# Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting

Francesco Iorio[1,5,*,#], Fiona M Behan[2,5,*], Emanuel Gonçalves[2], Charlotte Beaver[2], Rizwan Ansari[2], Rachel Pooley[2], Piers Wilkinson[2], Sarah Harper[2], Euan Stronach[3,5], Julio Saez-Rodriguez[1,4,5,^], Kosuke Yusa[2,^], Mathew J Garnett[2,5,^,#]

*Equal contribution, #Corresponding author, ^Senior authors

[1] European Molecular Biology Laboratory - European Bioinformatics Institute, Cambridge - UK

[2] Wellcome Trust Sanger Institute, Cambridge - UK

[3] GlaxoSmithKline, Stevenage - UK

[4] JRC-COMBINE, RWTH Aachen University, Aachen - Germany

[5] Open Targets, Cambridge - UK

[#] Contact Information: iorio@ebi.ac.uk, mg12@sanger.ac.uk

## Abstract

**Genome editing by CRISPR-Cas9 technology allows large scale screening of gene essentiality in cancer. A confounding factor when interpreting CRISPR-Cas9 screens is the high false-positive rate in detecting essential genes, particularly for those that are within copy number amplified regions of the genome. We have developed the computational tool *CRISPRcleanR* which is capable of identifying and correcting gene-independent responses to CRISPR-Cas9 targeting. CRISPRcleanR uses an unsupervised approach based on the segmentation of single-guide RNA (sgRNA) fold change values across the genome, without making any assumption on the copy number status of the targeted genes. Applying our method to newly generated genome-wide essentiality profiles from 15 cancer cell lines, we demonstrate that CRISPRcleanR reduces false positives when calling essential genes, correcting biases within and outside of amplified regions, while maintaining true positive rates. Established cancer dependencies and essentiality signals of amplified cancer driver genes are detectable post-correction. CRISPRcleanR reports sgRNA fold changes and normalised sgRNA read counts, and is therefore compatible with downstream analysis tools, and works with multiple sgRNA libraries. CRISPRcleanR is a versatile, open-source tool for the analysis of CRISPR-Cas9 knockout screens to identify essential genes.**

## Introduction

CRISPR-Cas9-based genome editing techniques are transforming the landscape of genetic studies[1,2]. The high efficiency and specificity of the CRISPR-Cas9 system to mutagenise genes through the introduction of DNA double strand breaks (DSB), either at the level of individual genes or at genome-wide scale, enables the systematic investigation of loss-of-function phenotypes. We and others have developed genome-wide pooled CRISPR knock-out (CRISPR-KO) screens, in which a pool of single guide RNAs (sgRNA) together with Cas9 are introduced into cells, and the phenotypic outcome is quantified by the abundance of each sgRNA between control and test samples[3–5]. Improvements continue to be made to screen design [6-7] and CRISPR-KO screens have already revealed novel genes and pathways involved in diverse biological processes[8,9].

One prominent application of CRISPR-KO screening is to systematically identify genes that are essential for cancer cell survival to identify possible strategies for the development of novel targeted therapies. Several groups have already performed screens and successfully identified novel drug targets[6,10]. A recent landmark study has reported gene essentialities in 342 cancer cell lines[11] and the number of cell lines screened is expected to increase further. This will empower association studies between gene essentialities and genomic/transcriptomic features to develop biomarkers for patient stratification of proposed therapies.

One drawback of the CRISPR-KO screening system is caused by its mode of action, namely DSB induction. DSBs induce a DNA damage response and can cause cell cycle arrest and, in some cases, cell death[12–14]. This is problematic when performing CRISPR-KO screens in cancer cells because of frequent copy number (CN) alterations in their genome, resulting in widespread Cas9 induced DNA damage. Consequently, DSBs at genes in amplified regions result in depletion of these genes in a pooled CRISPR-KO screen regardless of their essentialities, and thus they are erroneously called as essential genes. This can result in a high false-positive rate and correcting for this CN-associated effect is crucial for the interpretation of CRISPR-KO screening results. Solutions proposed thus far encompass scanning the dataset for biased regions and their removal from downstream analysis[15], resulting in the exclusion of potentially biologically relevant genes residing in

CN-amplified regions, or to apply a piecewise linear model to infer true gene dependencies based on CN profiles[11].

From our own CRISPR-KO screen dataset consisting of 15 human cancer cell lines, we identified a number of instances for which existing approaches for correcting bias in CRISPR-KO data were unsuitable or hampered further downstream analyses. To address this, we developed *CRISPRcleanR*, a computational approach implemented in an open-source R package, which identifies biased genomic regions from CRISPR-KO screen datasets in an unsupervised manner and corrects both read count and log fold change (logFC) of individual sgRNAs in such regions. Our method can reduce false positive calls while keeping the true positive rate of known essential genes largely unchanged and allows the detection of essential genes, even within focally amplified regions.

## Results

### *Gene-independent responses in CRISPR-KO screens*

We performed genome-wide CRISPR-KO screens on a panel of 15 human cancer cell lines, which are a subset of the Genomics of Drug Sensitivity in Cancer (GDSC) collection (**Supplementary Table 1**)[16,17]. This panel included six different tumour types with different mutational processes including high frequency of single-nucleotide variants (large intestine, lung, and melanoma) and CN variation (breast and ovary). We used the Sanger Institute CRISPR library (version 1.0) targeting 18,010 genes (90,709 sgRNAs; ~5 sgRNAs per gene)[6]. The screens showed high consistency between technical replicates in each cell line (median average correlation for sgRNA counts = 0.83) and readily discriminated between pre-defined fitness essential (FE) and non-essential genes (median area under the Receiver Operating Characteristic curve (AUROC) = 0.92 (**Supplementary Fig. 1**)[18]. Additionally, when considering average logFC values, a high true positive rate (TPR, or recall) was observed for known essential genes assembled from the Molecular Signature Database (MsigDB)[19] and from literature[20] (median TPR across gene sets and cell lines = 85% at 5% FDR).

From a comparison between our CRISPR data and CN profiles for each line, we confirmed a large negative effect for logFCs of sgRNAs targeting genes in CN-amplified regions, particularly with CN ≥ 8 (**Supplementary Fig. 2** and **Supplementary Table 2**). Notably, sgRNA targeting CN-amplified (CN ≥ 8) non-expressed genes (FPKM < 0.05) were significantly more depleted in six cell lines than the rest of the sgRNA in the whole library. For three cell lines (HT55, EPLC-272H, and MDA-MB-415), the negative effect on logFC of sgRNA in CN-amplified regions was even greater than the effect size for FE genes (**Supplementary Fig. 3** and **Supplementary Table 2**). Collectively, using independent data, our analysis confirms the systematic negative bias on sgRNA logFC values in particular regions of the genome, which are enriched in CN-amplified regions.

*Heterogeneous effect of copy number amplification on gene-independent responses to CRISPR-Cas9 targeting*

To gain greater insight into CN-associated biases, we performed a detailed analysis of the relationship between sgRNA logFC values and CN at the level of individual CN segments (**Fig 1a** and **Supplementary Fig 4**). For some cell lines, the negative bias on average logFC values within segments was positively correlated with increasing CN (EPLC-272H, NCI-H520, OVCAR-8, TOV-21G and SW48). In other cell lines the bias effect on average logFC was not observed (MDA-MB-436), plateaued (NCI-H2170), or fluctuated as CN varied (MDA-MB-453, HT55 and HuP-T3) (**Fig 1a** and **Supplementary Fig 4**). These effects were preserved when only considering sgRNA targeting non-expressed genes (**Fig 1b** and **Supplementary Fig 4**), demonstrating that this effect is most likely independent of true gene essentiality. In addition, we observe a wide range of average logFC values for segments of a given copy number (**Fig 1a, b**), and this is often larger than the variation between segments of different CN, indicating that CN does not capture all of the observed variance..

