Measuring error rates in genomic perturbation screens: gold standards for human functional genomics

Traver Hart[1], Kevin R. Brown[1], Fabrice Sircoulomb[3], Robert Rottapel[3,4,5], Jason Moffat[1,2,Ψ]

[1]Donnelly Centre and Banting and Best Department of Medical Research, University of Toronto, Toronto, Canada; [2]Department of Molecular Genetics, University of Toronto, Toronto, Canada; [3]Campbell Family Cancer Research Institute, Ontario Cancer Institute, Princess Margaret Hospital, University Health Network, Toronto, Canada; [4]Department of Medical Biophysics, University of Toronto, Toronto, Canada; [5]Division of Rheumatology, Department of Medicine, St. Michael's Hospital, Toronto, Canada

Author email addresses:

Traver Hart (traver.hart@gmail.com)

Kevin Brown (kr.brown@utoronto.ca)

Fabrice Sircoulomb (sircoulombf@gmail.com)

Robert Rottapel (rottapel@gmail.com)

Jason Moffat (j.moffat@utoronto.ca)

[Ψ]Correspondence should be addressed to Jason Moffat

Running Title: Measuring error rates in genomic perturbation screens

Abstract

Technological advancement has opened the door to systematic genetics in mammalian cells. Genome-scale loss-of-function screens can assay fitness defects induced by partial gene knockdown, using RNA interference, or complete gene knockout, using new CRISPR techniques. These screens can reveal the basic blueprint required for cellular proliferation. Moreover, comparing healthy to cancerous tissue can uncover genes that are essential only in the tumor; these genes are targets for the development of specific anticancer therapies. Unfortunately, progress in this field has been hampered by off-target effects of perturbation reagents and poorly quantified error rates in large-scale screens. To improve the quality of information derived from these screens, and to provide a framework for understanding the capabilities and limitations of CRISPR technology, we derive gold-standard reference sets of essential and nonessential genes, and provide a Bayesian classifier of gene essentiality that outperforms current methods on both RNAi and CRISPR screens. Our results indicate that CRISPR technology is more sensitive than RNAi, and that both techniques have nontrivial false discovery rates that can be mitigated by rigorous analytical methods.

Keywords

Lethal genes, essential genes, fitness genes, cancer, Bayes factor, shRNA, RNAi, CRISPR

Background

In the early 1900s, Lucien Cuénot observed unusual patterns of inheritance when studying coat color in mice and from his many crosses, never produced a single homozygous yellow mouse (Cuenot, 1905; Paigen, 2003). Not long after these observations, it was shown that Cuenot's crosses resulted in what appeared to be non-Mendelian ratios because he had discovered a lethal gene (Castle & Little, 1910). W.E. Castle and C.C. Little demonstrated that one-quarter of the offspring from Cuénot's crosses between heterozygotes died during embryonic development, ushering in embryonic lethality, or death, as a new phenotypic class for geneticists (Castle & Little, 1910). Consequently, the idea that organisms harbor sets of lethal or essential genes has taken shape over the past century. Over the past dozen years or so, systematic genomic studies in eukaryotic model systems have defined sets of lethal or essential genes under defined growth conditions, providing a nexus for biologists to study the essential molecular processes that occur during cell growth and proliferation.

The importance of defining essential genes is three-fold. First, it provides a blueprint for all components necessary for a cell to grow and divide under defined conditions. Second, it provides a parts list that can be deconstructed to uncover all the necessary cellular and molecular functions that proceed during cell growth and division under defined experimental conditions. Third, the list of essential genes and related functions provide a reference point for understanding disease. Indeed, the accurate identification of human disease genes is among the most important goals of biomedical research and there exists a complex relationship between disease genes and essential genes, particularly for cancer genomes. For example, a recent analysis has shown that the cumulative effects of copy number variants of cancer drivers and essential genes along a chromosome explain the recurring patterns of somatic copy number alterations of whole chromosomes and chromosome arms in cancer genomes (Davoli *et al.*, 2013).

Broadly speaking, a gene is defined as essential if its complete loss of function results in a complete loss of fitness. In single-celled organisms this is a fairly straightforward assessment; however, in metazoans a gene could be reasonably classified as essential if its loss of function resulted in sterility or failure to develop to adulthood. In practice, a prenatal lethal phenotype is typically the criterion for essentiality. Given the absence of

a set of well-established human essential genes, researchers have generally relied on orthology to infer essentiality. Lethal or essential gene sets have been generated under defined growth conditions for a number of eukaryotic model systems including the budding yeast *S. cerevisiae* (Giaever *et al.*, 2002; Winzeler *et al.*, 1999), the fission yeast *S. pombe* (Kim *et al.*, 2010), *C. elegans* (Kamath *et al.*, 2003), *D. melanogaster* (Boutros *et al.*, 2004; Dietzl *et al.*, 2007), *M. musculus* (White *et al.*, 2013) and others. Across model organisms, essential genes are more likely to be hubs in protein-protein interaction networks (Jeong *et al.*, 2001), a phenomenon driven to some degree by membership in large essential protein complexes (Hart *et al.*, 2007; Zotenko *et al.*, 2008). Moreover, model organism essentials are less likely to have paralogs (Makino *et al.*, 2009), consistent with the model of gene duplication buffering loss-of-function phenotypes (Gu *et al.*, 2003). Human orthologs of mouse knockouts which give rise to developmental lethal phenotypes are themselves enriched for developmental disease genes, even above the bias toward developmental genes in the mouse knockout set (Makino *et al.*, 2009). Furthermore, ubiquitously expressed human genes are very likely to contain a large proportion of essential genes and are different in their evolutionary conservation rates (i.e. higher nonsynonomous/synonomous substitution rates), DNA coding lengths, and gene functions compared with disease genes and other genes (Tu *et al.*, 2006).

Experimental assays of human gene essentiality are performed in cell lines. RNA interference has, to date, been the weapon of choice for genome-scale fitness screening, with roughly two hundred published cell line screens (Cheung *et al.*, 2011; Luo *et al.*, 2008; Marcotte *et al.*, 2012; Schlabach *et al.*, 2008; Silva *et al.*, 2008). Other approaches include gene traps in haploid human cells (Burckstummer *et al.*, 2013; Carette *et al.*, 2009) and, more recently, genome-scale gene editing approaches using lentiviral-based CRISPR technologies (Shalem *et al.*, 2013; Wang *et al.*, 2013). The RNAi screens to-date have typically been conducted in cancer cell lines or normal counterparts to elucidate not only which genes are essential, but also which genes are differentially essential in different contexts, with the ultimate goal of identifying genes or pathways that are tissue-, subtype-, or even tumor-specific (i.e. genotype- or context-dependent essential or lethal genes). With the widespread adoption of pooled library shRNA screens has come the understanding that there are caveats to this type of

genetic screening approach. In particular, off-target effects can lead to false positives (Echeverri *et al.*, 2006; Moffat *et al.*, 2007), if the unintended target of an shRNA hairpin is an essential gene. To mitigate these effects, analytical approaches have been developed that look for phenotypic consistency across multiple hairpins targeting a gene (Cheung *et al.*, 2011; Luo *et al.*, 2008; Marcotte *et al.*, 2012) and among the same hairpins in different screens (Shao *et al.*, 2013). Not surprisingly, different approaches can yield different results, and the degree to which false positives contaminate results is largely unknown (Kaelin, 2012).

No method currently exists to systematically evaluate these various approaches. Studies in other areas of functional genomics have relied on "gold standard" positive and negative reference sets (Jansen & Gerstein, 2004) to evaluate the sensitivity and specificity of, for example, protein-protein interactions (Hart *et al.*, 2006; Havugimana *et al.*, 2012). This approach applies equally well to gene essentiality studies, where negatives can outnumber positives by an order of magnitude. However, no such gold standards currently exist for screens using mammalian, and more specifically human, cell lines. The developmental essentials inferred by orthology contain many genes which are, by definition, essential for whole organism development but unlikely to be essential in any given cell line context. To our knowledge, no putative cell line nonessential reference set exists at all, though it is certainly an impossible task to prove that any gene is nonessential in all contexts.

In this study we derive gold-standard reference sets of human cell line essential and nonessential genes. We use them to train a Bayesian classifier of gene essentiality in pooled-library shRNA screens and, most importantly, to evaluate the error rates of individual screens. We demonstrate how to leverage this framework to evaluate the data quality of genome-scale fitness screens in human cell lines as well as the effectiveness of the analytical approaches applied to them. In addition, we develop models of gene essentiality that permit estimation of the number of core essential genes and total number of essential genes. Our method is applicable to new pooled screening methodologies such as gene traps with haploid cell lines (Carette *et al.*, 2009) or genome-scale pooled CRISPR approaches (Shalem *et al.*, 2013; Wang *et al.*, 2013). The

reference sets can be used to evaluate screen quality regardless of what analytical method is applied.

Results

Computational framework for predicting essential genes from reverse genetic screens.

Genetic screens in mammalian cells using pooled barcoding approaches have emerged as a powerful method for functional discovery. In particular, negative genetic selections have the potential to reveal entire genetic pathways that govern cell growth and proliferation (i.e. essential/lethal factors). In order to advance our ability to analyze and assess that quality of systematic genetic screens that are emerging, we developed an informatics approach that is applicable to any genome-scale genetic screen or set of screens for predicting essential/lethal factors. Using a compendium of shRNA screens across different human cancer cell lines (Marcotte *et al.*, 2012), we developed a Bayesian classifier to score essential/lethal factors. As cells harboring shRNA hairpins targeting essential/lethal factors drop out of a proliferating population, the corresponding shRNAs show strong negative fold-change relative to controls. The data for each cell line is comprised of microarray data for up to three replicates at an initial timepoint (T0) and each of two experimental timepoints, and we calculated fold-change for each observation relative to the mean of the control microarrays, resulting in a matrix of fold-changes for ~78,000 hairpins across nearly 400 cell lines/timepoints. The Bayesian classifier was developed to evaluate whether the distribution of fold-changes for hairpins targeting a given gene most closely matched the distribution of fold-changes of hairpins targeting training sets of essential genes or nonessential genes using two-fold cross-validation to prevent circularity (Figure 1). The classifier was trained on reference sets we generated and each screen's performance was evaluated against a withheld test set (Figure 1).

