A systematic genome-wide mapping of the oncogenic risks associated with CRISPR-Cas9 editing

Sanju Sinha[1,2,3], Kuoyuan Cheng[1,3], Mark DM Leiserson[3], David M Wilson III[4], Bríd M. Ryan[2], Joo Sang Lee[1,*], Eytan Ruppin[1,3,*]

[1]Cancer Data Science Lab, Center for Cancer Research, National Cancer Institute, National Institute of Health, Bethesda, MD 20892, USA; [2]Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute, National Institute of Health, Bethesda, MD, 20850, USA, [3]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, 20742, USA, [4]Laboratory of Molecular Gerontology, National Institute on Aging, Intramural Research Program, National Institutes of Health, Baltimore, MD, 21224, USA.

* Corresponding authors

Abstract

Recent studies have reported that CRISPR-Cas9 gene editing may bear cancer risk due to a selection against cells with a functional *p53*. Here, analyzing genome-wide CRISPR and RNAi screens we systematically chart the *p53*-related oncogenic risk of individual CRISPR knockouts and find supporting evidence in patients' tumors. Importantly, we find that CRISPR knockouts may select for mutated variants of the cancer drivers *VHL* and *KRAS,* comparable to that of *p53*.

Recent advances in CRISPR-Cas9 technology, which permits targeted gene disruption and editing, have demonstrated its potential to expand our understanding of fundamental biological processes and develop new strategies for therapeutic intervention, including cancer treatment. However, several clinical trials based on CRISPR technology have been deferred due to significant risks, including off-target effects [1,2,3] and potential immunogenicity [4]. Other studies have found that CRISPR knockout (CRISPR-KO) induces double-strand breaks (DSBs), leading to a DNA damage response, which is associated with a higher copy number of the respective knocked-out genes [5,6,7,8].

Recently, two studies in this journal showed that the DNA damage response following CRISPR-KO is mediated via *p53/TP53*, a known tumor-suppressor gene mutated in over half of

all of cancers [9,10]. Haapaniemi *et al.* performed genome-wide CRISPR screening in immortalized human retinal pigment epithelial (RPE1) cells [10], finding that a *p53*-mediated DNA damage response, followed by cell cycle arrest, is induced upon generation of DSBs by the Cas9 endonuclease, driving the selection of cells that have inactivated the *p53* pathway. Ihry *et al.* reported that the CRISPR-KO of even a single gene can induce a toxic DNA damage response via the *p53*-associated apoptotic pathway [9], with *p21/CDKN1A*, a canonical *p53* transcriptional target, playing a key role. These two studies both indicate that there exists a previously neglected oncogenic risk of CRISPR-Cas9 genome editing techniques, potentially selecting for *p53* mutated, tumorigenic cells [9,10]. These highly important findings were however performed in a small number of primary cell lines, calling for a broader and more comprehensive evaluation.

To expand the scope of this prior work we harnessed the power of data from genome-wide gene essentiality screens conducted in a large panel of cancer cell lines from different tissues of origin and with a variable *p53* status [5,11]. Our key idea was to search for individual genes whose silencing results in significantly higher viability (reduced essentiality) in *p53* mutated vs *p53* wild-type (WT) cell lines, as this testifies for genes whose knockout (KO) can lead to the selection of *p53* mutated cells. Importantly, we searched for genes that showed this effect only in the CRISPR-KO screens and not in shRNA gene knockdown (KD) screens. The latter serve as controls in the context of our effort that is focused on identifying the *CRISPR-specific* effects.

Analyzing the CRISPR essentiality data of the 248 cell-lines shared in both CRISPR and shRNA screens, we find 981 genes whose KO results in significantly higher cell viability in *p53* mutated (N=173) vs *p53* WT cell lines (N=75), while only 237 genes show the opposite trend (i.e. with significantly lower post-KO viability in *p53* mutant cell lines, Methods). In contrast, the respective gene numbers related to *p53* status were more balanced in the shRNA screens (~1500 each) (chi-squared test P<1.4E-284; the results are controlled for gene copy number as a potential confounding factor, Methods, **Figure 1a, top panel**). These differences are visualized in **Figure 1a (lower panel)**, which plots the median post-KO/KD cell viability values in *p53* mutant vs WT cell lines, underscoring the increased viability of mutated *p53* cells after their KO

(Kolmogorov-Smirnov P=1.3E-282). We repeated the analysis above focusing strictly on cell lines harboring known loss-of-function *p53* mutations (Methods, N=78), using the same WT cell-lines as controls (N=75), observing an even higher significance in the differences of median post-KO/KD cell viability (chi-squared test P=1.7E-292, **Table S2,** see **Methods**). These findings were further corroborated by a permutation test where the cell line's *p53* mutation status was shuffled 10,000 times (P<1E-4, Methods). Collectively, our results confirm and expand the findings of [9,10]. Reassuringly, the genes we identified with higher post-CRISPR-KO viability in *p53* mutated cell lines significantly overlap with those showing a similar effect in the CRISPR-KO screens performed in [10] (hypergeometric test P<0.03, Supplementary notes).

