by Western blot. **d**, An association between *WRN* dependency and MSI status was confirmed mining data from an independent RNAi study, Project DRIVE (*p*-value = 0.004). Each circle represents the *WRN* RNAi dependency score in a cancer cell line. **e**, Expression of wild-type mouse *Wrn* rescued the viability effect of *WRN* knockout in MSI cell line SW48. MSS cell line SW620 was a negative control. **f**, Western blots confirmed expression of protein using all variants of the *Wrn* vector.

Extended Data Figure 9: Validation of WRN dependency with RNA interference and *in vivo* validation of WRN dependency in MSI colorectal cancer cell line. a, siRNA depletion of WRN inhibited proliferation of HCT116 cells. Data are the mean ± SD of three independent experiments. P-value was determined using a non-parametric student t-test. b, siRNA-mediated depletion of WRN was verified by Western blot (one representative experiment of two performed). For western blot source data, see Supplementary Fig. 1. c, WRN knockout induced by doxycycline treatment in HCT116-WRN cells measured by Western blot for two separate clonal lines (one representative experiment of two performed) . For western blot source data, see Supplementary Fig. 1 d, Growth curves of HCT116 parental, HCT116-sgNon and HCT116-WRN cells grown in the absence (black line) or presence of doxycycline (2 μg/ml; yellow line), (n=10). Data are the mean ± SD of one representative experiment of two performed. e, Growth curves of HCT116-WRN (clone b) subcutaneous tumours from mice treated with doxycycline (50 mg/kg; yellow line) or vehicle (grey line). Tumour growth suppression was observed (p-value = 0.03, two-way ANOVA comparing doxycycline versus vehicle). The number of mice in each cohort are indicated. Values are mean ± SEM. f, Representative Ki-67 immunohistochemical assessment of HCT116-WRN (clone b) tumors explanted after one week of doxy (40x magnification; scale bar = 50 μm). f, Quantification of Ki-67 staining. Results are mean ± SD of 10 fields from three different samples (n = 30) and means were compared using a two-sided Welch's t-test.

Methods

CRISPR-Cas9 screening: Plasmids

All plasmids have previously been described(Tzelepis et al. 2016) and are available through Addgene (Cas9 vector - 68343; gRNA vector - 67974). Plasmids were packaged using the ViraPower Lentiviral Expression System (Invitrogen; Cat No. K4975-00) as per manufacturer's instructions.

CRISPR-Cas9 screening: Cell culture

Cell lines used in this study (Supplementary Table 1) were selected from the Genomics of Drug Sensitivity in Cancer (GDSC) 1,000 cell line panel(Iorio et al. 2016) and maintained as previously described. To control identify cross-contamination and sample swap a panel of 92 SNPs was profiled for each cell line before and following completion of the CRISPR-Cas9 screening pipeline. A separate set of HCT116 cell lines were used for WRN validation experiments: HCT116 parental, HCT116-Ch3, HCT116-Ch5 and HCT116-Ch3+Ch5 were a kind gift of Dr. Minoru Koi (University of Michigan). HCT116-Ch2 cells were a generous gift of Dr. Ajay Goel (Baylor Charles A. Sammons Cancer Center). HCT116-Ch2 and HCT116-Ch3 were maintained in 400 ug/mL G418 (Thermo Fisher Scientific; Cat No. 10131027); HCT116-Ch5 were maintained in 6 ug/mL blasticidin (Thermo Fisher Scientific; Cat No. A1113903); and HCT116-Ch3+Ch5 were maintained in the presence of 400 ug/mL G418 and 6 ug/mL blasticidin. All the cells were cultured in McCoy's 5A medium (Sigma; Cat No. M4892) with 10% FBS.

CRISPR-Cas9 screening: Generation of Cas9-expressing cancer cell lines

Cells were transduced with lentivirus containing Cas9 in T25 or T75 flasks at ~80% confluence in the presence of polybrene (8μg/ml). Cells were incubated overnight followed by replacement of the lentiviral containing medium with fresh complete medium. Blasticidin selection commenced 72 hours post-transduction at an appropriate concentration determined

for each cell line using a blasticidin dose response assay (blasticidin range 10-75μg/ml) and cell viability assessed using CellTiter-Glo 2.0 Assay (Promega; Cat No. G9241)). Cas9 activity was assessed as described previously(Tzelepis et al. 2016). Cell lines with Cas9 activity >75% progressed for sgRNA library transduction.