Source of reference essential genes (EGs). An effective reference set of EGs should include all genes that are essential across every cell line or context in which they have been studied. We used a linear algebra approach to find genes that were consistently essential across cell line screens previously performed in our lab (Marcotte *et al.*, 2012). Singular value decomposition (SVD) is a matrix factorization technique that yields a set

of orthogonal basis vectors that describe, in rank order, the major sources of variation in the data. Briefly, SVD was applied to half of the shRNA fold-change matrix, yielding one dominant left singular vector (LSV) that describes ~42% of the total variance in the matrix (Figure S1a in Additional File 1). The distribution of all shRNA projections onto this first LSV is shown in Figure S1b. shRNAs with strong positive projections show consistent dropout across effective shRNA screens, which had strong negative projections onto the corresponding right singular vector (Figure S1c), projections which correlated with the number of cell doublings at which each sample was measured (Figures S1d). For each gene, we found the median projection onto the first LSV of its cognate hairpins and measured whether hairpins with median rank or higher rank were enriched in the right tail by a hypergeometric test, yielding 179 genes at a false discovery rate (FDR) of 25% (Benjamini & Hochberg). This list was further filtered to 148 constitutively and invariantly expressed genes; that is, genes with mean $\log_2(\text{FPKM}) > 0$ across both the ENCODE (17 cell lines, (Tilgner *et al.*, 2012)) and Illumina BodyMap (16 tissues, EBI accession no. E-MTAB-513) RNA-seq datasets, and standard deviation of $\log_2(\text{FPKM})$ less than the mean standard deviation of all observed protein coding genes in each dataset (Figure S2 in Additional File 1).

Source of reference nonessential genes (NEGs). Defining a reference set of nonessential genes is less clear cut, as it is impossible to experimentally demonstrate nonessentiality in all contexts. However, we reasoned that genes that are not expressed in the majority of tissues and cell lines are reasonable candidates for such a reference set. To generate this set, we again turned to published RNA-seq data. We selected protein-coding genes that are probed by our shRNA library and have an expression level of less than 0.1 FPKM in 15 of 16 BodyMap tissues and 16 of 17 ENCODE cell lines, as genes expressed below this level are typically not biologically relevant (Hart *et al.*, 2013; Hebenstreit *et al.*, 2011). We label the resulting set of 927 putatively nonessential genes the *NEGs* set. While this set may include some genes that are essential in other cellular or organismal contexts, the net effect of a small number of "accidental essentials" in this set should be negligible. The seed and reference non-essential genes are listed in Additional File 2.

Bayes factor scores. Reference essentials and nonessentials were divided into equal sized training and testing sets for subsequent analyses, and each cell line in the withheld half of the shRNA fold-change matrix was analyzed independently. For each timepoint, the fold-change distributions for the essential and nonessential training sets, comprising 347 and 2,268 hairpins respectively, were determined. Then, for each gene, a Bayes Factor (BF) was calculated, representing the log likelihood that the observed fold-change for a given gene's cognate hairpins were drawn from either the essential or the nonessential reference distributions. Log BFs were summed across all time points for a final BF for each gene in each cell line. Additional File 3 contains a table of all calculated Bayes Factors.

F-measure. For each cell line, genes were rank ordered by BF and compared to the withheld reference test sets to evaluate precision vs. recall. Screen quality varied widely, with most screens showing moderate to high performance, though several outliers showed remarkably poor results (Figure 2a). We identified the point on the recall-precision curve for each screen where the BF crossed zero and calculated the F-measure (harmonic mean of recall & precision) of each screen at that point. We judged screens with F-measure >= 0.75 (n=48/68; Figure 2b) to be high-performing screens and retained them for downstream analyses. Screen performance measures are listed in Additional File 4.

Core essentials. Within this set of high-performing screens, we examined the frequency with which each gene was called essential (BF>0) (Figure 2c). Though 4,451 genes have a positive BF in at least one cell line, genes observed in few (1-4) screens are enriched for false positives. Repeated observation greatly improves the likelihood that a gene is truly essential. To identify likely global essential genes, and to avoid identifying cancer tissue/subtype-specific genes, we selected genes observed in at least half of the performing screens (n = 291 genes). We label these *core essentials*.

Cumulative analysis of EGs.

To identify the set of all EGs observed across all screens, we used a cumulative analysis approach. Most large-scale functional genomics screens try to differentiate a small number of true "hits" from a pool of negatives that can often be orders of magnitude

larger (Jansen & Gerstein, 2004). In such screens, even a tiny false positive rate applied across all true negatives can result in large FDRs for individual screens.

Researchers can attenuate the final FDR by conducting multiple repeats of screens and analyzing the frequency with which each hit is observed across repeats. By considering the cumulative distribution of hits across multiple screens, information about both the total number of true essentials in the population and the error rate in observing those hits can be calculated. In principle, a screen with zero FDR that is repeated to saturation will yield a cumulative observations curve that flattens to a slope of zero at the total number of hits in the screened population. In practice, repeated screens can saturate the true hits but random discovery of false positives yields a cumulative curve with positive slope as more and more false positives are accumulated. A variation of this cumulative observation analysis was used to evaluate the saturation of protein-protein interaction screens (Hart *et al.*, 2006).

To test this logic, we rank-ordered the top-performing 36 cell line screens in our compendium of pooled shRNA screens by F-measure and plotted the cumulative number of observed essential genes. The result is indeed a curve that flattens but with a positive slope (Figure 3a). To estimate both the number of essential genes and the average screen error rate, we conducted *in silico* simulations of the 36 screens, determined the synthetic cumulative observation curve for each set of simulations, and measured the curve's fit to our experimental observations. With fixed parameters of 15,687 genes assayed and 606 genes reported as essential in each screen (the mean number of genes in the top 36 screens with BF>0), we find that a model with a cellular population of 1,025 essential genes and an average screen FDR of 15% yields a cumulative essentials curve that mimics the observed curve very closely (Figure 3a). Running the model across a range of total essential population sizes and FDRs and calculating root-mean-squared deviation (RMSD) from the observed cumulative essentials curve shows models with 850-1,175 essential genes and FDRs of 14.0%-16.5% yield an RMSD that is less than 1.5x the minimum RMSD (Figure 3b). Notably, the FDR range for the best fit models is highly consistent with the average empirically-measured FDR of 13.8% across the top 36 screens. Moreover, while the top screens encompass several cancer subtypes from three tissues of origin, the model treats all 36

repeats as replicates. Tissue- or subtype-specific essential genes in the saturated region will be incorrectly treated as false positives using the cumulative approach; therefore FDR estimates derived in this manner are likely conservative.

Though the modeling approach can tell us approximately how many essential genes are in our cell lines *in toto*, it does not identify which genes are truly essential. To separate essential genes from false positives we rely on repeat observations of essential genes across multiple screens. Figure 3c shows a histogram of gene essentiality calls across the top 12 screens. Of the 2,130 unique hits in these screens, 945 (44%) are observed in a single screen, while only 392 (18%) are seen in six or more of the 12 cell lines.

To estimate the bin-wise FDR for this distribution, we again turn to the cumulative approach. Figure 3c also shows the distribution of essential gene calls in the 13th-24th ranked screens (blue) and the 25th-36th ranked screens (red). Assuming that the first 12 screens have achieved saturation, then all subsequent hits must be false positives. The second and third sets of 12 screens therefore model the frequency distribution of false positives and give an estimate of the expected number of false positives in each bin. Based on these estimates we conclude that hits in 3 or more of the top 12 screens are essential genes with an FDR of 6-11%. This set comprises 823 genes, which we label *total essentials* (see Additional File 5 for a complete list).

Given the diversity of tissues and subtypes in the cell lines studied, it is unlikely that all observations beyond the top 12 screens are false positives. Some fraction of subsequent hits may in fact be true subtype-specific essential genes. For example, the top 12 cell lines include five pancreatic, three ovarian, and four breast cancer cell lines, of which three are basal subtype and one, HCC-1954, is EGFR-high/Her2 amplified. Well-studied subtype-specific breast cancer oncogenes CDK4 and FOXA1 are not classified as essential in any of the top 12 screens, including HCC-1954, though this line does show a dependence on Her2/ERBB2 (BF=9.21). However, across all 48 performing screens, CDK4 and FOXA1 each show BF > 20 in 4 cell lines; two of the four CDK4 lines and three of the four FOXA1 lines are HER2+ breast cancer lines and the remainder are all luminal subtype. The net effect of these subtype-specific essentials in the analysis of

cumulative observations is to artificially inflate the imputed number of false positives in each bin, thus rendering our FDR estimates conservative.

Characteristics of EGs.

Core essential genes are expected to be essential in all cell lines and contexts, and must be constitutively expressed. Indeed, 219 of 291 core essentials (75.2%), as well as 483 of 823 total essentials (58.7%) showed high mean expression with low variation across a compendium of RNA-seq experiments (Figure S2), compared to 33.4% of other genes. These genes are also highly enriched for protein complexes; more than half of the set of total essentials encode subunits of annotated human protein complexes. Figure 4a shows the top nonoverlapping (i.e. minimal shared subunits) protein complexes that show strong enrichment for essential genes (comprising 231 genes), with most subunits detected as core essentials and coverage increased by the set of total essentials (see Additional File 6 for a complete list). These complexes represent the fundamental molecular functions of cellular life: transcription, translation, and replication. An additional 235 essential genes are also annotated as subunits of protein complexes, though the complexes do not meet our threshold for statistical significance. These complexes and their essential subunits are shown in Additional File 6. The remaining 357 essential genes not in any annotated protein complex also show enrichment for core cellular processes, including ribosome biogenesis (13 genes, 5.6-fold enrichment, P=3.6e-6); aminoacyl-tRNA synthetases (4 genes, 6.1-fold, P=2.8e-2), and protein tyrosine phosphatases (8 genes, 4.3-fold, P=2.7e-3). Essential genes not in complexes are generally not constitutively expressed; 117 of the 357 (32.8%) show constitutive and invariant expression compared to 33.4% of nonessentials, suggesting this may be a rich source of tissue-specific essentials.

Essential genes were divided into the categories described above (i.e. in enriched complexes, in other complexes, not in any complex; Figure 4b) to examine the fraction of genes in each of these categories that overlap/intersect with mouse essential genes or represent paralogs. For genes whose mouse orthologs have been knocked out, essentials in protein complexes were much more likely to have an essential mouse ortholog than other genes (Figure 4c). Furthermore, we find that essential genes in protein complexes are less likely to have paralogs than nonessential genes (Figure 4c).

In particular, essential genes in essential complexes are more likely to be singletons than other classes (Figure 4d).

Biological sources of variability in RNAi negative selection screens.

