Agreement between two large pan-cancer CRISPR-Cas9 gene dependency datasets

Authors

Joshua M. Dempster[1], Clare Pacini[2,3], Sasha Pantel[1], Fiona M Behan[2,3], Thomas Green[1], John Krill-Burger[1], Charlotte M Beaver[2], Victor Zhivich[1], Hanna Najgebauer[2,3], Felicity Allen[2], Emanuel Gonçalves[2], Rebecca Shepherd[2], John G. Doench[1], Kosuke Yusa[2], Francisca Vazquez[1], Leopold Parts[2], Jesse S. Boehm[1], Todd R. Golub[1], William C. Hahn[1], David E. Root[1], Mathew J Garnett[2,3], Francesco Iorio[2,3]*, Aviad Tsherniak[1]*

[1] Broad Institute of Harvard and MIT, 415 Main Street, Cambridge, MA 02142, USA
[2] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK
[3] Open Targets, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

* Corresponding authors: aviad@broadinstitute.org, fi1@sanger.ac.uk

Abstract

Genome-scale CRISPR-Cas9 viability screens performed in cancer cell lines provide a systematic approach to identify cancer dependencies and new therapeutic targets. As multiple large-scale screens become available, a formal assessment of the reproducibility of these experiments becomes necessary. Here we analyzed data from recently published pan-cancer CRISPR-Cas9 screens performed at the Broad and Sanger institutes. Despite numerous experimental differences, we found that the screen results are highly concordant across multiple metrics in that both common and specific dependencies were identified in cell lines jointly. Strong biomarkers of gene dependency found in one institute are recovered in the other. Through further analysis and replication experiments at each institute, we found that batch effects are driven principally by two key experimental parameters: the reagent library and the assay lengths employed in the two studies. These observations and analyses show that Broad and Sanger CRISPR-Cas9 viability screens produce robust and reproducible findings.

Introduction

The development of new oncology drugs remains challenging with a high degree of failure, despite sustained investment[1]. One factor that contributes to this challenge is that drug targets often lack robust genetic linkage between the trialed therapies and the disease they are designed for[2]. To tackle this problem, large-scale pharmacological screenings have been performed across panels of human cancer cell lines[3,4]. The drug-response datasets resulting from these studies have been integrated with multi-omic characterization of the screened *in vitro* models, unveiling established and novel associations between cell molecular features and drug sensitivities. Results from these analyses have laid a rich preclinical foundation for the development of selective cancer targeted therapies[5–7].

The advent of genome editing by CRISPR-Cas9 technology has allowed the extension of these studies beyond the domain of currently druggable targets with improved precision and scale[8,9]. Pooled CRISPR-Cas9 screens employing genome-scale libraries of single guide RNAs (sgRNAs) are being performed on growing numbers of cancer *in vitro* models[10–16]. This makes possible testing the extent to which inactivating each gene, in turn, impacts cancer viability in the context of a defined genomic/molecular make-up. The results of these screens can be used to identify and prioritize new cancer therapeutic targets[17]. However, fully characterizing genetic vulnerabilities in cancers is estimated to require thousands of genome-scale screens[18].

We have worked together in an international consortium aimed at defining all gene vulnerabilities and dependencies that could be exploited therapeutically in each [17]cancer cell. By leveraging multiple drug/CRISPR-Cas9 screening studies together with similar existing datasets from RNAi screens, the overall aim of this consortium is to generate and make publicly available a comprehensive *Cancer Dependency Map*[18,19]. This has the potential of transforming the way oncology therapeutic targets and biomarkers are discovered, and could significantly accelerate the development of new anti-cancer drugs. Despite the fact that we employ similar CRISPR-Cas9 screening strategies (detailed in Meyers *et al.*[20] and Behan *et al*.[17]), these involve distinct experimental aspects unique to each site.

We present a comparative analysis of datasets derived from the two largest independent CRISPR-Cas9 based gene-dependency screening studies in cancer cell lines published to date[17,21,22], which will be part of the Cancer Dependency Map effort. The main aim of this analysis is to assess the concordance of these datasets and the reproducibility of the analytical outcomes they yield when they are investigated individually. To perform this analysis, we designed a computational strategy including comparisons at different levels of data-processing and abstraction: from gene-level dependencies to molecular markers of gene dependencies, and genome-scale cell line profiles of gene dependencies. Lastly, we shed light on the differences in the experimental settings that give rise to batch effects across independent studies of this kind, discerning between biological and technical confounding factors.

Results

Overview of datasets and comparison strategy

We compared two sets of pooled genome-scale CRISPR-Cas9 drop out screens in cancer cell lines, generated at the Broad Institute and the Sanger Institute through the experimental pipelines summarized in **Fig. 1a** and detailed in **Supplementary Table 1** and **Supplementary text**. Datasets were filtered for the 147 cell lines and 16,733 genes screened independently by both institutes (**Supplementary Table 2**). We performed comparisons of gene dependency scores, quantifying the reduction of cell viability upon gene inactivation via CRISPR-Cas9 targeting; of profiles of individual gene dependency scores across cell lines (gene dependency profiles), as well as of individual cell lines across genes (cell line dependency profiles).

We calculated gene dependency scores using three different strategies. First, we considered fully processed gene scores, representing the dependency of each cell line on each gene, which are available for download from the Broad[23] and Sanger[17,24] Cancer Dependency Map web-portals (*processed* data). For the Broad data, processing involves calculating log fold change of sgRNA representation at the end of the assay versus plasmid library counts, then using CERES to separate the contributions of variable guide efficacy, gene copy number, and gene knockout effect to cell viability[20]. For the Sanger data, processing removes copy number effect with CRISPRCleanR from estimates of log fold change values for each guide before collapsing to gene level. Because the two institutes use very different data processing pipelines, we also examined minimally processed gene scores, generated by computing median sgRNA abundance log fold changes for each gene (*unprocessed* data). Lastly, we applied an established empirical Bayesian batch correction method (ComBat)[25] to the unprocessed gene scores to remove experimental batch effects between the institutes. ComBat aligns gene means and variances between the datasets, thereby eliminating simple batch effects. We refer to this form of the data as the *batch-corrected* gene scores.

Agreement of gene dependency scores

We found concordant gene scores across all genes and cell lines with Pearson correlations equal to 0.658, 0.627, and 0.765, respectively for processed, unprocessed and batch-corrected data (*p* values below machine precision in all cases, N = 2,465,631, **Fig. 1b**). The reproducibility of gene scores between the two institutes can be considered a function of two factors: the mean dependency across all cell lines for each gene, which is relevant to infer the role of a gene as a potential common dependency; and the patterns of scores across cell lines for each gene, which is relevant to identifying selective oncology therapeutic targets. We examined gene mean scores among all cell lines and found excellent agreement (**Supplementary Fig. 1a**), with Pearson correlation 0.784, 0.818, and 0.9997 respectively for processed, unprocessed and batch-corrected data (*p* below machine precision in all cases; N = 16,773). The near-unity correlation of gene score means in the batch-corrected data reflects the nature of the ComBat batch correction, which explicitly aligns gene means and variances between experimental conditions.

We further tested whether it was possible to recover consistent sets of common dependencies. To this end, we defined as “common dependencies” those genes that rank among the top dependencies in their 90th percentile least dependent cell line (**Fig. 1c**, Methods). For the unprocessed data, the Broad and Sanger jointly identify 1,031 common dependency genes (**Supplementary Table 3**). 260 putative common dependencies were only identified by the Sanger and 397 were only identified by the Broad, thus showing that even without batch correction, the majority of genes defined as common dependencies are

consistent across the two datasets (Cohen's kappa = 0.737, Fisher's exact test *p* below machine precision, *N* = 16,773, **Fig. 1d**).

