Resource

Integrated cross-study datasets of genetic dependencies in cancer

Clare Pacini[1,2], Joshua M. Dempster[3], Emanuel Gonçalves[1], Hanna Najgebauer[1,2,4], Emre Karakoc[1,2], Dieudonne van der Meer[1], Andrew Barthorpe[1], Howard Lightfoot[1], Patricia Jaaks[1], James M. McFarland[3], Mathew J. Garnett[1,2], Aviad Tsherniak[3], Francesco Iorio[1,2,5,*]

[1] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK
[2] Open Targets, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK
[3] Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA
[4] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge CB10 1SA, UK
[5] Human Technopole, Via Cristina Belgioioso 147, 20157 Milano - Italy

* Corresponding author: francesco.iorio@sanger.ac.uk

Abstract

CRISPR-Cas9 viability screens are being increasingly performed at a genome-wide scale across large panels of cell lines to identify new therapeutic targets for precision cancer therapy. Integrating the datasets resulting from these studies is necessary to adequately represent the heterogeneity of human cancers, and to assemble a comprehensive map of cancer genetic vulnerabilities that might be exploited therapeutically. Here, we integrated the two largest independent CRISPR-Cas9 screens performed to date (at Broad and Sanger institutes), by assessing and selecting methods for correcting technology-specific biases and batch effects arising from differences in the underlying experimental protocols. Our integrated datasets recapitulate findings from the individual ones, provide larger statistical power allowing novel cancer- and subtype-specific analyses, unveil additional biomarkers of gene dependency, and improve the detection of common essential genes. Finally, we provide the largest integrated resources of CRISPR-Cas9 screens to date and the basis for harmonizing existing and future functional genetics datasets and assembling large cross-study cancer dependency maps.

Cancer is a complex disease that can arise from multiple different genetic alterations. The alternative mechanisms by which cancer can evolve result in a large amount of heterogeneity between patients, with the vast majority of them still not benefiting from approved targeted therapies[1]. In order to identify and prioritize new potential therapeutic targets for precision cancer therapy, analyses of cancer vulnerabilities at a genome-wide scale and across large panels of *in vitro* cancer models are being increasingly performed[2–11]. This has been facilitated by recent advances in genome editing technologies allowing unprecedented precision and scale via CRISPR-Cas9 screens. To date, two large pan-cancer CRISPR-Cas9 screens have been independently performed by the Broad and Sanger institutes[2,12]. The two institutes have also joined forces in a collaborative endeavor with the aim of assembling a joint comprehensive map of all the intracellular genetic dependencies and vulnerabilities of cancer: the *Cancer Dependency Map (DepMap)*[13,14].

Despite the two generated datasets containing so far data from over 300 cell lines each, detecting all cancer dependencies has been estimated to require the analysis of thousands of cancer models[3]. Consequently, the integration of these two datasets will be key for the DepMap and other projects aiming at systematically probing cancer dependencies. This will provide a more comprehensive representation of heterogeneous cancer types as well as an increased sample size to support the use of statistical/machine-learning methods to identify molecular features associated with differential gene dependencies. This will form the basis for the development of effective new therapies with associated biomarkers for patient stratification[15]. Furthermore, designing robust standards and computational protocols for the integration of these types of datasets will also mean that future releases of data from CRISPR-Cas9 screens can be integrated and analyzed together, paving the way to even larger cancer dependency resources.

We have previously shown that two large pan-cancer CRISPR-Cas9 datasets independently generated at Broad and Sanger institutes are consistent on the domain of 147 commonly screened cell lines[16]. This holds despite several extensive differences in the experimental pipelines underlying the two datasets, including distinct CRISPR-Cas9 sgRNA libraries. Here we extend our previous analysis by investigating the integrability of full releases of the Broad/Sanger gene dependency datasets. This yields the most comprehensive integrative cancer dependency resources encompassing dependency profiles of 16,827 genes across 786 different cell lines, spanning 26 tissues and 42 cancer types. We compare different state-of-the-art data processing methods to account for

heterogeneous single-guide RNAs (sgRNAs) on-target efficiency, and to correct for gene independent responses to CRISPR-cas9 targeting[12,17,18], reporting strengths and caveats for each of them (**Figure 1a, 1b and 1c**).

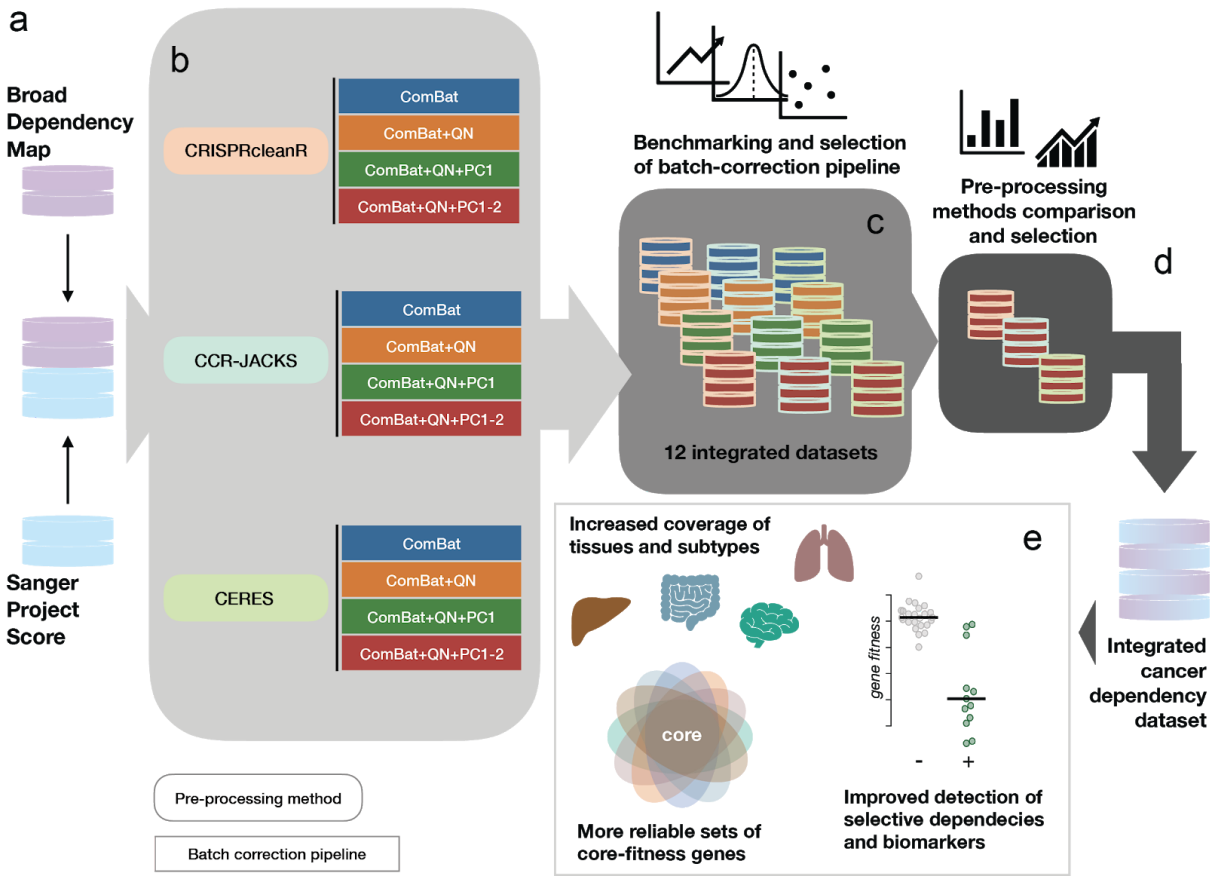

Figure 1: Schematic of the integration strategy. a. Broad and Sanger gene dependency datasets (raw count data of single-guide RNAs) are downloaded from respective web-portals. b. The datasets from each institute are pre-processed with three different methods, accounting for gene-independent responses to CRISPR-cas9 targeting (arising from copy number amplifications) and heterogeneous sgRNA efficiency, providing gene-level corrected depletion fold changes. Then, four different batch-correction pipelines are applied to the gene level fold changes across the two institute datasets for each of the pre-processing methods. c. Twelve different integrated datasets resulting from applying three different pre-processing methods (as indicated by the border colors) and four different batch-correction pipelines (as indicated by the fill colors) are benchmarked. d. Three different integrated datasets obtained by applying three pre-processing methods and the best batch-correction pipeline are benchmarked. e. Advantages provided by the final integrated datasets and conservation of analytical outcomes from the individual ones are investigated.