Having derived a set of performance metrics using essential genes at the screen and gene level, we sought to understand some of the drivers of variability, particularly in lentiviral-based pooled RNA interference screens across a large panel of human cancer cell lines. Fortunately, the pancreatic and ovarian cancer cell line screens have matching gene expression microarray data collected on the same array platform (Marcotte *et al.*, 2012). Measuring the correlation between gene expression and screen F-measure across 31 cell lines (one outlier removed), we found that AGO2 was the top ranked correlation among more than 10,000 expressed genes (Pearson correlation coefficient=0.59; Figure 5a). The AGO2 protein, coupled with short RNA, comprises the RNA-induced Silencing Complex (RISC), which catalyzes the cleavage of target mRNA and was expected to be an important predictor of RNAi efficiency. The relationship between AGO2 mRNA expression and shRNA screen quality was weaker in the breast cancer screens (Figure S3 in Additional File 1), which may reflect some combination of generally better performing screens in breast cancer cell lines – with corresponding lower variability – and the fact that the expression data were collected on a different microarray platform.

While AGO2 expression may help explain why some screens perform better than others, it does little to explain the variability within high-performing screens. Though the large number of genes observed infrequently in Figure 2c reflects the expected distribution of false positives across the screens, we expected a more pronounced peak at the right edge of the distribution from core essentials observed across most or all high-performing screens. We explored other molecular genetic data to explain this false negative rate among known essentials and derived absolute copy number for each gene across 30 pancreatic and ovarian cancer cell lines in our study (see Methods and Additional File 7). We calculated a Pearson correlation coefficient for each gene's copy number profile vs. its Bayes Factor profile across the same screens and observed that core essential genes show a negative correlation between copy number and essentiality (Figure 5b). Notably, the core essential genes largely encode members of essential protein

complexes, and our observation is consistent with a model whereby increased copy number yields protein levels in excess of stoichiometric requirements for protein complex function. The genes are likely no less essential, as complete knockout would probably still kill the cells, but the copy number amplification renders them less sensitive to RNAi perturbation. Binning core essential genes by absolute copy number and measuring the fraction in each bin that are successfully identified in the screens (Figure 5c) supports this model. In other words, as copy number increases, the likelihood that a core essential is accurately classified drops markedly. Based on the difference between the overall observed false negative rate and the false negative rate at copy number=2, we estimate that 15-20% of false negatives (core essentials not accurately classified as essential in a cell line) are attributable to copy number variation. This hypothesis could be tested by employing orthogonal genome editing technologies, such as CRISPR.

Leveraging gold-standard reference sets to improve analyses of CRISPR and shRNA screens.

We used matrix decomposition to generate a seed set of reference global essentials to train our Bayesian classifier, the application of which ultimately yielded 291 core essential genes across 48 high-performing shRNA screens of 68 cell lines. We filtered these genes for constitutive, invariant expression across the BodyMap and ENCODE RNA-seq samples, yielding a set of 217 genes we label Constitutive Core Essentials (*CCE*). We then divided the *CCE*, as well as the previously described *NEGs*, into equal-sized training and test sets (*CCE-train, CCE-test, NEG-train, NEG-test*) and used them as improved reference sets to train our classifier and evaluate both data quality and analytical approaches in other data sets as well as screens withheld from our initial set of 72 cancer cell lines.

Evaluating CRISPR Negative Selection Screens*.*

Gold-standard reference sets of essential and nonessential genes can be used to evaluate any large-scale assay of gene essentiality. Recently the CRISPR system has been adapted to induce targeted genetic modification of human cells (Cong *et al.*, 2013; Mali *et al.*, 2013), and has been applied in genome-scale pooled library positive selection screens for specific pathway members and negative selection screens for essential genes. For example, Shalem *et al.* (Shalem *et al.*, 2013) recently published negative

selection screens targeting 18,080 genes with ~65,000 guide sequences (gRNA) in two human cell lines, A375 melanoma cells and HUES62 embryonic stem cells. As with shRNA screens, a CRISPR gRNA targeting an essential gene will drop out of a population, resulting in a strong negative fold change for that gRNA. As expected, the fold change distributions of gRNA targeting training-set essential genes at the early (Figure 6a) and late (Figure 6b) experimental timepoints were left-shifted relative to the distributions of gRNA targeting nonessential genes. We used these distributions to train our Bayesian classifier and evaluated our results against the withheld test sets. Figure 6c shows the improvement that the Bayesian classifier offers over the approach used in the original study. It also highlights the poor performance of the HUES62 screen, which explains the sparse overlap between the two screens reported in the original study.

Concurrently, Wang *et al.* (Wang *et al.*, 2013) reported negative selection screens targeting 7,114 genes in two human cell lines, including the near-haploid KBM7 cell line. The performance curves of these screens, measured against CCE-test and NEG-test and shown in Figure 6d, are impressive but likely underestimate the actual error rates of these screens as core-essential ribosomal gRNA are overrepresented and the nonessential reference set is severely underrepresented among target genes (~6-fold depletion relative to the Shalem *et al.* library).

Taken together, these analyses offer two key insights into the differences between CRISPR and RNAi screens. The Bayes Factor analysis of the Shalem *et al.* screen classifies 805 targets as essential at zero FDR. These 805 genes represent 47% recall of the reference essential set; extrapolation suggests there may be well over 1,600 essential genes in this cell line. As this is more than double the number of high-confidence essentials detected in any shRNA screen, and 50% more than the total number of essentials suggested by the cumulative analysis of RNAi screens, it suggests that CRISPR screens may have substantially greater sensitivity than pooled library shRNA screens.

However, the error rate increases markedly after the top ~1500 hits. False discovery rates of genome-scale CRISPR screens are largely unexplored in the first-generation published screens, but our analysis indicates that nontrivial numbers of false positives

are indeed present in these screens. It is currently unknown whether these false positives arise from the technical variability inherent in large-scale sequencing screens or from the biological activity of off-target gRNA sequences. The gold-standard reference sets and analytical methods we describe here offer a framework for understanding the nature of these false positives and, in turn, for refining the design of CRISPR gRNA libraries and experimental protocols.

Improving analyses of shRNA pooled library screens.

Our lab recently published a study of shRNA-driven synthetic lethality with several query knockout genes in an isogenic HCT116 colon cancer cell line background (Vizeacoumar *et al.*, 2013). The HCT116 cell line is near diploid and thus does not suffer from SCNA-driven biological artifacts. We trained our Bayesian classifier with CCE-train and NEG-train, applying a uniform prior (P(essential)/P(nonessential); see Methods) of 0.1, to yield a posterior log odds (LOD) of essentiality for each gene in each screen. Recall and precision were evaluated against CCE-test and NEG-test and an F-measure was calculated at a point on the curve where the LOD score crossed zero (Figure 6a). All six screens had F-measures > 0.8, adding confidence to analyses of essentiality and differential essentiality gleaned from these screens.

We applied the same analytical approach to the compendium of 102 pooled library shRNA screens from Project Achilles (Cheung *et al.*, 2011), after filtering the reference sets for genes assayed by the 54k hairpin library used for those screens. Finding the point on the precision-recall curve where the LOD score crosses zero (Figure 6b, blue), we observe wide variability in the quality of the screens: only 41 of the 102 screens had an F-measure of 0.75 or greater (60 with F>=0.70).

The reference sets can be used independently to evaluate any method of analyzing essentiality screens. For example, we used CCE-test and NEG-test to evaluate the published results of the ATARiS algorithm as applied to the Achilles data (Shao *et al.*, 2012). After filtering the reference sets for genes with ATARiS solutions, genes from each screen were ranked by phenotype score, with the most negative score indicating strongest phenotype. Recall and precision were determined at a phenotype score of -1 (Figure 6b, red); generally only a few hundred genes have stronger scores. ATARiS

gave predictions for many cell lines that were worse than random, perhaps in part because it included all the Achilles data sets, including the lower quality ones. Filtering the input data for known performing screens may significantly improve scoring performance.

Using a different approach, Solimini *et al.* (Solimini *et al.*, 2013) analyzed the distribution of copy number changes of tumor suppressors ("STOP" genes) and essential genes ("GO" genes) across thousands of tumor-normal pairs. In the absence of a reference set of essential genes, the authors used two approaches to define GO genes: screening and theoretical. In the screening approach, the authors identified hairpins that dropped out in 5 of 9 library shRNA screens, yielding 1,127 genes with at least one hairpin. Though the single-hairpin approach is not widely accepted due to the frequency of off-target effects (Kaelin, 2012), the error rate is mitigated somewhat by requiring multiple observations across multiple screens. Evaluated against CCE-test and NEG-test, this set shows 49.5% recall and 16.9% FDR. The theoretical approach drew upon genes from selected core pathways in KEGG and yielded 545 genes with 54.1% recall and 1.7% FDR.

While the reference sets are broadly applicable to cancer genomics studies, the Bayesian approach used to classify essential genes can be readily extended to integrate other molecular data. We collected RNA-seq gene expression data on four pancreatic cancer cell lines withheld from our analysis of the COLT-cancer dataset, and rank-ordered and binned (n=500) genes by expression level. Within each bin, we plotted the mean expression (+/- s.d.) vs. the fraction of genes in the CCE-train reference set (Figure 7c). We then used a linear fit to this data to calculate an expression-based informative prior for each gene, replacing the uniform prior used above in the calculation of LOD score (see Methods). Figure 7c shows the relationship between gene essentiality and expression in the CAPAN-2 cell line, while Figure 7d shows the relative performance of four analytical approaches evaluated against CCE-test and NEG-test. Applying the gene-specific expression prior improved the performance of the screen in all cases (see Figure S4 for the other 3 screens) over the LOD score with the uninformative prior, and increased the margin of improvement over two current state of the art algorithms for library RNAi screens, GARP and RIGER. Thus the combination of the reference set of core essentials and the Bayesian classifier offer a best-in-class method for analyzing

such screens as well as a framework for integrating other molecular data to improve performance. Moreover, the core essentials offer a ready reference set against which to evaluate the relative performance of such screens.

The Daisy model of gene essentiality.