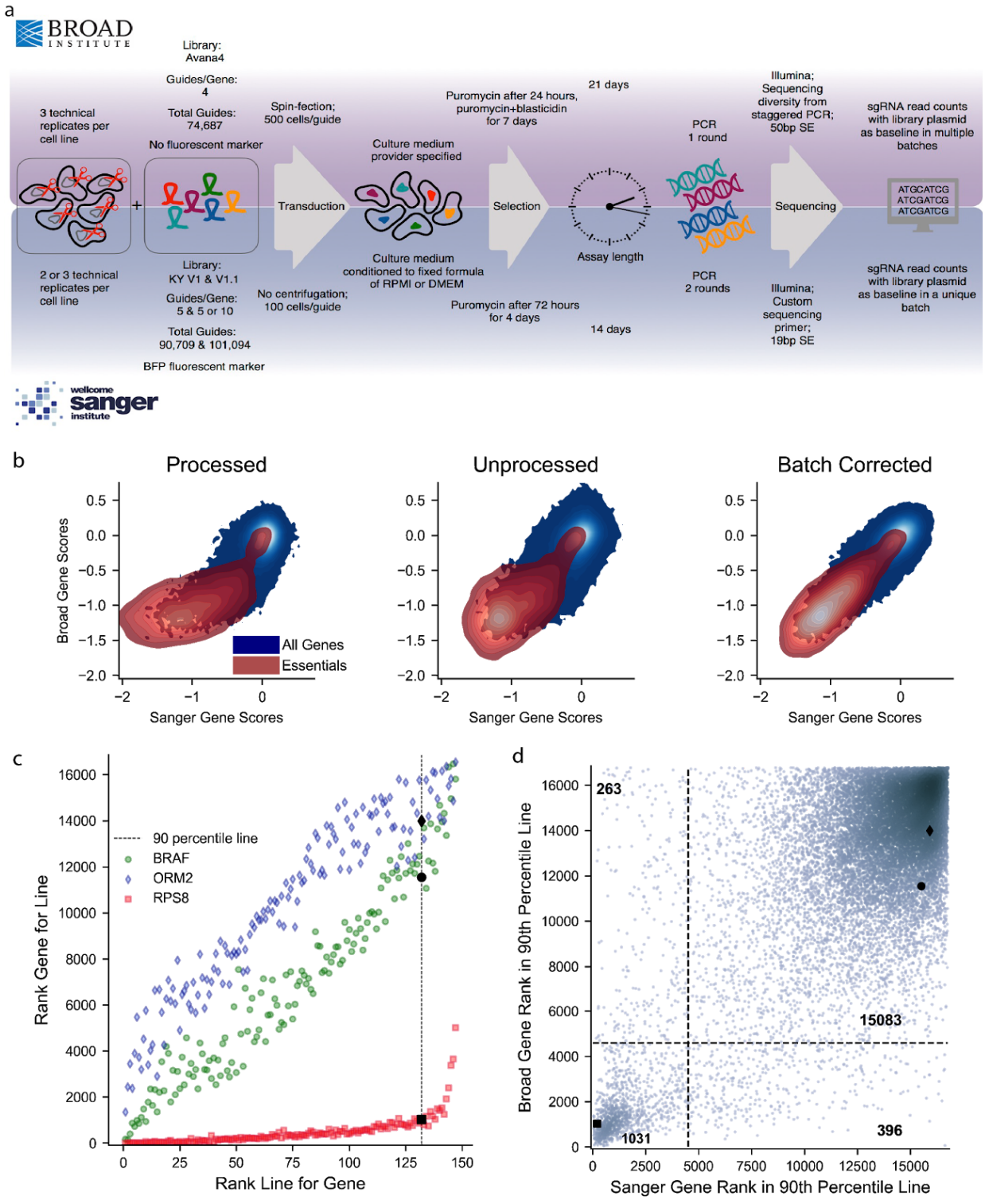

Figure 1: Comparison of experimental protocols and gene score results. **(a)** Comparison of experimental settings and reagents used in the experimental pipelines underlying the two compared datasets. **(b)** Densities of individual gene scores in the Broad and Sanger datasets, across processing levels. The distributions scores for

previously identified essential genes[26] are shown red. **(c)** Examples of the relationship between a gene's score rank in a cell line and the cell line's rank for that gene using Broad unprocessed gene scores, with gene ranks in their 90th percentile least dependent lines highlighted. For the common dependency RPS8, even its 90th percentile least dependent cell line still ranks the gene among the strongest of its dependencies. **(d)** Distribution of gene ranks for the 90th-percentile of least dependent cell lines for each gene in both institutes. Black dotted lines indicate natural thresholds at the minimum gene density along each axis.

Agreement of selective gene dependency profiles across cell lines

In the context of therapeutic target identification in oncology, consistently identifying the same cell lines as being dependent on a gene across studies is an important step for establishing that gene as a target of interest. In both studies, most genes show little variation in their dependency scores across cell lines. Thus we expect low shared variance even if most scores are numerically similar between the datasets[27]. Accordingly, we focused on two groups of genes for which the score variance across lines is of potential biological interest. The first set is composed of genes whose observed dependency profile is suggestive of real biological selectivity in at least one of the two unprocessed datasets. We call these 49 genes Strongly Selective Dependencies (SSDs) (**Supplementary Table 4**). The second gene set consists of 119 known cancer genes identified from literature by Vogelstein *et al.*[28] (**Supplementary Table 5**). The cancer gene set was not filtered for the presence of dependent cell lines

We evaluated agreement between gene score patterns using Pearson's correlations to test the reproducibility of selective viability phenotypes. **Fig. 2a** illustrates the score patterns for the example cancer genes MDM4 (R = 0.820, p = 6.91 x 10^{-37}), KRAS (R = 0.765, p = 1.66 x 10^{-29}), CTNNB1 (R = 0.803, p = 1.92 x 10^{-34}), and SMARCA4 (R = 0.664, p = 4.61 x 10^{-20}) with unprocessed data (N = 147). For SSDs and unprocessed data, the median correlation was 0.633 and 84% of SSDs showed a correlation greater than 0.4. For known cancer genes, we observed a median correlation of 0.342, with 42.9% of genes showing a correlation greater than 0.4 (**Fig. 2b**). Five SSDs showed a correlation below 0.2 (ABHD2, CDC62, HIF1A, HSPA5, C17orf64). As expected, correlation across datasets for all genes was lower (median R = 0.187, with 8.34% genes with R > 0.4).

For CRISPR-Cas9 screens to provide guidance for precision medicine, it should be possible to consistently classify cells as dependent or not dependent on selective dependencies. Therefore, we evaluated the agreement of the Broad and Sanger datasets on identifying the cell lines which are dependent on each SSD gene. We classified cell lines as dependent if their gene scores were less than -0.7, as scores higher than this are dominated by a single large group of insensitive gene scores centered

at zero (**Fig. 2c**). The area under the receiver-operator characteristic (AUROC) for recovering binary Sanger dependency on SSDs using Broad dependency scores was 0.940 in processed data, 0.963 in unprocessed data, and 0.965 in corrected data; to recover Broad binary dependency from Sanger scores, AUROC scores were 0.933, 0.859, and 0.967 respectively. The recall of Sanger-identified dependent cell lines in Broad data was 0.700 with precision equal to 0.252 for processed data, 0.847 and 0.347 for unprocessed data, and 0.756 and 0.603 for batch-corrected data (**Supplementary Fig. 1b**). Agreement is higher than could be expected by chance under all processing regimes (Fisher's exact $p = 4.16 \times 10^{-41}$ in processed, 1.37×10^{-86} in unprocessed, and 2.96×10^{-215} in batch-corrected data; $N = 7{,}203$). We observed a distinct group of Broad-exclusive dependencies in both processed and unprocessed data (lower right quadrants of **Fig. 2c**), which reduced agreement. A large proportion of these (56.2 % in processed data and 42.7% in unprocessed data) were due to the single gene HSPA5, which is an SSD in Sanger data but a common dependency in Broad data. Examining agreement among SSDs individually avoids the contamination caused by this gene. We found median Cohen's kappa for sensitivity to individual SSDs of 0.437 in processed, 0.661 in unprocessed, and 0.735 in batch-corrected data. In unprocessed data, 65% of SSDs had Cohen's kappa greater than 0.4, as opposed to 0.17% seen by chance (**Supplementary Fig. 1c**). On the basis of these results, the Broad and Sanger datasets identified a shared set of dependent cell lines for most selective genes.

Agreement of cell line dependency profiles

For the two datasets to be considered reproducible, the dependency profile of a given cell line should also be consistent across the two studies. Therefore, to evaluate the agreement of cell line dependency profiles we assembled a combined dataset of dependency profiles from both studies and computed all possible pairwise correlation distances between them. Gene sets used to define dependency profiles for each cell line included any gene showing significant dependency effect in the batch-corrected data for any cell line from either study. Significant dependency effect was defined using the threshold of -0.7 as above. A t-distributed stochastic neighbor embedding (tSNE)[29] visualization derived from these distance scores is shown in **Fig. 2d**. For the uncorrected data, we observed a perfect clustering of the dependency profiles by their study of origin, confirming a major batch effect that overwhelmed dependency profile similarity. Using the batch-corrected datasets improved agreement between dependency profiles across studies. Following correction, we observed a larger integration of cell lines from the two institutes as well as increased proximity of cell lines from one study to their counterparts in the other study (**Fig. 2e**).