Among the 981 genes whose CRISPR-KO results in higher cell viability when *p53* is mutated, 861 genes show this trend only in the CRISPR screens and not in the shRNA screens, and thus are termed *CRISPR-specific differentially essential positive (CDE+)* genes. Importantly, the mutational status of *p53* is significantly associated with the essentiality of CDE+ genes independent of copy number (**Figure S1**). One example of the most significant CDE+ gene, *TAF8*, is presented in **Figure 1b**. Notably, the CDE+ genes are enriched for pathways including homology-directed repair and the Fanconi anemia (FA) pathway (both with FDR-corrected hypergeometric P<0.03, **Table S1**), in accordance with the recent finding that FA genes are specifically involved in DSB repair during CRISPR-KO [13]. Furthermore, the CDE+ genes are preferentially located in chromosomal bands containing common fragile sites (CFSs) (hypergeometric P=2E-4, **Figure 1c**, **Table S3**), which are prone to replicative stress, fork collapse and DNA breaks that may cause genomic instability [14] (notably, this is CRISPR-specific as it not observed for the genes showing differential essentiality with regard to *p53* mutation status in the shRNA screens). As CRISPR-KO events can induce kilobase-scale structural alterations near the targeted site [15], it is possible that the formation of a CRISPR-induced DSBs in the vicinity of CFS-associated replicative intermediate can lead to excessive stress on the system promoting the activation of a *p53*-dependent cell death response. Such a phenomenon could underlie the selection pressure for *p53* inactivation and the apparent higher essentiality in *p53* WT vs *p53* mutated cell lines and is consistent with homology-directed repair events participating in the CRISPR-specific response. **Figure 1d** shows the median differences of post-KO viability between *p53* mutant vs *p53* WT cell lines for each CDE+ gene, which we

term its *CRISPR-KO risk*, and the top hits are marked, including genes located in CFSs and involved in DNA repair (Methods). As evident, the magnitude of the differential viability scores observed for top ranked CDE+ genes is high, at a level that is likely to result in a widely enriched *p53* mutated cell population after just a few cell generations starting from a heterogeneous population of mutated and WT cells.

In the inverse direction, 237 genes are more essential in CRISPR screens in *p53* mutated than WT cell lines. Among those, 185 genes show this trend only in the CRISPR screen and not in the shRNA screen, *termed CRISPR-specific differentially essential negative (CDE-)* genes. CDE- genes are enriched for pathways related to mitotic checkpoints, DNA replication and cell cycle regulation, cellular processes that frequently engage *p53* (**Table S1, Figure 1e**, FDR corrected hypergeometric P<0.1; these enrichments are CRISPR-specific as differentially essential genes in the shRNA screens are not significantly enriched for these pathways). In accordance with the findings of [9,10], the top ranked CDE- gene is *CDKN1A* (a.k.a. *p21*, Wilcoxon rank sum P<1.85E-08, **Figure 1d, S2**), a potent cyclin-dependent kinase inhibitor regulating cell cycle progression whose expression is tightly controlled by *p53*. CDE- genes are of potential interest from a translational point of view, as their CRISPR-KO may possibly be combined with a CRISPR-KO of a target CDE+ gene to counteract the potential oncogenic selective effects.