Furthermore, although in the majority of instances CN segments matched segments of equal sgRNA logFCs (**Fig 1c**), we identified several CN segments with discontinuous logFC patterns (**Fig 1d**). Additionally, regions of consistently depleted sgRNAs were identified also in diploid regions of the genome. For example, the cell

line CL-40 harbours two copies of chromosome 16, but several contiguous genes (of which many are not expressed) in region 16q23 exhibited a negative logFC across targeting guides (**Fig 1e**).

Our results indicate that biases observed in CRISPR-KO screens are often associated with CN alteration but are heterogeneous with poorly understood variation between segments of differing CN, and variation within segments of the same CN. This highlights the need for an unsupervised approach, not dependent on CN alone, to correct for biased regions in CRISPR-KO data.

### *Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting*

In order to detect biased regions in an unsupervised manner and correct corresponding sgRNA logFCs observed in CRISPR-KO screening data, we developed *CRISPRcleanR*, a computational approach implemented in an open-source R package. CRISPRcleanR applies a circular binary segmentation algorithm, originally developed for array-based comparative genomic hybridization assay [21,22], directly to the genome-wide pattern of sgRNA logFCs across individual chromosomes in a cell line. It then detects genomic segments containing multiple sgRNAs with sufficiently equal logFCs. If these segments contain sgRNAs targeting a minimum number of different genes (adaptively determined as detailed in the Results and Methods sections), then the sgRNA in the segment are most likely responding to CRISPR-Cas9 targeting in a gene-independent manner, and logFCs values are corrected via mean-centering.

CRISPRcleanR has several features that make it statistically robust, versatile and practical for downstream applications: (i) it works in an unsupervised manner, requiring no chromosomal CN information, *a priori* defined sets of essential genes, nor any user-defined parameter values; (ii) it implements a logFC correction, making depletion scores for all genes usable in follow up analyses; (iii) it examines logFC at the sgRNA level to gain resolution and to account for different levels of sgRNA on-target efficiency, and enables the subsequent use of algorithms to call gene depletion significance that require input data at the sgRNA level (e.g. BAGEL[23]); (iv) by applying an inverse transformation to corrected sgRNA logFCs, it computes corrected sgRNA counts, which are required as input for commonly used mean-variance modeling approaches, such as MAGeCK[24], to call gene

depletion/enrichment significance; (v) lastly, CRISPRcleanR corrects logFC values using data from an individual cell line and with invariant performances, unlike other computational correction approaches whose performances depend on the number of analysed cell lines[11]; as a consequence, CRISPRcleanR is suitable for the analysis of data from both small- and large-scale CRISPR-KO studies.

When applied to our data, CRISPRcleanR effectively corrected the bias in sgRNA logFCs over a wide range of chromosomal segments with variable CN alterations. Furthermore, this included detection and correction of different level of biases in sgRNA logFCs within an individual segment of equal CN (**Fig. 2a**, **b**). An immediate result of the application of CRISPcleanR to our data was that biases in particularly high CN regions were strongly attenuated over all the cell lines (**Fig. 2c**).

Overall, CRISPRcleanR reduced the recall of sgRNAs targeting CN-amplified regions, including sgRNAs targeting CN-amplified non-expressed genes, towards expectation when classifying the whole library of sgRNAs based on their logFCs (**Fig. 2d**). The correction was also consistently observed at the gene level (average logFCs of targeting sgRNAs) across all screened cell lines at a fixed 5% FDR, with a median reduction in recall equal to 72% and 88%, respectively for CN-amplified and CN-amplified non-expressed genes (**Fig 2e** and **Supplementary Table 3**). This reduction was also observed at the level of the area under the overall recall curves (AURCs), thus regardless any fixed depletion significance threshold. Specifically, we observed the median AURCs across all cell lines shifting from 0.74 to 0.51 and from 0.7 to 0.5 (thus, in both cases, to expectation), respectively, for CN-amplified and CN-amplified non-expressed genes (**Fig. 2e** and **Supplementary Table 3**). The reduction in AURC was independent of whether amplified genes in cell lines were identified using CN data from the GDSC or the cancer cell line encyclopedia (CCLE). In contrast, for the MsigDB known essential genes and the FE genes, the reduction was negligible at less than 2%, with median AURCs preserved at $\geq 0.82$.

Excluding from the essentiality profiles the sgRNAs targeting *a priori* known essential genes (taken from MSigDB) before CRISPRcleanR correction yielded very similar results as when imposing the constraint that, for a segment to be corrected, it must contain sgRNA targeting $n = 3$ different genes (**Supplementary Fig. 5**). This was determined by performing several correction attempts varying $n$ and considering or not considering FE and other MSigDB essential genes. Thus, CRISPRcleanR can be used in a completely unsupervised setting, without making any assumption on gene essentiality.

### *CRISPRcleanR preserves cell line essentiality profiles and detection of cancer dependencies*

We next determined whether the correction performed by CRISPRcleanR alters the overall essentiality profile of a given cell line. For each of the 15 cell lines, we compared the top-depleted sgRNA essentiality profiles using corrected and uncorrected logFC sgRNA values by computing precision/recall curves, using the uncorrected values as positive controls (**Fig. 3a,b**). We observed a median area under the precision/recall curve (AUPRC) of 0.92 (min = 0.81 for HCC-15, max = 0.96 for MDA-MB-436) when considering only the top 50 depleted sgRNA, and a median AUPRC of 0.96 when considering the top 2,500 depleted sgRNA (min = 0.88 for HCC-15, max = 0.98 for MDA-MB-453). Considering that an experiment typically yields ~6,000 sgRNAs called as significantly depleted, this indicates that the CRISPRcleanR correction, while reducing false-positive rates, does not have an unwanted impact on the overall essentiality profile of a cell line.

Since a major application of CRISPR-KO screens is the accurate identification of genes essential for cellular fitness, we investigated the ability of CRISPRcleanR to preserve the detection of expected cancer essentialities in individual cell lines. Given the bias associated with CN-amplified regions, we first focused on the ability to selectively detect cancer dependencies involving amplified cancer driver genes, as compared to co-amplified putative passenger genes present in the same CN segment. The proto-oncogene *MYC* is recurrently amplified across human cancers of different tissues. A large region of chromosome 8 carrying *MYC* is amplified in the cell line HT-29 and sgRNAs targeting *MYC,* as well as flanking genes, are reported as significantly depleted when using uncorrected logFCs (**Fig. 3c**). The logFC depletion is greater for *MYC* compared to other genes in this

region. Following CRISPRcleanR correction, the sgRNAs targeting *MYC* remained significantly depleted, whereas those targeting the co-amplified flanking genes were no longer significant (**Fig. 3c**). A similar essentiality was selectively preserved post-CRISPRcleanR correction in an amplified region of chromosome 16 that contains *ERBB2* in the NCI-H2170 cell line (**Fig. 3d**).

We next evaluated our sensitivity to detect dependencies present in four cell lines involving either cancer driver genomic alterations or previously described sensitivity to drug treatment. These included A2058 and HT-29 carrying the *BRAFv600e* mutation (**Fig. 3e**), MDA-MB-453 carrying the *PIK3CAh1047r* mutation and MDA-MB-415 known to be sensitive to the PI3Kβ inhibitor AZD6482 (**Fig. 3f**). For all four cell lines, the dependencies were confirmed with CRISPRcleanR-corrected logFCs of the corresponding targeting sgRNAs (**Fig. 3e,f**).

