Supplementary discussion

Heterogeneity in screening reagents

Heterogeneity of reagents has historically been associated with poor performance in RNAi-based screens[9,10]. That is, while a cell population with a given shRNA should have low phenotypic variability, two different shRNAs targeting the same gene are not expected to have the same knockdown efficiency[10]. How this affects the measured phenotype depends on the relationship between gene-dosage and fitness, which is unknown and unique to each gene[9] (**Supplementary Fig. 4**).

Cas9 screens also suffer from heterogeneity of sgRNA phenotypes[4,10] (**Supplementary Fig. 1-3**). In non-screening formats, this heterogeneity can be controlled for by creating clonal cell lines. Without single-cell selection, a population of cells with a given guide contains a mixture of true knockouts, heterozygotes, and wild-type cells[4,10]; the array of genotypes that exist during the course of a screen likely depends on the efficiency of guide cutting as well as the relative fitness between these subpopulations. Over time the proportion of each subpopulation, and thus the measurable fitness, will change as on-going knockouts remove alleles and selection favors particular subpopulations. Even sgRNAs with high cutting efficiency generate a significant number of functional alleles via in-frame indels[28,36]. This variability in both RNAi and Cas9 screens precludes a simple phenotype to measure for a given gene. Thus, to determine whether a gene is involved in a given process, we can instead measure the maximum phenotype possible, which may not correspond to the highest degree of knockdown or the null phenotype.

casTLE accounts for reagent heterogeneity

In order to calculate this maximum phenotype for each gene using a clear statistical framework, we have developed cas9 high-Throughput maximum Likelihood Estimator (casTLE) (**Supplementary Fig. 4**), which we can use for direct comparison of shRNA and CRISPR/Cas9 screens. For each gene, casTLE gives an effect size estimate for the gene perturbation as well as a p-value associated with that effect, combining the advantages of previous approaches. Distribution-based methods, such as MW, RIGER, and RSA, provide robust statistical testing by comparing the distribution of phenotypes for elements targeting a given gene to the distribution of phenotypes from non-targeting control elements or all other elements[1,19,20,37]; however, these

methods do not give a readily interpretable estimate of gene effect. Alternative, heuristic approaches have easily interpretable effect size estimation, for example by taking the phenotype of the single most active shRNA/sgRNA as a proxy for the phenotype of the gene or the median phenotype[23,37–39]; however, these methods lack a statistical framework to assign confidence to a hit. casTLE uses a semi-parametric approach to provide both a statistical framework and biologically-interpretable effect sizes. Intuitively, casTLE estimates the maximum possible phenotype such that targeting elements are most likely to be found between this phenotype and zero (**Supplementary Fig. 2, 3** and see also **Supplementary Methods**). Additionally, casTLE allows data from multiple screen types or from replicates of the same screen type to be compared and combined by finding a single effect size consistent with all data. Here we apply casTLE to compare the abilities of shRNA and Cas9 screens to identify essential genes, though the framework should be widely applicable to detect positive and negative effects for other phenotypes.

Validation of casTLE

In order to validate casTLE as a broadly applicable tool, we used it to re-analyze data from several published screens. These include an shRNA screen for modulators of ricin toxicity in K562 cells[1], a CRISPR/Cas9 deletion screen for LPS-induced TNF expression in primary mouse dendritic cells[16], and both CRISPRi and CRISPRa screens for modulators of the fusion toxin CTx-DTA[17] (**Supplementary Data 5-7**). For each of these, casTLE produced results broadly consistent with previous findings, with even higher correlation for validated hits from each screen when available (**Supplementary Fig. 5)**. Consistent with the previously validated results of the TNF CRISPR/Cas9 deletion screen, 18 of the 20 top depleted hits identified by casTLE in the CRISPR/Cas9 deletion screen were identified in the previous study, with 16 successfully validating and two failing to validate[16]. While the previous study reported difficulties in detecting enriched hits, we find that casTLE identified positive regulators of TNF expression (**Supplementary Fig. 5b**), and that the top enriched hit, TNFRSF9, has been recently implicated in TNF regulation[40]. These analyses of previous data demonstrate casTLE can be widely used across screening technologies and phenotypes, simultaneously detecting positive and negative regulators.

All individual analyses of replicates of the shRNA and CRISPR/Cas9 screens show that

casTLE performs well in the detection of essential genes (AUC of the ROC curve > 0.91). In addition, the same data was analyzed with the previously published statistical tools MAGeCK[18], Mann-Whitney[1], RSA[19], RIGER[20], and HiTSelect[21] as well as two commonly used heuristics, the highest effect and the median effect (**Supplementary Fig. 6, Supplementary Data 2-4**). Although some algorithms performed better in the high-error region, casTLE performs better than or on par with existing methods in the low-error region, which is most relevant for the selection of top hits for follow-up analysis. This quantitative analysis, along with the broad agreement with previous screen data, establishes casTLE as a valid tool for screen analysis.

SUPPLEMENTARY METHODS

1. CASTLE

We built cas9 high-Throughput maximum Likelihood Estimator (casTLE) that uses an Empirical Bayesian framework to account for multiple sources of variability. For each gene, we have the phenotypes of multiple targeting reagents, measured as a median normalized log ratio of counts. From this, and the phenotypes of negative controls, we obtain an effect size estimate for each gene and an associated log likelihood ratio. By shuffling targeting reagents, we can generate an expected negative distribution of log likelihood ratios, allowing hypothesis testing.

For a given gene, i, we have a set of elements, $j \in S_i$, each of which has an observed enrichment γ_{ij}. We can then define the relationship between the true phenotype of the element, ξ_{ij} and the observed enrichment, $P(\gamma_{ij}|\xi_{ij})$, by taking advantage of the non-targeting controls, which represent both measurement noise and off-target effects. We can fit this distribution with a Gaussian kernel as $N(\gamma)$ and use the shifted distribution $P(\gamma_{ij}|\xi_{ij}) = N(\gamma_{ij} - \xi_{ij})$.

The distribution of true effects, ξ_{ij}, for a given gene, i, is bounded by a maximum effect, I_i, and by zero. Due to ineffective reagents, a certain percentage $1 - \theta_i$ of true effects will be zero. The effective fraction θ_i is assumed to be uniformly distributed between I_i and 0. This gives us a distribution of true effects for a given gene parametrized by a maximum effect I_i and fraction effective θ_i.

Together, this gives an Empirical Bayesian framework, where our observables, γ_{ij} are drawn from the distribution $P(\gamma_{ij}|\xi_{ij}) = N(\gamma_{ij} - \xi_{ij})$, depending on the unknown, element-specific parameter ξ_{ij}, which itself is distributed according to the gene-specific hyperparameters I_i and θ_i. It is these hyperparameters we need to estimate, which we can do with a Maximum Likelihood approach.