To evaluate the agreement of cell lines across studies we examined how closely the gene score profile of each cell line screened in one institute matched its clone in the other institute. For each cell line gene

score profile in one institute, we ranked all the other cell line gene score profiles (from both institutes) based on their correlation distance to the profile under consideration. For batch-corrected data, 175 of 294 (60%) cell lines from one study have their clone in the other study as the closest (first) neighbor, and 209 of 294 (70%) cell lines having it among the 5 closest neighbors. The area under the normalized Recall curve (nAUC) averaged across all classification tests was equal to 0.91 for batch-corrected data, while for uncorrected data we observed near-random performance, with nAUC = 0.54 and no cell lines matching their counterpart within the 5 closest cell lines (**Fig. 2f**). The performance was comparable for all gene sets used with the best performance observed for the cancer gene sets (nAUC=0.98), the second best for the SSD genes (nAUC=0.94) and worst for all genes (nUAC=0.90). The percentage of cell lines matching closest to their counterparts in the other study was 57% for the full gene set, 68% for the cancer gene set and 43% for SSD genes. Further, the tSNE plots for each gene set showed similar improvement after correction (**Supplementary Fig. 2a-c**). These results show a consistent ability to match cell lines following batch correction and that variation in performance is consistent with a trade-off between signal and noise when integrating data sets.

The batch correction resulted in an increased agreement between numbers of significant (at 5% FDR) dependencies across cell lines between the two datasets with the median number of dependencies changing from 2,109 and 1,717 to 2,053 and 1,950, for Broad and Sanger respectively, **Supplementary Fig. 3a**. The screen agreement, quantified as the proportion of dependencies detected in both studies over those detected in at least one study across all cell lines on average, increased from 47.75% to 59.14%. Furthermore, the correlation between cell lines after correction rose above the correlation within each individual screen for each gene set considered (**Supplementary Fig 3d**).

ComBat is designed to remove systematic differences in dependency profiles between the two studies. As the agreement in dependency profiles after correction was 57% we next investigated whether differences in data quality between cell line pairs explained the residual discordance. First, we computed a quality score (QS) genome-wide for each cell line in each study. For the quality score, we used the True positive rate (TPR) where a true positive was defined as a known common dependency that was also identified as being significantly depleted. Depletion scores were based on the log FCs and the threshold for a significant depletion was set to give an overall 5% FDR based on known common dependencies and nonessential genes. Second, we compared the average quality scores for a cell line pair to their screen agreement using linear regression. We found that quality score is a strong predictor of screen agreement for both the uncorrected and batch-corrected data sets (p-values 2.06×10^{-35}, 4.74×10^{-35} and adjusted R-squared 0.65, 0.64 for uncorrected and batch-corrected respectively; **Supplementary Fig. 3b**).

These results show that applying an established batch correction procedure mitigates the disagreement between the two datasets. Following correction, the two studies show a good concordance at the level of overall dependency profiles suggesting that these screens are reproducible. Furthermore, part of the residual disagreement following correction can be explained by differences in the inherent quality of the individual screens within the two datasets.

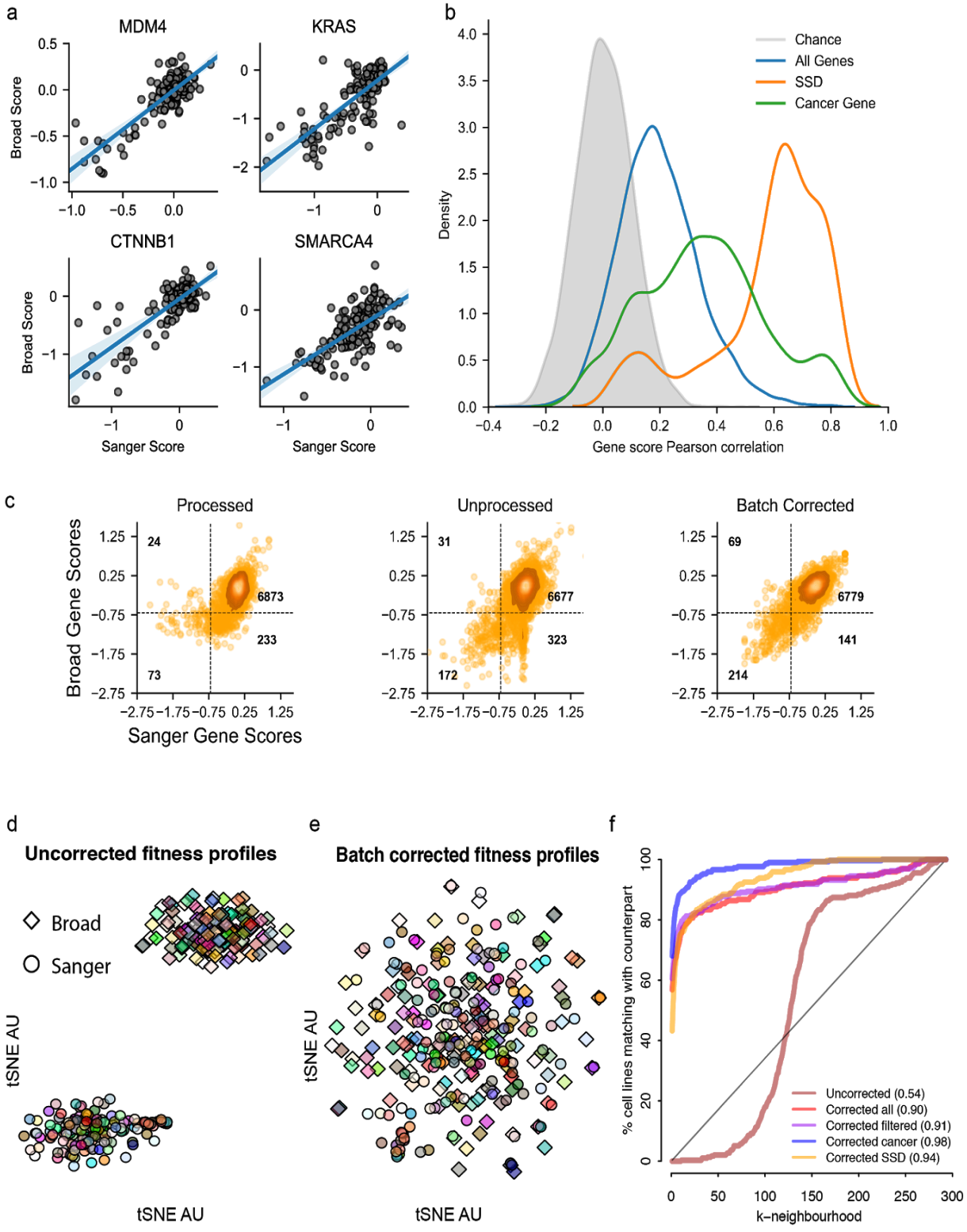

Figure 2: Reproducibility of gene and cell line dependency profiles. **(a)** Score pattern examples for selected known cancer genes. **(b)** Distribution of correlations of scores for individual genes in unprocessed data. **(c)** Gene

dependency scores for SSDs across all cell lines, with the threshold for calling a line dependent set at -0.7. **(d)** t-SNE clustering of cell lines in unprocessed data using correlation between gene scores. Colors represent the cell line whilst shape denotes the institute of origin. (**e)** The same as in (d) but for data batch-corrected using ComBat. **(f)** Recovery of a cell line's counterpart in the other institute's data before (Uncorrected) and after correction (Corrected) measured using Pearson correlation between cell lines based on different gene sets. First nearest neighborhood defined by ranking cell line similarity based on correlations between cell lines, nAUC values are shown in brackets.

Agreement of gene dependency biomarkers

A selective dependency is of limited therapeutic value unless it can be reliably identified by some characteristic of the tumor (*biomarker*). Following a similar approach to that presented in[27], we performed a systematic test for molecular-feature/dependency associations on the two datasets. Cell lines were split into two groups based on the status of 587 molecular features derived from Iorio *et al.*[7] and encompassing somatic mutations in high-confidence cancer driver genes, amplifications/deletions of chromosomal segments recurrently altered in cancer and hypermethylated gene promoters, microsatellite instability status and the tissue of origin of the cell lines (**Supplementary Table 6**). For each feature in turn, all SSD genes were sequentially t-tested for significant differences in dependency scores between the obtained two groups of cell lines.

This yielded 71 significant associations (FDR < 5%, ΔFC < -1) between molecular features and gene dependency when using the Broad unprocessed data, and 90 when using the Sanger unprocessed data (**Supplementary Table 7**). Of these, 55 (77% of the Broad associations and 61% of the Sanger ones) were found in both datasets (FET p-value = 9.08×10^{-133}, **Fig. 3a and Supplementary Tables 7-8**). The concordance between the associations identified by each study was proportional to the threshold used to define significance. This was assessed by considering for each study, in turn, the associations in a fixed quantile of significance and measuring the tendency of these associations to be among the most significant in the other study **Fig. 3b**. Further, the overall correlation between differences in gene depletion FCs was equal to 0.763, and 99.2% of associations had the same sign of differential dependency across the two studies.

The gene dependency associations identified by both institutes included both expected and potentially novel cases. Expected associations included an increased dependency on ERBB2 in ERBB2-amplified cell lines, and increased dependency on beta catenin in APC mutant cell lines. In addition, a potentially novel association between FAM72B promoter hypermethylation and beta catenin was identified (**Fig. 3c**).