CRISPR-Cas9 screening: Genome-wide sgRNA library and screen

Two genome-wide sgRNA libraries were used in this study: Human CRISPR Library v1.0 and Human CRISPR library v1.1. The Human CRISPR library v1.0 is described previously(Tzelepis et al. 2016) and targets 18,025 genes. Human CRISPR library v1.1 contains all sgRNAs from v1.0 plus an additional 1,004 non-targeting sgRNAs. Library v1.1 was synthesised using high-throughput silicon platform technology (Twist Bioscience), resulting in a more uniform representation of individual sgRNAs. A subset of genes in library v1.1 were targeted with an additional 5 sgRNAs. This subset of genes was comprised of kinases, epigenetic related genes and *a priori* known essential genes. Although, for consistency, all the computational analyses were focused on the set of overlapping sgRNAs across the two libraries only, and counts for the additional guides in library v.1.1 where removed at the preprocessing phase and not used in this study. The HT-29 cell line was screened with both libraries and resulting datasets kept separated for comparative analyses (results summarised in Extended Data Fig. 3k).

A total of 3.3 x 10^7 cells were transduced with an appropriate volume of lentiviral packaged whole-genome sgRNA library to achieve 30% transduction efficiency (100x library coverage). The volume was determined for each cell line using a titration of packaged library and assessing the percentage of BFP-positive cells by flow cytometry. Transductions were performed in technical triplicate (or duplicate for cell lines with a large cell size such as

glioblastoma). Due to the large number of screens performed, multiple batches of packaged library virus were prepared. Each batch was tested in HT-29 to ensure consistency of batch preparations. In addition, HT-29 was screened every 3 months to ensure the quality of data generated by the pipeline was consistent. Transduction efficiency was assessed 72 hours post transduction. Samples with transduction efficiency between 15-60% proceeded to puromycin selection. The appropriate concentration of puromycin for each individual cell line was determined from a dose response curve (puromycin range 1-5µg/ml) and cell viability assessed using CellTiter-Glo 2.0 Assay (Promega; Cat No. G9241)). The percentage BFP-positive cells was reassessed after a minimum of 96 hours of puromycin selection. For samples with <80% BFP-positive cells, puromycin selection was extended for an additional 3 days and the percentage BFP-positive cells were assessed again. Cells were maintained until day 14 post transduction with a minimum of 5.0 x 10^7 cells reseeded at each passage (500x library coverage). Approximately 2.5 x 10^7 cells were harvested, pelleted and stored at -80 ^oC for DNA extraction.

CRISPR-Cas9 screening: DNA extraction, sgRNA PCR amplification, Illumina sequencing and sgRNA counting

Genomic DNA was extracted from cell pellets using either the QIAsymphony automated extraction platform (Qiagen; QIAsymphony DSP DNA Midi Kit, Cat No. 937255) or by manual extraction (Qiagen; Blood & Cell Culture DNA Maxi Kit, Cat No. 13362) as per manufacturer's instruction. PCR amplification, Illumina sequencing (19-bp single-end sequencing with the custom primer on the HiSeq2000 v4 platform) and sgRNA counting were performed as described previously(Tzelepis et al. 2016).

CRISPR screen data analyses: low-level Quality Control (QC) assessment and filtering