The probability of observing γ_{ij} given I_i and θ_i can be separated into three regions. Without loss of generality, let $I_i > 0$. With probability θ_i, ξ_{ij} was drawn from a uniform distribution between $[0, I_i]$. If $\gamma_{ij} < 0$, then the most likely estimate of ξ_{ij} is 0. If $\gamma_{ij} \in [0, I_i]$, then the most likely estimate of ξ_{ij} is γ_{ij}. If $\gamma_{ij} > I_i$, then the most likely estimate of ξ_{ij} is I_i. In all regions, there is a $(1 - \theta_i)$ probability that $\xi_{ij} = 0$. Combining these gives us the probability of observing our data:

$$Pr(\gamma_{ij}|I_i,\theta_i) = \begin{cases} (1-\theta_i)N(\gamma_{ij}) + \theta_i N(\gamma_{ij}-0) & : \gamma_{ij} < 0 \\ (1-\theta_i)N(\gamma_{ij}) + \theta_i N(\gamma_{ij}-\gamma_{ij}) & : 0 \leq \gamma_{ij} \leq I_i \\ (1-\theta_i)N(\gamma_{ij}) + \theta_i N(\gamma_{ij}-I_i) & : \gamma_{ij} > I_i \end{cases}$$

or

$$Pr(\gamma_{ij}|I_i,\theta_i) = \begin{cases} N(\gamma_{ij}) & : \gamma_{ij} < 0 \\ (1-\theta_i)N(\gamma_{ij}) + \theta_i N(0) & : 0 \leq \gamma_{ij} \leq I_i \\ (1-\theta_i)N(\gamma_{ij}) + \theta_i N(\gamma_{ij}-I_i) & : \gamma_{ij} > I_i \end{cases}$$

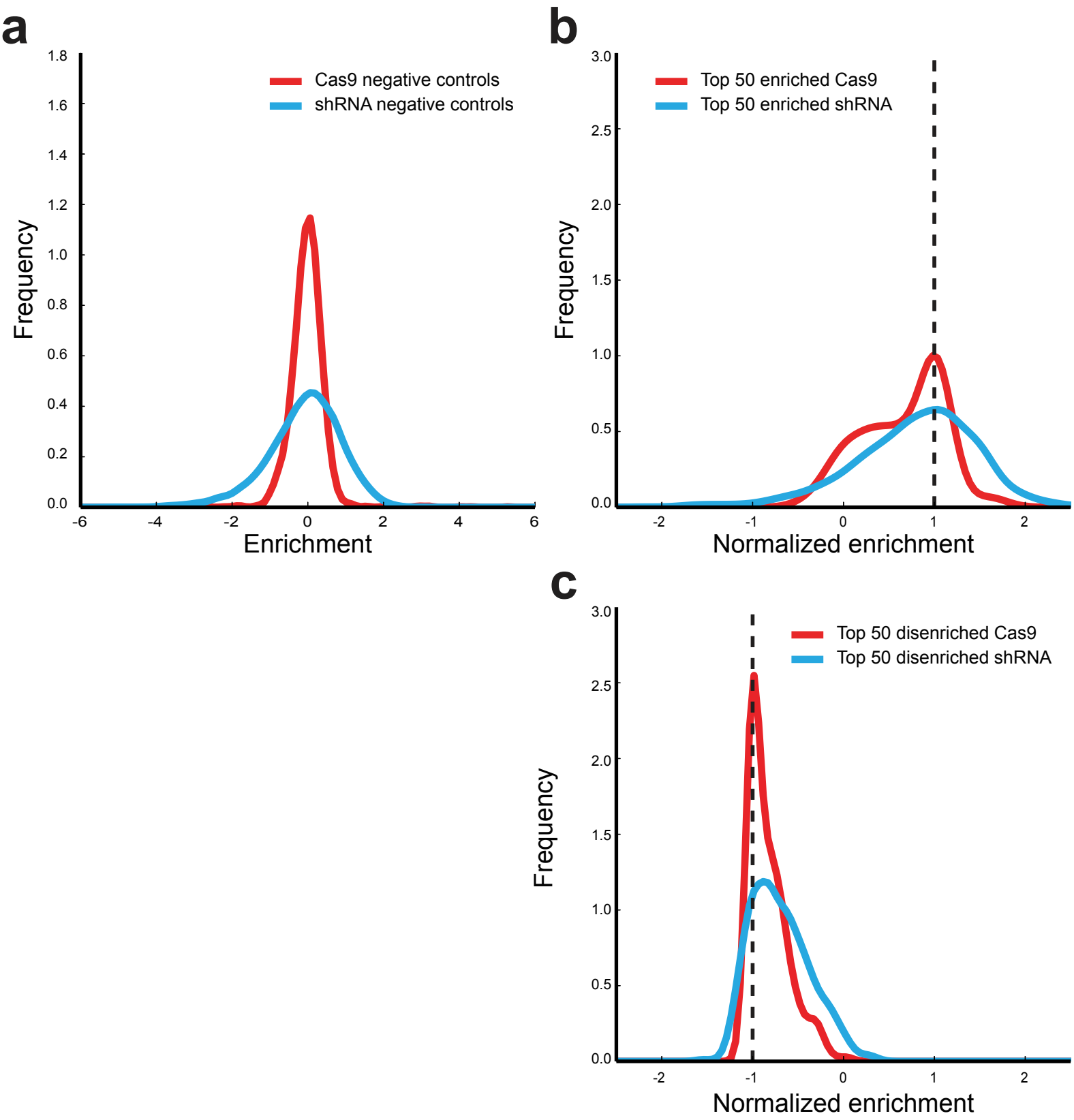
A grid search is performed over possible values of θ_i and I_i, and the likelihood of the model is calculated with the above probability function. θ_i is then marginalized to calculate a maximum likelihood estimate $\hat{I}_i$ and an associated 95% credible interval.

For hypothesis testing, a likelihood ratio test is performed against the following null model:

$$Pr(\gamma_{ij}|0,0) = N(\gamma_{ij})$$

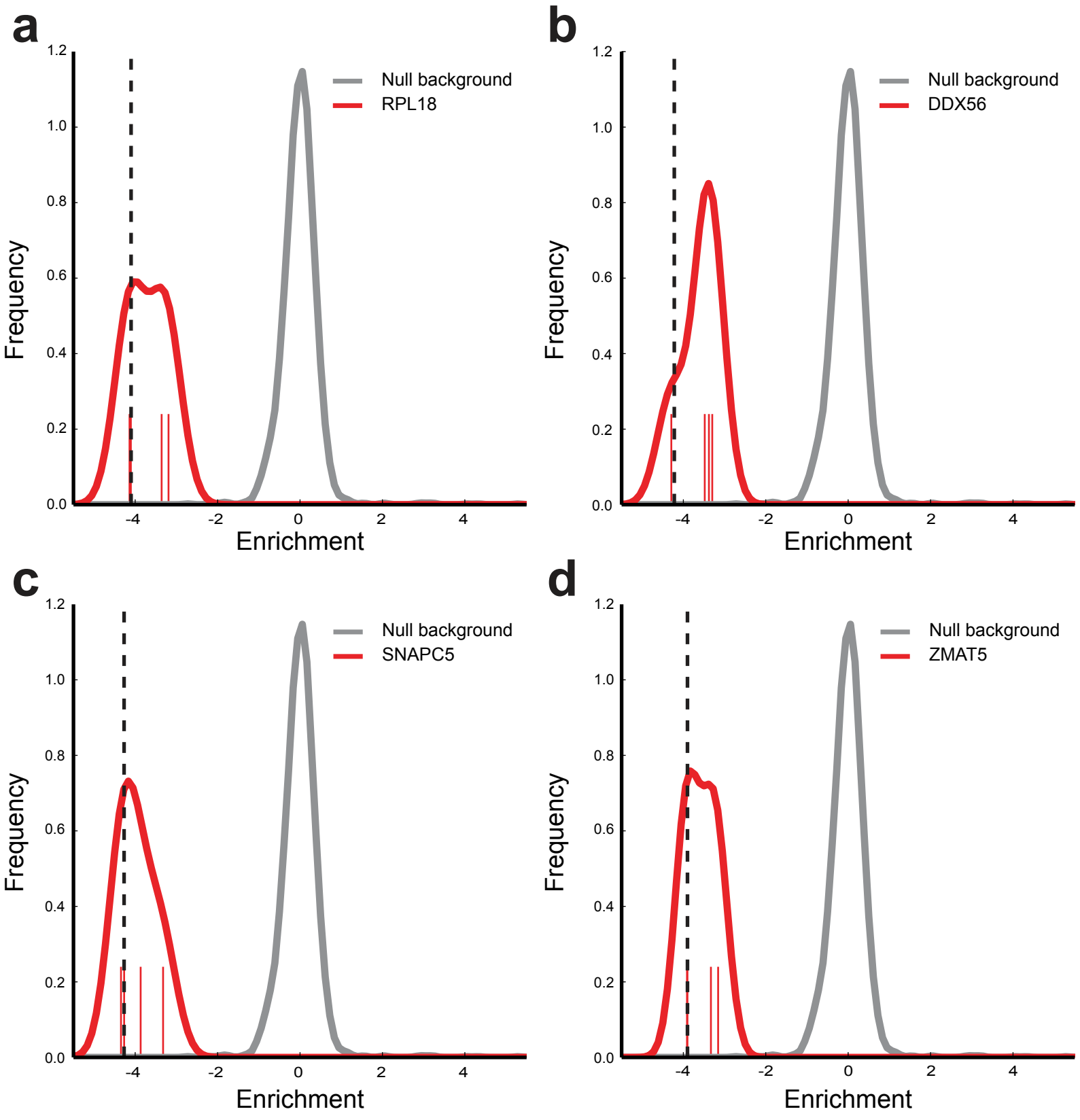
The distribution of the log-likelihood ratio is estimated by randomly drawing targeting elements and calculated the log-likelihood ratio as above. To combine data from replicates or from disparate screens, the grid search is performed to find the likelihood of (I_i, θ_{i1}) and (I_i, θ_{i2}). Both θ_{i1} and θ_{i2} are then marginalized to find the most likely I_i.