### *CRISPRcleanR is effective using multiple genome-wide sgRNA libraries*

To investigate the versatility of CRISPRcleanR, we assessed its performance across different libraries of sgRNAs. For the purpose of comparability we used our previously published dataset derived from screening the HT-29 cell line with the Brunello[25] and Whitehead[15] libraries, using the same lentiviral vector as our library[26]. It is noteworthy that despite all three libraries targeting 17,646 overlapping genes, fewer than 5% of the 19-mer gRNA in the libraries are overlapping in sequence. All three libraries showed matching patterns of biased logFCs in the same CN-amplified genomic region spanning *MYC* (**Fig 4a**). CRISPRcleanR corrected the sgRNA logFC values for this bias in all three libraries (**Fig 4a**). A similar reductions in recall for CN-amplified genes (mean = 40 ± 2.7 %), CN amplified non-expressed genes (45 ± 5.7 %), fitness essential genes (2 ± 0.47 %), and non-essential genes (mean = -3.8 ± 1.81 %) was observed across all three libraries (**Fig. 4b,c**). We conclude that CRISPRcleanR is suitable for correcting bias in CRISPR-KO screening datasets generated using multiple different sgRNA libraries.

*CRISPRcleanR corrects sgRNA treatment counts and can be used prior to mean-variance modeling*

MAGeCK is a widely used computational tool to call gene depletion or enrichment in CRISPR-KO screens and is based on mean-variance modelling of median-ratio normalised sgRNA read-counts[24]. To make CRISPRcleanR compatible with mean-variance modeling approaches such as MAGeCK, we designed an inverse transformation to derive corrected sgRNA treatment counts from CRISPRcleanR corrected sgRNA logFC values. To benchmark our transformation, we compared results obtained from executing MAGeCK using normalised uncorrected and CRISPRcleanR-corrected sgRNA counts by mean of recall estimation when classifying predefined gene sets. The inverse transformation had an effect on both the mean and variance of the sgRNA counts, with the greatest impact on sgRNAs targeting genes in CN-amplified regions, whose value was consistently shifted toward the corresponding value in the plasmid/control condition (**Fig. 5a,b**). Furthermore, we observed a strong reduction in recall when classifying sgRNAs targeting genes in the biased regions (CN ≥ 8), when considering as positive predictions the sgRNAs called significantly depleted by MAGeCK. The median reduction was 75% for CN-amplified genes and 80% for CN-amplified non-expressed genes at a 10% FDR, and 72% and 100% reductions at a 5% FDR; **Fig. 5c** and **Supplementary Table 4**). In contrast, the effect on the recall of FE and non-essential genes was negligible (median = 2.9% reduction) (**Fig. 5c**).

*Code availability and overview*

CRISPRcleanR is implemented as an R[27] package with full documentation, tutorials, built in datasets to reproduce the results in this manuscript, and is publically available at https://github.com/francescojm/CRISPRcleanR (submission to bioconductor in progress). It includes core functionality for processing raw sgRNA count files through to generating corrected sgRNA logFC values and corrected sgRNA counts for downstream analyses. CRISPRcleanR also includes functions to measure and visualise the extent and effect of the performed correction, the ability to detect CN-amplified non-expressed genes (which can be used as positive controls), and classification performances for *a priori* known sets of essential/non-essential genes pre- and post-correction of corresponding targeting sgRNAs.

## Discussion

In this study, we report CRISPRcleanR, a computational framework that detects genomic segments with gene-independent bias in CRISPR-KO data in an unsupervised manner, and applies segment-by-segment correction at sgRNA-level logFC and read counts. The correction substantially reduces false-positive calls without altering true essentiality profiles and, importantly, preserves known cancer gene dependencies within and outside biased segments. CRISPRcleanR works on multiple genome-wide sgRNA libraries, and resulting corrected sgRNA logFC and read counts are compatible with downstream analyses performed by methods such as BAGEL or MAGeCK to identify screen hits. CRISPRcleanR works equally efficiently whether analysing results from a single cell line or multiple cell lines in parallel.

Our motivation for developing CRISPRcleanR came from the observation that biases in gene essentialities observed in CRISPR-KO screens did not always show correlation to their CN status, although biased segments are frequently associated with CN alteration. Additionally, in most of the cell lines analysed, variation in the mean logFCs of segments with the same CN were often greater than those between segments with different CN. Some cell lines showed greater bias in segments with lower CN. We even identified multiple instances of discontinuous bias on sgRNA logFCs within a particular CN segment and biased responses within segments that are not CN-amplified. These observations argue for the development of methods such as CRISPRcleanR which are independent of CN for the analysis of CRISPR-KO screening data, and indicate that biased responses are not solely due to the amount of DNA damage and may also be caused by additional factors, such as local genomic structures.

CRISPRcleanR detects biased segments using sgRNA-level logFC in an unsupervised manner, eliminating the requirement for cell line CN information. This simplifies the analysis and is advantageous when reliable CN information is not available for a cell line; for example, when using a newly derived cancer cell model. In addition, cancer genomes are dynamic and continuously evolving, causing genetic variation between different clones of the same cell line. Genetic drift may occur during prolonged *in vitro* cell culture, due to different

growth conditions (e.g. media composition), or in response to selective pressure (e.g drug treatment) and genetic manipulation (e.g. gene-editing). Thus, the genomic heterogeneity of cancer cells, even within clones of the same cell line, may confound CN-based correction methods when relying on pre-existing CN data, and negatively impact identification of gene essentialities. Furthermore, the performance of different copy number calling algorithms is variable and depends on the underlying genomic data available, and as a result this can be a further confounding factor when using CN-based correction methods. CRISPRcleanR overcomes these limitations by effectively correcting for biases in CRISPR-KO screens without requiring additional information about the cell models screened, and without making assumptions about the underlying cause of bias. We conclude that CRISPRcleanR is a useful and versatile tool which will greatly facilitate the analysis of CRISPR-KO screens in oncology.

## Methods

### *Plasmids, cell lines and reagents*

Cells were maintained in culture media as indicated in **Supplementary Table 1** in a 5% $CO_2$ humidified incubator at 37 °C. The plasmids used in this study were from the mutagenesis toolkit described in [6] and are available through Addgene (Cas9 - 68343; CRISPR sgRNA library - 67989). Plasmids were packaged using the Virapower (Invitrogen) system as per manufacturer's instructions.

### *Genome-wide mutant library and screen*

Cells were first transduced with lentivirus carrying Cas9 in T75 flasks at ~80% confluence in the presence of polybrene (8 µg/ml). The following day, lentiviral containing medium was replaced with complete medium. Blasticidin selection was started on day 4 post transduction at a concentration determined from a titration in the parental cell line. Cas9 activity was assessed following selection using the Cas9 functional assay as described in[6] and a cut-off of 80% activity was applied (median = 89% activity across all cell lines). Cas9-expressing cells were maintained in blasticidin prior to transduction with the sgRNA library. Transduction with sgRNA library was carried out at ~80% confluency with 3.3 x $10^7$ cells in T150 or T525 (triple layer) flasks, depending on cell size and surface area required, in technical triplicates. Cells were transduced with a predetermined viral amount that gives rise to ~30% transduction, measured by BFP expression by cytometry, to ensure approximately 1 viral particle entering each cell based on a Poisson distribution. Based on these initial cell numbers and transduction efficiency, the coverage of the sgRNA library (i.e. the number of cells containing each sgRNA) in each replicate was 100x. Puromycin selection commenced at day 4 to select for cells that had successful lentiviral integration. Actual library transduction efficiency and puromycin selection was analysed using cytometry before and after puromycin selection, respectively. A minimum number of 5.0 x $10^7$ cells (500x coverage per sgRNA) were maintained at all times to ensure library representation was maintained. The cells were harvested 14 days post transduction and dry pellets were stored at -80 ºC.