We also considered gene expression to mine for biomarkers of gene dependency. We used the RNA-seq data from each institute for overlapping cell lines, which includes some sequencing files that have been used by both institutes and processed separately. We considered the intersection of the top two thousand most variable genes from either institute as potential biomarkers. The overlap between the most variable genes in each RNA-seq data set was high with 1,987 out of the 2,000 shared by both institutes. Further, clustering of the RNA-seq profiles found that each cell lines' transcriptome matched closest to its counterpart from the other institute (**Supplementary Fig. 4a**). To assess the relationship between gene expression and dependency, gene expression for the most variably expressed genes was correlated to the gene dependency profiles of the SSD genes. Systematic tests of each correlation showed significant associations between gene expression and dependency (**Fig. 3d**). As with the genomic biomarkers, we found a strong overall correlation between gene expression markers and SSD genes dependency across institutes, Pearson's correlation 0.804 and significantly high overlap between gene expression biomarkers identified by each institute (FET p-value below machine precision). We observed both positive and negative correlations; for example, ERBB2 dependency score was positively correlated with its expression, while ATP6V0E1 showed significant dependency when its paralog ATP6V0E2 had low expression **Fig. 3e**.

Taken together these results show that there is a significant agreement between potential biomarkers of gene dependency unveiled by the two studies. This agreement is proportional to the level of statistical significance of the detected biomarkers. We found good agreement when using both genomic and transcriptomic biomarkers. In both cases, there is a large and significant correlation between patterns of significance level between tested associations across the two studies.

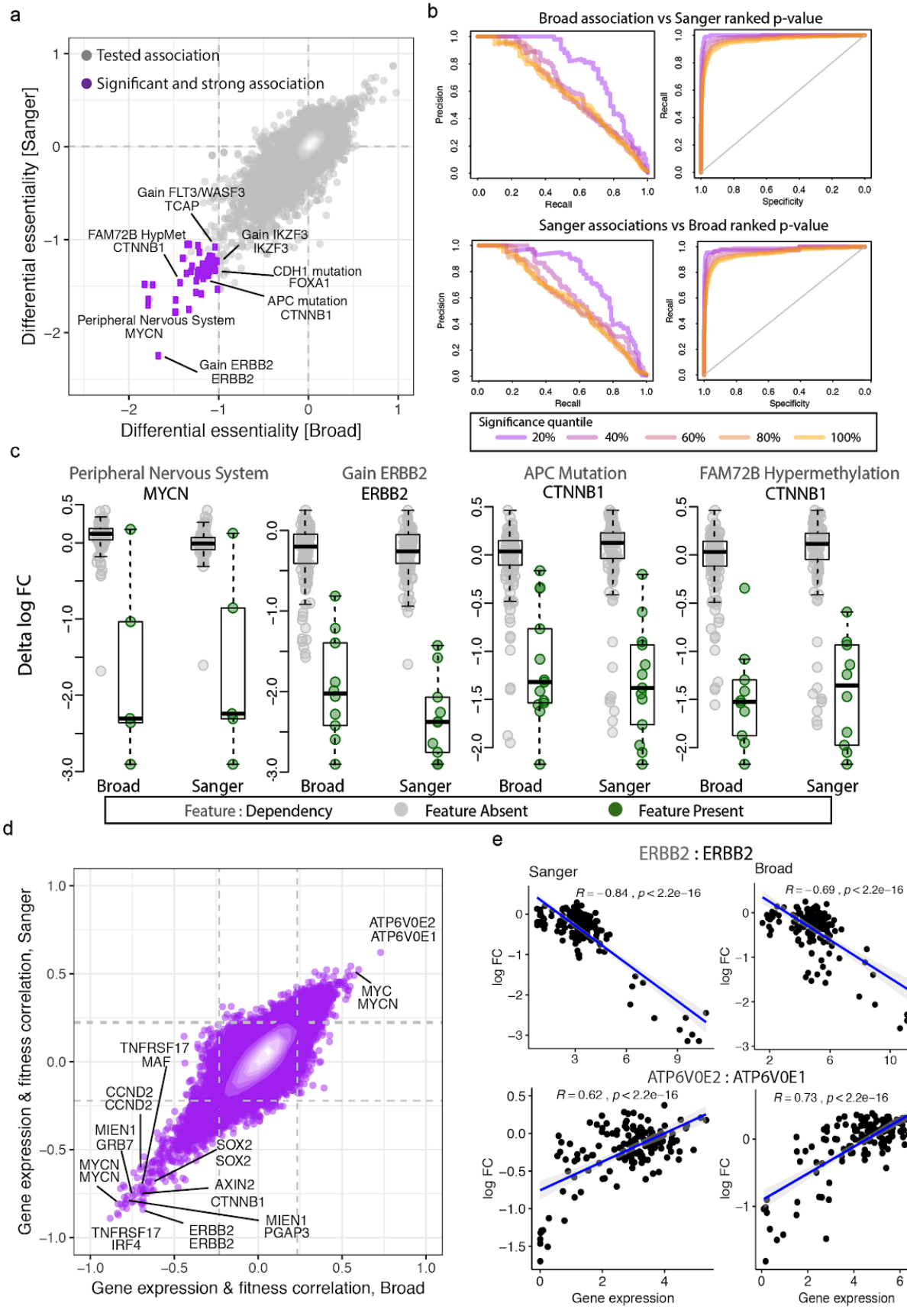

Figure 3: Reproducibility of biomarkers (a) Results from a systematic association test between molecular

features and differential gene dependencies (of the SSD genes) across the two studies. The feature is on the first line with the dependency underneath. **(b)** Precision/Recall and Recall/Specificity curves obtained when considering as true positives the associations falling in a fixed top quantile of significance in one of the studies and a classifier based on the p-values of the associations from the other study. **(c)** Examples of significant statistical associations between genomic features and differential gene dependencies across the two studies. **(d)** Results from a systematic correlation test between gene expression and dependency of SSD genes across the two studies. **(e)** Examples of significant correlations between gene expression and dependencies in both studies

Elucidating sources of disagreement between the two datasets

Despite the concordance observed between the Broad and Sanger datasets, our investigations so far have shown batch effects in the unprocessed data both in individual genes and across cell lines. Although the bulk of these effects are mitigated by ComBat, their cause is an important experimental question. We enumerated the experimental differences between institutes (**Fig. 1a**) to identify likely causes of batch effects. The choice of sgRNA can significantly influence the observed phenotype in CRISPR-Cas9 experiments, implicating the differing sgRNA libraries as a likely source of batch effect[30]. Additionally, previous studies have shown that some gene inactivations only produce depleting phenotypes in lengthy experiments[15]. Accordingly, we selected the sgRNA library and the timepoint of viability readout for investigation as causes of institute batch effects.

To elucidate the role of the sgRNA library, we examined the data at the level of individual sgRNA scores. The correlation log fold change for reagents targeting the same gene ("co-targeting") is related to the gene's average normLRT score (**Fig. 4a**), a reminder that most co-targeting reagents have low correlation simply because they target genes with little phenotypic variation. However, even among SSDs there is a clear relationship between sgRNA correlations within and between institutes (p = 4.9 x 10^{-10}, N = 49; **Fig. 4b**). In such cases, poor reagent efficacy at one or both institutes may explain the discrepancy. We estimated the efficacy of each sgRNA in both libraries using Azimuth 2.0[30] which uses only information about the genome in the region targeted by the sgRNA. We found that among genes identified as common dependencies in either dataset, mean sgRNA depletion indeed had a strong relationship to its Azimuth estimated efficacy (**Fig. 4c**). When we examined SSDs, we found that reagent efficacy likely explains some differences. For example, both the Avana and KY libraries have good mean estimated sgRNA efficacy (MESE) for the transcription factor TFAP2C, and the unprocessed gene scores from each institute show similar scales. However, for the translation initiation factor EIF3F, the Avana library has MESE of only 0.398 compared to 0.613 in KY. This gene is a common dependency in Sanger screens but nonscoring in Broad screens. The known cancer gene MDM2 is an example of the opposite case, where

the KY library has MESE of only 0.402, compared with 0.585 in Avana. Unsurprisingly, while the Sanger dataset identifies the same set of depleted cell lines after MDM2 knockout as the Broad, the depletion is weaker (**Fig. 4d**). Overall, reagent analysis indicates that poor reproducibility of a gene's profile can be identified in advance by examining sgRNA consistency and that such cases may be due to poor sgRNA efficacy in at least one library.