Furthermore, we show that our integration strategy accounts and corrects for technical biases whilst preserving gene dependency heterogeneity, and recapitulates established associations between molecular features and gene dependencies (**Figure 1d and 1e**). Finally, we highlight the benefits of the integrated dataset over the two individual ones in

terms of improved coverage of the genomic heterogeneity across different cancer types, gain of statistical power for biomarker identification, and increased reliability of unveiled sets of human core-fitness/common-essential genes (**Figure 1e**). Collectively, this study presents a robustly benchmarked framework to integrate independently generated CRISPR-Cas9 data-sets which provides the most comprehensive integrative resource to explore novel cancer dependencies and propose novel therapeutic targets.

Results

Overview of the integrated CRISPR-Cas9 screens

To establish the current scope of an integrated CRISPR-Cas9 dataset we compared numbers and tissue distributions of cell lines from the Project Score dataset (part of the Sanger DepMap)[19] and the 19Q3 release of the Project Achilles dataset (part of the Broad DepMap)[20]. Overall, 786 unique screened cell lines are accounted for in these two datasets (**Figure 2a**, **Supplementary Data 1**). Together these cell lines spanned 26 different tissues (**Figure 2b**) and for 65%, the number of cell lines covered increased when considering both datasets together. Similarly, the integrated dataset provided richer coverage of specific cancer types and clinically relevant subtypes (**Figure 2c**). These preliminary observations highlight the first benefit of combining these resources to increase statistical power for tissue-specific as well as pooled pan-cancer analyses.

Furthermore, between the two datasets there was an overlap of 156 cell lines screened by both institutes, encompassing 16 different tissue types (median = 9, min 1 for Soft Tissue and Kidney, max 25 for Lung, **Figure 2a and 2b**). This enables estimating batch effects due to differences in the experimental protocols underlying the two datasets[16], without biasing the correction toward cell line lineages specific to either institute.

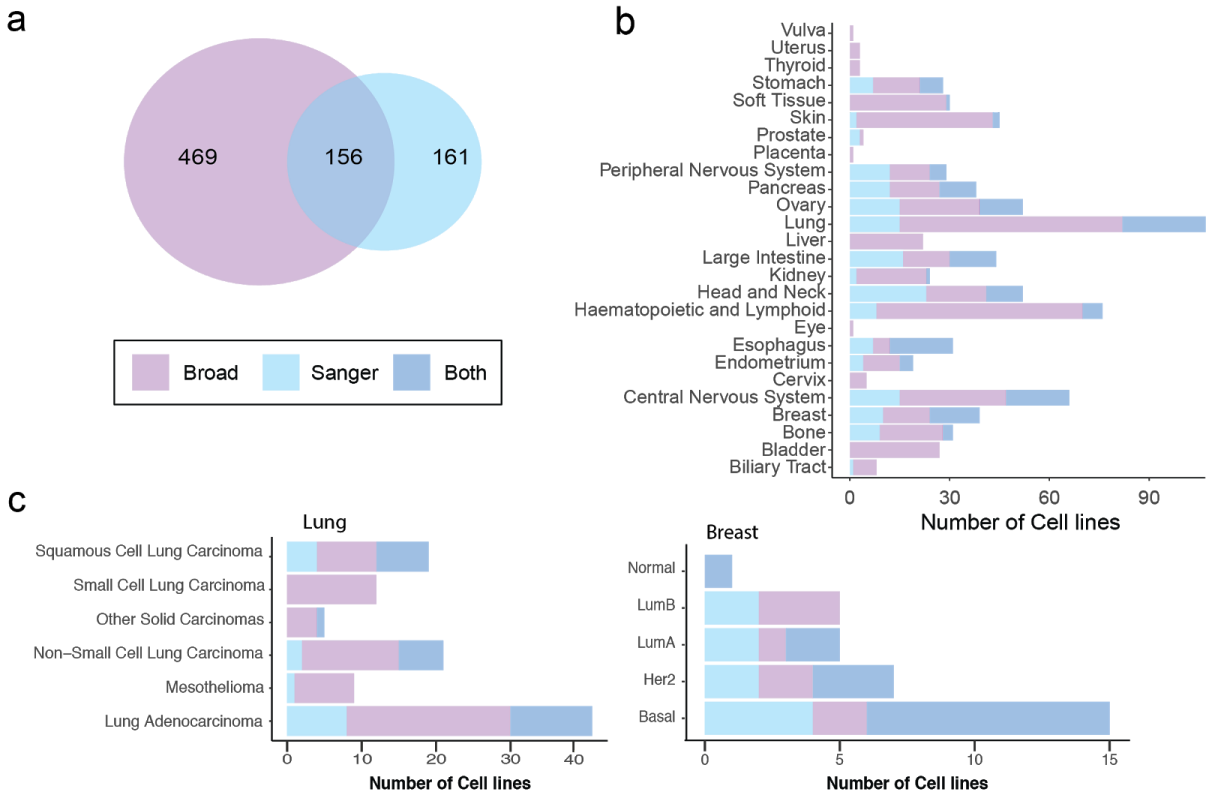

Figure 2. Overview of CRISPR-Cas9 screened cancer cell lines. a. Number of cell lines screened by the Broad and the Sanger institutes. b. Overview of the number of cell lines screened for each tissue type across the two datasets. c. Number of screened Lung cancer and Breast cancer cell lines split according to cancer types and PAM50 subtypes, respectively, across the two datasets.

Data Pre-processing

CRISPR screens have known nonspecific cutting toxicity that increases with copy number amplifications (CNAs)[21,22]. Another factor biasing the data derived from this type of screens is the heterogeneous levels of on-target efficiency across sgRNAs targeting the same gene in the employed libraries[23]. Multiple methods exist to correct for these biases. Here, we evaluate three: CRISPRcleanR, an unsupervised nonparametric CNA effect correction method for individual genome-wide screens[17]; a method resulting from using CRISPRcleanR with JACKS, a Bayesian method for inferring guide on target efficacy[18] (CCR-JACKS) by a joint analysis of multiple screens; and CERES, an algorithm that simultaneously corrects for CNA effects and accounts for differences in guide efficacy[12], also analyzing multiple screens jointly.

Batch effect correction

In addition to biases within datasets caused by copy number alterations and variable guide efficiencies, technical differences in screening protocols, reagents and experimental settings can cause batch effects between datasets. These might arise from factors that vary within institute screens (for example, differences in control batches and Cas9 activity levels) as well as between institutes (such as differences in assay lengths and employed sgRNA libraries). When focusing on the set of cell lines screened at both institutes, a Principal Component Analysis (PCA) of the cell line dependency profiles across genes (DPGs) in the collated dataset highlights a clear batch effect determined by the screening of origin, irrespective of the pre-processing method, consistent with previous results (**Figure 3a**)[16].

We quantile-normalized each cell line DPG to adjust for differences in screen quality then correct batch effects using ComBat[24]. Following this correction, the combined datasets showed reduced yet persistent residual non-linear batch effects clearly visible along the two first principal components (**Supplementary Figure 1**). Based on this, we considered four different batch correction pipelines and evaluated their use in our integrative strategy. In the first pipeline, we processed the combined Broad/Sanger DPG dataset using ComBat alone (ComBat); In the second, we applied a quantile normalization following ComBat correction (ComBat+QN) to account for different phenotype intensities across experiments, resulting in different ranges of gene dependency effects; In the third and fourth pipelines we also removed the first or first two principal components respectively (ComBat+QN+PC1) and (ComBat+QN+PC1-2).