Extraction of genomic DNA, PCR amplification of sgRNAs and Illumina sequencing of sgRNAs were carried out as described previously[3,6]. The number of reads for each sgRNA was determined using a script developed in-house.

*Data pre-processing and availability*

sgRNA counts were normalised assembling one batch per cell line, including the read counts from the matching library plasmid and all final read counts replicates, with a median-ratio method [28] to adjust for the effect of library sizes and read count distributions, after filtering out sgRNAs with less than 30 reads in the plasmid. Depletions/enrichments for individual sgRNAs were quantified as $\log_2$ ratio between post library-transfection read-counts and library plasmid read-counts. Finally, sgRNAs were averaged across replicates. This was performed executing the ccr.NormfoldChanges function of the CRISPRcleanR R package.

sgRNAs read count files are available for the HT-29, EPLC-272H, and A2058 cell lines as external data in the CRISPRcleanR R package, and they are used in the package vignette and documentation examples. sgRNA read counts for all cell lines used in this study have been deposited on BioStudies[29] (accession number S-BSST79).

*Transcriptional and copy number data*

Genome-wide substitute reads with fragments per kilobase of exon per million reads mapped (FPKM) for the 15 cell lines considered in this study were described in[30]. Genome-wide gene level copy number data derived from PICNIC analysis of Affymetrix SNP6 segmentation data (EGAS00001000978) for the cell lines in the Genomics of Drug Sensitivity 1,000 (GDSC1000) cancer cell line panel[16], were downloaded from the GDSC data portal (dataset version: July 4th 2016), http://www.cancerRxgene.org. This dataset is also available at ftp://ftp.sanger.ac.uk/pub/project/cancerrxgene/releases/release-6.0/Gene_level_CN.xlsx. For each gene, the minimum copy number of any genomic segment containing coding sequence was considered.

Additionally, gene level Gistic[31] values were obtained by processing Affymetrix SNP array data in the Cancer Cell Line Encyclopaedia[32] repository in cBioPortal[33] (http://www.cbioportal.org/study?id=cellline_ccle_broad#summary) .

*Identification of gene-independent responses in cancer cell lines*

For each cell line, segments of equal CN were identified by using CN data from the GDSC data portal[16,17] (as detailed below), and assigned a mean-logFC value by averaging across all of the sgRNAs targeting a segment.

A CN *bias starting point* was computed for each cell line as the copy number value $n > 2$ such that statistically significant differences, as quantified by a Welch's t-test, were observable between the mean-logFCs of segments of $n$ CNs and those of segments of 2 CN.

A CN *bias critical point* was computed for each cell line as follows. For each CN value $n = 3, \ldots, m\text{-}1$ (with $m$ = maximal segment CN value observed in the cell line under consideration), two univariate linear models were fitted, considering segment CN values as observations of the independent variable and the corresponding average segment mean-logFCs as those of the dependent one. The first model $P(n)$ was fitted using CN values in $\{2, \ldots, n\}$ and corresponding average segment mean-logFCs, while the second one $L(n)$ was fitted using CN values in $\{n+1, \ldots, m\}$ and corresponding average segment mean-logFCs. The *bias critical point* was then defined as the value $n$ providing the large absolute difference between the slopes of the corresponding fitted models $P(n)$ and $L(n)$.

*Calling significantly depleted sgRNAs and genes*

All sgRNA were ranked by average logFCs derived from screening an individual cell line. This ranked list was used to classify sgRNAs targeting genes from two gold-standard reference sets of FE and non-essential genes[18,23]: from now the *essential-sgRNAs* ($E$) and the non-essential-sgRNAs ($N$). For each rank position $k$, a set of predictions $P(k) = \{s \in E \cup N : \rho(s) \leq k\}$, with $\rho(s)$ indicating the rank position of $s$, was assembled and corresponding Precision (or Positive Predicted Value, $PPV(k)$) was computed as:

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

Subsequently the largest rank position $k^*$ corresponding to a 0.95 Precision (equivalent to a False Discovery Rate (FDR) = 0.05) was determined as

$$k^* = \max_{k=1}^{N} \{ 1 - PPV(k) \leq 0.05 \}.$$

Finally, a 5% FDR logFCs threshold $F^*$ was determined as the logFCs of the sgRNAs $s$ such that $k(s) = k^*$, and all the sgRNAs of the entire library with a logFC < $F^*$ were considered significantly depleted at this FDR level.

To call depletion significance at a gene level, the same procedure was followed but averaging logFCs of sgRNAs targeting the same gene prior to the analysis, and considering ranks and positive/negative sets of genes instead of sgRNAs.

For the follow up study on the effect of correcting sgRNA treatment counts (computed as detailed below) we used the test function of the MAGeCK python package, indicating none as the value of the parameter specifying the normalisation method to use prior to the analysis, as a median-ratio normalisation was already applied to the analysed count files prior CRISPRcleanR correction.

### *Receiver Operating Characteristic analyses*

Across the different analyses, standard ROC indicators were computed considering as prediction sets significantly depleted sgRNAs (or genes) at a fixed level of 5% FDR (computed as detailed in the previous section or output by the MAGeCK tool), or genome-wide profiles of essentiality (as ranked lists of sgRNA logFCs, in some instances averaged on a per gene basis) to compute overall indicator curves, and using different positive/negative control sets (detailed below). To this aim, we made use of functions included in the pROC R package[34].

For the positive controls, sets of *a priori* essential genes were assembled by downloading relevant gene signatures from the MSigDB[19] (**Supplementary Table 5**). A list of ribosomal protein gene was derived from[20]. The consensual signatures resulting from this curation are available as individual data objects in the CRISPRcleanR R package.

*Segmentation analysis and logFC correction*

Genome-wide essentiality profiles in the form of lists of sgRNAs logFC were sorted according to the genomic coordinates of the individual sgRNAs (library annotation and coordinates derived from [6]) using the function ccr.logFCs2chromPos of the CRISPRcleanR R package. Then, a circular binary segmentation algorithm[21,22] was applied using the ccr.GWclean function of the CRISPRcleanR R package, with a significance threshold to accept change-points $p = 0.01$, 10,000 permutations for $p$-value computation, a minimal number of 2 markers per region, and making use of the function segment from the DNAcopy R package with other parameters set to default values.

Subsequently, sgRNA included in a segment had their logFCs mean-centered (across that segment) if collectively targeting at least $n = 3$ different genes, without pre-filtering any essential gene (differently from the sliding window approach used in[12]).

This number was adaptively determined by executing different trials of segments' detection and correction varying $n \in \{2, 3, 5, 10\}$ and excluding/not-excluding from the analysis sets of *a priori* known essential genes assembled from MSigDB (as detailed in the previous section), collectively the *filter set*. Removing the filter set from a reference set of the FE genes yielded a *test set*. Areas under the recall curve (AURCs) were then computed evaluating the classification performances using as positive controls the test set, CN amplified genes, and CN amplified non-expressed genes (determined for each cell line) were then computed, across each trial using targeting sgRNAs' logFCs before/after correction. For each of the positive control sets, reduction of recall (recall $\Delta$) were computed by comparing AURCs obtained before/after CRISPRcleanR correction.

Results showed that $n = 3$ provided the largest reduction of recall (**Supplementary Fig. 6**) of CN amplified and CN amplified non-expressed genes, and the lowest reduction of recall of the test set. Most importantly, this was observed invariantly with respect to removing/not-removing the filter set prior the analysis.
As a conclusion, all the corrections presented in this manuscript were executed with this setting ($n = 3$ and without pre-filtering any gene). CRISPRcleanR package uses these settings by default, although offering to the user the possibility of changing them.