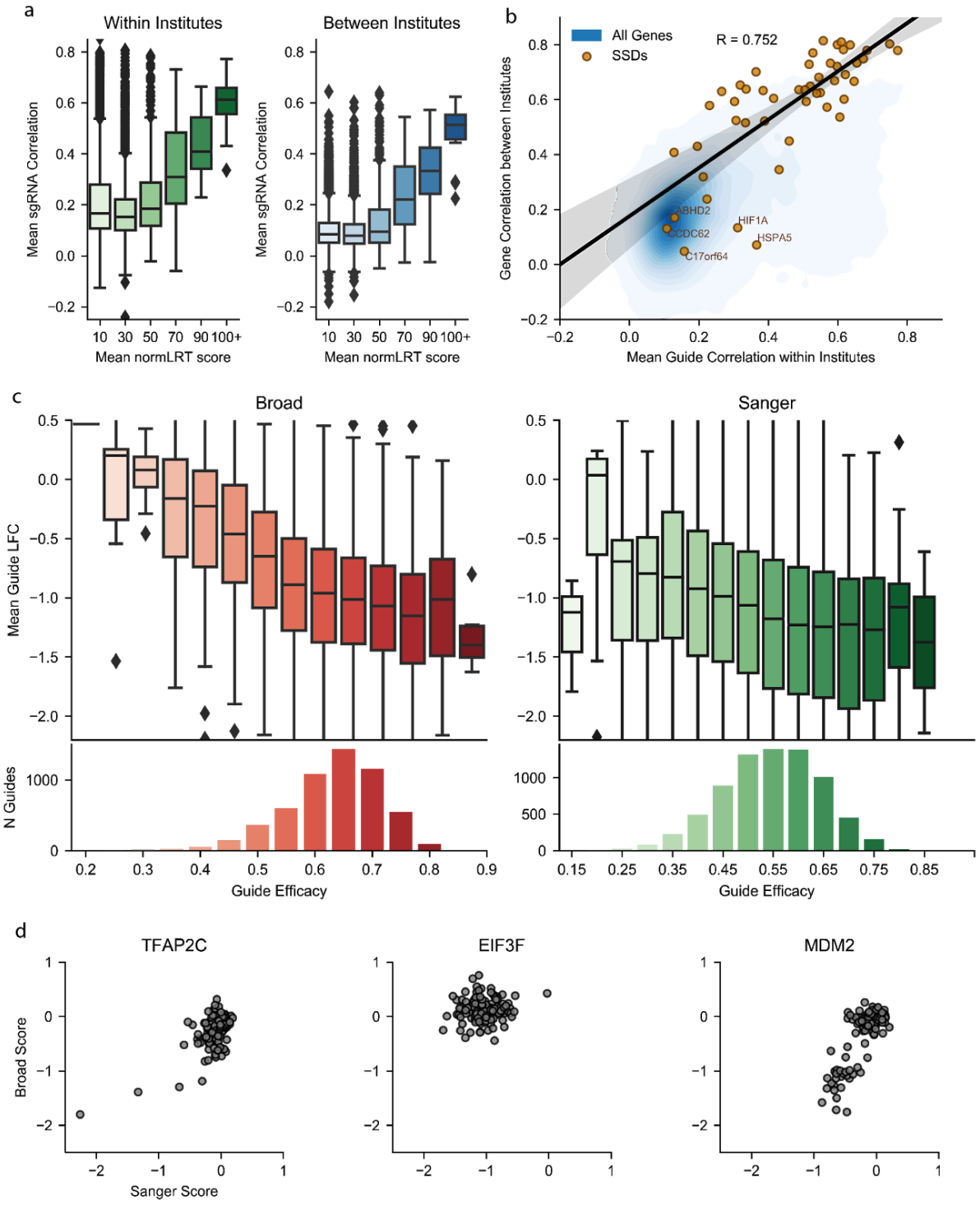

Figure 4: **Influence of reagent library**. **(a)** Distributions of sgRNA correlation for sgRNAs targeting genes with varying NormLRT scores within each institute and between them. Both axes are averaged between the institutes. **(b)** Relationship between sgRNA correlation within institutes and gene correlation between institutes. The linear trend is shown for SSD genes. **(c)** The mean depletion of guides targeting common dependencies across all

replicates vs Azimuth estimates of guide efficacy. **(d)** Comparison of Broad and Sanger unprocessed gene scores for genes matching (1) SSD with highest minimum MESE across both libraries, (2) common dependency in either dataset and greatest difference between KY and Avana MESE, (3) SSD with worst KY MESE.

We next investigated the role of different experimental timepoints on the screens. Given that the Broad has a longer assay length (21 days versus 14 days) we expected differences to be observed between late dependencies across the institutes. Therefore, we compared the distribution of gene scores for genes known to exert a loss of viability effect upon inactivation at an early- or late-time (early or late dependencies) [15]. While early dependencies have similar score distributions in both datasets (median average score -0.781 at the Sanger and -0.830 at the Broad), late dependencies are more depleted at the Broad with median average score -0.402 compared to -0.269 for the Sanger screens (**Fig. 5a**). Similarly, gene set enrichment analysis (GSEA) of the mean time point change in cell dependency profiles shows that late dependencies are strongly enriched for greater dependency at a later time point, confirming that the longer time point for the Broad screens was the reason these genes are systematically more depleted in Broad data **Fig. 5b**.

Unlike differences in sgRNA efficacy, timepoint effects are expected to lead to uniformly greater signal (typically depletion) in the Broad data and be related to the biological role of late dependencies. We tested this hypothesis by first filtering the unprocessed data to remove genes with significantly different efficacy across libraries so as to remove possible confounding results due to library differences (Methods). Second, we functionally characterized, using gene ontology (GO), genes that were exclusively detected as depleted in individual cell lines (at 5% FDR), in one of the two studies. Results showed 29 gene ontology categories significantly enriched in the Broad-exclusive dependencies genes (Broad-exclusive GO terms) for more than 50% of cell lines (**Supplementary Fig. 5 and Supplementary Table 9)**. The Broad-exclusive enriched GO terms included classes related to mitochondrial and RNA processing gene categories and other gene categories previously characterized as late dependencies[15]. In contrast, no GO terms were significantly enriched in the Sanger-exclusive common dependencies in more than 30% of cell lines.

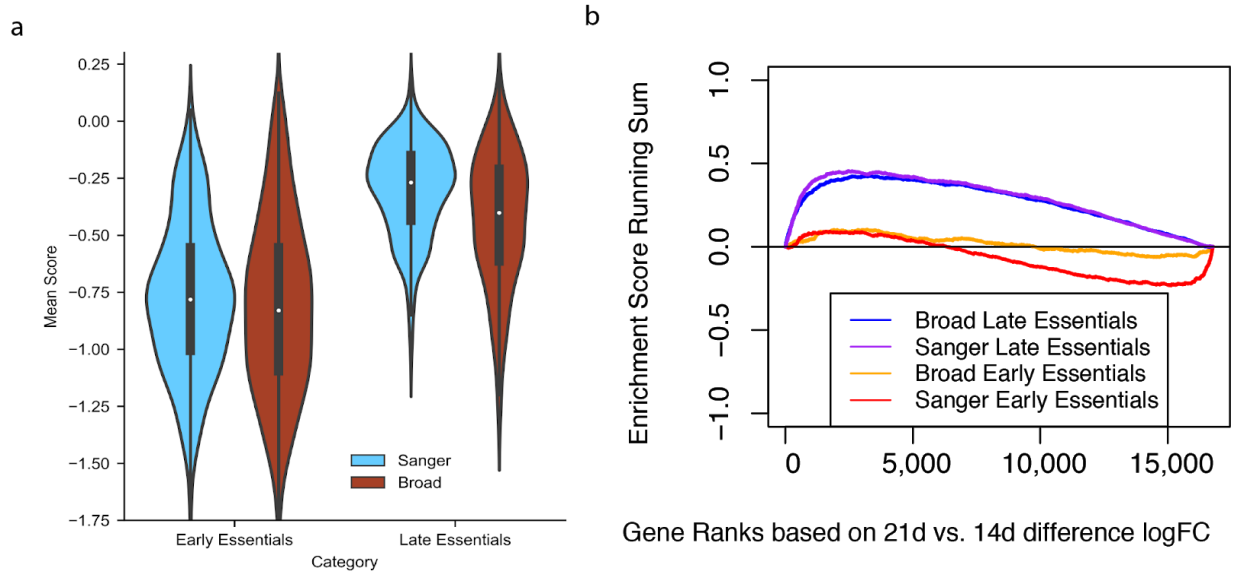

Figure 5: Influence of Time point. (a) Distribution of early and late common dependency scores in the Broad and Sanger datasets. **(b)** GSEA curves for the scores of late and early common dependencies between the Broad and Sanger datasets. **(c)** GO enrichment annotations of Broad-exclusive common dependencies not accounted for by estimated efficacy.