The mouse knockout data highlight an important factor in the study of gene essentiality. The definition of essentiality is context-dependent: a mouse (or human) gene may reasonably be classified as essential if its complete loss of function results in a phenotype ranging from prenatal to juvenile lethality or even sterility, with the onus on the researcher to explicitly define the term. Cell line assays of gene essentiality necessarily sample only the genes required for the proliferation of that cell line in cultured conditions; genes which may be required for organismal health may not be expressed in a given cell line and thus will not be detectable. Nevertheless, there is a core set of ubiquitously expressed, ubiquitously essential genes that should be detectable in virtually any cell-line screen. This gives rise to the "daisy model" of gene essentiality (Figure 8a), where each petal represents a cell line- or tissue-specific context in which a gene's activity might be required. Petals will overlap to varying degrees but all will share the core set of essential genes. The core essentials described here represent our effort to define this set of universally essential genes.

The link between gene essentiality and genetic predisposition to disease has long been a topic of active study. We took the set of mouse knockout essentials and divided them into core and peripheral essentials based on our definition of the core. Analyzing these sets for disease gene enrichment reveals that peripheral essentials are strongly enriched for disease genes (P = 1.5e-40; Figure 8b) while core essentials show no enrichment beyond random expectation. This is consistent with previous findings (Chavali *et al.*, 2010; Dickerson *et al.*, 2011; Lohmueller *et al.*, 2008), which gives rise to a model wherein core essential genes are less tolerant to genetic variation than peripheral essentials. The recent publication of large-scale human population genetic studies allows us to test these hypotheses. Figure 8c shows the rate of putative deleterious mutation observed in 2440 exomes (Tennessen *et al.*, 2012), by gene class. Core essentials are much less likely to show deleterious variants than nonessential genes,

and an even larger fraction of core essentials have no observed variant (Figure 8c, inset).

Discussion

In this study, we have generated a global set of essential genes in human cell lines based on experimental data. Drawn from genes that show consistent strong anti-proliferative effects across a panel of pooled library shRNA screens in cancer cell lines, these essential genes are highly enriched for conserved protein complexes that carry out the fundamental work of the cell: transcription, translation, DNA replication, and protein degradation. Consistent with previous studies, these genes are more likely to be essential in mouse knockout studies and less likely to have a human paralog than other genes. We label these genes *core essentials* as they are likely essential across all cell lines, tissue types, and developmental states.

We exploit the difference between core and peripheral (or *context-specific*) essentials in two ways. First, at the organismal level, we show that peripheral essentials, including human homologs of mouse essential genes, are more likely to be disease genes and demonstrate that core essentials show lower incidence of putative deleterious mutation in a normal human population. This finding explains a longstanding observation that human disease genes are enriched for whole-organism essentials but tend not to be housekeeping genes. That is, hypomorphic alleles of peripheral essentials cause a partial loss of fitness (i.e. disease) but hypomorphic alleles of core essentials are fatal. Cumulative analysis of RNAi screens suggests a total population of ~1,000 human cell-line essential genes, while preliminary analysis of genome-scale CRISPR screens suggest roughly double this number, perhaps reflecting reduced sensitivity of RNAi methods against lower-expression genes.

Second, we derive the "daisy model" of gene essentiality from the difference between core and context-specific cell line essentials, wherein each petal represents the set of essential genes in one cell line, tissue, or genomic context. Petals will overlap to varying degrees but all contexts share the common core essentials. While the focus of essentiality studies in cancer cell lines is to find context-specific essentials that can

provide highly specific therapeutic targets, the degree to which a screen recapitulates the shared core essentials is a critical measure of its accuracy.

We used the core essentials, in conjunction with a set of putative nonessentials derived from the Illumina BodyMap and ENCODE studies of gene expression in human tissues and cell lines, as gold-standard reference sets to train a Bayesian predictor of gene essentiality in pooled library shRNA screens, and to test our algorithm as well as several previously published algorithms and data sets. Our algorithm substantially outperforms other methods on the data sets we tested, particularly when coupled with sample-matched gene expression data. We also demonstrate that our method is applicable to other pooled library negative selection screens using CRISPR genome-editing technology, and look forward to the onslaught of genome-scale screens that will emerge using this technology.

Our analyses reveal that copy number amplification in cancer cell lines can substantially decrease a core essential gene's sensitivity to RNAi perturbation. This is most likely driven by the encoded protein's membership in a protein complex: genomic amplification leads to over-expression and protein abundance beyond the stoichiometric requirements for complex function. Interestingly, the converse is also true: hemizygosity increases sensitivity. A recent study found that partial loss of some genes in tumors resulted in increased vulnerability to perturbations of those genes -- the so-called CYCLOPS genes (Nijhawan *et al.*, 2012). As CYCLOPS genes are enriched subunits of core essential complexes, our findings likely extend the CYCLOPS concept to all core essential complexes. That is, copy number losses among essential subunits may render cancer cells more susceptible to pharmacological compounds targeting these complexes.

Broadly speaking, the reference sets of cell line essential and nonessential genes we provide represent a useful yardstick against which cancer functional genomics studies can be measured. Lack of such suitable yardsticks has led to critical errors in the field, including high profile reports of synthetic lethal interactions with common oncogenes (Scholl *et al.*, 2009) that were later disproven (Babij *et al.*, 2011; Luo *et al.*, 2012; Weiwer *et al.*, 2012) (and also do not appear in our data), and has led to a reassessment of

shRNA methodologies (Kaelin, 2012). Such gold-standard reference sets will become increasingly important as the CRISPR genome-scale genetic perturbation technology matures (Cong *et al.*, 2013; Mali *et al.*, 2013). Our analysis of available data indicates that CRISPR screens can be more sensitive than RNAi methods in detecting essential genes, but that CRISPR library screening is also subject to a nontrivial false discovery rate—a finding that is largely ignored in the current literature. Progressively improving performance against an established set of benchmarks is the best way to validate such new technologies and their accompanying analytical methods, to ensure their widespread adoption, and to unlock the biological discovery that their application enables.

Materials and methods

Using Matrix Decomposition to find a seed set of putative essentials. The 72 pooled library shRNA screens were divided into three sets: group one (n=34), group two (n=34), and withheld (n=4; see sample key in Additional File 1). For screens in group one, all repeats from all timepoints were combined into a fold change matrix of ~78000 hairpins by ~200 arrays. Singular value decomposition was performed on the matrix; the top singular value was found to explain >40% of the total variance of the matrix (see Figure S1). Hairpins with strong positive projections onto the first left singular vector (U1) showed strong negative fold-change across most of the 34 samples in the group one matrix.

We used a statistical filter to find genes enriched for hairpins with strong U1 projections. For each gene, hairpins were rank-ordered by U1 projection, and the median hairpin projection *p* was determined. Then the enrichment P-value was calculated by the hypergeometric test:

$$P(\text{enrichment}) = \text{hypergeometric}(X >= x \mid n, m, N)$$

where x is the rank of the median hairpin for the gene; n is the number of hairpins targeting the gene; m is the total number of hairpins in the population with U1 projection >= p, and N is the total number of hairpins in the experiment.

Q-values were calculated from P-values by the method of Benjamini & Hochberg and genes with q<0.25 were selected as putative seed essentials. This list was further filtered for genes with constitutive, invariant gene expression across two sets of RNA-seq data, the ENCODE set of 17 human cell lines and the Illumina BodyMap set of 16 healthy human tissues (see RNA-seq analysis, below).

RNA-seq analysis. We used Tophat v1.4.1 to align RNA-seq reads to the hg19 human transcriptome defined in the Gencode v14 GTF file, using default Tophat parameters. We used Cufflinks in quantitation-only mode with the same GTF file to generate FPKM values for each gene. FPKM values were filtered for protein-coding genes (as defined by HGNC, www.genenames.org) and log-transformed (adding 0.01 as a pseudocount). The mean log(FPKM) of technical or biological repeats was used, where applicable (e.g. biological repeats in ENCODE and technical repeats at 2x50 and 1x75 read type for BodyMap).

For ENCODE (GEO accession GSE30567) and BodyMap (EBI accession E-MTAB-513), constituent, invariant genes were defined as genes with mean expression in each data set > 0 and standard deviation < mean standard deviation across all protein-coding genes. Genes must be constituent & invariant in both data sets. The reference set of putative nonessential genes is defined as protein coding genes with FPKM < 0.1 in 15 of 16 BodyMap tissues *and* FPKM < 0.1 in 16 of 17 ENCODE cell lines. The set is filtered for genes that are assayed by the pooled shRNA library.

Calculating the Bayes Factor. Seed essentials from SVD of group one and nonessentials from gene expression were divided into equal-sized sets for training and testing, and used to train and evaluate the classifier for each cell line in group two (and vice versa). Each cell line was assayed at two timepoints. For each timepoint, an empirical distribution of the fold-changes of all hairpins targeting essential genes in the training set was calculated using the scipy.stats.gaussian_kde function in Python. The process was repeated for nonessential genes. Then, for each gene, the Bayes Factor is calculated as follows:

$$BF = \frac{\Pr(data \mid essential)}{\Pr(data \mid nonessential)} = \prod_{i,j} \frac{\Pr(fc_{i,j} \mid essential)}{\Pr(fc_{i,j} \mid nonessential)}$$

across hairpin observations *i* and timepoints *j*, where Pr(x) is the likelihood function.

Log-transforming the equation yields:

$$\log(BF) = \sum_{i,j} \log(\Pr(fc_{i,j} \mid essential)) - \log(\Pr(fc_{i,j} \mid nonessential)$$

For a typical gene with 5 cognate hairpins with three biological repeats, the log(BF) is the sum of 15 values at each of two timepoints.

Using priors to calculate posterior log odds. A Bayes Factor can be extended to a posterior odds ratio by multiplying by an appropriate ratio of priors:

$$OR = \frac{\Pr(data \mid essential)}{\Pr(data \mid nonessential)} x \frac{\Pr(essential)}{\Pr(nonessential)}$$

$$\log(OR) = \log(Pr(data|essential)) - \log(Pr(data|nonessential)) + \log(prior)$$

Where indicated in the main text that a posterior log odds ratio (LOD score) was calculated (e.g. the withheld group, the HCT116 screens, and the Achilles screens), a

uniform prior of 0.1 was applied (by adding $\log_2$(prior) = -3.32 to each logBF), representing a background expectation that 10% of assayed genes are essential.

For samples in the withheld group, we also calculated a specific prior for each gene based on its expression level. We generated $\log_2$(FPKM) values for all protein-coding genes as described above. Genes were rank-ordered by expression level and binned (n=500). For each bin, we calculated the mean expression level of genes in the bin and the $\log_2$ of the fraction of genes in the CCE-train reference set, adding a pseudocount of 0.001 to prevent infinities (see Figure 7c). A linear fit was applied to bins with mean expression > 1. This linear fit was used to calculate an expression-based prior, with the $\log_2$(fraction essential genes in bin) approximating the log-prior described above.