*Comparison of results across different libraries*

Data from the mutagenesis of the HT-29 cell lines with the Brunello and Whitehead libraries were downloaded from the supplementary material of [26] and processed as described in the section *Data pre-processing and availability*. Correction outcomes were computed as detailed in *Receiver Operating Characteristic analyses*.

*Correction of sgRNA counts*

For each individual single guide RNA, a corrected treatment count $t_i$ was computed observing that:

$$N = E(\log_2 \frac{t_i}{c})$$

with $N$ = CRISPRcleanR corrected logFCs for the sgRNA under consideration, $i = 1, \ldots, n$, where $n =$ number of treatment replicates, and $c =$ counts of the sgRNA in the plasmid.

This implies

$$N = \frac{\sum_{i=1}^{n} \log_2 \frac{t_i}{c}}{n}$$

$$\implies nN = \sum_{i=1}^{n} \log_2 t_i - \sum_{i=1}^{n} \log_2 c = \sum_{i=1}^{n} \log_2 t_i - n \log_2 c$$

$$\implies nN + n \log_2 c = \sum_{i=1}^{n} \log_2 t_i$$

Assuming, for simplicity that all the $t_i$ are the same $= t$,

$$nN + n \log_2 c = n \log_2 t$$

$$\implies 2^{N + \log_2 c} = t$$

$$\implies t = c2^N = E(t)$$

To derive the corrected counts for the individual replicates (which are obviously different from each other) from their mean, we have kept constant the proportions seen in the uncorrected counts with respect to the sum of the counts across replicates:

$$E(t) = \frac{\sum_{i=1}^{n} t_i}{n}$$

$$\implies nE(t) = \sum_{i=1}^{n} t_i$$

$$\implies t_i = nE(t)\frac{t_i^*}{T^*} = nc2^N\frac{t_i^*}{T^*}$$

where $t_i^*$ is the count of the sgRNA under consideration before correction in the $i$-th replicate and $T^*$ is their overall sum across replicates.

## Competing interests

The authors declare no competing interests.

## Author's contributions

FI conceived the study, designed algorithms, processed data, performed validation tests, implemented and documented the R package, wrote and revised the manuscript; FMB contributed to the design of the study, performed screens and related experiments, curated data, contributed to documenting the R package, wrote and revised the manuscript; EG contributed ideas for the design of the study, curated data, revised the manuscript; CMB, RA, RP, PW, SH performed screens and related experiments; ES contributed ideas for the design of the study, co-supervised it and revised the manuscript; JSR contributed ideas for the design of the study and supervised it, revised the manuscript; KY conceived the study and co-supervised it, designed the sgRNA library and the experimental setting of the screen, wrote and revised the manuscript; MJG conceived the study, designed the experimental setting of the screen, wrote and revised the manuscript, supervised the study.

## Acknowledgements

We thank David R. Wille, Vivek Lyer, Leopold Parts, Felicity Allen, Gabriele Picco, and Ultan McDermott for a number of insightful discussions. We thank Giuseppe Iorio for a number of acute and critical observations on the mathematics underlying this study. This work was supported by OpenTargets (015), CRUK (C44943/A22536), SU2C (SU2C-AACR-DT1213) and the Wellcome Trust (102696)

## Figure Legends

### Figure 1: Heterogeneous gene-independent responses to CRISPR-Cas9 targeting

(A) Average LogFC values of sgRNA within segments of equal CN (excluding FE and histones) for three cell lines. Each circle represents a CN segment of the indicated copy number. Asterisks mark the CN at which a significance difference (Welchs t-test, $p < 0.05$) is initially (starting point) and continuously (critical point) observed compared to logFC values at CN = 2. Box-plots show the median, inter-quartile ranges and 95% confidence intervals.

(B) Same as for A but considering only non-expressed genes (FPKM < 0.05).

(C, D, E) Segments of equal gene copy number and segments of equal sgRNA logFCs for selected chromosomes in three cell lines.

### Figure 2: Unsupervised detection of segments of equal sgRNA logFCs and their correction.

(A and B) Example segments of equal gene copy number and equal sgRNA logFC values detected and corrected by CRISPRcleanR in two cell lines.

(C) logFC values of sgRNAs of the entire library for all cell lines grouped according to the copy number of their targeted gene before (left) and after (right) CRISPRcleanR correction. Box-plots show the median, inter-quartile ranges and 95% confidence intervals.

(D) Recall curves of sgRNA when classified as targeting amplified genes, amplified non-expressed genes, FE genes, and non-essential genes before and after CRISPRcleanR correction, for an example cell line (EPLC−272H).

(E) Assessment of CRISPRcleanR correction comparing recall (top row) or AURC (bottom row) of sgRNAs targeting genes in six predefined gene sets (5% FDR) based on their uncorrected or corrected logFCs.

### Figure 3: CRISPRcleanR retains overall profiles of essentiality and known dependencies

(A) Example precision/recall curves for the indicated number of top depleted sgRNAs after CRISPRcleanR correction, classified based on their un-corrected sgRNAs logFCs.

(B) Area under the precision/recall curves defined as for A for all cell lines. Box-plots show the median, inter-quartile ranges and 95% confidence intervals.

(C) A CN-amplified region of chromosome 8 in HT-29 cell line including *MYC* and 3 surrounding up-streaming/down-streaming genes. Expanded view of sgRNAs targeting *MYC* and its surrounding genes, with each gene identified by a different colour. The heatmaps (first 7 columns) show ranked positions of the sgRNAs targeting the 7 considered genes (blue bars) before (top heatmap) and after (bottom heatmap) CRISPRcleanR correction. The last two columns show rank positions for the sgRNAs targeting FE genes (second last column) and non-essential genes (last column).

(D) Same as for C but considering a region on chromosome 16 in the NCI-H2170 cell line, including ERBB2 and four flanking upstreaming/downstreaming genes.

(E and F) Rank position of sgRNAs targeting the indicated genes before (top) and after (bottom) CRISPRcleanR correction. FE and non-essential genes are shown for comparison.

**Figure 4: CRISPRcleanR is effective with multiple different sgRNA libraries**

(A) Segments within chromosome 8 of equal gene copy number juxtaposed to segments of equal sgRNA logFCs before and after CRISPRcleanR in HT-29 cells screened with three different sgRNA libraries.

(B) Recall curves for three sgRNA libraries when classifying sgRNAs targeting amplified genes, amplified non-expressed genes, FE genes, and non-essential genes using sgRNA logFCs before (first column of plots) and after (second column of plots).

(C) Correction of the area under the recall curve for sgRNAs targeting genes in six predefined sets, based on their uncorrected/corrected logFCs, across the three different libraries (one circle per library).

**Figure 5: CRISPRcleanR corrected sgRNA counts and downstream analysis with MAGeCK**

(A and B) Normalised counts of sgRNAs of the transfected libraries versus the control plasmid for FE and non-essential genes (first two rows of plots), CN amplified genes (third row) and CN non-expressed genes (fourth row), for two example cell lines before (first and third column) and after (second and fourth column) CRISPRcleanR correction. Essentialities for CN-amplified cancer driver genes such as *MYC*, *ERBB2* and *CCND1* are retained post correction.

(C) Comparison of recall using MAGeck for sgRNAs targeting genes in six predefined gene sets when using as input CRISPRcleanR uncorrected and corrected sgRNAs counts.