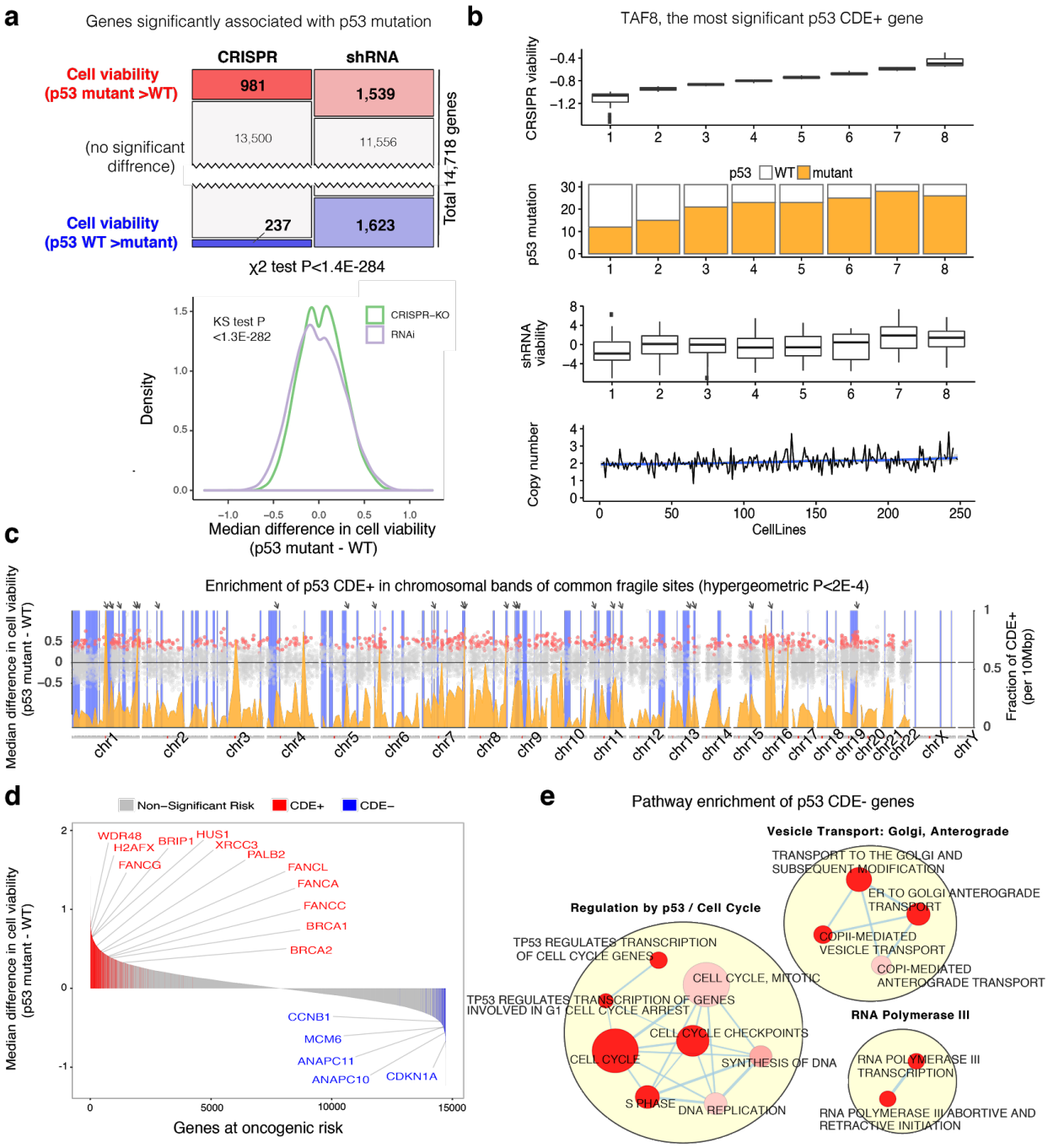

Figure 1. A genome wide view of *p53*-associated oncogenic risk. (a) Upper panel: number of genes whose essentiality is significantly associated with *p53* mutation status in CRISPR and shRNA screens (one sided Wilcoxon rank-sum has been performed with FDR threshold of 0.1). Lower panel: density plot showing the distribution of differences of median differences in cell viability values after gene KO/KD in *p53* mutant vs *p53* WT cell lines. **(b)** The distribution of

TAF8 (the top CDE+ gene of *p53*) post KO/KD viability scores in the CRISPR (top panel) screen as a function of *p53* mutation status. The cell lines were binned into 8 bins of equal size, which are ordered by their viability after the CRISPR-KO of *TAF8*. The fraction of *p53* mutant cell lines in each bin is plotted on the second panel. The distribution of *TAF8* post shRNA-KD viability scores is displayed on the third panel as a control. The fourth, lowest panel visualizes that the copy number of *TAF8* is about the same in the different bins. **(c)** Enrichment of *p53* CDE+ genes in common fragile sites (CFSs). The x-axis denotes the chromosomal position; the scatter plot (y-axis on the left-hand side) shows the difference of median post-CRISPR-KO cell viability values in *p53* mutant vs *p53* WT cell lines for *p53* CDE+ genes (red dots) and all other genes (grey dots); the density plot (colored orange, y-axis on the right-hand side) shows the fraction of *p53* CDE+ genes among all genes per DNA segments of 10 Mbp along the genome; the vertical blue bars indicate the chromosomal bands of CFSs, and prominent sites where peaks of high CDE+ gene density coincide with CFSs are marked by arrows on the top. **(d)** The distribution of CRISPR-risks across the genome. Significant CDE+ genes that are part of FA pathway are marked in red and significant CDE- genes that are part of cell cycle regulation in blue. **(e)** Visualization of the pathways enriched for *p53* CDE- genes. Only significantly enriched pathways (FDR<0.1) specific to CRISPR (and not in genes showing differential essentiality in the shRNA screens are shown). Pathways are depicted as nodes whose sizes correlate with pathway lengths and colors represent enrichment significance (the darker, the more significant). Pathway nodes are connected and clustered based on their functional similarities, and higher-level functional terms are given for each of the clusters (Methods). For clarity, only the largest clusters are shown.

To gain further insights into the genes and pathways that are specifically involved in response to CRISPR-KO vs those involved in response to shRNA-KD, we analyzed the LINCS l1000 [16] data, where expression was measured before and after CRISPR-KO and shRNA-KD screens across 53 genes in six cell lines. We identified a *CRISPR-specific consensus differential expression signature (CDES)*, comprised of genes that are most frequently differentially expressed specifically following CRISPR-KO but not shRNA-KD (Methods). The top pathways that CDES genes are enriched for include cell cycle, mitotic M/G1 phases and G2-M checkpoints (all with FDR-corrected Kolmogorov-Smirnov P<0.025, see Methods, **Table S4**), consistent

with the previous findings that cell cycle arrest is triggered in response to CRISPR mediated DNA damage [10].

Our analysis in cancer cell lines has shown that CDE+ genes of *p53* are (i) enriched for homology-directed DNA repair (a.k.a., homologous recombination) and FA pathway, and (ii) preferentially located on CFSs in the genome, two elements that can play a role in gene copy number variation. Given that the CRISPR-KOs of CDE+ genes preferentially reduce the viability of *p53* WT cells, we hypothesized that copy number alterations in CDE+ genes will analogously reduce the fitness of *p53* WT tumors. To test this hypothesis, we analyzed the somatic copy number alteration (SCNA) and patient survival data of 7,547 samples from the cancer genome atlas (TCGA) [17]. As control, we used genes whose essentiality is not associated with *p53* mutational status. We find that the SCNA (both amplifications and deletions, Methods) in CDE+ genes is significantly lower in *p53* WT tumors compared to that of control genes (Wilcoxon rank sum $P<2.7E-5$), suggesting that copy number variations in CDE+ genes are selected against specifically in *p53* WT tumors, as *p53*-mediated responses are detrimental. As further controls, this is not the case for *p53* mutant tumors or for CDE- genes (Wilcoxon rank sum $P>0.5$). In addition, we observe that the magnitude of SCNA events of CDE+ genes is associated with enhanced patient survival specifically in *p53* WT tumors and not in *p53* mutated tumors (hypergeometric $P<0.046$, Methods), further testifying that these amplification/deletion events reduce the fitness of these WT tumors. Taken together, these observations suggest that the CRISPR risks of CDE+ genes identified *in vitro* may also influence the evolution of patients' tumors.