To assess the performance of different batch correction pipelines we first calculated a weighted Pearson's (wPearson) correlation (Methods) distance score between all possible pairs of 312 cell line DPGs, i.e. from 156 cell lines screened in both studies, similarly to our previous work[16]. Secondly, we fixed each cell line DPG in turn and ranked all others based on how similar they were to the fixed one in decreasing order, according to the wPearson scores. For each rank position *k* we determined the number of cell line DPGs from one of the two original datasets (over the total) that had a DPG derived from the same cell line (a matching DPG) from the other dataset in its *k* most similar DPGs, i.e. its *k-neighborhood* (**Figure 3b**). Among the tested pre-processing methods, the best performances were obtained when removing the first two principal components following ComBat and quantile normalization, i.e. ComBat+QN+PC1-2. Across pre-processing methods, the best performances were found for CERES (88.5%) followed by CRISPRcleanR (78.2%) and CCR-JACKS (63.5%) (**Figure 3b**).

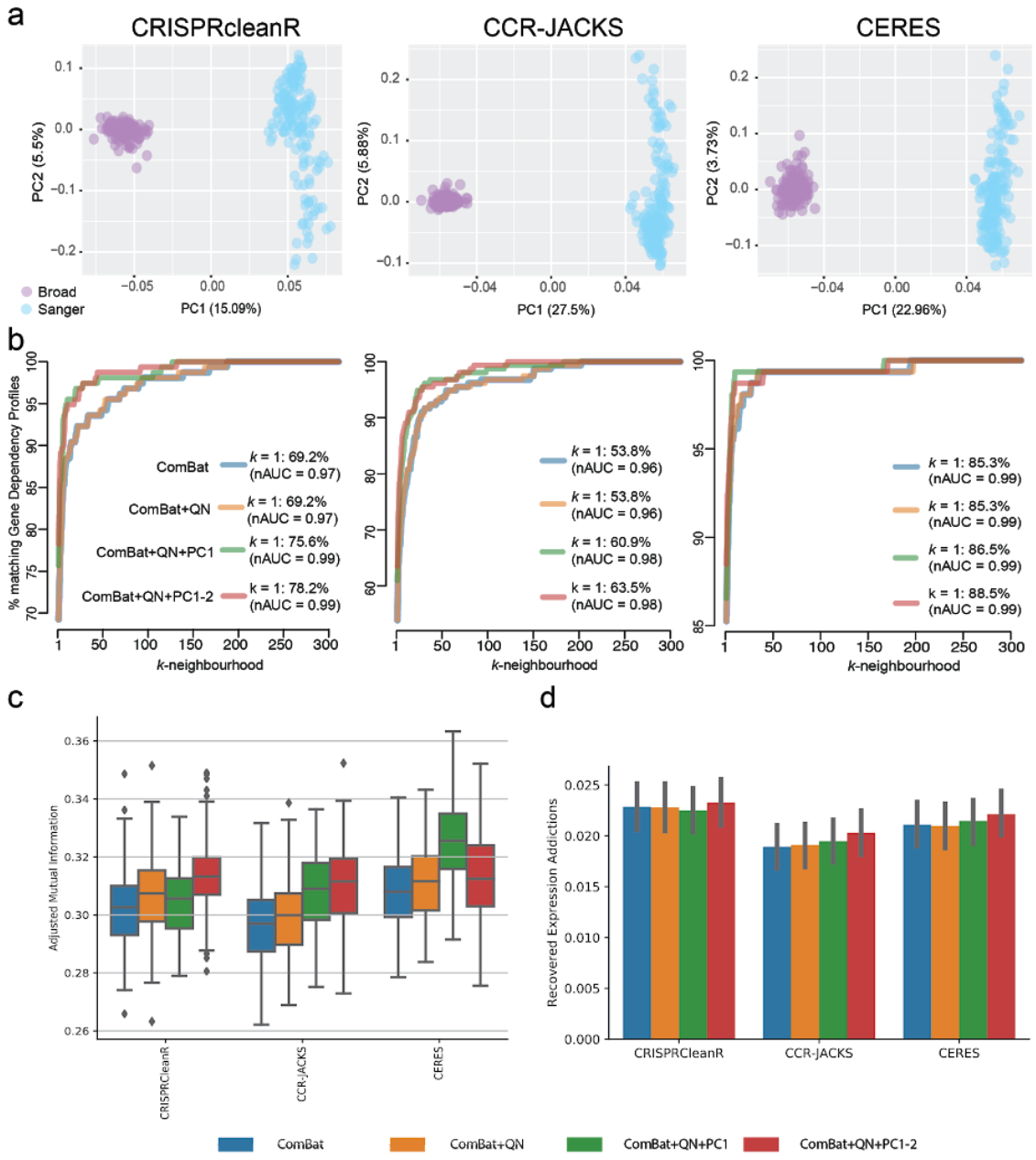

Figure 3: Batch effect assessment and correction. a. Principal component plots of the dependency profile across genes (DPGs) for cell lines screened in both Broad and Sanger studies and pre-processing methods. b. Percentages of DPGs in the Broad/Sanger combined dataset that are derived from one studies by screening a given cell line and have among their *k* most correlated DPGs, i.e. the *k-neighborhood*, a DPG derived from the other study by screening the same cell line. Results are shown across different pre-processing methods (in different plots) and different batch correction pipelines (as indicated by the different colors). Correlations between DPGs are computed using a weighted Pearson correlation metric, in which the contribution of each gene is proportional to the average selectivity of its dependency signal across cell lines. c. Agreement between cell line clusters based on DPGs correlation and tissue lineage labels of corresponding cell lines, across pre-processing methods and batch-correction pipelines. d. Fraction of gene expression addictions identified (across cell lines) for each of the 12 different versions of the Broad/Sanger integrated datasets.

Evaluation of data integration methods extended to 786 cell lines

To further assess the batch correction pipelines we considered their performances extended to a fully integrated dataset including also DPGs of cell lines exclusive to each study. To that end, we generated integrated batch corrected datasets across all 786 cell lines using each of the three pre-processing methods and four different batch correction pipelines as outlined in the previous section. Next we applied the batch correction vectors (Methods) estimated by ComBat on the DPGs of the 156 overlapping cell lines to the whole set cell lines, including those exclusive to each of the two starting individual datasets.

To evaluate batch effects corrections pipelines we considered different criteria. One of this is estimating the extent of conserved similarity between screens of cell lines derived from the same lineage. We evaluated this by comparing unsupervised clusterings of the batch corrected cell line DPGs to the lineage labels of the cell lines. To this aim, we performed one hundred *k*-means clusterings of each of the 12 datasets, with *k* equal to the number of tissue lineages screened in at least one study. We then calculated the adjusted mutual information (AMI, Methods) between each DPG clustering and corresponding cell line lineage labels. We observed higher than chance AMI between the obtained *k* clusters and the tissue lineages of the cell line DPGs, regardless of the starting batch corrected dataset (largest single-sample *t*-test *p*-value of $3.36\ 10^{-131}$, $N = 100$, **Figure 3c**).

Under each pre-processing method the removal of one or two principal components resulted in an increased AMI between cell line DPGs clusters and tissue lineages. For Broad-exclusive screens, we additionally tested the extent of AMI between cell line DPGs clusters and tissue lineages before/after integration, considering each integration method. All

methods yielded substantial improvements over the original, uncorrected data, with batch-correcting increasing AMI between 0.0757 (ComBat+QN) and 0.165 (ComBat+QN+PC1-2); the largest two-sample *t*-test *p*-value was $2.23\ 10^{-88}$, $N = 200$.

Further, we evaluated to what extent our data integration strategies retain the biological signals present in the (uncorrected) individual datasets. For a given cell line in each of the Broad and Sanger individual datasets, we identified significant dependency genes based on a fitness effect threshold corresponding to a false discovery rate (FDR) of 5% (Methods) pre/post-batch-correction We observed strong agreement between sets of significant dependency genes overall cell lines, pre-processed methods and batch correction pipelines, with median agreement above 80% in all cases (**Supplementary Figure 2a**).

Next, we evaluated gene dependency false-positive rates in integrated datasets. For each cell line DPG, we defined a set of putative negative controls composed of genes not expressed at the basal level in that cell line (Methods). False positives rates were calculated as the sum of negative controls called as significant dependencies (in the top 15% most depleted genes) over their total number across DPG. There was little difference in false-positive rates across the four different batch correction pipelines, with a slight improvement when two principal components were removed (**Supplementary Figure 2b**). Similar results were observed for true positive rates, when calculating the Recall of a reference set of common essential genes[11] within the top 15% most depleted genes for each cell line (**Supplementary Figure 2c**).