## Supplementary Material

Supplementary Figures 1 to 5
Supplementary Tables 1 to 5
Supplementary Text 1: Legends of Supplementary Figures and Tables

**References**


1. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339,** 823–826 (2013).
2. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–823 (2013).
3. 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).
4. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343,** 84–87 (2014).
5. 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).
6. Tzelepis, K. *et al.* A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell Rep.* **17,** 1193–1205 (2016).
7. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34,** 184–191 (2016).
8. Park, R. J. *et al.* A genome-wide CRISPR screen identifies a restricted set of HIV host dependency factors. *Nat. Genet.* **49,** 193–203 (2017).
9. Manguso, R. T. *et al.* In vivo CRISPR screening identifies Ptpn2 as a cancer immunotherapy target. *Nature* **547,** 413–418 (2017).
10. 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).
11. Meyers, R. M. *et al.* Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* (2017). doi:10.1038/ng.3984
12. Wang, T. *et al.* Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. *Cell* **168,** 890–903.e15 (2017).

13. Aguirre, A. J. *et al.* Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. *Cancer Discov.* **6,** 914–929 (2016).

14. Munoz, D. M. *et al.* CRISPR Screens Provide a Comprehensive Assessment of Cancer Vulnerabilities but Generate False-Positive Hits for Highly Amplified Genomic Regions. *Cancer Discov.* **6,** 900–913 (2016).

15. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350,** 1096–1101 (2015).

16. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* (2016).

17. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483,** 570–575 (2012).

18. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163,** 1515–1526 (2015).

19. 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).

20. Yoshihama, M. *et al.* The human ribosomal protein genes: sequencing and comparative analysis of 73 genes. *Genome Res.* **12,** 379–390 (2002).

21. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Comput. Appl. Biosci.* **23,** 657–663 (2007).

22. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5,** 557–572 (2004).

23. Hart, T. & Moffat, J. BAGEL: a computational framework for identifying essential genes from pooled library screens. *BMC Bioinformatics* **17,** 164 (2016).

24. Li, W. *et al.* MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15,** 554 (2014).

25. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34,** 184–191 (2016).

26. Ong, S. H., Li, Y., Koike-Yusa, H. & Yusa, K. Optimised metrics for CRISPR-KO screens with second-generation gRNA libraries. *Sci. Rep.* **7,** 7384 (2017).

27. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2016).

28. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

29. McEntyre, J., Sarkans, U. & Brazma, A. The BioStudies database. *Mol. Syst. Biol.* **11,** 847 (2015).

30. Garcia-Alonso, L. *et al.* Transcription factor activities enhance markers of drug response in cancer. *Cancer Research (to Appear)* (2017). doi:10.1101/129478

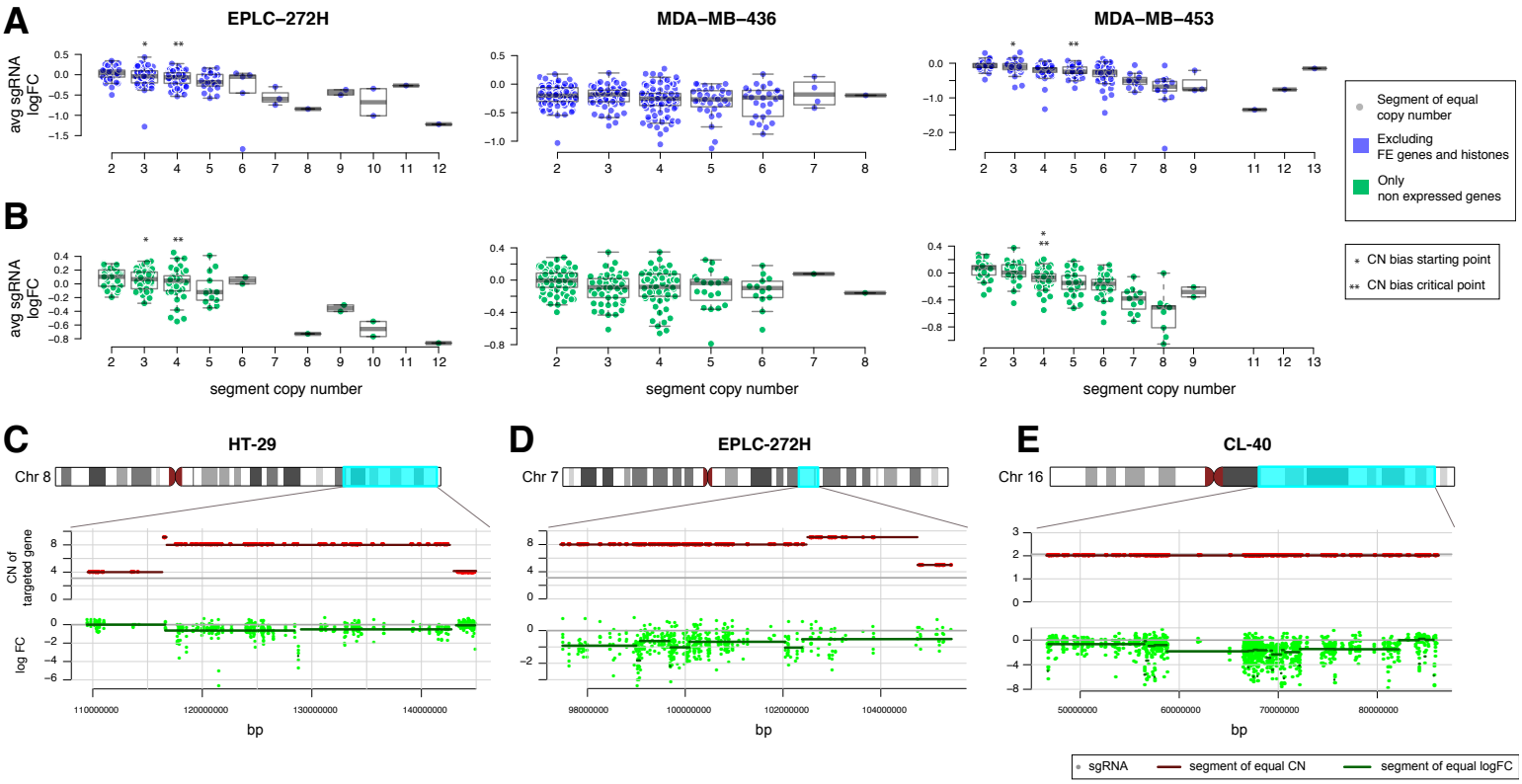
31. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12,** R41 (2011).

32. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607 (2012).