Experimental Verification of Batch Effects

To verify that batch effects between the datasets can be removed by changing library and the readout timepoint, the Broad and Sanger undertook replication experiments where these factors were systematically permuted. The Broad sequenced cells collected from its original HT-29 and JIMT-1 screens at the 14-day timepoint and conducted an additional screen of these cell lines using the KY1.1 library with readouts at days 14 and 21. The Sanger used both the Broad's and the Sanger's clones of HT-29 to conduct a new KY screen and an Avana screen with readouts at days 14 and 21. Principal component analysis (PCA) of the concatenated unprocessed gene scores, including replication screens, showed a clear institute batch effect dominating the first principal component. By highlighting replication screens, we found that this effect is principally due to library choice, with timepoint playing a smaller role (**Fig. 6a**, **Supplementary Fig. 6a**). Changing from Sanger to Broad versions of HT-29 had minimal impact. We examined the change in gene score profile for each screen caused by changing either library or timepoint while keeping other conditions constant. We found that the gene score changes induced by either library or timepoint alterations were consistent across multiple conditions (**Fig. 6b**). Sanger-exclusive common dependencies were strongly enriched for genes that became more depleted with the KY library, and Broad-exclusive common dependencies were enriched among genes more

depleted with the Avana library (**Supplementary Fig. 6b**). Late dependencies were strongly enriched among genes that became more depleted in the later timepoints, while early dependencies were not (**Supplementary Fig. 6c**). We compared the deviations in gene score between Broad and Sanger screens under different conditions, first comparing Broad original and replication screens of HT-29 (Fig. 6c) and JIMT-1 (**Supplementary Fig. 6d**) to the original Sanger screens of the same cell line. Matching Sanger's library and timepoint reduces the variance of gene scores in HT-29 from 0.0486 to 0.0252 and in JIMT-1 from 0.0556 to 0.0260. Specifically, matching library and timepoint removes most of the average gene score change (batch effect) between institutes, as indicated by the low correlation of the remaining gene score differences in the replication screens with the average gene score change. We next compared Sanger original and replication screens of HT-29 to the Broad original HT-29 screen. Matching library and timepoint successfully detrended the data in this case as well; however, the Sanger Avana screens of HT-29 contained considerable excess noise, causing these screens to have higher overall variance from the Broad than the original screens (0.0486 vs 0.115). Nonetheless, the replication experiments confirm that the majority of batch effects between institutes are driven by library and timepoint.

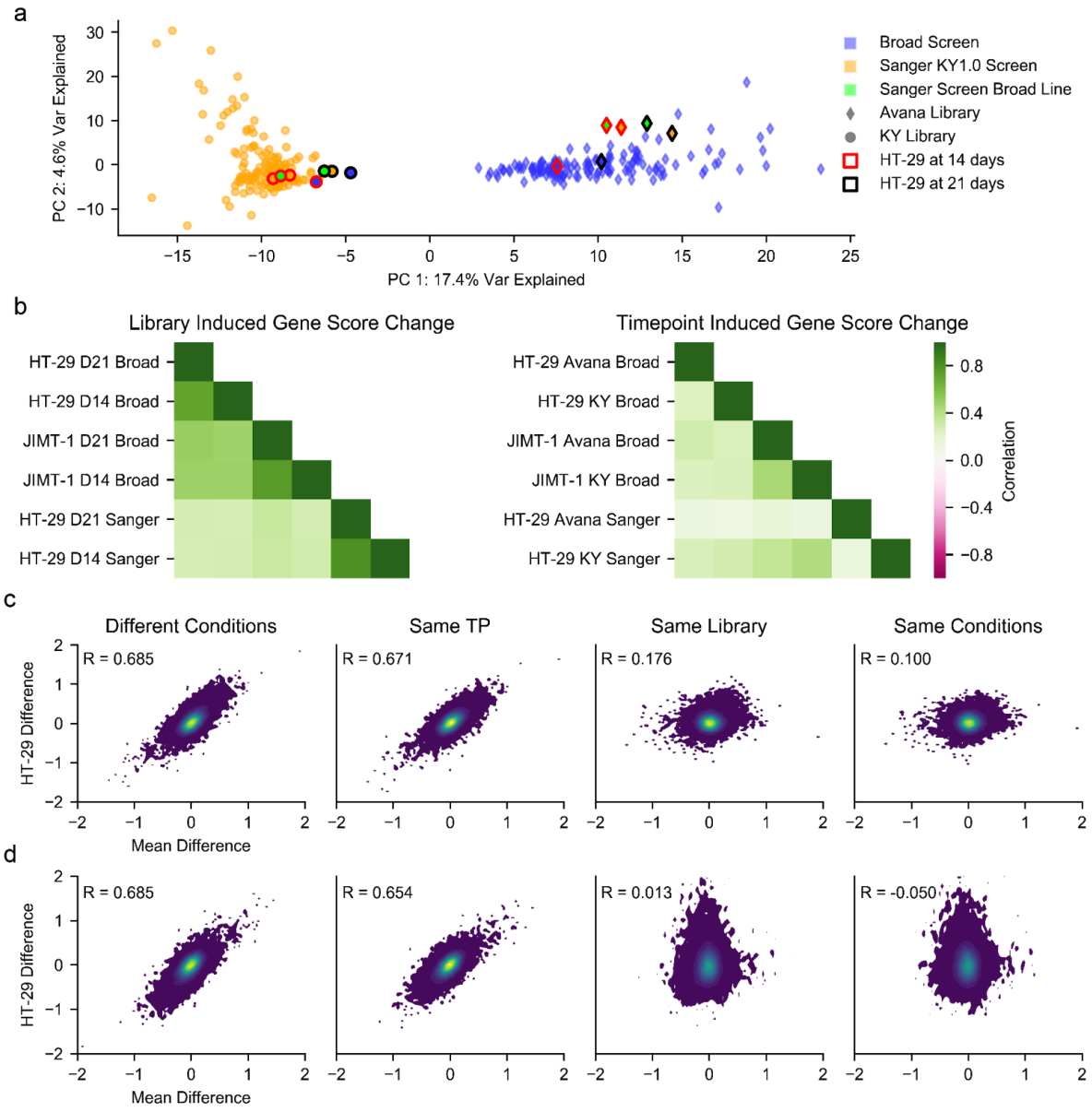

Figure 6: Results of each institute's replication experiments. **(a)** PCA plot of the first two principal components of the concatenated original and replication screens from each institute with HT-29 screens highlighted. Axes are scaled to the variance explained by the component. **(b)** Correlations of the changes in gene score caused when changing a single experimental condition. **(c)** The difference in unprocessed gene scores between Broad screens of HT-29 and the original Sanger screen (Sanger minus Broad), beginning with the Broad's original screen and ending with the Broad's screen using the KY library at the 14-day timepoint. Each point is a gene. The horizontal axis is the mean difference of the gene's score between the Sanger and Broad original unprocessed datasets. **(d)** A similar plot taking the Broad's original screen as the fixed reference and varying the Sanger experimental conditions (Broad minus Sanger).

Discussion

Providing sufficient experimental data to adequately sample the diversity of human cancers requires high-throughput screens. However, the benefits of large datasets can only be exploited if the underlying data is reliable and robustly reproducible. In this work, we survey the agreement between two large, independent CRISPR-Cas9 knock-out datasets, generated at the Broad and Sanger institutes.

Our findings illustrate a high degree of consistency in estimating gene dependencies between studies at multiple levels of data processing, albeit with the longer duration of the Broad screens leading to stronger dependencies for a number of genes. The datasets are concordant in identifying common dependencies and identifying mean dependency signals. Their agreement is also striking in the more challenging task of identifying which cell lines are dependent on selective dependencies.

One of the aims of the studies underpinning the two compared datasets is the identification of novel and highly selective oncology therapeutic targets, prioritizing them based on the availability of a molecular marker of their dependency. As a consequence, we compared the two datasets at the level of gene dependency markers, as this provides one of the most relevant and impactful measures of their concordance. We found that, when analyzed individually, the two datasets revealed common gene dependency biomarkers, with significantly correlated corresponding patterns of significance and differential dependency.

Although we found that the overall agreement was significantly high at all the tested levels, we found that genes with highly selective dependency profiles in at least one dataset are nonetheless not well-correlated between datasets. We tested whether it was possible to identify genes with poor correlation in advance, knowing the properties of its distribution in one dataset and the concordance of its targeting sgRNAs in both datasets. We found that the great majority of low-correlation SSDs can indeed be identified in advance. This provides some hope that in those cases where two gene profiles disagree, it may be possible to adjudicate which dataset to trust using sgRNA consistency and the properties of the gene profile.

Another source of observed disagreement is due to diffuse batch effects visible when the whole profiles of individual cell lines are compared. Such effects can be readily corrected with standard methods without compromising data quality, thus making possible integration and future joint analyses of the two compared datasets. Furthermore, much of this batch effect can be decomposed into a combination of two experimental choices: the sgRNA library and the duration of the screen. The effect of each choice on the

mean depletion of genes is readily explicable and reproducible, as shown by screens of two lines performed at the Broad using the Sanger's library and screen duration. Consequently, identifying high-efficacy reagents and choosing the appropriate screen duration should be given high priority when designing CRISPR-Cas9 knock-out experiments.