Evaluating precision and recall for each screen. For each screen, the applicable reference sets were divided into equal-sized training and testing sets. Training sets were used to generate the empirical distributions of essential and nonessential hairpin fold-changes, as described above, and (where applicable) to calculate the expression-based prior. Withheld testing sets were used to evaluate the performance of each screen.

Genes from each evaluated screen were rank ordered by Bayes Factor or LOD score, which ever was applicable. Then, for each gene, the cumulative precision and recall were calculated as Recall = TP / (TP + FN) and Precision = TP / (TP + FP), where TP = true positives, the number of genes in the essentials test set with BF/LOD score greater than the current gene; (TP+FN) = the total number of essentials in the test set; and FP = false positives, the number of genes in the nonessentials test set with BF/LOD score greater than the current gene.

The F-measure was calculated as a single, global metric for screen quality. The F-measure is the harmonic mean of precision and recall calculated at BF/LOD = 0:

$$F = 2\frac{(precision\ \ x\, recall)}{(precision + recall)}$$

To evaluate the ATARiS results, we used phenotype scores from Achilles_102lines_gene_solutions.gct (downloaded from http://www.broadinstitute.org/achilles/datasets/2/download). For each screen, genes were rank ordered by phenotype score and precision and recall were calculated as

above, using the CCE-test and NE-test reference sets. F-measure was calculated at phenotype score = -1.

Absolute copy number. SNP analysis was performed at the University Health Network Microarray Center (Toronto, ON, CA) using Illumina (Illumina, San Diego, CA) HumanOmni1 BeadChip according to manufacturer's instructions. Normalized LogR ratio (LRR) and B allele frequency (BAF) signals for each probe were exported from the Illumina BeadStudio utility. Export files were then processed with the Genome Alteration Print (GAP) algorithm (Popova *et al.*, 2009). Projections of LRR and BAF profiles were created and pattern recognition was performed for each samples. Parameters were set as followed: germHomozyg.mBAF.thr > 0.97 and p_BAF=0 (no normal contamination). Each pattern was visually inspected and corrected when the grid was off the segment center clusters. Output files produced by GAP were processed in order to obtain segments defined by copy number change only. Briefly, adjacent segments with identical absolute copy number were merged and the LRR values were averaged. Gene level absolute copy number and LRR was obtained using the CNTools package.

Competing interests

The authors declare that they have no competing interests.

Additional data files

Additional files are available upon email request

Additional File 1 contains supplementary figures S1-S4 and figure legends.

Additional File 2 is a table of reference sets used in the analysis.

Additional File 3 is a table of all Bayes Factors calculated for the shRNA screens

Additional File 4 is a table of screen performance metrics, including F-measure

Additional File 5 contains tables of essential genes and the number of screens in which they were called essential.

Additional File 6 is a table of essential protein complexes and their essential subunits

Additional File 7 is a table of absolute gene copy number by cell line.

Author contributions

TH and JM conceived of the ideas and wrote the manuscript with input from KRB, FS, and RR.

Acknowledgements

The authors would like to thank all the members of the Moffat and Rottapel labs for helpful discussions. This work was supported by the Ontario Ministry of Research and Innovation's GL2 Program and the Selective Therapies Program at the Ontario Institute for Cancer Research. JM is a Tier II Canada Research Chair in Functional Genetics and a Research Fellow at the Canadian Institute for Advanced Research.

References

Babij C, Zhang Y, Kurzeja RJ, Munzli A, Shehabeldin A, Fernando M, Quon K, Kassner PD, Ruefli-Brasse AA, Watson VJ, Fajardo F, Jackson A, Zondlo J, Sun Y, Ellison AR, Plewa CA, San MT, Robinson J, McCarter J, Schwandner R *et al.* (2011) STK33 kinase activity is nonessential in KRAS-dependent cancer cells. *Cancer Res* **71:** 5818-5826

Boutros M, Kiger AA, Armknecht S, Kerr K, Hild M, Koch B, Haas SA, Paro R, Perrimon N, Heidelberg Fly Array C (2004) Genome-wide RNAi analysis of growth and viability in Drosophila cells. *Science* **303:** 832-835

Burckstummer T, Banning C, Hainzl P, Schobesberger R, Kerzendorfer C, Pauler FM, Chen D, Them N, Schischlik F, Rebsamen M, Smida M, de la Cruz FF, Lapao A, Liszt M, Eizinger B, Guenzl PM, Blomen VA, Konopka T, Gapp B, Parapatics K *et al.* (2013) A reversible gene trap collection empowers haploid genetics in human cells . *Nat Methods***:** 1548-7105

Carette JE, Guimaraes CP, Varadarajan M, Park AS, Wuethrich I, Godarova A, Kotecki M, Cochran BH, Spooner E, Ploegh HL, Brummelkamp TR (2009) Haploid genetic screens in human cells identify host factors used by pathogens. *Science* **326:** 1231-1235

Castle WE, Little CC (1910) On a Modified Mendelian Ratio among Yellow Mice. *Science* **32:** 868-870

Chavali S, Barrenas F, Kanduri K, Benson M (2010) Network properties of human disease genes with pleiotropic effects. *BMC systems biology* **4:** 78

Cheung HW, Cowley GS, Weir BA, Boehm JS, Rusin S, Scott JA, East A, Ali LD, Lizotte PH, Wong TC, Jiang G, Hsiao J, Mermel CH, Getz G, Barretina J, Gopal S, Tamayo P, Gould J, Tsherniak A, Stransky N *et al.* (2011) Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc Natl Acad Sci U S A* **108:** 12372-12377

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science* **339:** 819-823

Cuenot L (1905) Les races pures et leurs combinaisons chez les souris. *Archives de Zoologie Experimentale et Generale* **4:** 123-132

Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ, Elledge SJ (2013) Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155:** 948-962

Dickerson JE, Zhu A, Robertson DL, Hentges KE (2011) Defining the role of essential genes in human disease. *PLoS One* **6:** e27368

Dietzl G, Chen D, Schnorrer F, Su KC, Barinova Y, Fellner M, Gasser B, Kinsey K, Oppel S, Scheiblauer S, Couto A, Marra V, Keleman K, Dickson BJ (2007) A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. *Nature* **448:** 151-156

Echeverri CJ, Beachy PA, Baum B, Boutros M, Buchholz F, Chanda SK, Downward J, Ellenberg J, Fraser AG, Hacohen N, Hahn WC, Jackson AL, Kiger A, Linsley PS, Lum L, Ma Y, Mathey-Prevot B, Root DE, Sabatini DM, Taipale J *et al.* (2006) Minimizing the risk of reporting false positives in large-scale RNAi screens. *Nat Methods* **3:** 777-779

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A *et al.* (2002) Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418:** 387-391

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* **421:** 63-66

Hart GT, Lee I, Marcotte ER (2007) A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics* **8:** 236

Hart GT, Ramani AK, Marcotte EM (2006) How complete are current yeast and human protein-interaction networks? *Genome Biol* **7:** 120

Hart T, Komori HK, Lamere S, Podshivalova K, Salomon DR (2013) Finding the active genes in deep RNA-seq gene expression studies. *BMC Genomics* **14:** 778

Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, Wang PI, Boutz DR, Fong V, Phanse S, Babu M, Craig SA, Hu P, Wan C, Vlasblom J, Dar VU, Bezginov A, Clark GW, Wu GC, Wodak SJ *et al.* (2012) A census of human soluble protein complexes. *Cell* **150:** 1068-1081

Hebenstreit D, Fang M, Gu M, Charoensawan V, van Oudenaarden A, Teichmann SA (2011) RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol Syst Biol* **7:** 497

Jansen R, Gerstein M (2004) Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Current opinion in microbiology* **7:** 535-545

Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* **411:** 41-42

Kaelin WG, Jr. (2012) Molecular biology. Use and abuse of RNAi to study mammalian gene function. *Science* **337:** 421-422

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J (2003) Systematic

functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* **421:** 231-237

Kim DU, Hayles J, Kim D, Wood V, Park HO, Won M, Yoo HS, Duhig T, Nam M, Palmer G, Han S, Jeffery L, Baek ST, Lee H, Shim YS, Lee M, Kim L, Heo KS, Noh EJ, Lee AR *et al.* (2010) Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. *Nat Biotechnol* **28:** 617-623

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, Clark AG, Bustamante CD (2008) Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451:** 994-997

Luo B, Cheung HW, Subramanian A, Sharifnia T, Okamoto M, Yang X, Hinkle G, Boehm JS, Beroukhim R, Weir BA, Mermel C, Barbie DA, Awad T, Zhou X, Nguyen T, Piqani B, Li C, Golub TR, Meyerson M, Hacohen N *et al.* (2008) Highly parallel identification of essential genes in cancer cells. *Proc Natl Acad Sci U S A* **105:** 20380-20385

Luo T, Masson K, Jaffe JD, Silkworth W, Ross NT, Scherer CA, Scholl C, Frohling S, Carr SA, Stern AM, Schreiber SL, Golub TR (2012) STK33 kinase inhibitor BRD-8899 has no effect on KRAS-dependent cancer cell viability. *Proc Natl Acad Sci U S A* **109:** 2860-2865

Makino T, Hokamp K, McLysaght A (2009) The complex relationship of gene duplication and essentiality. *Trends in genetics : TIG* **25:** 152-155

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM (2013) RNA-guided human genome engineering via Cas9. *Science* **339:** 823-826

Marcotte R, Brown KR, Suarez F, Sayad A, Karamboulas K, Krzyzanowski PM, Sircoulomb F, Medrano M, Fedyshyn Y, Koh JL, van Dyk D, Fedyshyn B, Luhova M, Brito GC, Vizeacoumar FJ, Vizeacoumar FS, Datti A, Kasimer D, Buzina A, Mero P *et al.* (2012) Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Discov* **2:** 172-189

Moffat J, Reiling JH, Sabatini DM (2007) Off-target effects associated with long dsRNAs in Drosophila RNAi screens. *Trends Pharmacol Sci* **28:** 149-151

Nijhawan D, Zack TI, Ren Y, Strickland MR, Lamothe R, Schumacher SE, Tsherniak A, Besche HC, Rosenbluh J, Shehata S, Cowley GS, Weir BA, Goldberg AL, Mesirov JP, Root DE, Bhatia SN, Beroukhim R, Hahn WC (2012) Cancer vulnerabilities unveiled by genomic loss. *Cell* **150:** 842-854