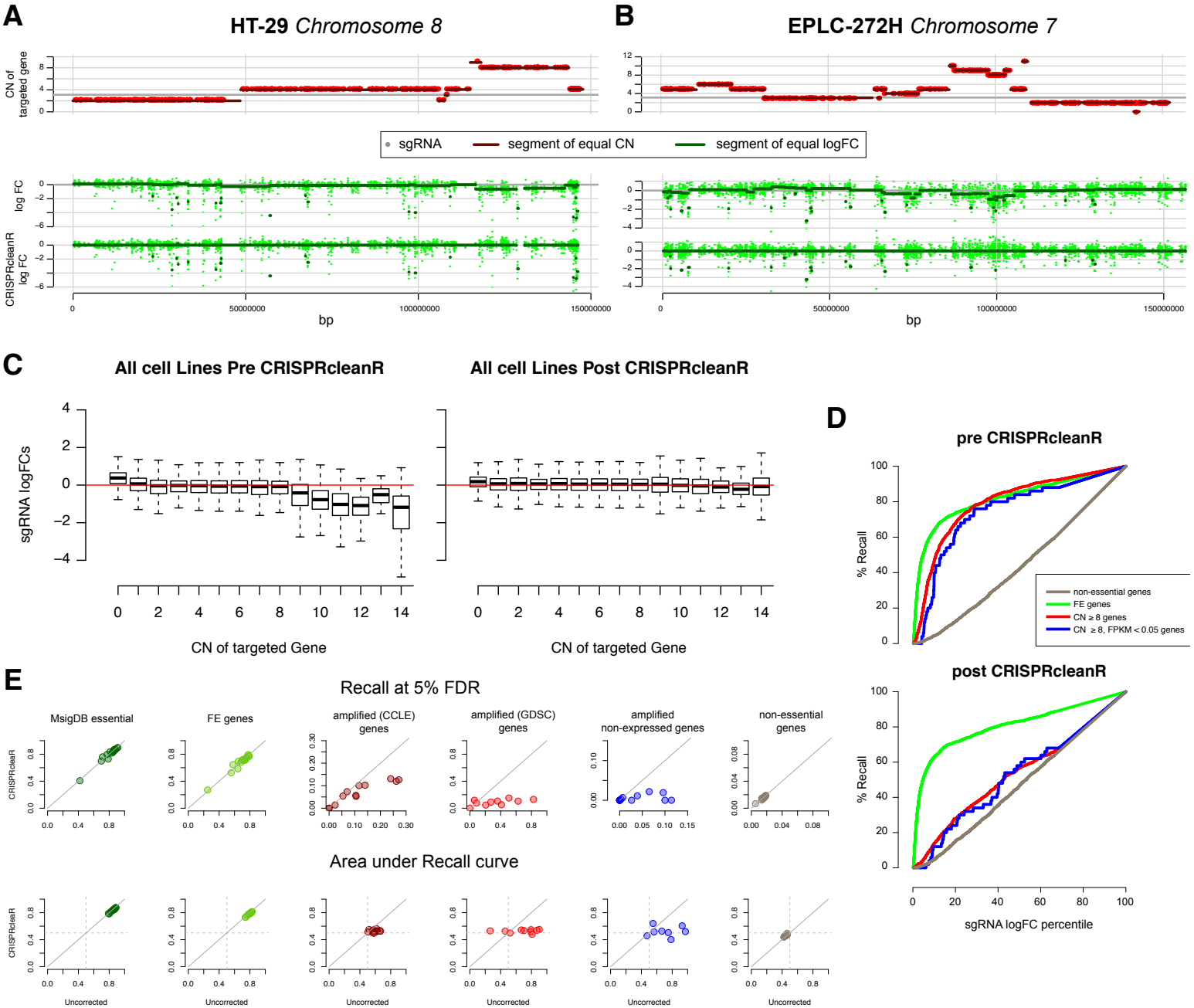
33. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6,** l1 (2013).

34. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12,** 77 (2011).