Supplementary Figure 1



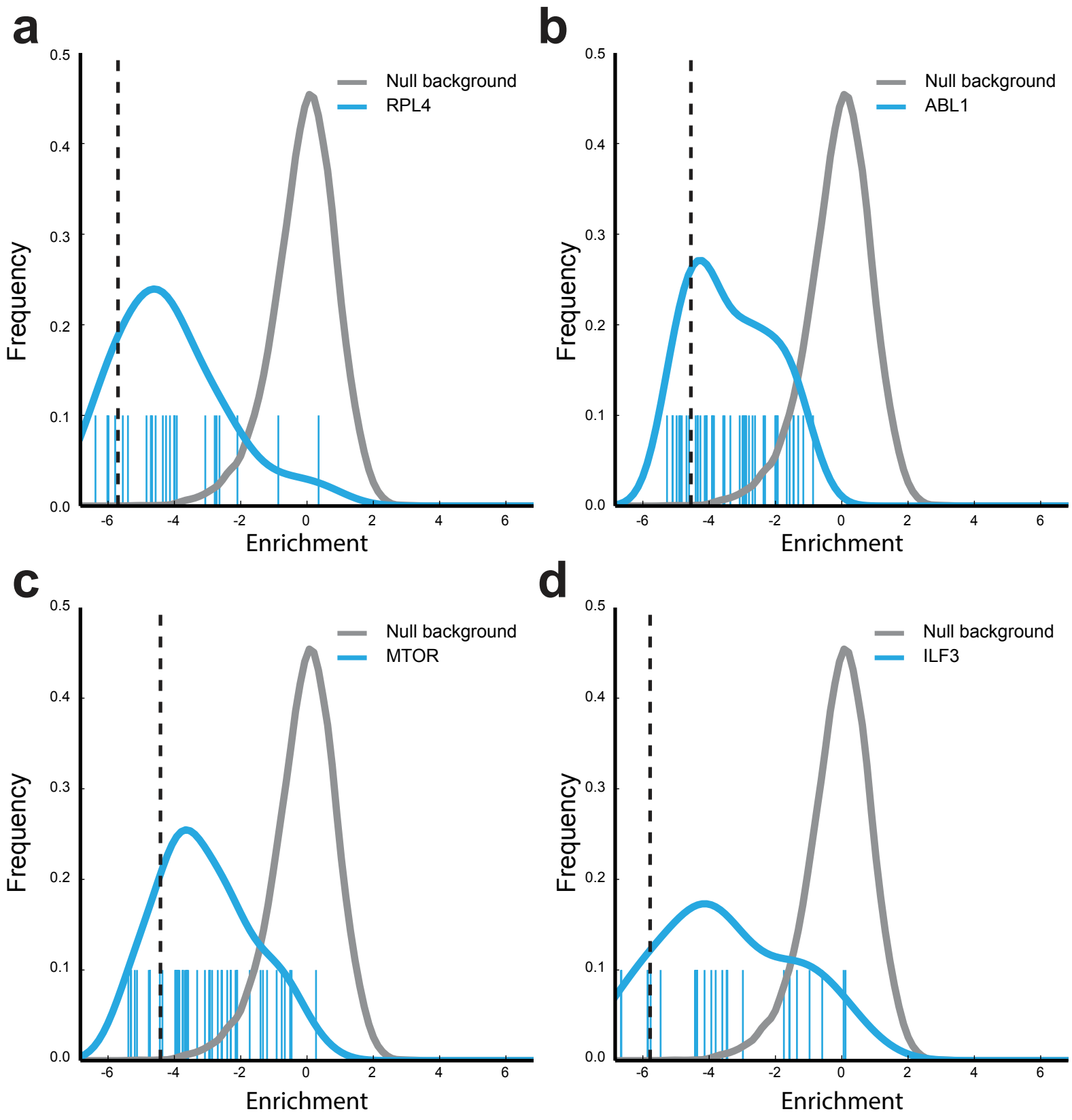
Supplementary Figure 1. Distribution of targeting and control elements. (a) Distribution of negative controls for a single replicate of Cas9 and shRNA screens. Enrichments are calculated as a median-normalized log ratio of counts. (b,c) Distribution of targeting elements is shown in meta-gene plots for the top 50 (b) enriched and (c) disenriched genes found in a single replicate of the Cas9 and shRNA screens as identified by casTLE. To normalize, the enrichment of each individual element was divided by the effect size estimate for the gene generated by casTLE. The dotted line is placed at the estimated effect size and normalized to one.