Paigen K (2003) One hundred years of mouse genetics: an intellectual history. I. The classical period (1902-1980). *Genetics* **163:** 1-7

Popova T, Manie E, Stoppa-Lyonnet D, Rigaill G, Barillot E, Stern MH (2009) Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biol* **10:** R128

Schlabach MR, Luo J, Solimini NL, Hu G, Xu Q, Li MZ, Zhao Z, Smogorzewska A, Sowa ME, Ang XL, Westbrook TF, Liang AC, Chang K, Hackett JA, Harper JW, Hannon GJ, Elledge SJ (2008) Cancer proliferation gene discovery through functional genomics. *Science* **319:** 620-624

Scholl C, Frohling S, Dunn IF, Schinzel AC, Barbie DA, Kim SY, Silver SJ, Tamayo P, Wadlow RC, Ramaswamy S, Dohner K, Bullinger L, Sandy P, Boehm JS, Root DE, Jacks T, Hahn WC, Gilliland DG (2009) Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. *Cell* **137:** 821-834

Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelson T, Heckl D, Ebert BL, Root DE, Doench JG, Zhang F (2013) Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science*

Shao DD, Tsherniak A, Gopal S, Weir BA, Tamayo P, Stransky N, Schumacher SE, Zack TI, Beroukhim R, Garraway LA, Margolin AA, Root DE, Hahn WC, Mesirov JP (2012) ATARiS: Computational quantification of gene suppression phenotypes from multi-sample RNAi screens. *Genome Res*

Shao DD, Tsherniak A, Gopal S, Weir BA, Tamayo P, Stransky N, Schumacher SE, Zack TI, Beroukhim R, Garraway LA, Margolin AA, Root DE, Hahn WC, Mesirov JP (2013) ATARiS: computational quantification of gene suppression phenotypes from multisample RNAi screens. *Genome Res* **23:** 665-678

Silva JM, Marran K, Parker JS, Silva J, Golding M, Schlabach MR, Elledge SJ, Hannon GJ, Chang K (2008) Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science* **319:** 617-620

Solimini NL, Liang AC, Xu C, Pavlova NN, Xu Q, Davoli T, Li MZ, Wong KK, Elledge SJ (2013) STOP gene Phactr4 is a tumor suppressor. *Proc Natl Acad Sci U S A* **110:** E407-414

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337:** 64-69

Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigo R (2012) Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* **22:** 1616-1625

Tu Z, Wang L, Xu M, Zhou X, Chen T, Sun F (2006) Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics* **7:** 31

Vizeacoumar FJ, Arnold R, Vizeacoumar FS, Chandrashekhar M, Buzina A, Young JT, Kwan JH, Sayad A, Mero P, Lawo S, Tanaka H, Brown KR, Baryshnikova A, Mak AB,

Fedyshyn Y, Wang Y, Brito GC, Kasimer D, Makhnevych T, Ketela T *et al.* (2013) A negative genetic interaction map in isogenic cancer cell lines reveals cancer cell vulnerabilities. *Mol Syst Biol* **9:** 696

Wang T, Wei JJ, Sabatini DM, Lander ES (2013) Genetic Screens in Human Cells Using the CRISPR/Cas9 System. *Science*

Weiwer M, Spoonamore J, Wei J, Guichard B, Ross NT, Masson K, Silkworth W, Dandapani S, Palmer M, Scherer CA, Stern AM, Schreiber SL, Munoz B (2012) A Potent and Selective Quinoxalinone-Based STK33 Inhibitor Does Not Show Synthetic Lethality in KRAS-Dependent Cells. *ACS medicinal chemistry letters* **3:** 1034-1038

White JK, Gerdin AK, Karp NA, Ryder E, Buljan M, Bussell JN, Salisbury J, Clare S, Ingham NJ, Podrini C, Houghton R, Estabel J, Bottomley JR, Melvin DG, Sunter D, Adams NC, Sanger Institute Mouse Genetics P, Tannahill D, Logan DW, Macarthur DG *et al.* (2013) Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. *Cell* **154:** 452-464

Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, Davis K, Dietrich F, Dow SW, El Bakkoury M, Foury F, Friend SH, Gentalen E, Giaever G *et al.* (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285:** 901-906

Zotenko E, Mestre J, O'Leary DP, Przytycka TM (2008) Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput Biol* **4:** e1000140

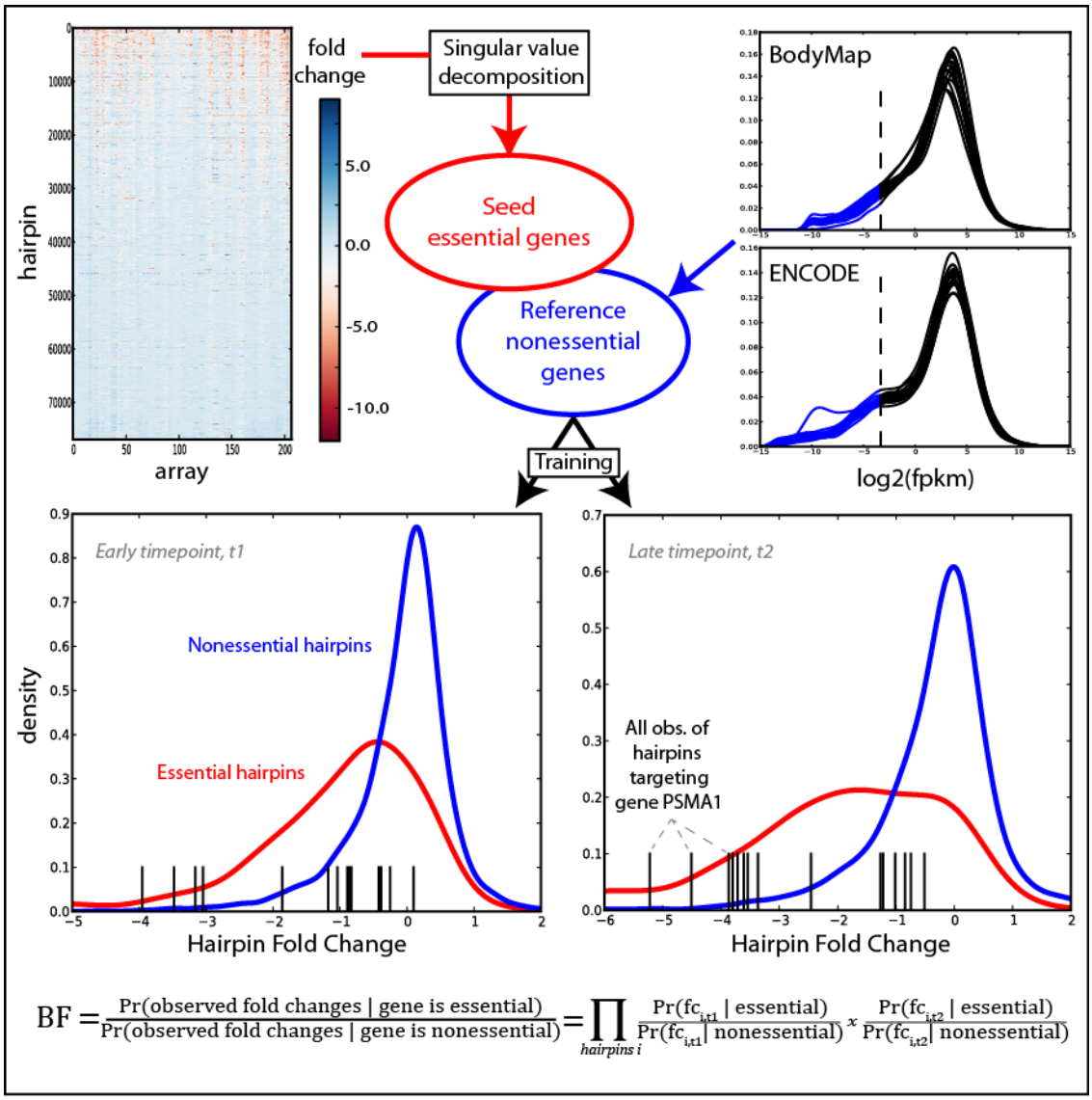

Figure 1. Analytical overview. Half of the matrix of shRNA hairpins was decomposed using linear algebra techniques to find a set of reference essential genes. Reference nonessentials were derived from low-expression genes across a compendium of RNA-seq experiments. For each cell line/timepoint in the 2nd half of the shRNA data, the empirical distributions of training essentials and nonessentials were determined, and for each remaining gene a Bayes Factor (BF) is calculated which measures which distribution its cognate hairpin data most closely matches.

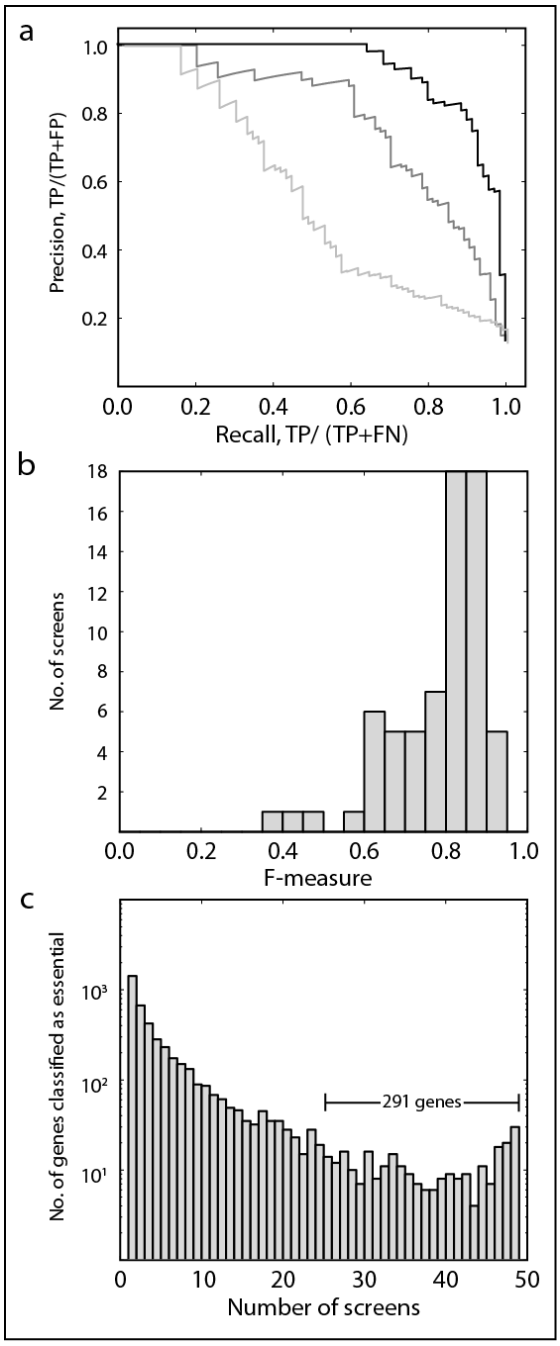

Figure 2. Screen quality and core essentials. (a) For each screen, genes are ranked by BF and evaluated against a test set of reference essentials and nonessentials, and a precision vs. recall (PR) curve is calculated. Three screens representing the variability in global performance are shown. (b) Distribution of F-measures of the 68 screens used in this study. Screens with F-measure > 0.75 (n=48) were considered high-performing and were retained for downstream analyses. (c) Histogram of essential gene observations across the 48 performing cell lines. Genes essential in 24/48 lines (n=291) were considered core essentials. Genes observed in only 1-3 cell lines are highly enriched for false positives.