To identify additional cancer drivers like *p53*, whose mutational status impacts CRISPR-KO outcome in a differential and specific manner, we applied the analysis described above to each gene in the list of cancer driver genes provided by Vogelstein *et al.* [18] (focusing on those genes that are mutated in at least 10 of the screened cell lines (N=61)). We identified the tumor suppressor *VHL* (Von Hippel-Lindau) and the oncogene *KRAS* as such *master CDE regulators* (or "master regulators" for short, Methods, **Figure 2a**, **Table S7**). That is, like *p53*, each of these two genes have a CDE+/- skewed distribution in the CRISPR compared to the shRNA screens. The CDE+ and CDE- genes of each of these master regulators are listed in **Table S1**. Some

genes are classified as CDE+ of multiple master regulators, and thus their CRISPR-KO may impose considerable risk of selecting for more than one mutated cancer gene: *SREK1* is both a *p53* and *VHL* CDE+ gene; fourteen genes are CDE+ genes of both *KRAS* and *p53*; and ten genes are CDE+ genes of both the *KRAS* and *VHL* (**Table S6**).

As evident from **Figure 2b**, the two additional master regulators uncovered by our analysis have considerably higher median risk levels than *p53* (that is, the CRISPR-KO of their CDE+ genes is likely to drive higher levels of selection than KO of the CDE+ genes of *p53*). Notably, for *VHL*, although the absolute number of its CDE+ genes is smaller than that of *p53*, the ratio of its CDE+ to CDE- genes is strikingly higher than that observed for *p53*, while the parallel distribution observed in the shRNA-KD screens is balanced **(Figure 2c)**. Just like *p53* the mutational status of *VHL* is significantly associated with the essentiality of CDE+ genes independent of copy number (**Figure 2d**). Notably, CDE+ genes of *VHL* are also enriched in chromosomal bands of CFSs (**Figure 2e and Table S3**, hypergeometric P=0.024). Indeed, *VHL* can act as a positive regulator of *p53* in DNA damage-induced cell cycle arrest or apoptosis [19], possibly accounting for its role as a master regulator.

For *KRAS*, a major oncogene that is involved in cell signaling, we find high numbers of CDE genes, while only very few *KRAS* mutation-associated genes are identified in the shRNA-KD screens **(Figure 2c)**. Indeed, most of its mutations in the cell-line collection we studied are missense hotspot mutations, suggestive of gain of function. Mutated *KRAS* is known to activate various DNA repair pathways and may override the trigger of cell death upon DNA damage [20,21], supporting its role as a master regulator. Consistently, the CDE- genes of *KRAS* are significantly enriched for cell cycle checkpoints and DNA DSB repair pathways (FDR-corrected Kolmogorov-Smirnov P<0.05, Methods, **Figure 2f, Table S1**). As observed for *VHL and p53* **(Figure 2d and S1**, respectively**)**, the effects of *KRAS* are also independent of gene copy number (**Figure S1**). Reassuringly, mining the data from CRISPR and shRNA screens performed in the same isogenic *KRAS*-mutated cell line [22], the inferred *KRAS* CDE- genes significantly overlap with the differentially essential genes identified in the CRISPR (and not in the shRNA) screens (hypergeometric P<0.05, Methods).