Supplementary Figure 2



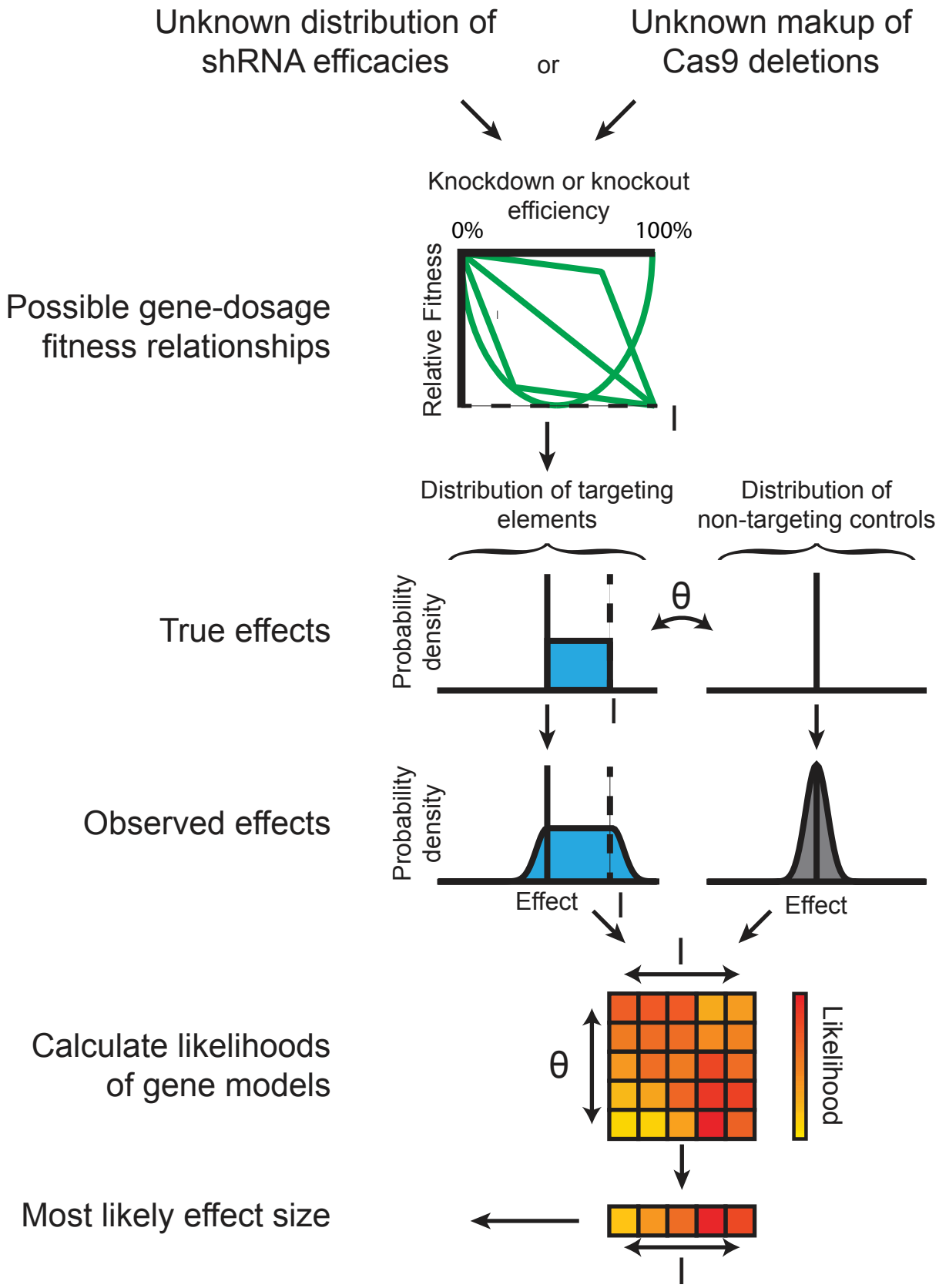
Supplementary Figure 2. Distribution of targeting sgRNAs for top disenriched genes. (a-d) Enrichment of targeting elements and estimated effect size is shown for the top four disenriched genes from Cas9 data from a single replicate. Enrichments are calculated as a median-normalized log ratio of counts. Gray lines represent the smoothed distribution of non-targeting controls. Red vertical lines represent enrichment of individual targeting guides towards indicated genes. Vertical dotted line represents effect size estimate from casTLE. Red distribution is a smoothed distribution of guides targeting the genes indicated.

Supplementary Figure 3



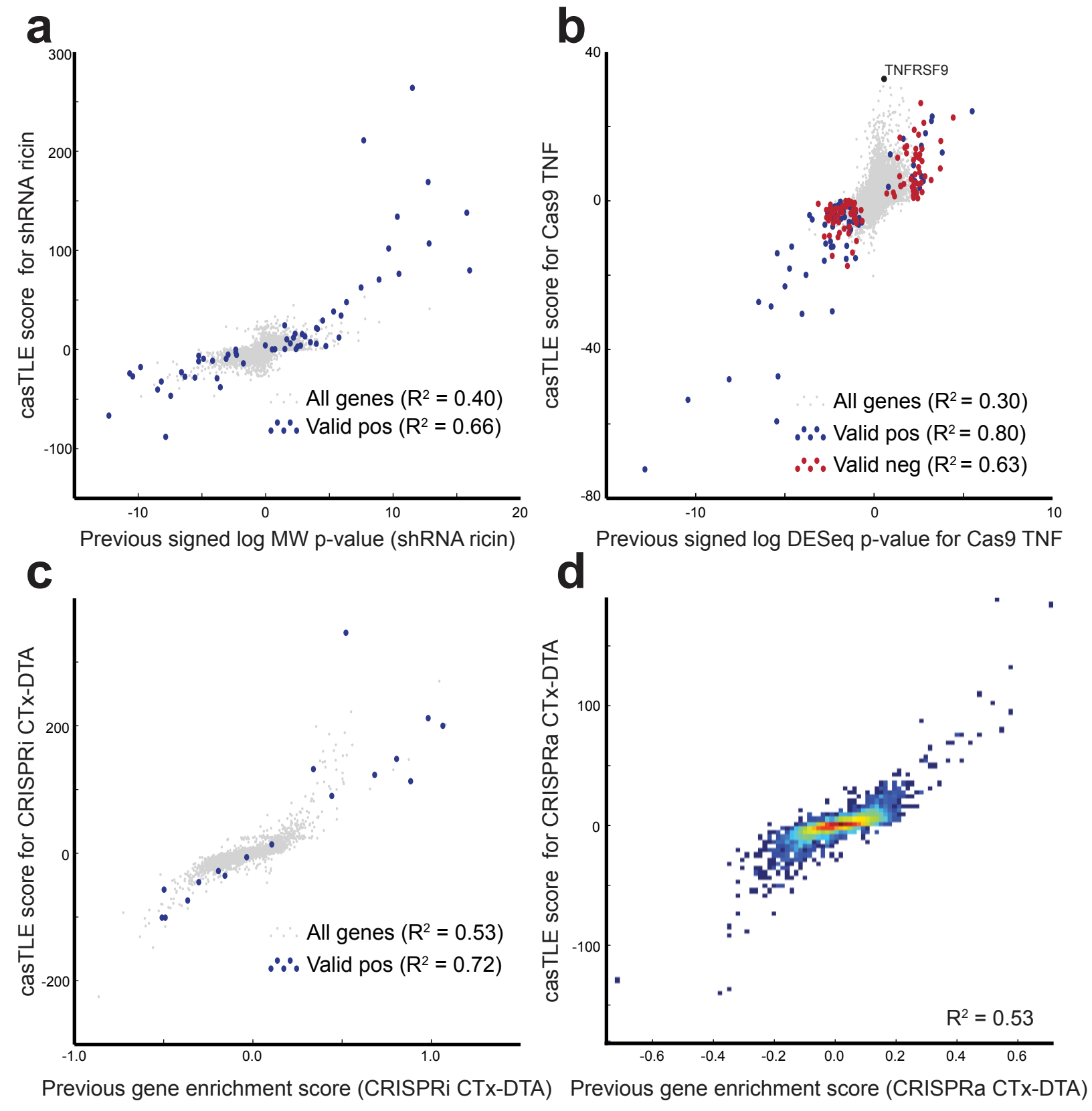
Supplementary Figure 3. Distribution of targeting sgRNAs for top disenriched genes. (a-d) Enrichment of targeting elements and estimated effect size is shown for the top four disenriched genes from shRNA data from a single replicate. Enrichments are calculated as a median-normalized log ratio of counts. Gray lines represent the smoothed distribution of non-targeting controls. Blue vertical lines represent enrichment of individual targeting hairpins towards indicated genes. Vertical dotted line represents effect size estimate from casTLE. Blue distribution is a smoothed distribution of hairpins targeting the genes indicated.