To perform initial low-level QC, the Pearson's correlation between replicate sgRNA genome-wide treatments counts was assessed for each cell line (results shown in Extended Data Fig. 1c). The resulting correlation scores were generally high (median = 0.8) but not sufficiently distinguishable from expectation (median correlation between replicates of any pair of randomly selected cell lines). Thus, to define a reproducibility threshold, we developed an approach based on(Ballouz and Gillis 2016). Specifically, we selected a set of 838 most informative sgRNAs. These were defined within sets of sgRNAs targeting the same gene, as those with an average pair-wise Pearson's correlation between corresponding patterns of logFCs across all screened cell lines > 0.6. We next computed average gene-level logFC profiles for 308 genes targeted by these informative sgRNAs for each individual technical replicate, and then computed all possible pairwise Pearson's correlation scores between the resulting profiles. This allowed the estimation of a null distribution of replicate correlations (plotted in gray in Extended Data Fig. 1d). We then defined a reproducibility threshold R value of 0.68, where the estimated probability mass function of the correlation scores computed between replicates of the same cell line (considering the identified 308 genes only) was at least twice that of the null mass probability function (Extended Data Fig. 1d). Of the 332 screened cell lines with at least two technical replicates, 305 had an average replicate correlation higher than this threshold, thus passed the reproducibility assessment, and for 7 cell lines there were no replicates. Excluding the least reproducible replicate for 14 cell lines not passing the first reproducibility assessment allowed their average replicate correlation to exceed the threshold defined above, thus resulting in a set of 326 cell lines passing low-level QC assessment (Supplementary Table 1).

CRISPR screen data analyses: screening performance assessment

We considered the genome-wide profiles of gene-level logFCs (averaged across targeting sgRNAs and replicates) of each cell line as a classifier of predefined sets of essential and non-essential genesHart and Moffat 2016), respectively *E* and *N*, by means of Receiver Operating Characteristic (ROC) indicators (results in Extended Data Fig. 1g and Supplementary Table 1). In addition, we measured the depletion signal magnitude observed in each screened cell line by evaluating the median logFC, and the discriminative distance between their distributions (as measured by the Glass's Δ), for predefined essential and non-essential genes(Hart and Moffat 2016), and ribosomal protein genes(Yoshihama et al. 2002). Two of 326 cell lines were removed because they had an area under the ROC curve, area under the Precision/Recall curve, and both Glass's Δs values 3 standard deviations lower than the average. Based on our low-level QC and screening performance, the final analysis set was composed of 324 cell lines (Supplementary Table 1). Further details on these analysis are included in the Supplementary Information.

CRISPR screen data analyses: sgRNA count preprocessing and CRISPR-bias correction

The analysis set of 324 cell lines was further processed using *CRISPRcleanR*(Iorio et al. 2018)(https://github.com/francescojm/CRISPRcleanR). sgRNAs with less than 30 reads in the plasmid and sgRNAs belonging to library v1.1 only were removed from further analysis. The remaining sgRNAs were assembled into one file per cell line, including the read counts from the matching library plasmid and all replicates, followed by normalisation using a median-ratio method to adjust for the effect of library sizes and read count distributions(Anders and Huber 2010). Depletion/enrichment logFCs for individual sgRNAs were quantified between post library-transduction read-counts and library plasmid read-counts at the individual replicate level. This was performed using the *ccr.NormfoldChanges*

function of CRISPRcleanR. Next we performed a correction of gene independent responses to CRISPR-cas9 targeting(Aguirre et al. 2016) using the *ccr.GWclean* function of CRISPRcleanR with default parameters.

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

The CRISPRcleanR-corrected sgRNAs-level logFCs values (corrected logFCs) were used as input to an in-house R implementation of the BAGEL method(Hart and Moffat 2016) to call significantly depleted genes (code publicly available at https://github.com/francescojm/BAGELR). Our BAGEL implementation computes gene-level Bayesian factors (BFs) by the sgRNAs on a targeted-gene basis, instead of summing them. Additionally, it uses reference sets of predefined essential and non-essential genes(Hart and Moffat 2016). However in order to avoid their status (essential/non-essential) being defined *a priori*, we removed any high-confidence cancer driver genes as defined in(Iorio et al. 2016). The resulting curated reference gene sets are available as built-in data objects in the R implementation of BAGEL (*curated_BAGEL_essential.rdata* and *curated_BAGEL_nonEssential.rdata*). A statistical significance threshold for gene-level BFs was determined for each cell line as in(Hart et al. 2015). Each gene was assigned a scaled BF computed by subtracting the BF at the 5% FDR threshold defined for each cell line from the original BF, and a binary fitness score equal to 1 if the resulting scaled BF was > 0. Further details on these analysis are included in the Supplementary Information.