In contrast to common essential genes, interesting potential novel therapeutic targets are genes that show a pattern of selective dependency, i.e. those exerting a strong reduction of viability upon CRISPR-Cas9 targeting but on a subset of cell lines. Furthermore, these selective dependencies are often associated with molecular features that might explain their dependency profiles (biomarkers). We first tested the ability of each dataset to recall gene expression based addictions – cases where the dependency signal of a gene is correlated with the basal expression of that gene, across cell lines (Methods). Genome-wide, ComBat+QN+PC1-2 recovered the greatest number of significant gene expression addictions across all preprocessing methods (**Figure 3d**). Finally we tested the ability of each dataset to conserve expected dependency relations, between paralogs, gene pairs coding for interacting proteins, or members of the same complex using gene pairs annotation from publicly available databases[25, 26, 27] (Methods). Again the

ComBat+QN+PC1-2 batch correction pipeline recovered the greatest number of expected gene dependency relations (**Supplementary Figure 2d**).

Taken together these results show that over the set of benchmarked batch correction pipelines, using ComBat with quantile normalization and removing the first one or two principal components outperformed any other batch correction pipelines, across all three tested pre-processing methods (CRISPRcleanR, CCR-JACKS and CERES). In addition, removing two principal components (ComBat+QN+PC1-2) showed the best ability to recover gene expression based vulnerabilities and other expected gene-dependency relationships. In conclusion, ComBat+QN+PC1-2 appeared to be the most robust batch correction pipeline, which we selected for follow-up analyses.

Comparison of pre-processing methods

In contrast to the batch correction pipeline, determining the best performing pre-processing method across batch-correction-pipelines was not equally immediate. Therefore, we performed further comparative analysis to assess the performance of the three pre-processing methods (CERES, CRISPRcleanR and CCR-JACKS) once the best batch correction pipeline was determined (ComBat+QN+PC1-2). CRISPRcleanR is an unsupervised method that processes each screen individually. In contrast, JACKS and CERES borrow information across screens to different degrees; CERES assumes shared profiles of gene essentiality across screens, whilst JACKS makes a weaker assumption of shared guide efficiency across profiles.

Despite differences in underlying algorithms, we expect most biological signals to be invariant to the pre-processing method used. Accordingly, we found significant correlation between median gene dependency scores across cell lines and method pairs (Spearman's correlation = 0.80 for CERES vs. CCR-JACKS, 0.96 for CRISPRcleanR vs. CCR-JACKS, 0.83 for CERES vs. CRISPRcleanR, p-value < 2.2e-16 in all cases, n=16,827 **Supplementary Figure 3a**).

Comparisons of binary dependency matrices (Methods) from the integrated datasets also revealed a high overlap of dependencies found between each processing method, with more than 60% of significant dependencies found across all the methods and 75.4% found in at least two methods (**Supplementary Figure 3b**). The largest difference was found

between CRISPRcleanR and CERES (17.1% of dependencies were exclusive to one method). However, even in this case, 90% of dependencies found by CRISPRcleanR were also found by CERES, and 75% of dependencies found by CERES were identified by CRISPRcleanR (**Supplementary Figure 3b**).

Pharmacological and CRISPR screens are being increasingly integrated as a means to elucidate drug mechanism-of-action[28]. Here we used drug sensitivity analysis to compare the Recall of significant associations between a drug sensitivity profile and its nominal targets' CRISPR dependency profile, across cell lines. Thus, we systematically calculated Spearman's correlation between the dependency profile of a drug target across different pre-processing methods with the profile of cell line sensitivity to that drug. As indicators of drug sensitivity, we used viability reduction indicators (IC50 values) from two studies[29,30] for 307 unique compounds mapped to their targets[28], across 486 cell lines included in our integrated datasets. Next we compared the resulting drug-response/target-dependency correlation patterns across pre-processing methods. We found a significant and large effect-size correlation between these patterns across each pair of pre-processing methods, and a significantly large overlap between associations identified by each pre-processing method (Fisher's exact test largest p-value = 1.11e-60). These results confirm that the vast majority of associations between drug-response and target-gene essentiality are invariantly identified across pre-processing methods (**Figure 4a**) (**Supplementary Data 2**).

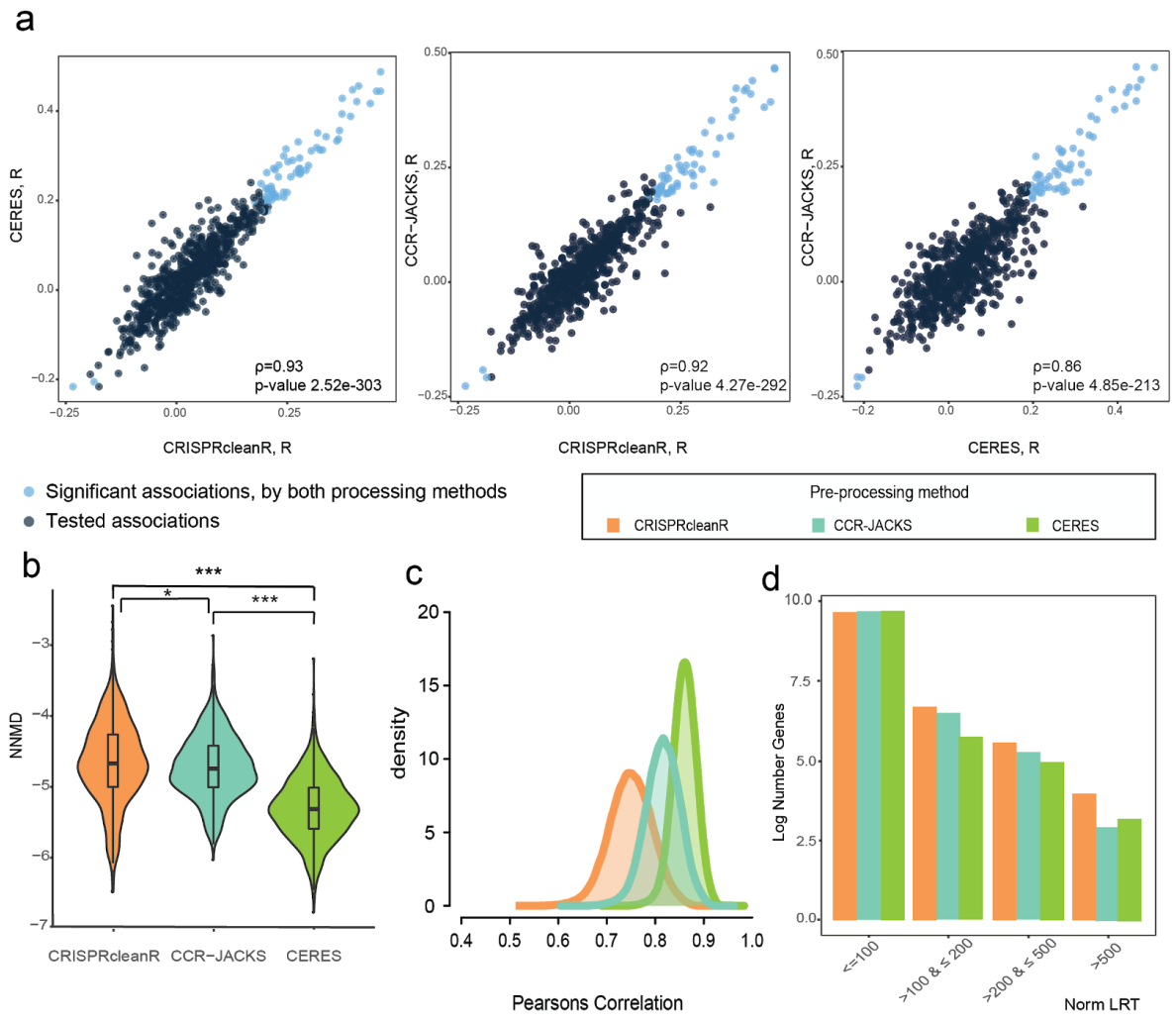

Figure 4: Comparison of pre-processing methods CERES, CCR-JACKS and CRISPRcleanR. a. Spearman correlation (R) scores between drug IC50 values and drug's target dependency profiles across cell lines compared between different pairs of pre-processing methods. Significant scores (at 5% FDR) identified under both pre-processing methods are shown in light blue. b. Null-normalized mean difference (NNMD, a measure of separation between dependency scores of prior-known essential and non-essentials genes): defined as the difference in means between dependency scores of essential and non-essential genes divided by standard deviation of dependency scores of the non-essential genes. Lower values of NNMD indicate better separation of essential genes and non-essential genes. Significant differences in NNMD between pre-processing methods are shown with Wilcox test p-values < 0.05 (*) and < 1e-5 (***). c. Distribution of Pearson's correlations scores between all possible pairs of cell line dependency profiles across genes, for each pre-processing method. d. Log distribution of significant fitness genes stratified by their normLRT scores, under each pre-processing method.