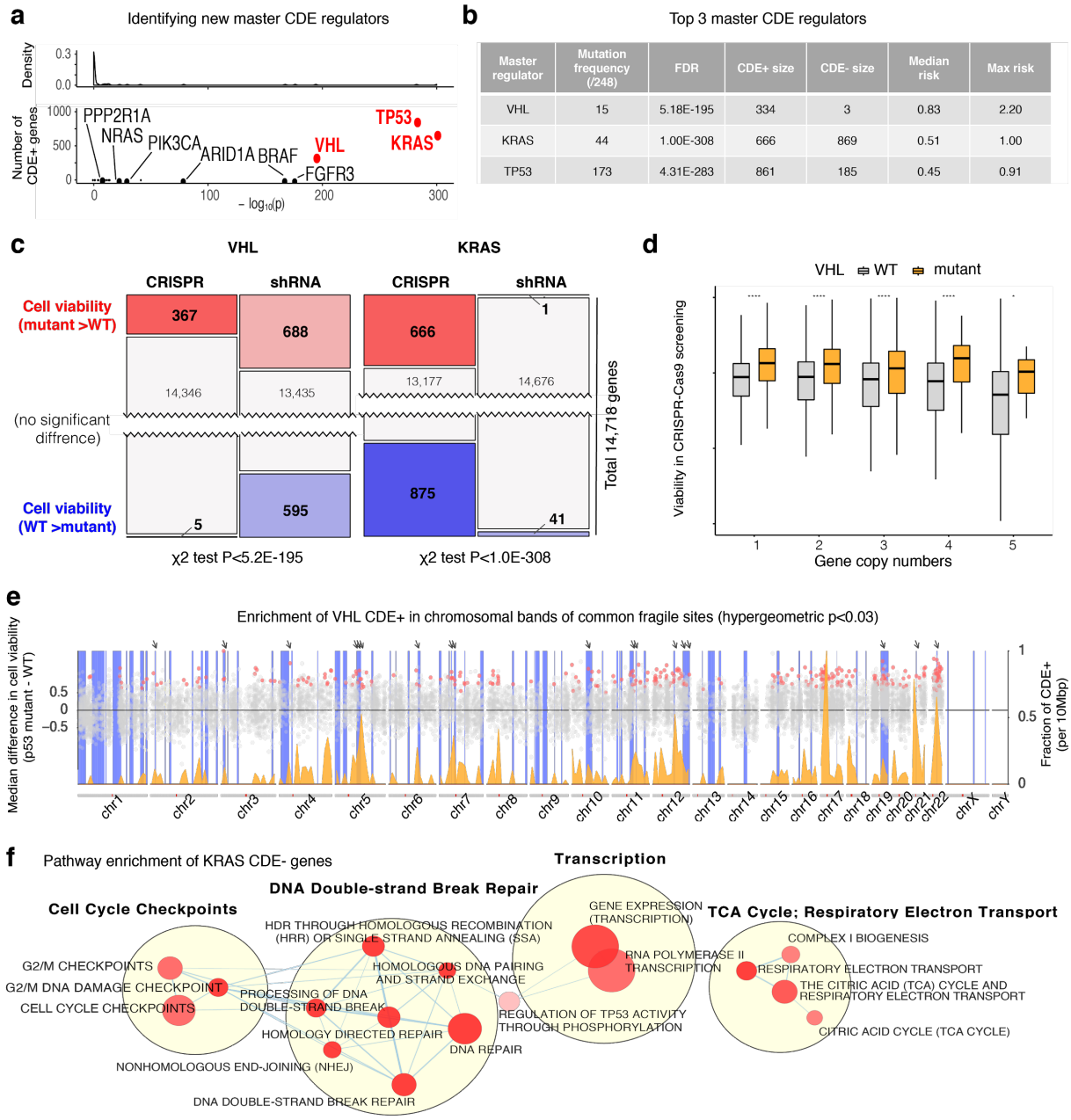

Figure 2. Beyond *p53*: *VHL* and *KRAS* as master CDE regulators. (a) Top panel: A density plot showing that almost all cancer drivers are not identified as significant master CDE regulators. The x-axis denotes their significance (negative log-transformed P value of the chi-squared test for being a master regulator, see Methods). Bottom panel: Displays the identity of the nine genes having significant P values and more than one CDE+ gene; only the top three most significant genes (marked in red) have a considerable number of CDE+ genes, and hence we focus on them in the main text. **(b)** A table describing key properties of the three master

regulators. The first column denotes their names, the second column displays the frequency of their non-synonymous mutations among 248 cell lines, the third column displays the FDR corrected Fisher's exact P value denoting the significance of the CRISPR-specificity of the association of gene essentiality with their mutation status. The fourth and fifth columns denote the number of CDE+ and CDE- genes identified for each of the master regulators. The sixth and seventh columns denote the median and maximal risk score (difference in cell viability after CRISPR-KO of the CDE+ genes between the mutant and WT cell lines of the respective master regulators). **(c)** The number of genes whose essentiality is significantly associated with *VHL* (left) and *KRAS* (right) mutational status in CRISPR and shRNA screens (one sided Wilcoxon rank-sum has been performed with FDR threshold of 0.1). **(d)** The effect of *VHL* mutation status on CRISPR-KO viability (y-axis) is significant across a wide range of CDE+'s copy numbers (x-axis). CN=1 denotes cases with copy number less than or equal to 1 and CN=5 denotes cases with copy number greater than or equal to 5. The number of stars at the top of boxplot represents the significance of the difference (one-sided Wilcoxon rank sum test). *P<0.05, **P<0.01, ***P<0.001, ****P<0.0001. **(e)** The enrichment of *VHL* CDE+ genes in common fragile sites (CFSs). The x-axis denotes the chromosomal position; the scatter plot (y-axis on the left-hand side) shows the difference of median post-CRISPR-KO cell viability values in *VHL* mutant vs WT cell lines for *VHL* CDE+ genes (red dots) and all other genes (grey dots); the density plot (colored orange, y-axis on the right-hand side) shows the fraction of *VHL* CDE+ genes among all genes per DNA segments of 10 Mbp along the genome; the vertical blue bars indicate the chromosomal bands of CFSs, and prominent sites where peaks of high CDE+ gene density coincide with CFSs are marked by arrows on the top. **(f)** Visualization of pathways enriched for *KRAS* CDE- genes. Only significant pathways (FDR<0.1) specific to CDE and not to the genes showing differential essentiality in the shRNA screens are included. Pathways are shown as nodes whose sizes correlate with pathway lengths and colors represent the significance of their enrichment (the darker the more significant). Pathway nodes are connected and clustered based on their functional similarities, and higher-level functional terms are given for each of the clusters (Methods). For clarity, only the largest clusters are shown.

In sum, our results extend the scope of the previous findings in human primary cells [9,10] to a genome-wide analysis of these effects across hundreds of cancer cell lines from diverse

tissues of origin. By analyzing shRNA-KDs as controls, we could identify gene-KOs with truly CRISPR-specific selection effects (that is, those that occur only CRISPR and not in shRNA screens). Those include both the CDE+ genes whose KOs imposes oncogenic risk and CDE- genes whose KOs may counteract the latter. Despite the noisy nature of pooled CRISPR/RNAi screens [23], our results support the earlier findings that the selection of mutant *p53* cells is mediated via cell-cycle regulation [9]. Beyond that, our findings also point to the role of DNA repair genes and their CFS locations. Our analysis is based on data from cancer cell lines, and yet shows a significant overlap with the findings in human non-cancerous cells reported earlier. While the results obtained in cancer cell lines should obviously be interpreted with care, the heterogeneity and prevalence of mutations available in the latter expand our ability to detect the underlying associations. Indeed, going beyond *p53*, our analysis reveals two novel master CDE regulators, *KRAS* and *VHL*, suggesting that genomic editing by CRISPR-Cas9 may induce tumorigenic alterations in these important cancer driver genes as well. Interestingly, we find evidence that the CRISPR-related KO effects identified *in vitro* may have an echo that is traceable *in vivo*, modulating the evolution of patients' tumors. Our work demonstrates yet again how big data analysis may be employed to complement experimental screens and extend their findings.