In addition, CRISPRcleanR-corrected sgRNA treatment counts were derived from the corrected sgRNA-level logFCs (using the *ccr.correctCounts* function of CRISPRcleanR) and used as input to MAGeCK(Li et al. 2014) for computing depletion significance using mean-variance modeling. This was performed using the MAGeCK python package (version 0.5.3), specifying in the command line call that no normalisation was required (as already performed

by CRISPRcleanR). At the end of this stage, the following gene-level depletion score matrices were produced in each cell line: raw logFCs, copy number bias-corrected logFCs, BFs, scaled BFs, binary fitness scores and MAGeCK depletion FDRs. All these scores are summarised for each cell line and available at https://cog.sanger.ac.uk/cmp/download/essentiality_matrices.zip, together with all the sgRNAs raw count files (available at https://cog.sanger.ac.uk/cmp/download/raw_sgrnas_counts.zip).

High-level data analyses: Adaptive Daisy Model (ADaM) to identify core fitness genes

We designed the adaptive Daisy model (ADaM), an heuristic algorithm for the identification of core fitness (CF) genes, implemented it in an R package and made it publicly available at https://github.com/francescojm/ADaM. ADaM is based on the Daisy Model(Hart et al. 2015), but it adaptively determines the minimal number of cell lines *m* from a given tissue in which a gene should exert significant fitness effect in order for that gene to be considered a CF genes for that tissue. ADaM is detailed in the Supplementary Material. In addition, we applied the same method to determine the minimal number *k* of cancer-types for which a gene should be predicted as CF gene in order to be considered as a pan-cancer CF gene.

High-level data analyses: Characterisation of ADaM pan-cancer CF genes

Reference sets of essential (*E*) and non-essential (*N*) genes were extracted from(Hart and Moffat 2016). Other reference gene sets (used while characterising the ADaM pan-cancer CF genes, detailed below) were derived from the Molecular Signature Database (MSigDB(Subramanian et al. 2005)) and post-processed as detailed in(Iorio et al. 2018). A more recent set of *a priori* known essential genes was derived from(Hart et al. 2017). The

pan-cancer CF genes not belonging to any of the aforementioned gene sets were tested for gene family enrichments (via hypergeometric test) by deriving gene annotations using the BioMart R package(Durinck et al. 2005), and biological pathway enrichments using a comprehensive collection of pathways gene sets from Pathway Commons(Cerami et al. 2011) (post-processed to reduce redundancies across different sets as detailed in(Iorio et al. 2018)). All enrichment *p*-values were corrected with the Benjamini-Hochberg method. Results are in Supplementary Table 4.

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

We compared the pan-cancer core fitness genes predicted by ADaM with the BAGEL reference set of essential genes(Hart and Moffat 2016), and a more recently proposed larger set of essential genes(Hart et al. 2017) in terms of size, estimated precision (number of included true positives / number of included genes) and recall (number of included true positives / total number of true positives). In these comparison we used gold-standard essential genes involved in cell essential processes (downloaded from the MSigDB(Subramanian et al. 2005) and post-processed as detailed in(Iorio et al. 2018)). In addition, we estimated false discovery rates for the three gene sets (number of included false positives / total number of false positives) considering as false positives genes predicted to be strongly context-specific essential (thus not core-fitness essential) according to an independent publication(McDonald et al. 2017), and using three different confidence levels, as further detailed in the Supplementary Information.

High-level data analyses: Basal expression of cancer-type specific CF genes in normal tissues

Basal gene median reads per kilobase of transcript per million mapped reads in normal human tissues were download from the GTEx Portal(GTEx Consortium 2013), log transformed and quantile normalised on a tissue type basis.