Supplementary Figure 4



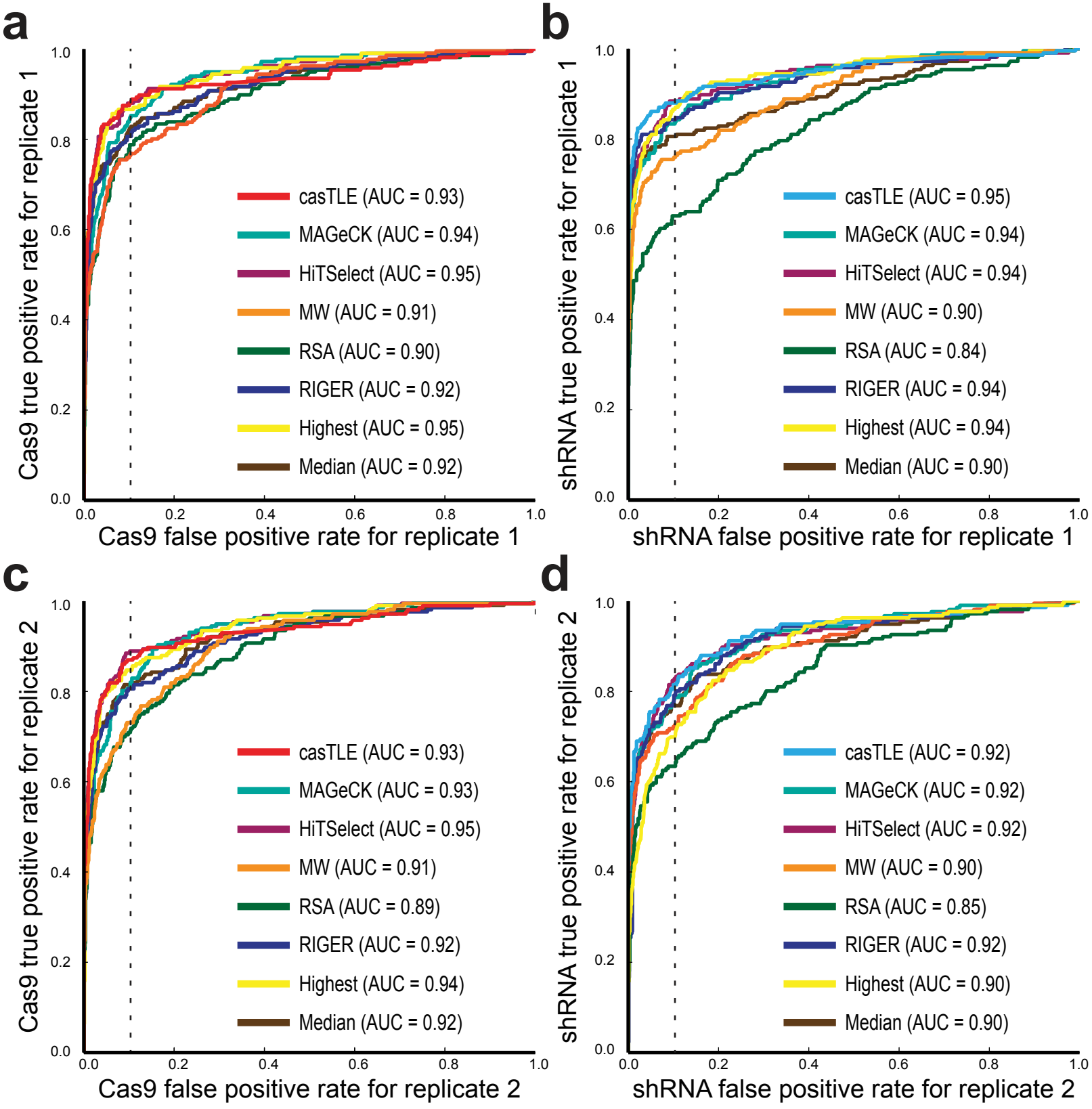
Supplementary Figure 4. casTLE provides a statistical framework to account for high-throughput screens. The unknown relationship between gene dosage and measured phenotype as well as the unknown distribution of shRNA and Cas9 efficacies restricts the predicted effect size of reagents to a bounded region, marked as the blue shaded region, between 0 and the maximum effect I, marked by the dotted line. Some fraction (1-θ) of the reagents have no on-target effect at all. The phenotype observed is thus the true effect obscured by noise, which is estimated using the distribution of non-targeting controls. The likelihood of models for different values of I and θ are calculated and by marginalizing θ the most likely effect size is selected. A likelihood ratio is then calculated by comparing to a null model where I is zero.

Supplementary Figure 5



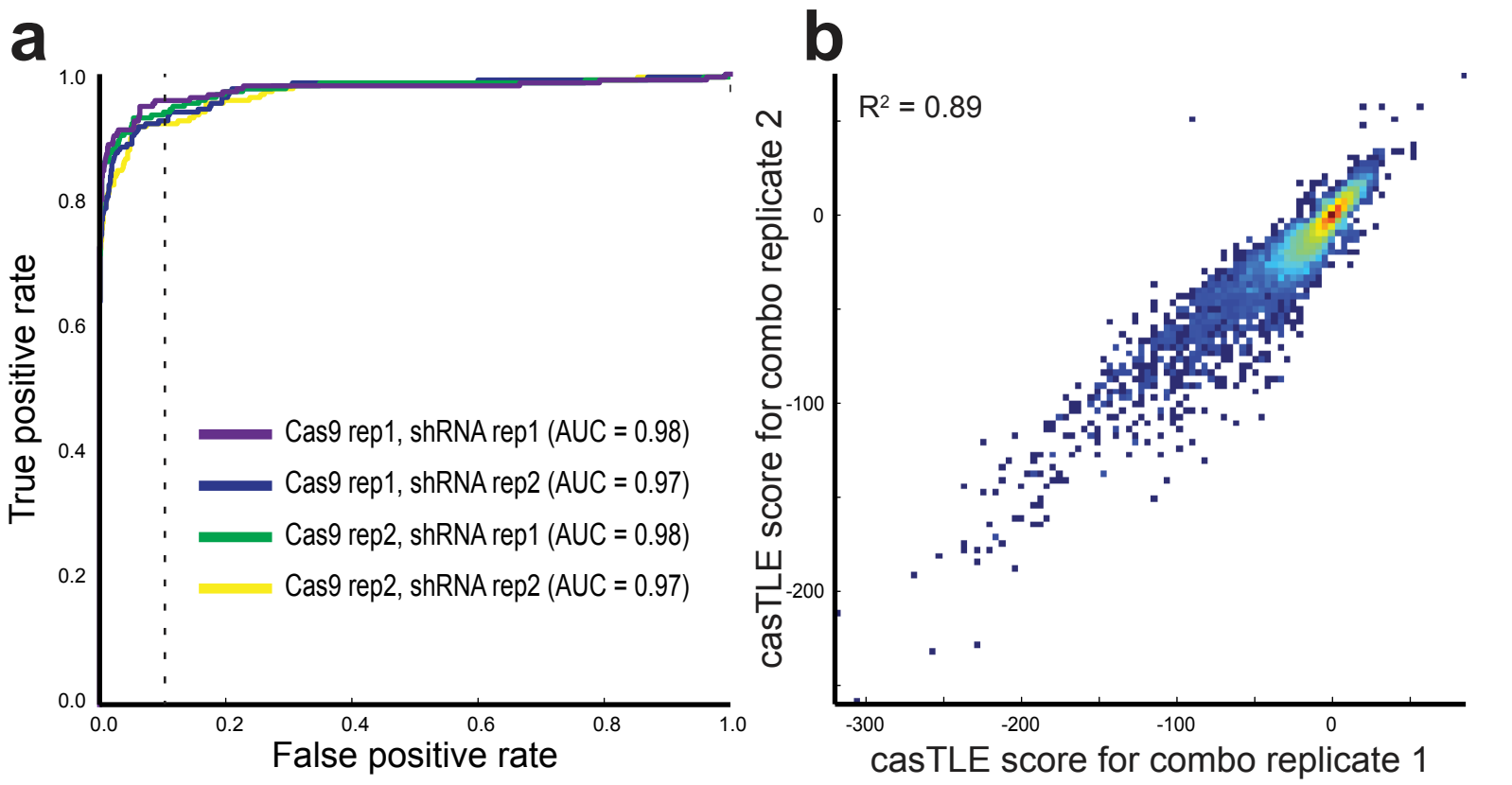
Supplementary Figure 5. Reanalysis of previous screens. (a) Results are shown for a previously published shRNA screen for ricin sensitivity reanalyzed with casTLE and compared to published results based on a MW test[1]. (b) Previous CRISPR/Cas9 deletion screen for LPS-induced TNF expression in primary mouse bone-marrow derived dendritic cells, analyzed with casTLE and the published DESeq results[16]. (c) Previous CRISPRi screen for sensitivity to the fusion toxin CTx-DTA, analyzed with casTLE versus the average of the top three sgRNA effects[17]. (d) Previous CRISPRa screen for sensitivity to the fusion toxin CTx-DTA, analyzed with casTLE versus the average of the top three sgRNA effects[17].