Methods

Collection and Preprocessing of Data

"Unprocessed" Gene Scores

Read counts for the Broad were taken from avana_public_19Q1[31] and filtered so that they contained only replicates corresponding to overlapping cell lines and only sgRNAs with one exact match to a gene. Read counts for Sanger were taken from Behan *et al.*[17] and similarly filtered, then both read counts were filtered to contain only sgRNAs matching genes common to all versions of the data. In both cases, reads per million (RPM) was calculated and an additional pseudo-count of 1 added to the RPM. Log fold change was calculated from the reference pDNA. In the case of the Broad, both pDNA and screen results fall into distinct batches, corresponding to evolving PCR strategies. Cell lines sequenced with a given batch were matched to pDNA profiles belonging to the same batch. Multiple pDNA profiles in each batch were median-collapsed to form a single profile of pDNA reads for each batch. Initial gene scores for each replicate were calculated from the median of the sgRNAs targeting that replicate. Each replicates initial gene scores for both Broad and Sanger were then shifted and scaled so the median of nonessential genes in each replicate was 0 and the median of essential genes in each replicate is -1[26]. Replicates were then median-collapsed to produce gene- by cell-line matrices.

"Processed" Gene Scores

Broad gene scores are taken from avana_public_19Q1 gene_effect[31] and reflect CERES[20] processing. The scores were filtered for genes and cell lines shared between institutes and with the unprocessed data, then shifted and scaled so the median of nonessential genes in each cell line was 0 and the median of essential genes in each cell line was -1[26]. Sanger gene scores were taken from the quantile-normalized averaged log fold-change scores at [] and globally rescaled by a single factor so that the median of essential genes across *all* cell lines is -1[26].

"Batch-Corrected" Gene Scores

The unprocessed sgRNA log FCs were mean collapsed by gene and replicates. Data were quantile normalised for each institute separately before processing with ComBat using the R package sva. One batch factor was used in ComBat defined by the institute of origin. The ComBat corrected data was then quantile normalized to give the final batch-corrected data set.

Alternate Conditions

Screens with alternate libraries, cell lines, and timepoints were processed similarly to the "Unprocessed" data above.

Gene Expression Data

Gene expression $\log_2$(Transcript per million+1) data was downloaded for the Broad from Figshare for the Broad data set. For the Sanger dataset, we used the read per kilobase million (RPKM) expression data from the iRAP pipeline. We added a pseudo-count of 1 to the RPKM values and transformed to $\log_2$. Gene expression values are quantile normalized for each institute separately. For the Sanger data, Ensembl gene ids were converted to Hugo gene symbols using BiomaRt package in R.

Guide Efficacy Estimates

On-target guide efficacies for the single-target sgRNAs in each library were estimated using Azimuth 2.0[30] against GRCh38.

Comparison of All Gene Scores

Gene scores from the chosen processing method for both Broad and Sanger were raveled and Pearson correlations calculated between the two datasets. 100,000 gene-cell line pairs were chosen at random and density-plotted against each other using a Gaussian kernel with the width determined by Scott's rule[32]. All gene scores for essential genes were similarly plotted in **Fig. 1b**.

Comparison of Gene Means

Cell line scores for each gene in both Broad and Sanger datasets with the chosen processing method were collapsed to the mean score, and a Pearson correlation calculated.

Gene Ranking, Common Essential Identification

For each gene in the chosen dataset, its score rank among all gene scores in its 90th percentile least depleted cell line was calculated. We call this the gene's 90th percentile ranking. The density of 90th-percentile rankings was then estimated using a Gaussian kernel with width 0.1 and the central point of minimum density identified. Genes whose 90th-percentile rankings fell below the point of minimum density were classified as essential.

Identification of Selective Gene Sets

Selective dependency distributions across cell lines are identified using a "Likelihood Ratio Test" as described in McDonald et al [19]. For each gene, the log-likelihood of the fit to a normal distribution and a skew-t distribution is computed using the R packages MASS [33] and sn [34], respectively. In the event that the default fit to the skew-t distribution fails, a two-step fitting process is invoked. This involves keeping the degrees of freedom parameter (nu) fixed during an initial fit and then using the parameter estimates as starting values for a second fit without any fixed values. This process repeats up to 9 times using nu values in the list (2, 5, 10, 25, 50, 100, 250, 500, 1000) sequentially until a solution is reached. The numerical optimization methods used for the estimates do not guarantee the maximum of the objective function is reached. The reported LRT score is calculated as follows:

LRT = 2*[ln(likelihood for Skewed-t) - ln(likelihood for Gaussian)]

Genes with NormLRT scores greater than 100 and mean gene score greater than -0.5 in at least one institute's unprocessed dataset were classified as SSDs. The cancer gene set was taken directly from Vogelstein *et al.*[28]

Binarized Agreement of SSDs

SSD gene scores in both Broad and Sanger datasets with the chosen processing method were binarized at -0.7, with scores falling below this threshold indicating the sensitivity of the cell line on the chosen gene. Cohen's kappa was calculated for each gene individually. Fisher's exact test, precision, recall, and AUROC scores were calculated globally for all SSD sensitivities in the three data versions.

Cell line agreement Analysis

To obtain the two dimensional visualisations of the combined dataset before and after batch correction and considering different gene sets, we computed the sample-wise correlation distance matrix and used

this as input into the t-statistic Stochastic Neighbor Embedding (t-SNE) procedure [29], using the *tsne* function of the homonym R package, with 1,000 iterations, a perplexity of 100 and other parameters set to their default value.

To evaluate genome-wide cell line agreement we considered a simple nearest-neighbor classifier that, for each dependency profile of a given cell line in one of the two studies, predicts its matching counterpart in the other study. This prediction was based on the correlation distance between one profile and all the other profiles. To estimate the performance of this classifier, we computed a Recall curve for each of the 294 dependency profiles in the tested dataset. Each of these curves was assembled by concatenating the number of observed true-positives amongst the first *k* neighbors of the corresponding dependency profile (for *k* = 1 to 293). We then averaged the 294 resulting Recall curves into a single curve and converted it to percentages by multiplying by 100/294. Finally we computed the area under the resulting curve and normalized it by dividing by 293. We considered the area under this curve (nAUC) as a performance indicator of the k-nn.
For the comparison of cell line profiles agreement in relation to initial data quality. First, to estimate the initial data quality we calculated True Positive Rates (TPRs, or Recalls) for the sets of significant dependency genes detected across cell lines, within the two studies. To this aim, we used as positive control a reference set of a priori known essential genes [16]. We assessed the resulting TPRs for variation before/after batch correction, and for correlations with the inter-study agreement.

Biomarker Analysis

We used binary event matrices based on mutation data, copy number alterations, tissue of origin and MSI status. The resulting set of 587 features were present in at least 3 different cell lines and fewer than 144. We performed a systematic two-sample unpaired Student's t-test (with the assumption of equal variance between compared populations) to assess the differential essentiality of each of the SSD genes across a dichotomy of cell lines defined by the status (present/absent) of each CFE in turn. From these tests we obtained p-values against the null hypothesis that the two compared populations had an equal mean, with the alternative hypothesis indicating an association between the tested CFE/gene-dependency pair. P-values were corrected for multiple hypothesis testing using Benjamini-Hochberg. We also estimated the effect size of each tested association by means of Cohen's Delta, i.e. difference in population means divided by their pooled standard deviations. For gene expression analysis we calculated the Pearson correlation across the cell lines between the SSD gene dependency profiles and the gene expression profiles from each institute. Significance of the correlation was assessed using the t-distribution (n-2 degrees of freedom) and p-values multiple hypothesis corrected using the q-value method.

For the agreement assessment via ROC indicators (Recall, Precision and Specificity), for each of the two studies in turn we picked the most significant 20, 40, 60, 80 and 100% associations as true controls and evaluated the performance of a rank classifier based on the corresponding significance p-values obtained in the other study.

Rank-based dependency significance and agreement quantification

To identify significantly depleted genes for a given cell line, we ranked all the genes in the corresponding essentiality profiles based on their depletion logFCs (averaged across targeting guides), in increasing order. We used this ranked list to classify genes from two sets of prior known essential (*E*) and non-essential (*N*) genes, respectively [16].