Statistical analyses: Analysis of Variance to identify genomic correlates with gene fitness We performed a systematic analysis of variance (ANOVA) to test associations between gene-level logFC fitness effects and the presence of 484 cancer driver events (CDEs; 151 SNVs and 333 CNVs)(Iorio et al. 2016) or microsatellite instability (MSI) status at pan-cancer as well as individual cancer-type levels. Ten cancer-types with at least 10 screened cell lines were analysed (Breast Carcinoma, Colorectal Carcinoma, Gastric Carcinoma, Head and Neck Carcinoma, Lung Adenocarcinoma, Neuroblastoma, Oral Cavity Carcinoma, Ovarian Carcinoma, Pancreatic Carcinoma, Squamous Cell Lung Carcinoma). The remaining cancer-types were collapsed on a tissue basis (annotation in Supplementary Table 1) and resulting tissues with at least 10 cell lines were included in the analysis (Bone, Central Nervous System, Esophagus, Hematopoietic and Lymphoid). A total of 14 analyses (referred for simplicity as cancer-type specific ANOVAs in the main text and in what follows) plus a pan-cancer analysis including all screened cell lines were performed. Each ANOVA was performed using the analytical framework described in(Iorio et al. 2016) and implemented in a Python package(Cokelaer et al. 2017), publicly available at https://github.com/CancerRxGene/gdsctools. Only genes not belonging to any set of prior known essential genes (defined in the previous sections) and not predicted by ADaM to be core-fitness essentials were included in the analyses. For all the tested gene

fitness/CDE associations, effect size estimations versus pooled standard deviation (quantified through the Cohen's d), effect sizes versus individual standard deviations (quantified through two different Glass's Δs, for the CDE-positive and the CDE-negative population respectively), CDE *p*-values and all the other statistical scores were obtained from the fitted models. An association was tested only if at least three cell lines were contained in the two sets resulting from the dichotomy induced by the CDE-status (i.e. at least 3 CDE-positive and 3 CDE-negative cell lines). The *p*-values from all ANOVA were corrected together using the Tibshirani-Storey method(Storey and Tibshirani 2003). Subsequently, for the MSI status was also tested for statistical associations with differential gene fitness effects pan-cancer and for cancer-types with at least 3 MSI cell lines. We used the following statistical significance and effect size threshold to categories associations between gene fitness effects and genomic markers:

Class A marker: a *p*-value threshold of 10^{-3} with a false discovery rate (FDR) threshold equal to 25% (or 5% for MSI) and both Glass's Δs > 1. Different FDR thresholds were used for associations with CDEs or MSI due to the number of tests performed in the former being 6 orders of magnitude larger than the latter.

Class B marker: a FDR threshold of 30% with at least one Glass's Δ > 1 for pan-cancer associations.

Class C marker or weaker: an ANOVA *p*-value threshold of 10^{-3} with at least one Glass's Δ > 1 for pan-cancer associations, and a simple *t*-test (for mean depletion logFC difference assessment between CDE-positive/CDE-negative cell lines) *p*-value threshold of 0.05 (with at least one Glass's Δ > 1 for pan-cancer associations).

The additional constraint of on the Glass's Δs (quantifying the effect size with respect to the standard deviations of the two involved sub-populations of samples) was considered for the pan-cancer markers in order to account for the significantly larger number of samples analysed in the pan-cancer setting, which might result in largely significant *p*-values even for small effect size associations. Further details on this analysis are reported in the Supplementary Information.

Target priority scores and target tractability

Computation of the target priority scores and their significance is detailed in the supplementary information. To estimate the likelihood of a target to bind a small molecule (SM), or the likelihood of a target to be accessible to an antibody, we made use of a genome-wide target tractability assessment pipeline(Brown et al. 2018). The *in silico* pipeline integrates data from public sources, and assigns human protein-coding genes into hierarchical qualitative buckets. Predicted tractability and confidence in the data increase from Bucket 10 to Bucket 1, with targets in Bucket 1 being considered the most tractable. Of note, targets in lower Buckets (i.e. Buckets 10 to 8) are considered to have uncertain tractability, and should

not be ruled out as "intractable" without a deep tractability assessment. Further details are provided in the Supplementary Information.

Other analyses: Characterisation of target protein families and enrichments

To characterise protein families and compute statistical enrichments we made use of the Panther online tool(Mi et al. 2013).

Other analyses: *GPX4* differential expression analysis

RNA-Seq voom(Law et al. 2014) transformed gene-expression measurements were obtained from(Garcia-Alonso et al. 2018). For the *GPX4* analysis, cell lines were divided into two groups according to their loss-of-fitness response to *GPX4* knockout (using a BAGEL FDR < 5% as significance threshold for gene depletion) and gene expression fold-changes were calculated between the *GPX4* essential and non-essential cell lines. Differential gene-expression was statistically assessed using R package Limma(Ritchie et al. 2015). Gene-set enrichment analysis was performed with ssGSEA(Subramanian et al. 2005) and cancer hallmark gene sets were used to identify significant enrichment amongst the top differential expressed genes. 10,000 random permutations were performed for each signature to calculate empirical p-values, and Benjamini-Hochberg FDR correction was applied.