Although we found good agreement across pre-processing methods we also investigated possible sources of discrepancies between the three integrated datasets. An aspect to consider is that as CERES and JACKS both borrow information across screens they are expected to better identify consistent signals across cell line DPGs (i.e. for essential

and non-essential genes), especially for DPGs derived from lower quality experiments, or reporting weaker depletion phenotypes[18,22].

We evaluated the separation of prior known essentials and non-essentials genes[11] based on their dependency score (DS) in individual DPGs, across pre-processing methods. We calculated, for each cell line and under each pre-processing method, a null-normalized mean differences (NNMDs)[31] defined as the difference between the mean DS of essential genes and that of non-essential genes divided by the standard deviation of the DSs of the non-essential genes. Results showed that CERES (median NNMD=-5.92) yielded a greater separation of prior known essentials and non-essentials genes compared to CCR-JACKS (median NNMD=-5.24) and CRISPRcleanR (median NNMD=-5.15, **Figure 4b**).

Whilst joint analysis of multiple screens improves true positive rates, this may reduce the heterogeneity of the DPGs. To investigate this, we computed Pearson's correlation scores between all pairs of cell line DPGs for each of the integrated datasets. We observed generally larger background correlations for CERES and CCR-JACKS compared to CRISPRcleanR (**Figure 4c,** Kolmogorov-Smirnov test p-values below machine precision), concordant with these two methods making DPGs more similar to each other by borrowing information across screens.

A ubiquitous increase in correlation between DPGs could result in a loss of selective gene dependencies. As a measure of selective dependency of a gene we calculated the likelihood ratio that its dependency scores across cell lines have a skewed-t over a normal distribution (normLRT)[4], with larger normLRT values indicative of selective dependency genes. Consistent with our hypothesis, we observed a greater number of genes with large normLRT values for CRISPRcleanR compared to CERES or CCR-JACKS (p-values = 1.95e-52 for CRISPRcleanR versus CERES, 7.13e-29 for CCR-JACKS versus CERES, for 1.85e-5 CRISPRcleanR versus CCR-JACKS proportions test, **Figure 4d**).

We investigated whether a reduction in DPG heterogeneity and context-specific dependencies translated into a smaller number of potential dependency-biomarkers. This analysis was performed by contrasting cell lines based on the status of each of 884 clinically relevant cancer functional events (CFEs[32]) encompassing mutations in cancer driver genes, amplifications/deletions of chromosomal segments recurrently altered in cancer, hypermethylated gene promoters and microsatellite instability status. For each CFE, we performed a Student's t-test for each selective gene dependency (SGD, Methods) across

two groups of cell lines based on the status of the considered feature (present/absent). This resulted in a total number of 366,860 tests.

To evaluate how similar the outcomes were from this systematic analysis, we assembled a vector of dependency-delta values (difference in means across the two groups) for each pre-processing method and calculated Spearman's correlation scores between each pair of vectors/methods. Interestingly, the largest correlation was found between CRISPRcleanR and CERES (0.81), followed by that between CRISPRcleanR and CCR-JACKS (0.77) then between CCR-JACKS and CERES (0.65, all p-values <2.2e-16, n=344,418). Although these correlations were all significant, the number of significant associations varied between pre-processing methods. At 5% FDR significance threshold CRISPRcleanR identified 1,195 differential dependencies, compared to 1,089 for CERES and 932 when using CCR-JACKS (p-values proportions test, 0.028 CRISPRcleanR versus CERES, 1.28e-08 CRISPRcleanR versus CCR-JACKS and 5.1e-04 CERES versus CCR-JACKS) (**Supplementary Data 3**).

An integrated dataset improving the detection of biomarkers, common essential genes and overall genetic dependencies

We set out to estimate the advantage of a larger integrated dataset of genetic dependencies. First, we considered the increase in coverage of gene dependencies in the integrated dataset. To evaluate this, we quantified the number of significant dependencies (at 5% FDR) found in a fixed number of cell lines *n* (with *n* = 1, 3, 5 or n ≥ 10) of the integrated datasets, as well as in the individual original Broad and Sanger datasets. This number was always larger for the integrated datasets with respect to the individual ones, regardless of the pre-processing method used, with the largest number provided by CRISPRcleanR (**Figure 5a**) and very similar results provided by CERES and CCR-JACKS (**Supplementary Figure 4**).

Second we evaluated the reliability of the integrated datasets to predict common essential genes using two methods: the 90th-percentile method[16] and the Adaptive Daisy Model (ADaM)[2]. We created a consensus list of predicted common essentials combining results from four lists of common essentials outputted by these two methods (ADaM and 90th percentile) and the two integrated datasets processed using either CERES or CRISPRcleanR (selected as they consistently provided the most orthogonal across all pre-processing methods' comparisons). The majority of genes called common essentials according to at least one of these lists was found consistently in all four (1,055 out of 1,973, **Figure 5b**). We assigned to each of the 1,973 common essential genes a tier based on the amount of supporting evidence of their common essentiality. Tier 1, the highest confidence set comprised the 1,055 genes found in all four gene lists. Tier 2 had 499 genes found in at least two lists. Finally, Tier 3 was assigned to the lowest confidence group and contained 419 genes that were found only in one list (**Supplementary Data 4**).

We compared the common essential genes identified using the integrated datasets to two existing sets of common essential genes from recent publications: Behan[2] and Hart[11]. We considered genes in Tier 1 only (Integrated Tier1), in Tier 1 or Tier 2 (Integrated Tier2) and in Tier 1, or 2 or 3 (Integrated Tier3). For each gene set we calculated the Precision and Recall rates of known core-fitness gene sets, such as Ribosomal protein genes, genes involved in DNA replication and Spliceosome. All Integrated Tiers of common essentials showed greater Recall of known core-fitness gene sets compared to Behan and Hart. To estimate a false discovery rate (FDR) we calculated the proportion of genes in each common essential gene set that was likely to be context-specific essential, according to the likelihood

of its dependency signal across cell lines (calculated from an orthogonal RNAi dataset) being drawn from skewed t-distribution[4]. The increase in the number of genes in the common essential sets resulted in a small decrease in precision particularly for Tier 2 and 3. However, FDR was lower than that of Behan and Hart across all Integrated Tiers (**Figure 5c**).

We next asked whether the integrated datasets unveiled additional significant gene dependencies and CFE/gene-dependency statistical interactions compared to either one of the Broad or Sanger (individual) datasets. We performed a systematic identification of potential biomarkers of gene dependency and found that 884 CFEs could be tested, i.e. they were present in sufficiently large numbers of cell lines, in comparison to 754 for the Sanger dataset and 805 for the Broad dataset. Thus highlighting the benefit of an integrated dataset, with greater diversity of molecular profiles represented compared to each of the individual datasets.

We compared the number of identified biomarkers (at 1% FDR) when testing the CRISPRcleanR integrated dataset with respect to those identified when analyzing the individual (Sanger/Broad) datasets. Analyzing the Sanger dataset alone unveiled 420 significant CFE/gene-fitness associations, 465 were identified while testing the Broad dataset alone, whereas the integrated dataset allowed finding 1,017 significant CFE/gene-fitness associations. Notably, 138 associations that were significant (at 1% FDR) when analyzing both the individual uncorrected datasets were all retained in the integrated dataset. For these 138 associations the effect sizes were significantly larger in the integrated datasets compared to either of the individual datasets (**Figure 5d**) (paired Wilcox test, p-value 3.36e-13 for Sanger and 3.183e-06 for Broad). Taking the union of the results obtained from the individual Broad and Sanger datasets, 747 associations were significant at 1% FDR, of these 562 (75.2%) were also identified in the integrated dataset. Finally, we identified 455 associations that were significant (1% FDR) in the integrated dataset only.