Supplementary Figure 6



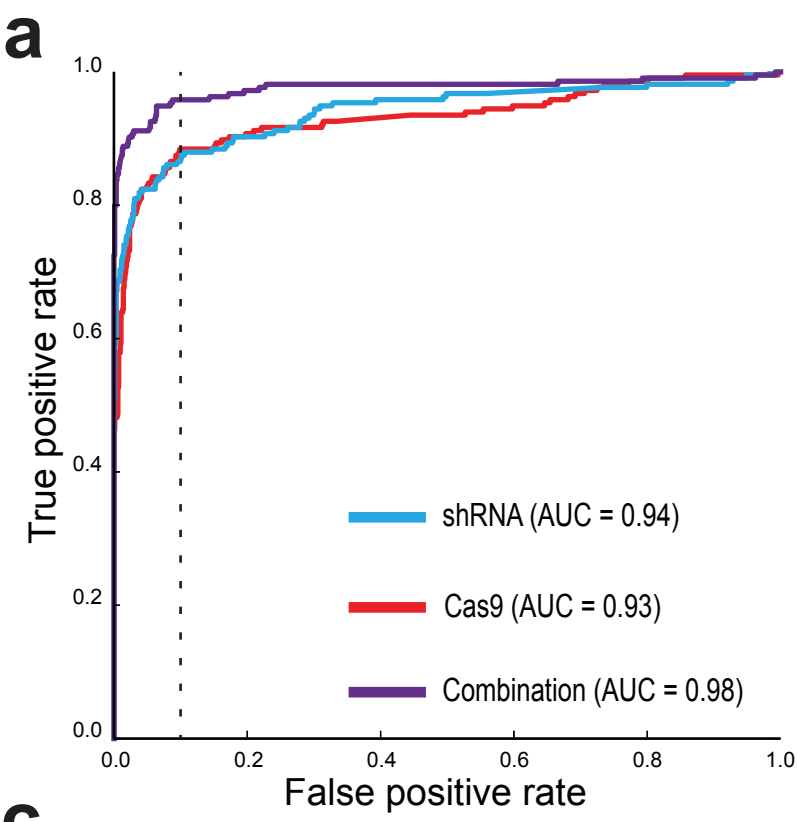
Supplementary Figure 6. Comparison of casTLE to other methods. (a-d) ROC curves indicate screen performance in identifying essential genes from changing composition between the plasmid library and two weeks growth. True positive rates and false positive rates are calculated using a previously established gold standard set of essential and nonessential genes[15]. Genes are ranked by likelihood to be essential using the indicated methods, including casTLE. Highest effect heuristic was calculated by ranking the genes according to their most disenriched element. Data is shown from single replicates of the (a,c) Cas9 and (b,d) shRNA screens for (a,b) replicate 1 and (c,d) replicate 2.

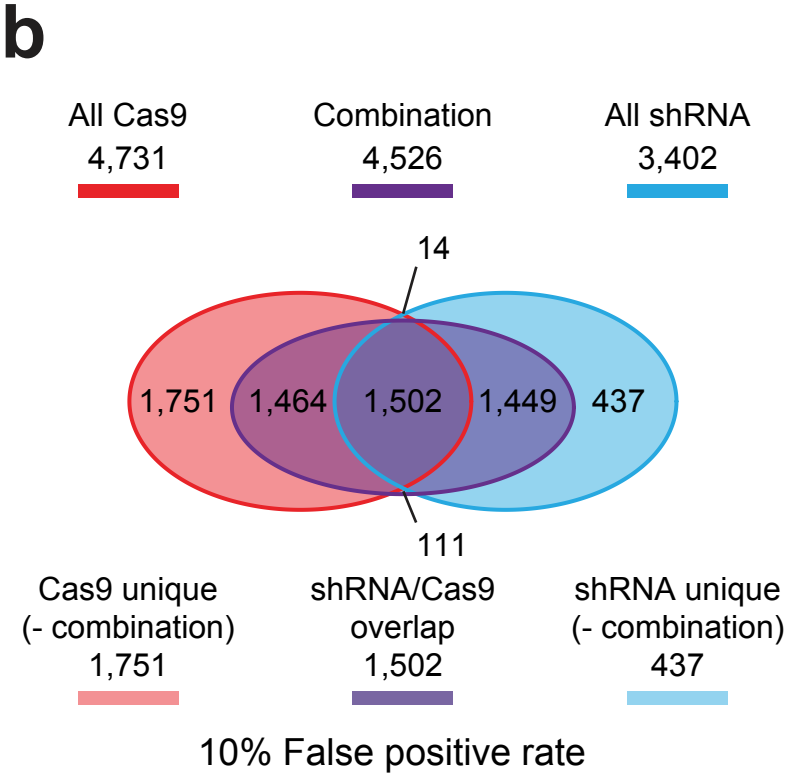
Supplementary Figure 7

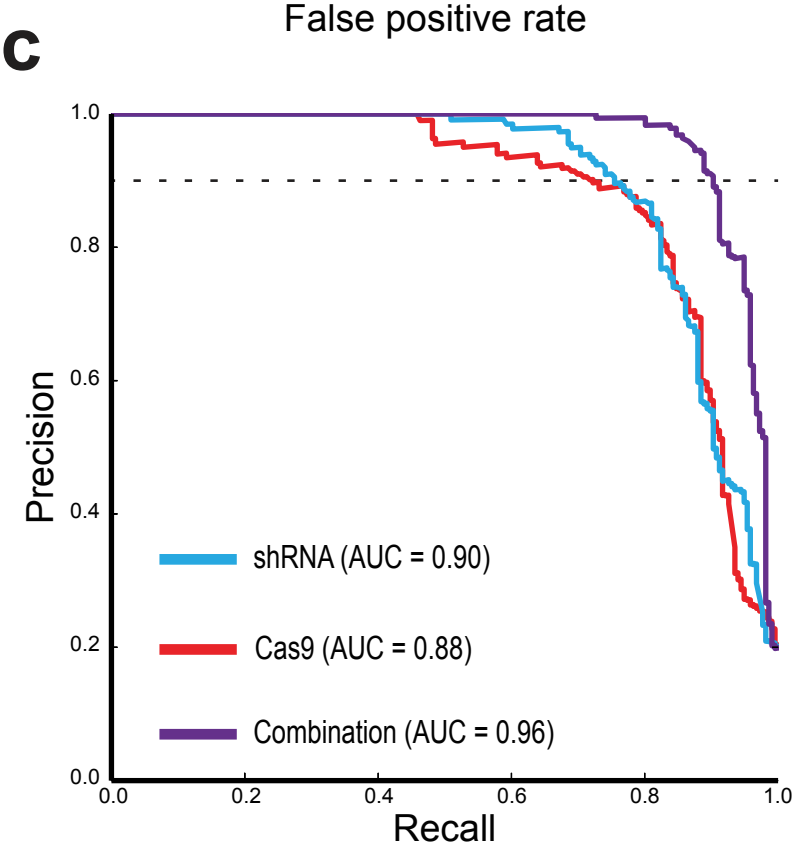


Supplementary Figure 7. Performance of combination of shRNA and Cas9 data. (a) ROC curves from combination of different replicates of Cas9 and shRNA using casTLE. ROC curves indicate screen performance in identifying essential genes from changing composition between the plasmid library and two weeks growth. True positive rates and false positive rates are calculated using a previously established gold standard set of essential and nonessential genes[15]. (b) Combination score has high reproducibility. A large positive casTLE score indicates a high confidence increase in growth rate, while a highly negative casTLE indicates a high confidence decrease in growth rate, i.e. gene essentiality. The graphs compare replicate measurements of likelihood ratio between plasmid and T14 of the combination score based on replicates 1 for Cas9 and shRNA and replicates 2 for Cas9 and shRNA. Density is in log scale.