WRN dependency in MSI cell lines: Co-competition Assay

The sequence of sgRNAs targeting *WRN* and cell lines used in validation experiments are described in Supplementary Table 10. This included 2 sgRNA from the original screen and 2 independent sgRNA. The sgRNAs were cloned into pKLV2-U6gRNA5(BbsI)-PGKpuro2ABFP-W (Addgene #67974). Cell lines were transduced at ~50% efficiency as

described above in 6-well plates. A co-competition score was determined as the ratio of the percentage BFP-positive cells (i.e. sgRNA positive cells) on day 14 compared to day 4, as measured by flow cytometry. A co-competition score less than 1 indicates a relative reduction in BFP-positive cells resulting from targeting of a loss of fitness gene.

WRN dependency in MSI cell lines: Clonogenic assay

Cell lines were transduced with lentivirus encoding sgWRN at ~100% efficiency as described above in 6-well plates (2,000 cells/well), typically for 15 - 21 days. Cells were fixed using 100% ice-cold ethanol for 30 minutes followed by Giemsa staining overnight at room temperature.

WRN dependency in MSI cell lines: Western blot analysis

Cells were transduced at ~100% as described above in 10 cm dishes. Day 5 post transduction, cells were lysed with 200µl RIPA buffer supplemented with protease and phosphatase inhibitors and lysates used for SDS-PAGE and immunoblot analysis. Antibodies used for were: WRN (Cell Signalling Technologies, #4666; dilution 1:2000), MLH1 (Cell Signalling Technologies, #3515, dilution 1:1000) and MSH3 (Santa Cruz Biotechnology, sc-271080, dilution 1:1000).

WRN dependency in MSI cell lines: WRN rescue experiment

SW620 and SW48 cells (2 x 10^5 cells) were transfected by nucleofection (Lonza 4D Nucleofector Unit X) with Cas9/sgRNA RNP targeting human *MAVS* gene (used as a non-essential knockout control) or *WRN*, together with overexpression of 200 ng pmGFP control or 200 ng mouse *Wrn* cDNA (NM 011721, Origene Cat# MR226496). From each sample post nucleofection, 5000 cells were seeded in a 96-well and allowed to grow for 5 days, after which cells were collected for either CellTiter-Glo assay (Promega Cat# G9241) or blotted

with a WRN antibody (ThermoFisher Cat# PA5-27319). β-actin (Cell Signaling Cat#4970) was used as Western blot loading control. sgRNA sequences used are listed below:

WRN gRNA sequences:

WRN g1: TGGCCACCATTATACAATAG

WRN g2: CATTCATTACGGTGCTCCTA

MAVS gRNA sequences:

MAVS g1: CCCTGGCCCGTTCCACCCCC

MAVS g2: AGGCTGGAGGTCGCACCTGC

CellTiter-Glo data was read on an Envision Multiplate Reader and data analysis was performed using GraphPad Prism 7 software. T-test was performed using the multiple t-test module in Prism 7.

WRN dependency in MSI cell lines: RNA interference

A pool of four siRNA targeting WRN were used Dharmacon; L-010378-00-0005). HCT116 were grown and transfected with siRNA using RNAiMAX (Invitrogen) transfection reagent following manufacturer's instructions. Each experiment included: mock control (transfection lipid only), ON-TARGETplus Non-targeting Control Pool (Dharmacon, D-001810-10-05) as negative control, and polo- like kinase 1 (Dharmacon), which served as positive control. siRNA sequences are listed in Supplementary Table 10.