Performing this systematic statistical inference on cell lines from individual tissue lineages unveiled 32 additional significant associations in the integrated dataset (when considering only CFE/gene-dependency pairs testable also in the individual datasets) with respect to those unveiled by the analysis of the Sanger datasets alone, and 48 with respect to the Broad dataset (**Supplementary Data 5**). Examples included decreased dependency on MDM4 in TP53 mutant Lung cell lines for the Sanger dataset, and increased dependency

on ERBB2 in ERBB2 amplified Breast cancer cell lines for the Broad dataset (**Figure 5e**). Furthermore, 13 tissue-specific significant associations identified in the integrated dataset were tested but not found significant in either the Broad or the Sanger dataset. Two examples of these associations are reported in **Figure 5f**.

Finally, we assessed whether we were able to identify a greater number of gene addiction relationships in the integrated dataset compared to the individual ones. For each tissue, we calculated the Pearson correlation between a gene's expression and dependency profile across cell lines. We called significant associations those with an FDR less than 25%. Results showed an increase in the number of associations identified in the integrated dataset when compared to the individual datasets across multiple tissue types (**Figure 5g**).

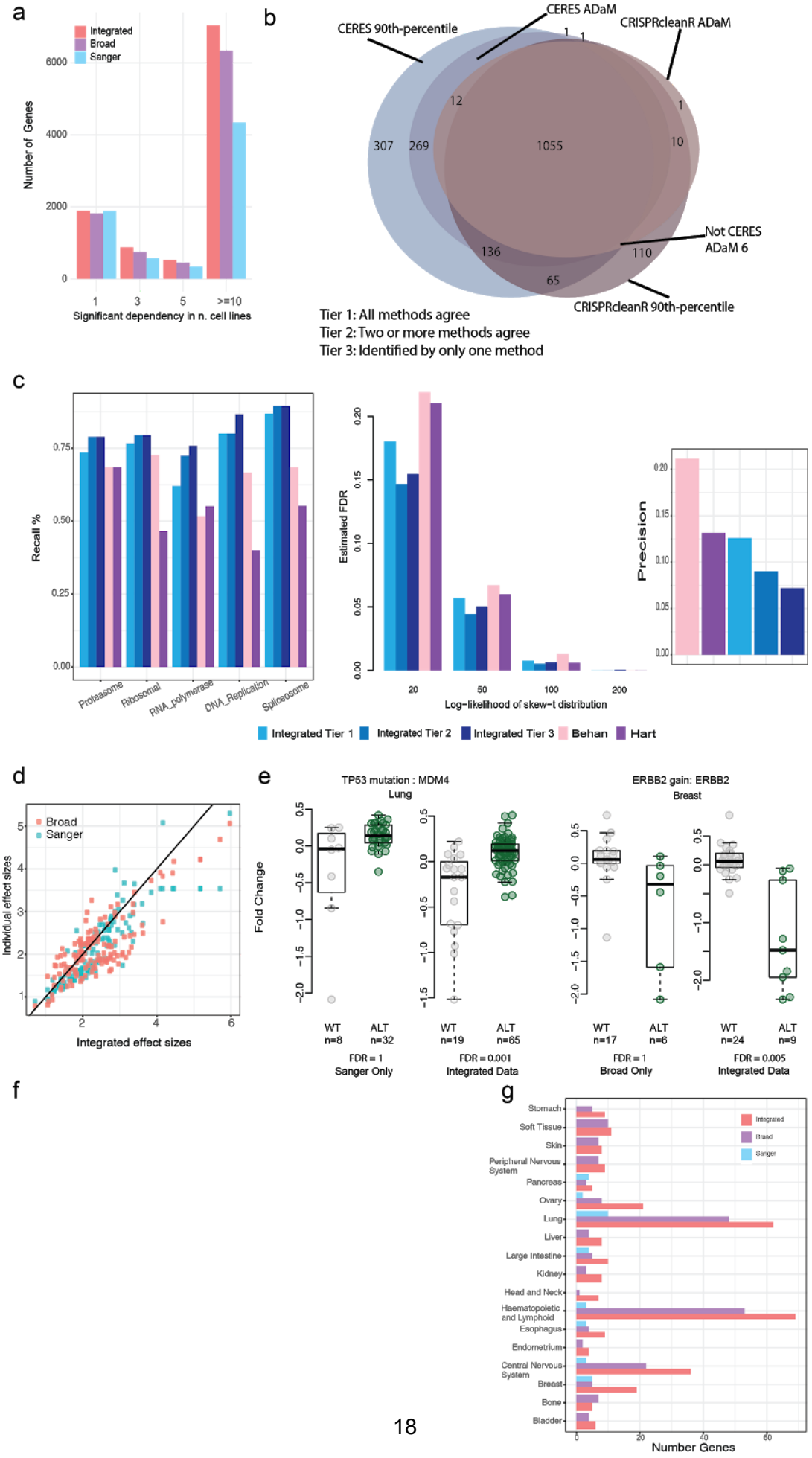

Figure 5: Advantages of an integrated dataset. a. Number of significant gene dependencies in fixed number of cell lines across individual datasets and the integrated one, using CRISPRcleanR as the pre-processing method. b. Venn diagram showing genes called common essential when using two different detection algorithms (ADaM and 90th-percentile) applied to integrated datasets resulting from two different preprocessing methods. c. Recall of essential genes sets for the integrated dataset, across different tiers, compared to two previously published gene sets (Behan and Hart). False discovery rates considered as negative hits genes that are putative context-specific essential according to a normLRT metric from an independent RNAi dataset, across levels of reliability. Precision values for each dataset using reference sets of essential genes as true hits. d. Scatter plot of the effect sizes for the 138 associations found in both individual datasets and the integrated dataset. The x-axis shows the effect sizes for the integrated dataset. The effect sizes for the individual datasets are shown on the y-axis. The black line is at intercept zero and slope 1. Most associations between dependencies and biomarkers have greater effect sizes in the integrated dataset compared to the individual datasets. e. Examples of significant associations between genes and features, found in the integrated dataset compared to the individual dataset. f. Examples of significant associations found in the integrated dataset that were not significant in either of the individual datasets. g The number of gene expression-based addiction associations that were significant at FDR 25% in the integrated dataset compared to each individual dataset.

Discussion

The integration of data from different high-throughput functional genomics screens is becoming increasingly important in oncology research to adequately represent the diversity of human cancers. Harmonizing CRISPR-Cas9 screens performed independently and/or using distinct experimental protocols, requires correction and benchmarking strategies to account for technical biases, batch effects and differences in data-processing methods. Here, we proposed a strategy for the integration of CRISPR-Cas9 screens and evaluated methods accounting for biases within and between two dependency datasets generated at the Broad and Sanger institutes.

Our results show that established batch correction methods can be used to adjust for linear and non-linear study-specific biases making integration possible. Following integration, dependency profiles of the same cell line screened by different institutes show strong concordance, as do cell lines from the same tissue lineage. Significant gene dependencies and associations with biomarkers unveiled by statistical analyses of the two individual datasets are retained and enriched by novel hits.

Our analyses and assessment yielded two final integrated datasets of cancer dependencies across 786 cell lines (with associated quality metrics). In contrast to existing

databases of CRISPR-Cas9 screens[33,34], our integrated datasets are corrected for batch effects allowing for their joint analysis. In addition, our integrated datasets cover a greater number of genetic dependencies, and the increased diversity of screened models allows additional associations between biomarkers and dependencies to be tested and called as significant.

The integrated datasets were outputted by two orthogonal pre-processing methods (CRISPRcleanR and CERES) for correcting gene independent responses to CRISPR-Cas9 targeting arising from gene copy number amplifications. We found these pre-processing methods highly concordant under different aspects including significant gene dependencies, dependency scores and their relationship with drug response. Further, evaluation of possible discrepancies between these methods showed that CERES (which borrows information across screens) yields a final dataset better able to identify prior known essential and non-essential genes based on their dependency scores, whilst CRISPRcleanR (a per sample method) retains more of the heterogeneity in the data.