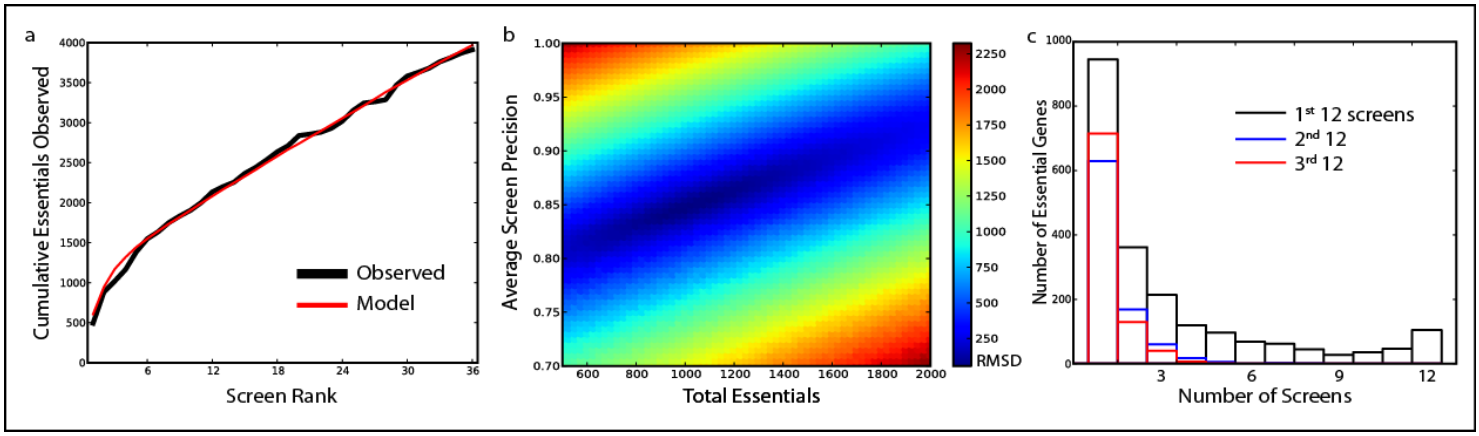

Figure 3. The cumulative model of essential genes. (a) The top 36 cell lines were rank ordered by F-measure and the cumulative count of classified essential genes was plotted (black curve). Simulated repeat experiments sampling a population of 1,025 essential genes at 15% FDR yield a similar cumulative count (red curve). (b) In simulated repeat experiments across parameter space, models sampling 875-1,175 essential genes at 13.5-16.5% FDR (1-Precision) yielded cumulative observation curves similar to what was observed experimentally. (c) Histogram of observations of essential genes in top-ranked 12 screens (black), genes exclusive to the next set of 12 (blue), and exclusive to the 3rd set of 12 (red). Genes observed in at least 3 of the top 12 screens are classified as global essentials.

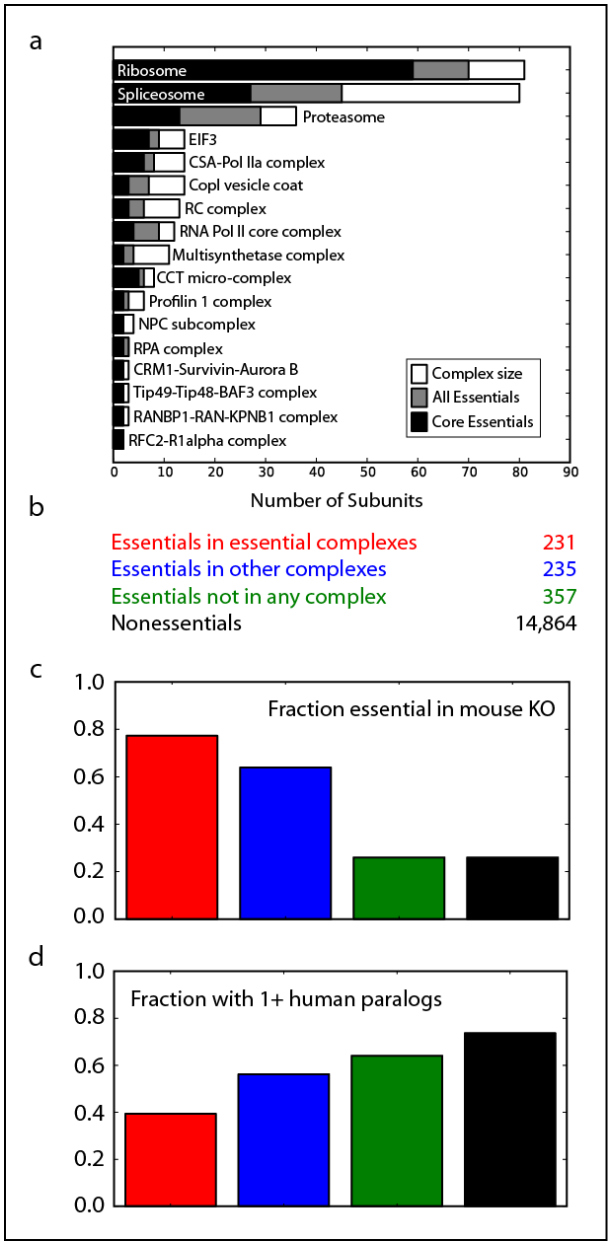

Figure 4. Characteristics of essential genes. (a) Essential genes are highly enriched for core protein complexes. Seventeen representative nonoverlapping complexes are shown, with the core essentials (black) and total essentials (gray) shown relative to the total number of subunits in the complex. (b) Total essentials are separated into categories: those in complexes enriched for essential genes, those in other complexes but which fail enrichment tests, and those not annotated to be in any protein complex. The remaining genes are classified as nonessential. (c) Fraction of genes in each category whose mouse orthologs are also essential; colors as in (b). (d) Fraction of genes in each category with one or more human paralogs; colors as in (b).

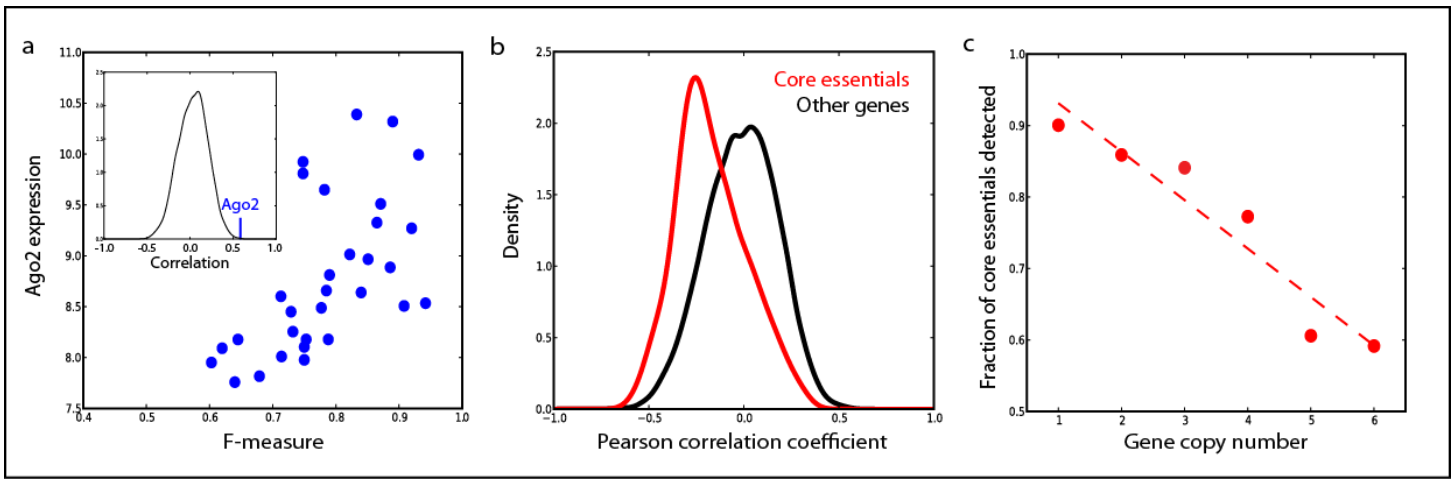

Figure 5. Biological drivers of variation in RNAi screen efficacy. (a) Plotting Ago2 gene expression (measured by microarrays; y-axis) vs. cell line F-measure (x-axis) for pancreatic and ovarian cancer cell lines reveals strong correlation (Pearson's r=0.59). Inset, distribution of correlations of expressed genes (n=10,673) vs. F-measure; Ago2 is the top ranked gene. (b) The Pearson correlation coefficient of absolute copy number vs. Bayes Factor was determined for all genes across 30 pancreatic and ovarian cancer cell lines. Core essential genes show a negative correlation between copy number and essentiality. (c) Core essential genes were binned by absolute copy number across the 30 samples. In each bin, the fraction of core essentials that were accurately classified in the corresponding screens is plotted. High copy number among core essentials reduces sensitivity to RNAi.