Methods

CRISPR and shRNA essentiality screen data

We obtained CRISPR-Cas9 essentiality screen (or dependency profile) data in 436 cell lines from *Meyers et al.* [5] for 16,368 genes, whose expression, CNV and mutation data are available via CCLE portal [24]. We obtained the shRNA essentiality screen data in 501 cell-lines from DepMap portal [14] for 16,165 genes, whose expression, CNV and mutation data is available publically via CCLE portal [24]. The 248 cell-lines and 14,718 genes that appear in both datasets were used in this analysis (**Table S5, S6**). For mutation data, only non-synonymous mutations were considered. Synonymous (silent) mutations were removed from the pre-processed MAF files downloaded from CCLE portal [24].

Identifying CRISPR specific differentially essential genes of a master CDE regulator

For a given master CDE regulator (e.g. *p53*), we checked which gene's essentiality (viability after knockout) is significantly associated with the mutational status of the master regulator using a Wilcoxon rank sum test in the CRISPR and shRNA datasets, respectively (FDR<0.1). *CRISPR-specific differentially essential (CDE) genes* denotes two sets of genes depending on the direction of the association with the mutation. CDE+ genes are those whose CRISPR-KO is significantly more viable when the master regulator is mutated while their shRNA silencing is not, whereas CDE- genes are those whose CRISPR-KO is significantly more viable when the master regulator is WT while their shRNA silencing is not. We filtered out any candidate CDE genes whose copy number was also significantly associated with the given mutation to control for potentially spurious associations coming from copy number (we removed genes showing significant association (FDR<0.1)) – the exact procedure used is described below in the section titled "*Identifying master CDE regulators of CRISPR-KO*").

Identifying CDEs associated with p53-loss-of-function mutations

Out of a total of 248 cell lines that we analyzed, 173 cell lines (69.7%) have *p53* non-synonymous mutations. In addition to identifying CDEs by considering all non-synonymous mutation, we additionally employed a more conservative approach where we aimed to consider only *p53 loss-of-function* (LOF) mutations in the CDE identification process. To this end, we considered a mutation to be LOF if it was classified as non-sense, indel, frameshift, or among the 4 most frequent non-functional hotspot mutations (R248Q, R273H, R248W and R175H within the DNA-binding domain, determined as pathogenic by COSMIC [26]). Using this definition we obtained new mutation profiles for *p53* and identified CDE genes via the same method described in the section titled "*Identifying CRISPR specific differentially essential genes* of a master CDE regulator."

Identifying master CDE regulators of CRISPR-KO

To identify additional master regulators like *p53*, we considered 121 cancer driver genes identified by Vogelstein *et al.* [18], whose nonsynonymous mutation is observed in at least 10 cell lines (N=61). We determined whether each of these genes is a master CDE regulator as follows: for each of the 61 candidate genes, we tested the association between the essentiality of each of genes in the genome (reflected by post-KO cell viability) with the mutational status of the

candidate master regulator gene using a Wilcoxon rank sum test. We then counted the number of genes, whose essentiality is: (i) significantly positively associated with the candidate master regulator mutational status (FDR-corrected p-value<0.1, median essentiality of WT>mutant of the cancer gene), (ii) significantly negatively associated with the candidate master regulator mutational status (FDR-corrected p-value<0.1, median essentiality of WT<mutant of the cancer gene), and (iii) not associated (FDR-corrected p-value>0.1) with the candidate master regulator mutation status; we performed this computation separately for the CRISPR and the shRNA screens, respectively. This computation results in a 3-by-2 contingency table for each candidate master regulator gene. We then checked whether the distribution of the above three counts in CRISPR dataset significantly deviates from that in shRNA dataset via a Fisher's exact test on the contingency table. If each of the values in the contingency table was greater than 30, we used the chi-squared approximation of the Fisher's exact test. We further filtered out any candidate CDE genes whose copy number was also significantly associated with the given mutation to control for potentially spurious associations coming from copy number (we removed genes showing significant association (FDR<0.1)). We performed this procedure for all 61 candidate genes one by one and selected those with FDR corrected Fisher's exact test <0.1. We further filtered out the candidate master regulators whose mutation profile is correlated with *p53* mutation profile via a pairwise Fisher test of independence (FDR<0.1). We finally report the master regulators that have substantial number of CDE+ genes (N>300).

Pathway enrichment analysis of CDE+/CDE- genes

We analyzed the CDE+/CDE- genes of each of the master CDE regulators for their pathway enrichment with the pathway annotations from the Reactome database [25] in two different ways. First, we tested for significant overlap between our CDE genes with each of the pathways with hypergeometric test (FDR<0.1). We repeated the test for genes showing differential essentiality in the shRNA screens (shRNA-DE genes), and only reported those significant pathways specific to CDE genes but not shRNA-DE genes. Second, we ranked all the genes in the CRISPR-KO screen by the differences in their median post-KO cell viability values in mutant vs WT cells, and the standard GSEA method [27] was employed to test whether the genes of each Reactome pathway have significantly higher or lower ranks vs the rest of the genes (FDR<0.1). We repeated the GSEA analysis with the genes ranked by differential post-KD cell

viability in the shRNA screen, and only reported significant pathways specific to CRISPR but not shRNA screen. We confirmed that for *p53*, the GSEA method was able to recover the top significant pathways identified by the hypergeometric test (e.g. those in **Figure 1e**), although extra significant pathways were identified (**Table S1**), consistent with the higher sensitivity of the GSEA method [27]. For *p53* and *KRAS* CDE- genes respectively, the enriched pathways were clustered based on the Jaccard index and the number of overlapping genes with Enrichment Map [28], and the largest clusters were visualized as network diagrams with Cytoscape [29].