Supplementary Figure 8

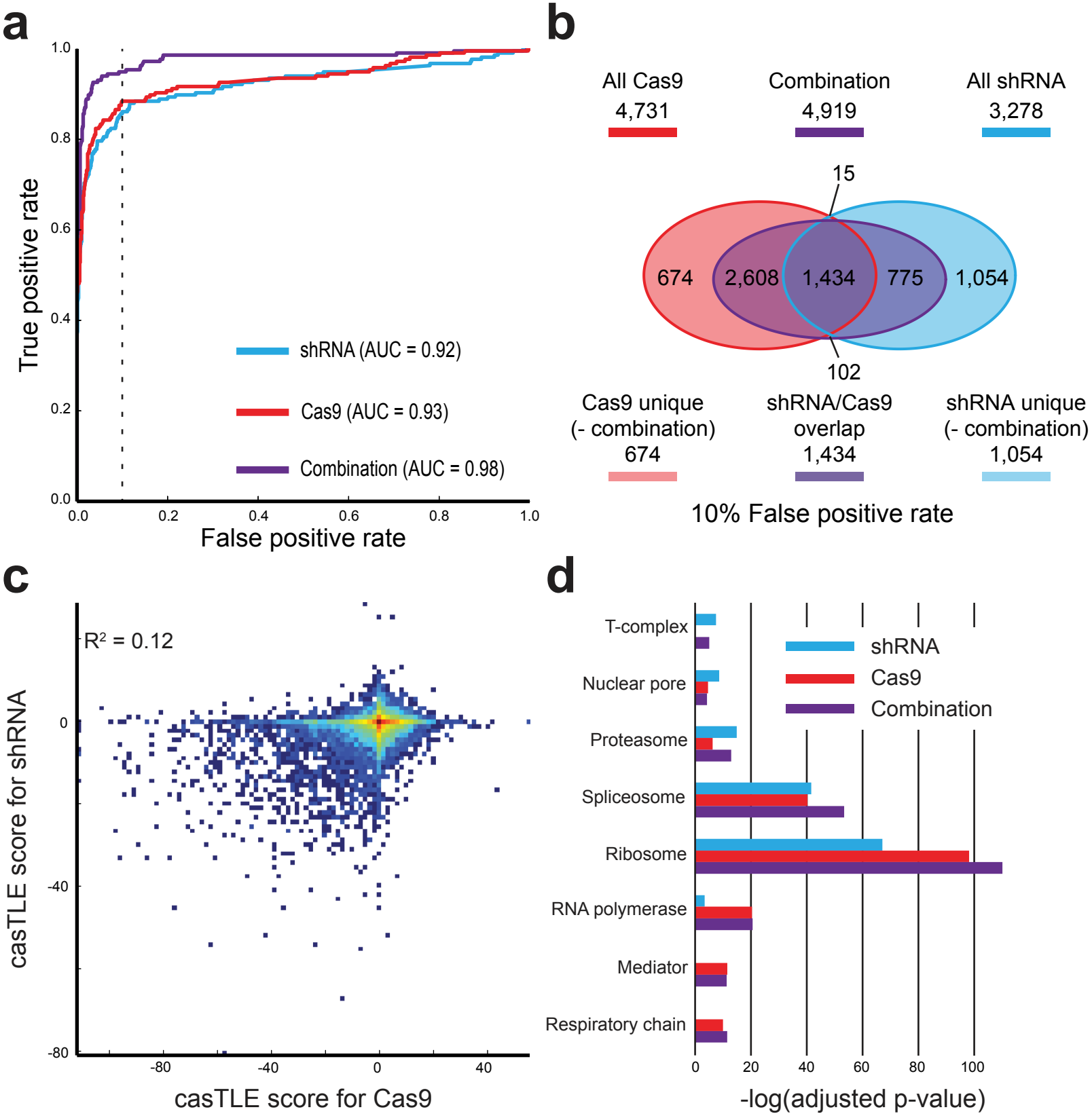






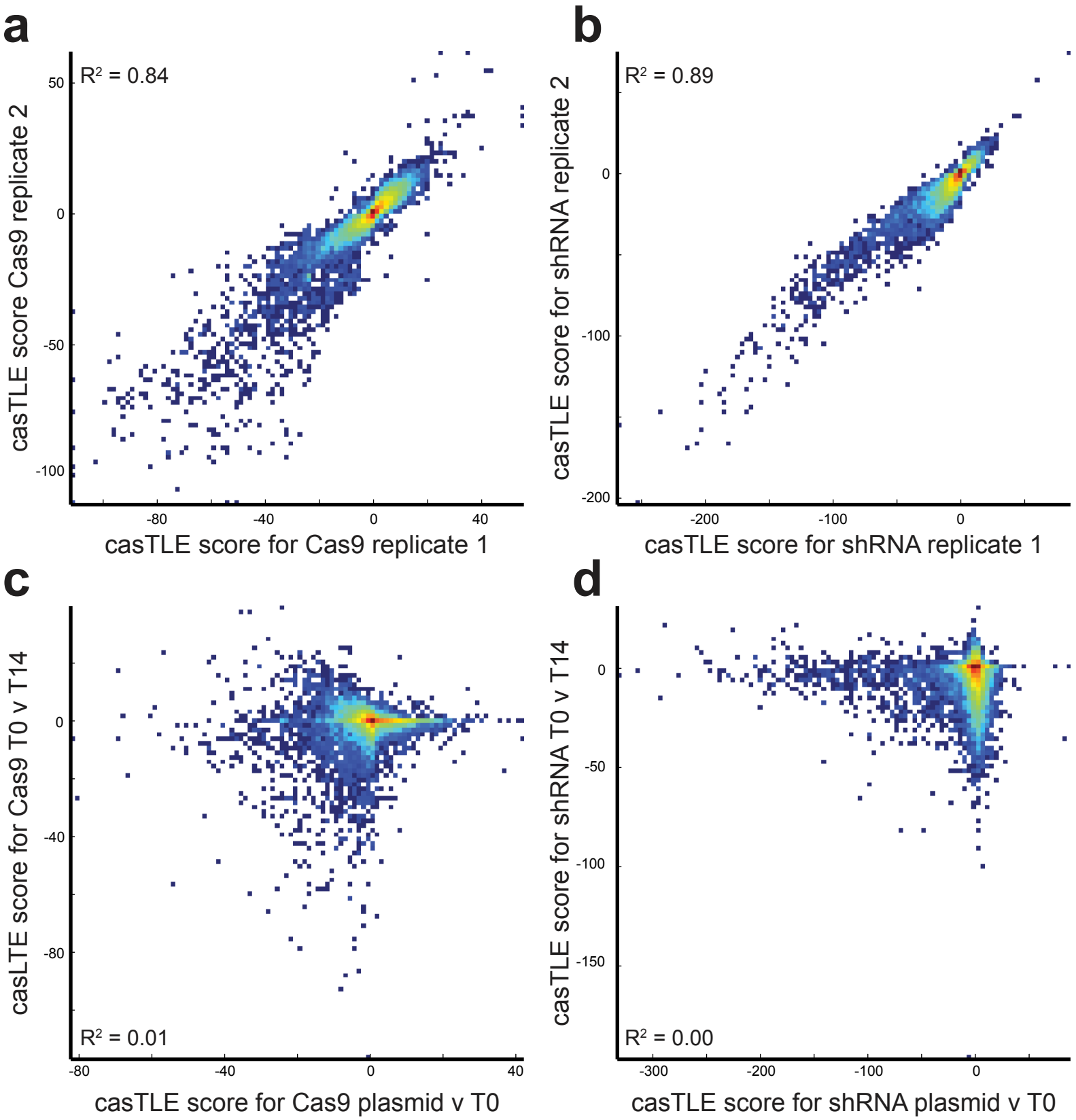
Supplementary Figure 8. Comparison of casTLE combination to casTLE analysis of single screens. (a) ROC curves indicate screen performance in identifying essential genes by comparing the library composition between the plasmid library and cells after two weeks growth. ROC curves for Cas9 (red) and shRNA (blue) screens based on duplicate data combined using casTLE. Alternatively, data from single replicates of both Cas9 and shRNA screens were combined using casTLE (purple). (b) The number of essential genes at 10% false positive rate and their overlap based on the duplicate data from Cas9 and shRNA screens, as well as combination of a single replicate from both screens. False positive rate was estimated using gold standard nonessential genes. (c) Precision recall curve for Cas9, shRNA, and combination data using casTLE.