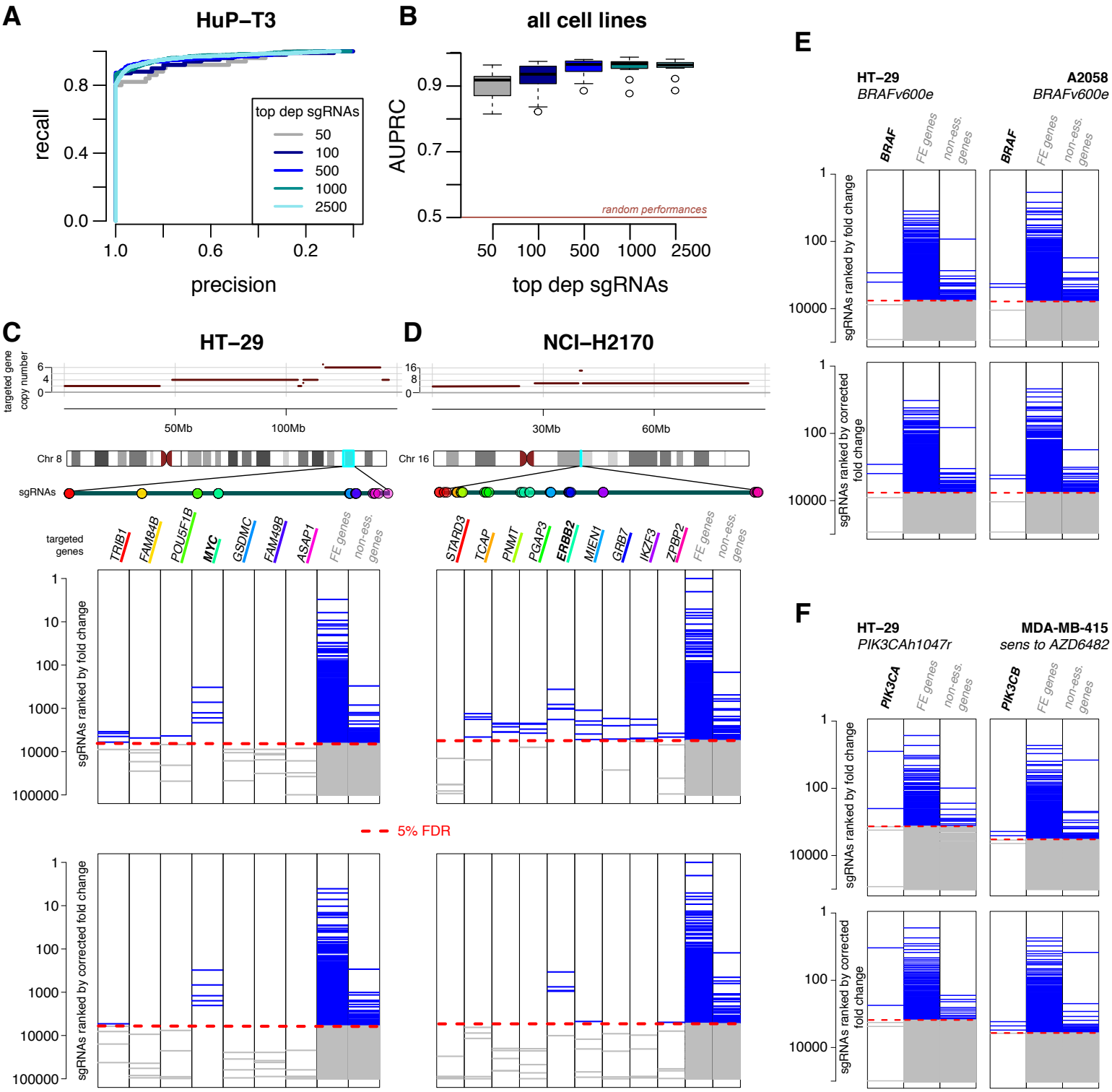
# Figure 1

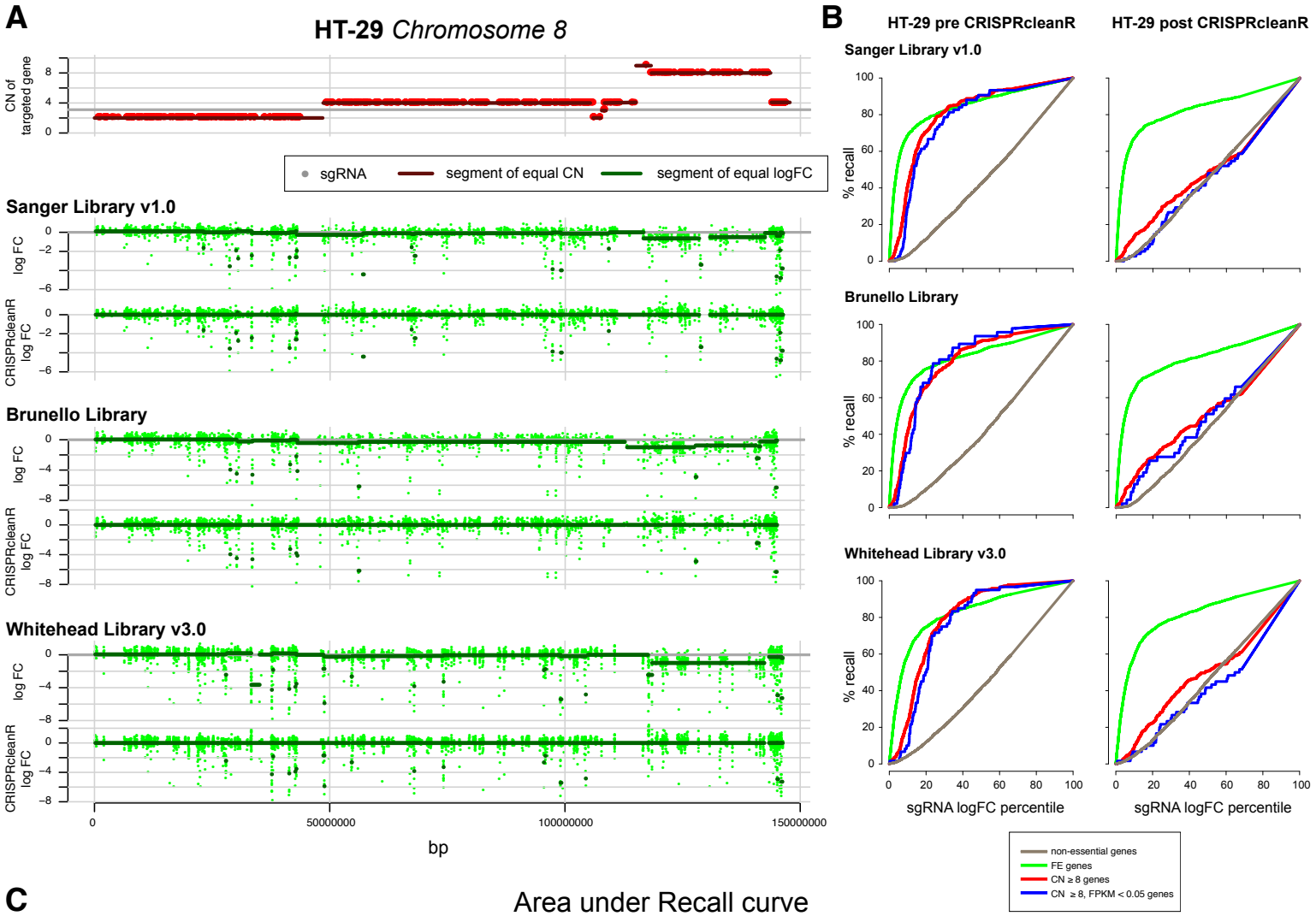

A
EPLC−272H
MDA−MB−436
MDA−MB−453
avg sgRNA logFC
B
segment copy number
Segment of equal copy number
Excluding FE genes and histones
Only non expressed genes
CN bias starting point
CN bias critical point
C
HT-29
Chr 8
CN of targeted gene
log FC
bp
D
EPLC-272H
Chr 7
E
CL-40
Chr 16
sgRNA
segment of equal CN
segment of equal logFC

# Figure 2


A
HT-29 Chromosome 8
B
EPLC-272H Chromosome 7
CN of targeted gene
sgRNA
segment of equal CN
segment of equal logFC
log FC
CRISPRcleanR log FC
bp
C
All cell Lines Pre CRISPRcleanR
All cell Lines Post CRISPRcleanR
sgRNA logFCs
CN of targeted Gene
D
pre CRISPRcleanR
post CRISPRcleanR
% Recall
sgRNA logFC percentile
non-essential genes
FE genes
CN ≥ 8 genes
CN ≥ 8, FPKM < 0.05 genes
E
Recall at 5% FDR
MsigDB essential
FE genes
amplified (CCLE) genes
amplified (GDSC) genes
amplified non-expressed genes
non-essential genes
Area under Recall curve
CRISPRcleaR
Uncorrected

# Figure 3


A
HuP–T3
recall
precision
top dep sgRNAs
50
100
500
1000
2500
B
all cell lines
AUPRC
top dep sgRNAs
random performances
C
HT–29
targeted gene copy number
Chr 8
sgRNAs
targeted genes
TRIB1
FAM84B
POU5F1B
MYC
GSDMC
FAM49B
ASAP1
FE genes
non-ess. genes
sgRNAs ranked by fold change
sgRNAs ranked by corrected fold change
D
NCI–H2170
Chr 16
STARD3
TCAP
PNMT
PGAP3
ERBB2
MIEN1
GRB7
IKZF3
ZPBP2
5% FDR
E
HT–29
BRAFv600e
A2058
BRAFv600e
BRAF
F
HT–29
PIK3CAh1047r
MDA-MB-415
sens to AZD6482
PIK3CA
PIK3CB


A
HT-29 Chromosome 8
CN of targeted gene
sgRNA
segment of equal CN
segment of equal logFC
Sanger Library v1.0
log FC
CRISPRcleanR log FC
Brunello Library
Whitehead Library v3.0
bp
B
HT-29 pre CRISPRcleanR
HT-29 post CRISPRcleanR
% recall
sgRNA logFC percentile
non-essential genes
FE genes
CN ≥ 8 genes
CN ≥ 8, FPKM < 0.05 genes

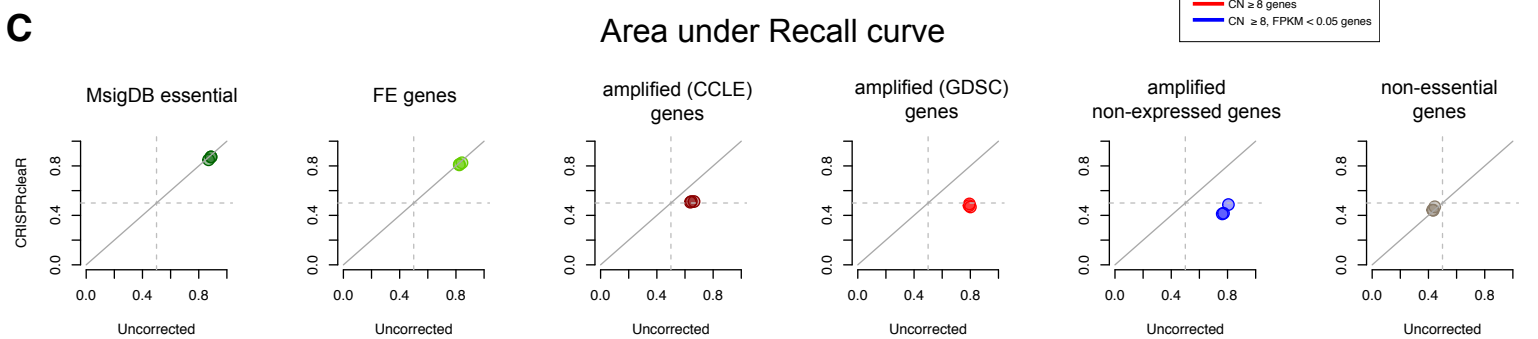

C
Area under Recall curve
MsigDB essential
FE genes
amplified (CCLE) genes
amplified (GDSC) genes
amplified non-expressed genes
non-essential genes
CRISPRcleaR
Uncorrected

**Figure 4**

# Figure 5

**A** **NCI-H2170** **B** **MDA-MB-415**

**Targeted Genes**

sgRNA counts | corrected sgRNA counts | sgRNA counts | corrected sgRNA counts

*fitness essential*

*non essential*

*amplified (CN ≥ 10)*

- *MYC*
- *ERBB2*
- *CCND1*

*amplified non expressed (CN ≥ 10, FPKM < 0.05)*

log2 E(replicate treatments)+1

log2 (plasmid+1)

**C**

## Recall at 10% MAGeCK FDR

MsigDB essential | FE genes | amplified (CCLE) genes | amplified (GDSC) genes | amplified non-expressed genes | non-essential genes

CRISPRcleaR

## Recall at 5% MAGeCK FDR

Uncorrected