To study the potential enrichment of CDE genes in common fragile sites (CFSs), we obtained the chromosomal band locations of CFSs from[14], and we defined the CFS gene set as the set of all genes located within these chromosomal bands (obtained from Biomart [30]). We tested for a significant overlap between our CDE genes and the CFS gene set with hypergeometric test, and also confirmed the lack of significant overlap with the corresponding shRNA-DE genesets.

Identification of CRISPR-specific consensus differential expression signature (CDES)

The Library of Integrated Network-Based Cellular Signatures (LINCS) l1000 Phase I and II datasets were obtained from the GEO database (GSE92742 and GSE70138). We selected the subset of level 5 data (computed gene expression signatures as z-scores based on the standard l1000 protocol) [16] for assays where both knockdown by RNA interference (RNAi) and knockout by CRISPR/Cas9 were performed for the same gene in the same cell line. This resulted in z-scores of expression level changes for 978 measured genes for perturbation of 53 genes in 6 cell lines. Consensus gene-level signatures were computed from shRNA or sgRNA-level signatures using the weighted-average method described in [23]. For each signature of gene perturbation in the same cell line, the RNAi z-scores were subtracted from the CRISPR z-scores as a way to obtain a gene signature specific to the CRISPR/Cas9 technology. The obtained CRISPR-specific signatures for each of 53 gene KO perturbations were then averaged (mean) to reduce gene-specific effect, resulting in a *CRISPR-specific consensus differential expression signature* (CDES) for each cell line that best represents the gene expression changes in response to the CRISPR/Cas9 technology specifically (i.e. Cas9-induced DNA breaks and related cellular events). A further CDES was also obtained by again averaging (mean) the CDES across the cell

lines. The CDES's were then analyzed for functional enrichment of the top altered genes using the GSEA method [27] with the geneset annotation from Reactome [25].

Enrichment of KRAS CDE- genes in the KRAS isogenic essentiality screens

To evaluate the relevance of our large-scale analysis in a single isogenic cell line, we analyzed the CRISPR-KO and shRNA-KD screen of 1,155 essential genes performed in KRAS mutated isogenic DLD1 cell line [22]. We first identified the differentially essential genes that show higher essentiality in KRAS mutant vs WT cells (top 40-percentile) for each of CRISPR and shRNA screens. We then determined a subset of genes that are specifically showing this trend in CRISPR screen, and checked whether this subset has significant overlap with our KRAS CDE- geneset using hypergeometric test.

Testing the clinical relevance of copy number alterations of CDE genes in p53 mutated vs WT tumors

Given that the CRISPR-KOs of CDE+ genes preferentially reduce the viability of *p53* WT cells, we hypothesized that copy number alterations in CDE+ genes will analogously reduce the fitness of *p53* WT tumors. To test this hypothesis, we downloaded the somatic copy number alteration (SCNA) and patient survival data of 7.547 samples in 26 tumor types from cancer genome atlas (TCGA) [17] from UCSC Xena browser (https://xenabrowser.net/). We confirmed that in these tumor types *p53* is mutated in more than 5% of the samples.

Our first goal was to study if the SCNA (both amplifications and deletions) in CDE+ genes is significantly lower specifically in *p53* WT tumors compared to that of control genes, and not in *p53* mutant tumors. As the control for our analysis, we used genes whose essentiality is not associated with *p53* mutational status. To this end, we computed the copy number alterations (genomic instability (GI)) of a given geneset, which aims to quantify the relative amplification or deletion of genes in a tumor based on SCNA. Given s_i, the absolute of log ratio of SCNA of gene i in a sample relative to normal control, GI of the sample is given as in [31]:

$$GI = \frac{1}{N} \sum_{1}^{n} I\left(s_i > 1 \right)$$

where I is the indicator function, and *i* runs through the given subset of genes. We then checked whether the GI of CDE+ geneset is significantly lower than that of control non-CDE genes in *p53* WT but not in *p53* mutant tumors (using a Wilcoxon rank sum test).

Our second goal was to study if the magnitude of these SCNA levels of CDE+ genes is associated with *enhanced* patient survival, specifically in *p53* WT tumors, as this would further testify that such amplification/deletion events reduce the fitness of these tumors. To this end, we used the following Cox proportional hazard model to identify the genes whose genomic alteration is associated with patient survival specifically in *p53* WT tumor, while controlling for various confounding factors including the effect of cancer types, genome-wide genomic instability, sex, age, and race, tumor purity. We considered both amplification and deletion by taking the absolute value of SCNA levels,

$$\lambda_g(t) = \lambda_{g0}(t) exp\ (\beta_I I(p53^+, GI_i^+) + \beta_p purity + \beta_a age + \beta_{p53} p53 + \beta_{GI} GI_i + \beta_{GGI} GGI,$$

where *g* is an indicator variable over all possible combinations of patients' stratifications based on cancer-type, race, and sex. $\lambda_g(t)$ is the hazard function (defined as the risk of death of patients per unit time) and $\lambda_{g0}(t)$ is the baseline-hazard function at time *t* of the gth stratification. The model contains six covariates: (i) $I(p53^+, GI^+)$, an indicator variable that has value 1 if *p53* is not mutated and the absolute value of SCNA level of the given gene *i* is greater than 50-percentile across all TCGA samples in the patient's tumor, and otherwise 0, (ii) purity, denoting tumor purity [32], (iii) age, denoting patient's age, (iv) P53, denoting the *p53* mutational status of the patient, (v) GI_i, the absolute value of SCNA levels of the given gene *i*, and (vi) GGI, quantifying the genome-wide genomic instability (GGI) of the sample, as computed above. We tested the enrichment of CDE+ genes among the genes whose absolute SCNA levels are significantly associated with better patient survival specifically in *p53* WT and not *p53*-mutated tumors using a hypergeometric test.