Therefore, using results from both pipelines may provide the best overall data-driven functional genetic landscape underlying the construction of a comprehensive Cancer Dependency Map. In particular, using results from both pipelines resulted in a reliable resource of human core-fitness genes, which were detected with increased reliability and with associated levels of supporting evidence.

Further, the data integration strategies outlined here can be used with future and additional CRISPR-cas9 datasets to increase coverage and understanding of cancer dependencies. This will be of paramount importance for oncological functional genomics, for the identification of novel cancer therapeutic targets, and for the definition of a global cancer dependency map.

Data availability

The final integrated datasets are available for download at https://depmap.org/broad-sanger/ and https://score.depmap.sanger.ac.uk/downloads (submission on FigShare in progress).

Code availability

Scripts and software packages implementing the integration pipeline described in this manuscript and needed to reproduce results and figures are available at https://drive.google.com/drive/folders/1MyVwzK2kehu3gHbKGMlG8o2bqyMFxGhc?usp=sharing (deposition on GitHub and FigShare in progress).

Acknowledgments

This work was partially funded by Open Targets project grant OTAR0255.

Author Contributions

CP conceived the study, designed, implemented and performed analyses, assembled figures, curated data, wrote the manuscript. JMD conceived the study, designed, implemented and performed analyses, assembled figures, and contributed to manuscript writing. EG performed analyses, assembled figures, revised the manuscript. HN assembled figures, revised the manuscript. EK, DvdM, AB, HL, PJ contributed to data curation. JMM, MJG, and AT revised the manuscript and contributed to study supervision. FI conceived the study, designed analyses, contributed to figure production, wrote the manuscript, acquired funds and supervised the study.

Competing interests

MJG, and FI receive funding from Open Targets, a public-private initiative involving academia and industry. MJG receives funding from AstraZeneca and performs consultancy for Sanofi. FI performs consultancy for the joint CRUK - AstraZeneca Functional Genomics Centre. AT is a consultant for Tango Therapeutics and Cedilla Therapeutics. JMD, JM and AT receive funding from the Cancer Dependency Map Consortium, but no consortium member was involved in or influenced this study. All the other authors declare no competing interests.

Methods

Preprocessing data

Sanger data processed with CRISPRcleanR were obtained from the Score website (https://score.depmap.sanger.ac.uk/). The CRISPRcleanR corrected counts were used as input into JACKS, for the CCR-JACKS processing method.

Raw counts and the copy number profiles for the Sanger dataset downloaded were processed with CERES[35]. The Broad data processed with CERES (unscaled gene effect) version 19Q3 scores were downloaded from the Broad DepMap portal[35]. The raw counts for Broad data 19Q3 were processed with CRISPRcleanR and the CRISPRcleanR corrected counts processed with JACKS. Gene names were matched across the Broad and Sanger datasets by updating both to the current version of HUGO gene symbols from the HGNC website. Genes with no NAs for any cell line under any processing method were retained for analysis. The 786 cell lines processed by both CERES and CRISPRcleanR were used for analysis. Tissue annotations for each cell line were obtained from the Cell Model Passports (https://cellmodelpassports.sanger.ac.uk/)[36].

Batch correction pipelines

The dependency profiles across genes (DPGs) overlapping cell lines from each institute were first quantile normalized using the preprocessCore package in R[37]. Screen quality adjustments were made by fitting a spline to the average gene fold change across cell line DPGs. Each DPGs was then adjusted to remove the difference between the fitted spline and the diagonal. The overlapping cell lines were then batch corrected using three different methods. A standard least squares model was fitted in R. The ComBat correction was performed using the sva package in R[38].

Batch correction pipelines' assessment and weighted Pearson correlation metric

Cell lines' rank neighborhoods were based on a weighted Pearson correlation metric. The weights were defined as the absolute mean (over the Broad and Sanger datasets) of a gene dependency signal skewness across the 156 overlapping cell lines for the Broad and Sanger datasets. This upweights genes with a variable and sufficiently selective fitness profile whilst downweighting those that show weak/no-signal or unselective dependencies.Then for each query DPG we ranked all the others based on how similar they were to the fixed one in decreasing order, according to the wPearson scores. For each position k in the resulting rank we then defined a *k-neighborhood* of the query DPG composed of all the other DPGs whose rank position was $\leq k$. Finally we determined the number of cell line DPGs that had

the DPG derived from screening the same cell line in the other dataset (a matching DPG) in its *k-neighborhood*. The final rank for each cell line was defined based on the minimum rank obtained for each cell line when considering the DPG for that cell line from the Broad data compared to all DPGs, and similarly the DPG for the cell line in the Sanger dataset compared to all DPGs.

Batch correction extended to 786 cell lines

The ComBat estimates, pooled mean, variance and empirical Bayes adjustments (mean and standard deviation) for each batch based on the analysis of 156 cell lines common to both initial dataset were computed. The ComBat correction using these estimates was then applied to all 982 screens, i.e. the union of the two initial datasets. Particularly, each individual cell line DPG was shifted and scaled gene-wise using the batch correction vectors outputted by ComBat.

Additional adjustments were then applied to the 982 screens including quantile normalization, and the removal of either the 1st principal component of the joint datasets or the first twos. Finally, DPGs for overlapping cell lines passing a similarity threshold (detailed below) were averaged. Across the three pre-processing methods the number of cell lines that matched their counterparts exactly after ComBat correction ranged from 54% - 85%. Suggesting that under all pre-processing methods there remained cell lines whose DPGs diverged between studies. For each of the cell lines that matched their counterpart as the first neighbor we considered their distances (1-wPearson) as a measure of the variability in distance profiles between DPGs of the same cell line across institutes. We called divergent DPGs those with a distance greater than the 95 percentile of distances from matching cell lines. For 19 cell lines with divergent DPGs across all three processing methods we selected the DPG from the screen with the highest quality to be included in the integrated datasets. As a quality metric we used the Null-normalized mean difference (NNMD, defined in the main text) and took its consensual value across the three datasets (resulting from applying CERES, CCR-JACKS and CRISPRcleanR).

Agreement between dependency profile clusterings and cell line tissue labels

We selected 500 genes with the highest variance in the CERES ComBat integrated dataset and performed repeated 100 k-means clusterings cell lines using the high variance genes for each pre-processing and batch-correction method. For each clustering, we calculated the adjusted mutual information between the obtained clusters and the cell line tissue labels as

specified in the annotation provided by the sample_info file of the DepMap_public_19Q4 dataset[39] using sklearn's python function adjusted_mutual_info_score (https://scikit-learn.org/stable/).

Recall of known gene relationships

We assembled a set of functionally related gene pairs using paralogs identified by EnsemblCompara[25], protein-protein interactions identified by Li et al[26], and CORUM complex comemberships[27]. For a given dataset, for each pair of related genes, we calculated a Pearson correlation coefficient between those genes' dependency scores across cell lines. We then binned each gene that appeared in the list of known gene relationships according to its mean gene score using 20 equally spaced bins. For pairs of genes in the related genes pairs, we chose one as the query gene and replaced its related partner with another randomly selected gene of similar gene mean, i.e. belonging to the same bin, excluding genes known to be related to the query gene. We calculated Pearson's correlation coefficients between these randomly selected gene pairs to generate a null distribution, from which we calculated empirical *p*-values and Benjamini-Hochberg FDRs for known related gene pairs. Ensuring that the pairs of genes used in the null distribution have similar distributions of mean gene effect as the pairs of known related genes is necessary because variable screen quality can produce a high but artificial correlation between any pair of common essential genes, and CORUM is highly biased towards common essentials. This is discussed further in the comparisons of batch corrections in Dempster et al[31].

Unexpressed false positives

We defined a gene as unexpressed in a cell line if the log2(Transcripts per million +1) of its DepMap expression was less than 0.01[39]. Any score of an unexpressed gene in a cell line was called a false positive if it fell in the bottom 15% of gene scores for that cell line.

Recall of prior-known essential genes at 5% FDR

For each dataset, we defined a true positive distribution from the (flattened) scores of the DepMap common essentials[39] and a null distribution from the scores of unexpressed genes. We then calculated a one-sided p-value for each essential gene's score and a Benjamini-Hochberg FDR. We report the number of essential gene scores with FDR less than 0.05.