Rescue WRN dependency in HCT116 derivatives

HCT116 parental and Ch2, Ch3, Ch5, Ch3+Ch5 derivatives were transduced at high MOI to efficiently express Cas9. All the lines displayed Cas9 activity > 80%. To assess WRN dependency,

In vivo validation: *WRN* knockout using an inducible CRISPR-Cas9 system

To generate Inducible HCT116-WRN cells we cloned sgWRN4 into pRSGT16H-U6Tet-(sg)-CMV-TetRep-TagRFP-2A-Hygro vector (Cellecta). Cas9-expressing HCT116 cells were transduced and selected with 500 μg/ml of hygromycin (Thermo Fisher Scientific). To obtain cell populations that both uniformly express Cas9 and contain the inducible WRN-targeting sgRNA, we generated single-cell clones by serial dilution. To measure growth rate of HCT116-WRN cells after conditional induction of WRN knockout, cells were grown in flask in the presence or absence of 2 μg/ml doxycycline for 24h and then seeded in 96 well plates, with or without the same concentration of doxy. Cell growth was monitored every 36 hours using an automated IncuCyte-FLR 4X phase-contrast microscope (Essen Instruments, Ann Arbor, MI). Average object summed intensity was calculated (10 technical replicate wells for each condition, 1 image per well) using the IncuCyte software (Essen Instruments).

In vivo validation: Mouse xenograft studies

Female non-obese diabetic/severe combined immunodeficient (NOD-SCID) mice (Charles River Laboratories) were used in all *in vivo* studies. All animal procedures were approved by the Ethical Committee of the Institute and by the Italian Ministry of Health (authorization 806/2016-PR). The methods were carried out in accordance with the approved guidelines. Mice were purchased from Charles River Laboratories (Calco, Italy), maintained in hyperventilated cages, and manipulated under pathogen-free conditions. In particular, mice were housed in individually sterilised cages, every cage contained a maximum of 7 mice and optimal amounts of sterilised food, water, and bedding. HCT116 xenografts were established by subcutaneous inoculation of 2×10^6 cells into the right posterior flank of 5- to 6- week-old

mice. Tumour size was evaluated without blinding by caliper measurements, and the approximate volume of the mass was calculated using the formula (d/2)2 × D/2, where d is the minor tumor axis and D is the major tumor axis. When tumors reached an average size of approximately 250-300 mm^3, animals with the most homogeneous size were selected and randomised by tumour size. Doxycycline (D9891, Sigma-Aldrich) was dissolved in water and administered daily at a 50 mg/kg concentration by oral gavage. 8-10 mice for each experimental group were used to allow reliable estimation of within-group variability. Operators allocated mice to the different treatment groups during randomisation but were blinded during measurements. The maximal tumour volume permitted in our in vivo experiments was 3500 mm^3 and this limit was never exceeded. In vivo procedures and related biobanking data were managed using the Laboratory Assistant Suite (LAS), a web-based proprietary data management system for automated data tracking(Baralis et al. 2012).

***In vivo* validation: Immunohistochemistry**

Formalin-fixed, paraffin-embedded tissues explanted from cell xenografts were partially sectioned (10 μm thick) using a microtome. 4-μm paraffin tissue sections were dried in a 37°C oven overnight. Slides were deparaffinized in xylene and rehydrated through graded alcohol to water. Endogenous peroxidase was blocked in 3% hydrogen peroxide for 30 minutes. Microwave antigen retrieval was carried out using a microwave oven (750 W for 10 minutes) in 10 mmol/L citrate buffer, pH 6.0. Slides were incubated with monoclonal mouse anti-human Ki67 (1:100; Dako) overnight at 4°C inside a moist chamber. After washings in TBS, anti-mouse secondary antibody (Dako Envision+System-horseradish peroxidase–labeled polymer, Dako) was added. Incubations were carried out for 1 hour at room

temperature. Immunoreactivities were revealed by incubation in DAB chromogen (DakoCytomation Liquid DAB Substrate Chromogen System, Dako) for 10 minutes. Slides were counterstained in Mayer's hematoxylin, dehydrated in graded alcohol, cleared in xylene, and the coverslip was applied by using DPX. A negative control slide was processed with secondary antibody, omitting primary antibody incubation. Immunohistochemically stained slides for Ki67 were scanned with a ×40 objective. Ten representative images selected from three cases were then analyzed using ImageJ (Wayne Rasband, NIH, Bethesda, MD), which segmented cells with positive and negative nuclei. The percentage of positive cells area was calculated as brown area divided by the sum of brown and blue area. The software interpretation was manually verified by visual inspection of the digital images to ensure accuracy.