References

1. Kim, Daesik, et al. "Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells." Nature methods 12.3 (2015): 237.

2. Tsai, Shengdar Q., et al. "GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases." Nature biotechnology 33.2 (2015): 187.
3. Fu, Yanfang, et al. "High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells." Nature biotechnology 31.9 (2013): 822.
4. Charlesworth, Carsten Trevor, et al. "Identification of pre-existing adaptive immunity to Cas9 proteins in humans." BioRxiv (2018): 243345.
5. Meyers, Robin M., et al. "Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells." Nature genetics 49.12 (2017): 1779.
6. Wang, Tim, et al. "Identification and characterization of essential genes in the human genome." Science 350.6264 (2015): 1096-1101.
7. Aguirre, Andrew J., et al. "Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting." Cancer discovery 6.8 (2016): 914-929.
8. Munoz, Diana M., et al. "CRISPR screens provide a comprehensive assessment of cancer vulnerabilities but generate false-positive hits for highly amplified genomic regions." Cancer discovery 6.8 (2016): 900-913.
9. Ihry, Robert J., et al. "p53 inhibits CRISPR–Cas9 engineering in human pluripotent stem cells." Nature medicine (2018): 1.
10. Haapaniemi, Emma, et al. "CRISPR–Cas9 genome editing induces a p53-mediated DNA damage response." Nature medicine (2018): 1.
11. Tsherniak, Aviad, et al. "Defining a cancer dependency map." Cell 170.3 (2017): 564-576.
12. Moldovan, George-Lucian, and Alan D. D'Andrea. "How the fanconi anemia pathway guards the genome." Annual review of genetics 43 (2009): 223-249.
13. Richardson, Chris D., et al. "CRISPR–Cas9 genome editing in human cells occurs via the Fanconi anemia pathway." Nature Genetics (2018): 1.
14. Lukusa, T, et al. Human chromosome fragility. Biochim Biophys Acta. 1779.1 (2008): 3-16.
15. Kosicki, M. et al. Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. Nature Biotechnology (2018).
16. Subramanian, Aravind, et al. "A next generation connectivity map: L1000 platform and the first 1,000,000 profiles." Cell 171.6 (2017): 1437-1452.

17. Cancer Genome Atlas Research Network et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics 45(10)(2013):1113-20.
18. Vogelstein, Bert, et al. "Cancer genome landscapes." science 339.6127 (2013): 1546-1558.
19. Roe, Jae-Seok, et al. Phosphorylation of von Hippel-Lindau protein by checkpoint kinase 2 regulates p53 transactivation. Cell Cycle 10:22 (2011), 3920-3928.
20. Hahnel, PS, et al. Targeting components of the alternative NHEJ pathway sensitizes KRAS mutant leukemic cells to chemotherapy. Blood (2014), 123: 2355–2366.
21. Jinesh, GG, et al. Molecular genetics and cellular events of K-Ras-driven tumorigenesis. Oncogene (2017), 37(7): 839-846.
22. Martin, Timothy D., et al. A Role for Mitochondrial Translation in Promotion of Viability in K-Ras Mutant Cells. Cell Reports (2017), 20(2): 427-438.
23. Smith, I. et al. Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map. PLOS Biology (2017), 15(11), e2003213.
24. Barretina, Jordi, et al. "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity." Nature 483.7391 (2012): 603.
25. Joshi-Tope, G., et al. "Reactome: a knowledgebase of biological pathways." Nucleic acids research 33.suppl_1 (2005): D428-D432.
26. Forbes, Simon A., et al. "COSMIC: exploring the world's knowledge of somatic mutations in human cancer." Nucleic acids research 43.D1 (2014): D805-D811.
27. Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102(43) (2005): 15545–1.
28. Merico, Daniele, et al. “Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation.” PLoS ONE 5(11) (2010).
29. Shannon, P. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.” Genome Research 13(11) (2003): 2498-2504.
30. Smedley, Damian, et al. “The BioMart Community Portal: an Innovative Alternative to Large, Centralized Data Repositories.” Nucleic Acids Research 43(W1) (2015).
31. Bilal, Erhan, et al. "Improving breast cancer survival analysis through competition-based multidimensional modeling." PLoS computational biology 9.5 (2013): e1003047.

32. Salas, Lucas A., et al. "Integrative epigenetic and genetic pan-cancer somatic alteration portraits." Epigenetics 12.7 (2017): 561-574.

Acknowledgements

We acknowledge and thank the National Cancer Institute for providing financial and infrastructural support. We thank Curtis Harris, Ze'ev Ronai, Sridhar Hannenhalli and the members of Cancer Data Science Lab for insightful feedback. E.R. gratefully acknowledges support from I-CORE Center of Excellence in Gene Regulation in Complex Human Disease, Grant No. 41/11, the Israel Science Foundation (grant No 41/11). S.S and K.C. are supported by NCI-UMD Cancer Research Training Fellowship.