For each rank position *k*, we determined a set of predicted genes $P(k) = \{s \in E \cup N : \varrho(s) \leq k\}$, with $\varrho(s)$ indicating the rank position of *s*, and the corresponding precision *PPV(k)* as:

$$PPV(k) = |P(k) \cap E| / |P(k)|$$

Subsequently, we determined the largest rank position k^* with $P(k^*) \geq 0.95$ (equivalent to a False Discovery Rate (FDR) ≤ 0.05). Finally, a 5% FDR logFCs threshold F^* was determined as the logFCs of the gene s such that $\varrho(s) = k^*$, and we considered all the genes with a logFC < F^* as significantly depleted at 5% FDR level. For each cell line, we determined two sets of significantly depleted genes (at 5% FDR): *B* and *S*, for the two compared datasets, respectively. We then quantified their agreement using the Jaccard index [35] $J(B,S) = |B \cap S| / |B \cup S|$, and defined their disagreement as $1 - J(B,S)$. Summary agreement/disagreement scores were derived by averaging the agreement/disagreement across all cell lines.

sgRNA Correlations

Broad and Sanger log fold-changes for their original screens were median-collapsed to guide by cell line matrices. For each gene present in the unprocessed gene scores, a correlation matrix between all the sgRNAs targeting that gene in each guide by cell line matrix was computed. The mean of the values in this matrix for each institute, excluding the correlations of sgRNAs with themselves, was retained. The mean sgRNA correlation within institutes was then calculated from the mean of the Broad and Sanger sgRNA correlation matrix means. The mean sgRNA correlation between institutes for each gene was calculated from the mean of all possible pairs of sgRNAs targeting that gene with one sgRNA chosen from Sanger and one from Broad.

Relating sgRNA Depletion and Efficacy

We chose the set of genes found to be essential in at least one unprocessed dataset. The log fold-change of guides targeting those genes in each dataset was calculated and compared to the guide's estimated on-target efficacy.

Timepoint Gene Ontology Analysis

We tested for enrichment of GO terms associated with genes showing a significant depletion in only one institute. To rule out the differences due to library, genes with significantly different guide efficacies were filtered from the analysis. Using the Azimuth scores average (mean) efficacy scores for each gene at each institute were calculated. A null distribution of differences in gene efficacy was estimated using genes not present in either institute specific sets (which were defined as depleted in at least 25% of cell lines). Institute specific genes greater than 2 standard deviations from the mean of the null distribution were removed.

For the filtered gene set prior known essential and non-essential gene sets from [36] were used to find significant depletions for each cell line and institute at 5% FDR. For each cell line, the genes identified as significantly depleted in only Broad or only Sanger were functionally characterized using Gene Ontology (GO) enrichment analysis [37]. To this aim, we downloaded a collection of gene sets (one for each GO category) from the Molecular Signature Database (MsigDB) [38], and performed a systematic hypergeometric test to quantify the over-representation of each GO category for each set of study-exclusive dependency genes, per cell line. We corrected the resulting p-values for all the tests performed within each study using the Benjamini-Hochberg procedure [39], and considered a GO category enriched in a cell line if the corrected p-value resulting from the corresponding test was < 0.05.

Principal Component Analysis of the Batch Effect and Alternate Conditions

The Broad and Sanger unprocessed gene scores and the gene scores for the alternate conditions tested by both institutes were concatenated into a single matrix with a column for each screen. Principal components were found for the transpose of this matrix, where each row is a screen and each column a pseudogene. Components 1 and 2 were plotted for all original screens and the alternate screens for either HT-29 (Fig. 6a) or JIMT-1 (**Supplementary Fig. 6a**). The aspect ratio for the plot was set to match the relative variance explained by the first two principal components.

Consistency of Timepoint and Library Effects on Gene Scores

To evaluate library differences, we took all screens that had been duplicated in each library with all other conditions (timepoint, clone, and screen location) kept constant. For each of these screens, we subtracted the gene scores of the version performed with the KY library from the version performed with the Avana library to create library difference profiles. For the case of Sanger's day-14 KY screen of the Sanger HT-29 clone, two versions exist, the original and an alternative that was eventually grown out to 21 days. We used the alternate version of this screen to be consistent with the day 21 results. A correlation matrix of library difference profiles was then calculated and is plotted in the left of Fig. 6b. The procedure was repeated for timepoint differences, creating timepoint difference profiles by subtracting day 14 results from day 21 results for pairs of screen readouts that differed in timepoint but not library, clone, or screen location.

Mitigating Differences in Gene Scores by Matching Experimental Conditions

For the cell line HT-29, we took Sanger's original screen as a baseline. We then subtracted from this baseline from four Broad HT-29 screens: the original (Avana library at day 21), then with the Avana library at day 14, the KY library at day 21, and the KY library at day 14, generating four arrays indexed by gene which form the y-axes in the succession of plots in Fig. 6c. We also computed the mean score of each gene across all original Broad screens and subtracted it from the mean score of each gene across all the original Sanger screens to form the x-axis of all four plots. For each condition the standard deviation of the HT-29 screen differences (y-axes) was computed along with the correlation of the HT-29 screen differences with the mean differences (x-axis). The plots themselves are Gaussian kernel density estimates. We repeated this process for JIMT-1 (**Supplementary Fig. 6d**) and then for HT-29 while swapping the roles of Broad and Sanger (Fig. 6d). For the Sanger alternate condition screens we used the Sanger clone of HT-29, and for its day 14 KY screen we used the Sanger's original HT-29 screen.

Replication Experiments

The replication screens at Broad and Sanger were performed using the normal current protocol of the respective institution[20] except with respect to the specifically noted changes to library (and the associate primer sequences required for post-screen amplification of the sgRNA barcodes) and the timepoint.

Data Availability

The data used for this manuscript have been posted to Figshare[40].

Acknowledgements

We thank Scott Younger for his help with design and executions of Broad's replication experiments.

Bibliography

1. Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* **32**, 40–51 (2014).
2. Cook, D. *et al.* Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discov.* **13**, 419–431 (2014).
3. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
4. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
5. Seashore-Ludlow, B. *et al.* Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov.* **5**, 1210–1223 (2015).
6. Basu, A. *et al.* An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* **154**, 1151–1161 (2013).
7. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740–754 (2016).
8. Evers, B. *et al.* CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. *Nat. Biotechnol.* **34**, 631–633 (2016).
9. Morgens, D. W., Deans, R. M., Li, A. & Bassik, M. C. Systematic comparison of CRISPR/Cas9 and RNAi screens for essential genes. *Nat. Biotechnol.* **34**, 634–636 (2016).

10. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
11. Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M. D. C. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat. Biotechnol.* **32**, 267–273 (2014).
12. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).
13. Wang, T. *et al.* Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. *Cell* **168**, 890–903.e15 (2017).
14. Shi, J. *et al.* Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat. Biotechnol.* **33**, 661–667 (2015).
15. Tzelepis, K. *et al.* A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell Rep.* **17**, 1193–1205 (2016).
16. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**, 1515–1526 (2015).
17. Behan, F. M. *et al.* Prioritisation of oncology therapeutic targets using CRISPR-Cas9 screening. *Nature* [In Press] (2019).
18. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564–576.e16 (2017).
19. McDonald, E. R., 3rd *et al.* Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. *Cell* **170**, 577–592.e10 (2017).
20. Meyers, R. M. *et al.* Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).
21. Meyers, R. M., Bryan, J. G., McFarland, J. M. & Weir, B. A. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nature* (2017).
22. DepMap Achilles 19Q1 Public. (2019). doi:10.6084/m9.figshare.7655150.v1
23. DepMap, B. Cancer Dependency Map. *DepMap Portal* (2018). Available at:

https://depmap.org/portal/. (Accessed: 5th February 2019)

24. DepMap, S. Project Score https://score.depmap.sanger.ac.uk/, part of the Cancer Dependency Map at Sanger. *Sanger DepMap Portal* (2019). Available at: https://score.depmap.sanger.ac.uk/. (Accessed: 9th April 2019)
25. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
26. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**, 1515–1526 (2015).
27. Cancer Cell Line Encyclopedia Consortium & Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528**, 84–87 (2015).
28. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
29. Bushati, N., Smith, J., Briscoe, J. & Watkins, C. An intuitive graphical visualization technique for the interrogation of transcriptome data. *Nucleic Acids Res.* **39**, 7380–7389 (2011).
30. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184 (2016).
31. DepMap, B. DepMap Achilles 19Q1 Public. *Figshare* (2019). Available at: https://figshare.com/s/362d32844d53eb5753c5. (Accessed: 4th March 2019)
32. Ramsay, P. H. & Scott, D. W. Multivariate Density Estimation, Theory, Practice, and Visualization. *Technometrics* **35**, 451 (1993).
33. Ripley, B. D. Modern applied statistics with S. *Statistics and Computing, fourth ed. Springer, New York* (2002).
34. Azzalini, A. The R package sn: The skew-normal and related distributions, such as the skew-t (version 1.5). *URL http://azzalini. stat. unipd. it/SN* **15**, (2017).
35. Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaud. sci. nat.* **37**, 547–579 (1901).
36. Iorio, F. *et al.* Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. *BMC Genomics* **19**, 604 (2018).

37. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

38. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545 (2005).

39. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).

40. Agreement between two large pan-cancer CRISPR-Cas9 gene dependency datasets. (2019). doi:10.6084/m9.figshare.7970993.v1