Recovery of expression addictions

For each dataset, we calculated the Pearson correlation of a gene's dependency score across cell lines with its own expression (using data from the Broad DepMap portal[39]). We then calculated its correlation with the nearest neighboring gene's expression to generate a null distribution that includes whatever residual copy number bias may exist. Using these two distributions we calculated one-sided *p*-values for a gene's score to be negatively correlated with its own expression, followed by Benjamini-Hochberg FDRs. We report the number of significant negative correlations at an FDR of 0.25 as recovered expression addictions.

Binary depletion calls

This was computed by considering each cell line DPG as a rank-based classifier of essential/non-essential genes[11] (with gene rank positions determined by their fitness effect, i.e. average depletion fold-change of targeting single guide RNAs abundance at the end of the assay with respect to plasmid counts).

The fitness effect threshold was then fixed as that corresponding to the largest rank position *r* guaranteeing a false discovery rate (FDR) < 5%, when the predicted essential genes are those with a rank position ≤ *r*. This allowed us to assign to each gene in each cell line, in each of the two datasets, a binary dependency score.

To identify significantly depleted genes for a given cell line, we ranked all the genes in the cell line DPG in increasing order based on their depletion log fold-changes. We used this ranked list to calculate the precision curve using a set of prior known essential (*E*) and non-essential (*N*) genes, respectively.

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

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

We then determined the largest rank position k^* with $P(k^*) \geq 0.95$ (equivalent to a FDR ≤ 0.05). The 5% FDR logFCs threshold F^* was defined as the logFCs of the gene s such that $r(s) = k^*$. We called all genes with a logFC $< F^*$ as significantly depleted at 5% FDR.

Binary dependency matrices were defined as gene x cell lines matrices with non null entries corresponding to significant dependency genes at 5% FDR, for each cell line, i.e. column.

Selective dependencies

NormLRT and likelihood of normal distribution was calculated in R using the MASS package[40]. For the t-distribution the MASS package was used to calculate the likelihood, if the fitting procedure failed different degrees of freedom were used iteratively until a solution was found. The degrees of freedom used in order were 2,5,10,25,50 and 100.

Systematic association test between molecular features and gene dependencies

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 gene across a dichotomy of cell lines defined by the status (present/absent) of each CFE in turn. We tested genes whose NormLRT values were greater than 200 in any integrated dataset. 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 using Cohen's Delta (ΔFC), i.e. the difference in population means divided by their pooled standard deviations.

Identification of common essential genes via the 90th Percentile method

This method finds for each gene the cell line on the boundary of its 90th percentile of least dependent cell lines. It then calculates the rank of that gene in that cell line, by sorting all the genes based on their dependency score in increasing order. A mixture of two normal distributions is then fitted to the rank positions of all genes. Those genes with ranks below the crossover point of these two distributions are labeled as common essentials.

ADaM method

Binary depletion matrices for the integrated datasets were used with the ADaM method as described in [2]. The ADaM method determines the number of cell lines required to call a gene common essential. This threshold is calculated by maximizing the tradeoff between true positive rate (using a set of known prior essential genes) and the deviance from the null expected rate (calculated using random permutations of the binary depletion matrix).

Common essential genes were identified for each tissue separately and then were used as input into ADaM to determine pan-cancer common essential genes.

False-positive rates of common essentials

To estimate false-positive rates of the common essential genes we used the sets of negative controls as in Behan et al[2] derived from an independent study of RNAi screen[4]. In McDonald et al, gene essentiality was analyzed across a large dataset of RNAi screens across hundreds of cancer cell lines. For each gene, its tendency to follow a skewed Student-t distribution (indicative of a selective dependency) was estimated. As negative controls, we used sets of context-specific essential genes (as false positives) at different levels of likelihood of a skewed Student-t distribution.

References

1. Prasad, V. Perspective: The precision-oncology illusion. *Nature* **537**, S63 (2016).
2. Behan, F. M. *et al.* Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature* **568**, 511–516 (2019).
3. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564–576.e16 (2017).
4. 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).
5. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
6. 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).
7. 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).
8. Steinhart, Z. *et al.* Genome-wide CRISPR screens reveal a Wnt-FZD5 signaling circuit as a druggable vulnerability of RNF43-mutant pancreatic tumors. *Nat. Med.* **23**, 60–68 (2017).
9. Shi, J. *et al.* Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat. Biotechnol.* **33**, 661–667 (2015).
10. Tzelepis, K. *et al.* A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell Rep.* **17**, 1193–1205 (2016).
11. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**, 1515–1526 (2015).

12. 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).
13. Wellcome Sanger Institute. Cancer Dependency Map. *https://depmap.sanger.ac.uk/*.
14. Broad Institute of Harvard and MIT. Cancer Dependency Map. https://depmap.org/.
15. Feng, F. Y. & Gilbert, L. A. Lethal clues to cancer-cell vulnerability. *Nature* vol. 568 463–464 (2019).
16. Dempster, J. *et al.* Agreement between two large pan-cancer genome-scale CRISPR knock-out datasets. *Nature Communications* **In Press**,.
17. Iorio, F. *et al.* Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. *BMC Genomics* **19**, 604 (2018).
18. Allen, F. *et al.* JACKS: joint analysis of CRISPR/Cas9 knockout screens. *Genome Res.* **29**, 464–471 (2019).
19. Project Score. https://score.depmap.sanger.ac.uk/.
20. Project Achilles. https://figshare.com/articles/DepMap_19Q3_Public/9201770.
21. Aguirre, A. J. *et al.* Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. *Cancer Discov.* **6**, 914–929 (2016).
22. Gonçalves, E. *et al.* Structural rearrangements generate cell-specific, gene-independent CRISPR-Cas9 loss of fitness effects. *Genome Biol.* **20**, 27 (2019).
23. Doench, J. G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
24. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
25. Aken, B. L. *et al.* Ensembl 2017. *Nucleic Acids Res.* **45**, D635–D642 (2017).
26. Li, T. *et al.* A scored human protein-protein interaction network to catalyze genomic

interpretation. *Nat. Methods* **14**, 61–64 (2017).

27. Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes--2009. *Nucleic Acids Res.* **38**, D497–501 (2010).
28. Goncalves, E., Segura-Cabrera, A., Pacini, C. & Picco, G. Drug mechanism-of-action discovery through the integration of pharmacological and CRISPR screens. *bioRxiv* (2020).
29. Yang, W. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–61 (2013).
30. Picco, G. *et al.* Functional linkage of gene fusions to cancer cell fitness assessed by pharmacological and CRISPR-Cas9 screening. *Nat. Commun.* **10**, 2198 (2019).
31. Dempster, J. M., Rossen, J., Kazachkova, M. & Pan, J. Extracting Biological Insights from the Project Achilles Genome-Scale CRISPR Screens in Cancer Cell Lines. *BioRxiv* (2019).
32. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740–754 (2016).
33. Lenoir, W. F., Lim, T. L. & Hart, T. PICKLES: the database of pooled in-vitro CRISPR knockout library essentiality screens. *Nucleic Acids Res.* **46**, D776–D780 (2018).
34. Rauscher, B., Heigwer, F., Breinig, M., Winter, J. & Boutros, M. GenomeCRISPR - a database for high-throughput CRISPR/Cas9 screens. *Nucleic Acids Research* vol. 45 D679–D686 (2017).
35. DepMap, B. Project SCORE processed with CERES. (2019) doi:10.6084/m9.figshare.9116732.v1.
36. van der Meer, D. *et al.* Cell Model Passports—a hub for clinical, genetic and functional datasets of preclinical cancer models. *Nucleic Acids Res.* **47**, D923–D929 (2019).
37. Bolstad, B. M. preprocessCore: A collection of pre-processing functions. 2016. *R*

package version **1**,.

38. Leek, J. T. *et al.* sva: Surrogate Variable Analysis. R Package Version 30. 2017.

39. DepMap, B. DepMap 19Q4 Public. (2020) doi:10.6084/m9.figshare.11384241.v2.

40. Ripley, B. *et al.* Package 'mass'. *Cran R* **538**, (2013).