Supplementary Figure 9



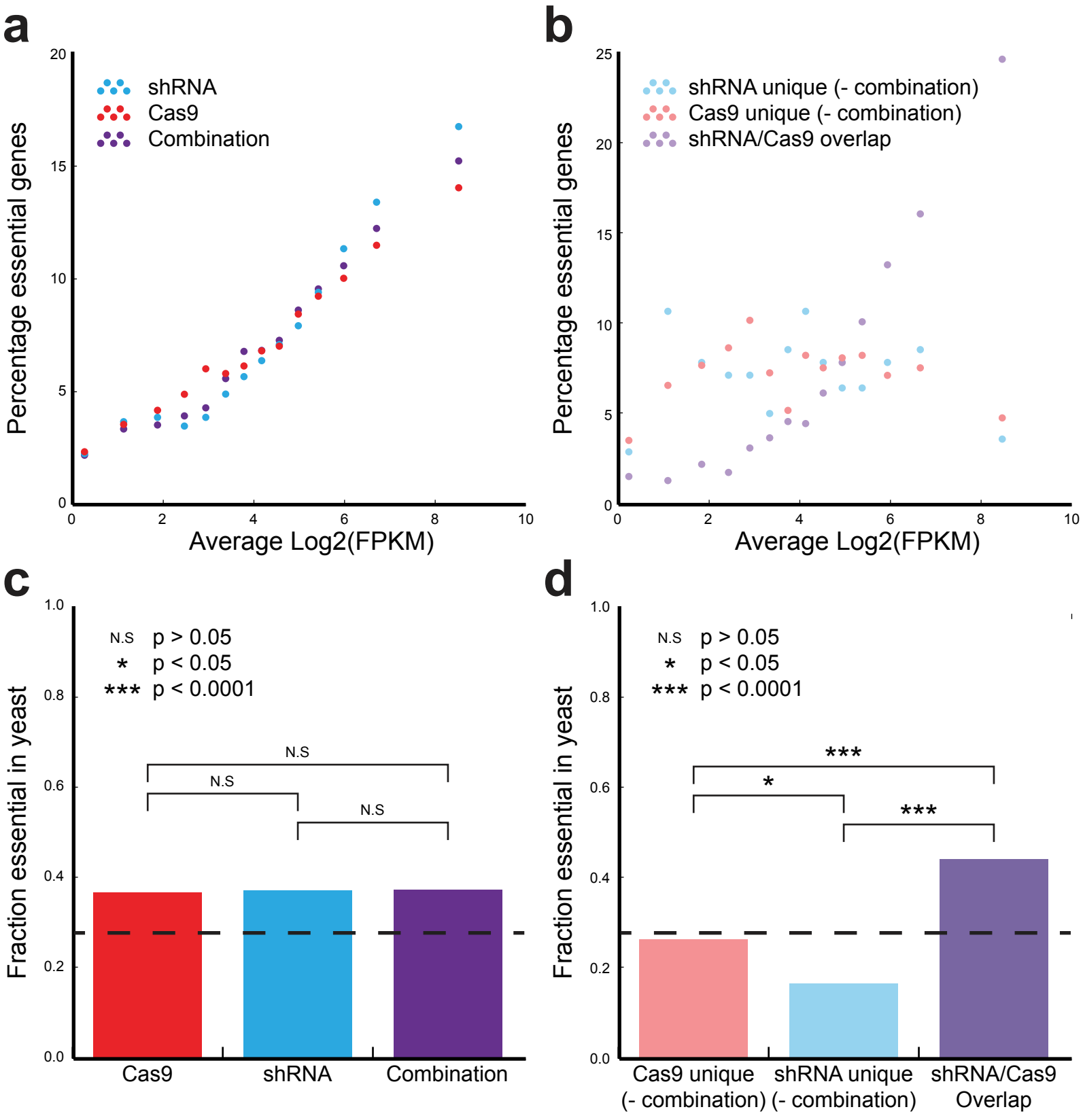
Supplementary Figure 9. Comparison to an in silico 4 shRNA per gene library. Results from the 25 shRNA library were downsampled by only including four hairpins per gene, selected by previous computational ranking. (a) ROC curves indicate screen performance in identifying essential genes by comparing the library composition between the plasmid library and cells after two weeks growth. (b) The number of essential genes at 10% false positive rate and their overlap based on the duplicate data from Cas9 and shRNA screens, as well as combination of a single replicate from both screens. (c) Comparison of casTLE scores derived from casTLE between single replicates of Cas9 and shRNA data. (d) Adjusted p-values for select GO terms for shRNA and Cas9 screens as well as for data from both screens combined with casTLE.

Supplementary Figure 10



Supplementary Figure 10. Screen reproducibility and time-dependence of phenotypes. (a,b) shRNA and Cas9 screens have high reproducibility. A large positive casTLE score indicates a high confidence increase in growth rate, while a highly negative casTLE score indicates a high confidence decrease in growth rate, i.e. gene essentiality. The graphs compare replicate measurements of casTLE scores between plasmid and T14 for (a) Cas9 and (b) shRNA screens. Density is in log scale. (c,d) Time dependence of phenotypes. casTLE scores in different time-frames for (c) Cas9 and (d) shRNA screens.

Supplementary Figure 11



Supplementary Figure 11. Analysis of gene expression and yeast essential homologs. Genesets are defined for Cas9, shRNA, and Combination by a 10% FPR cutoff. Genesets are defined for Cas9-combo and shRNA-combo by the genes present in Cas9 or shRNA set and not in the Combination set. Overlap set is defined as genes present in both the Cas9 and shRNA set (See Supplementary Fig. 8b). (a,b) ~7,000 genes with detectable expression in K562 were binned by expression. The fraction of genes identified as essential in each bin is reported versus the average expression level of the bin. (c,d) Fraction of genes that are homologs of essential yeast genes versus genes that are homologs of nonessential yeast genes. P-values calculated using Fisher's exact test.