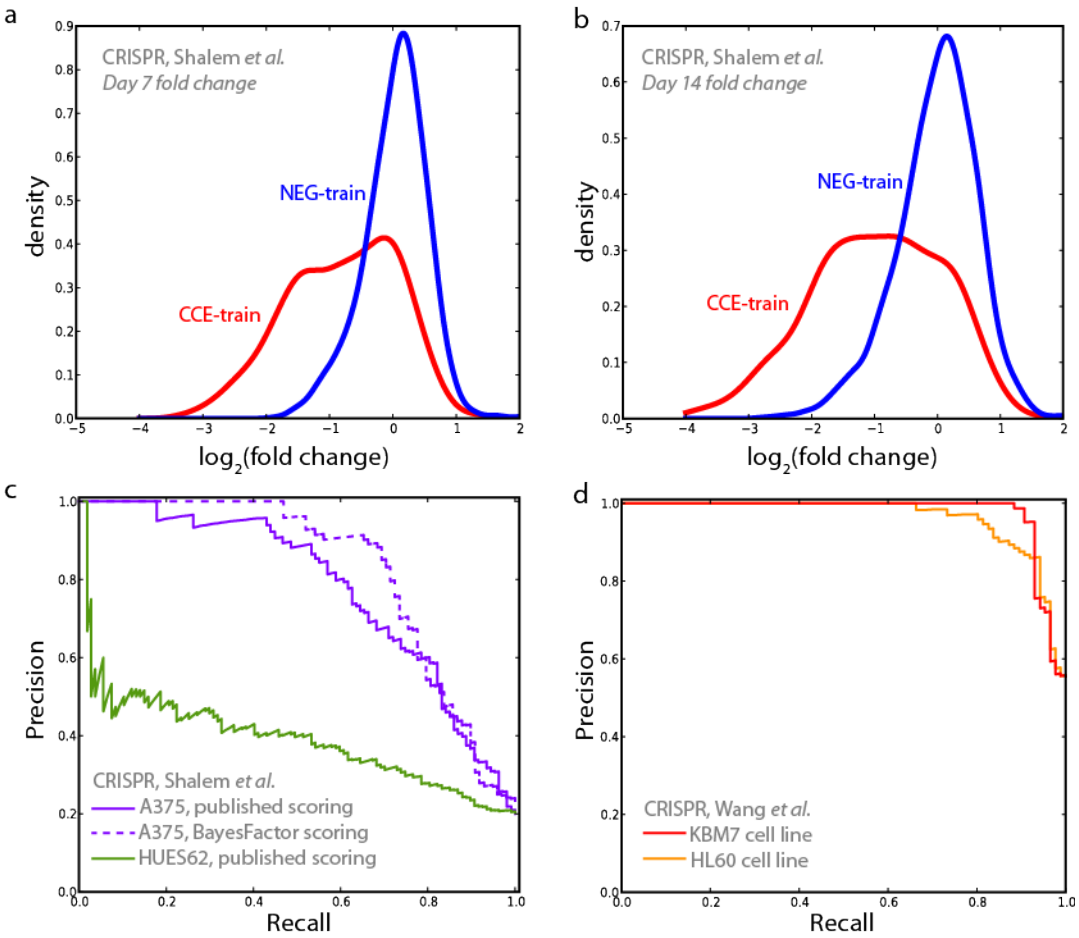

Figure 6. Evaluating CRISPR negative selection screens. (a,b) The fold change distributions of gRNA targeting reference essential and nonessential genes in Shalem *et al.* (Shalem *et al.*, 2013) are similar to those shown by shRNA hairpins (see Fig 1) and enable the application of the Bayes Factor approach. (c) Published results from Shalem *et al.*, evaluated against CCE-test and NE-test. Dashed line shows that Bayes Factor approach more accurately captures essential genes in the A375 screen, the only screen for which raw data is available. (d) Whole screen results from Wang *et al.* (Wang *et al.*, 2013), evaluated against the same sets. NE-test genes are underrepresented in the Wang *et al.* gRNA library, which gives the appearance of an artificial boost in precision when compared to the Shalem *et al.* results.

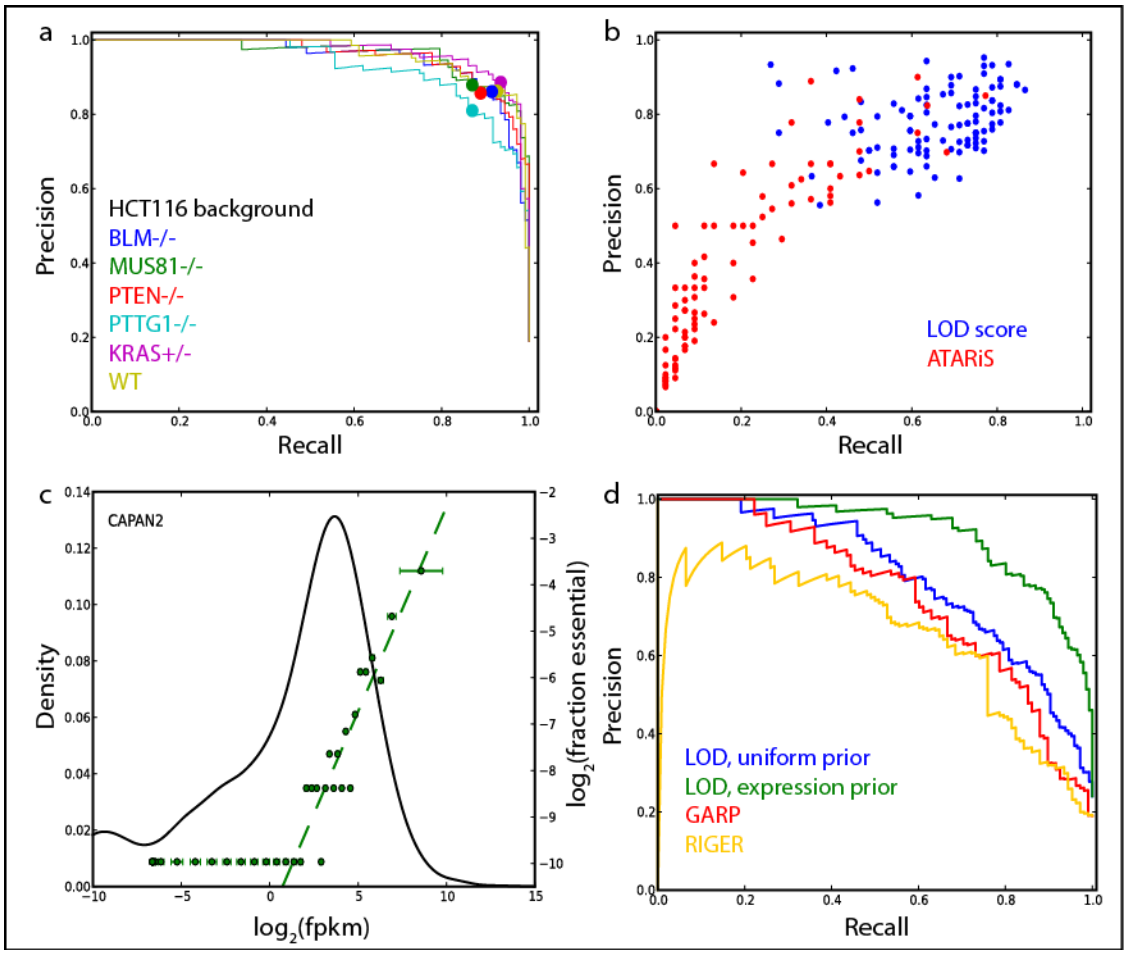

Figure 7. Evaluating other shRNA data and methods. (a,b) Evaluating other RNAi data sets. (a) LOD scores were calculated for the pooled library shRNA screens in the HCT116 background in (Vizeacoumar *et al.*, 2013) and evaluated against CCE-test and NE-test. Recall, TP/(TP+FN); Precision, TP/(TP+FP). All six screens showed very high accuracy. The filled circle indicates the point on the curve where LOD=0. (b) LOD scores were calculated for the pooled library shRNA screens in 102 cancer cell lines in (Cheung *et al.*, 2011). Blue, points represent recall & precision at LOD=0 as measured against CCE-test and NE-test. Red, recall & precision for the same cell lines and same reference sets from ATARiS gene solutions at phenotype score=-1. (c) Integrating gene expression into the Bayesian classifier. For RNAi screens with matched gene expression data (in this example, PDAC cell line CAPAN-2, black curve), genes are binned by expression level and the fraction of reference essentials in each bin (right Y axis) is plotted against the mean expression of genes in the bin (green points). A linear fit on the log-log plot (green dashed line) can be integrated into the Bayesian classifier as an informative prior. (d) Integrating expression data improves the performance of the classifier (green) over the base algorithm (blue). Both forms show better performance than other algorithms such as GARP (red) and RIGER (gold).

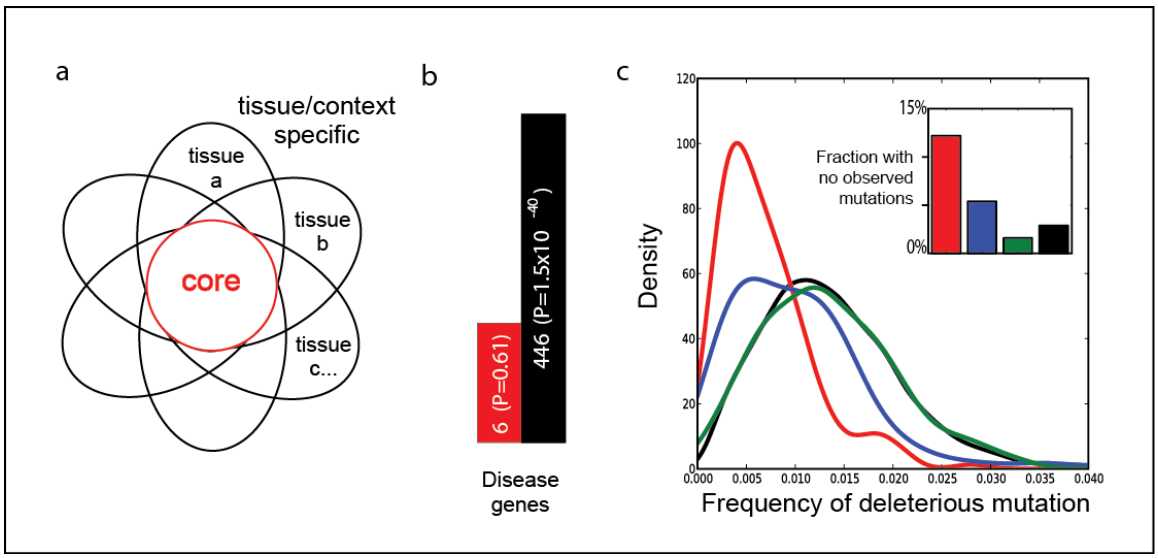

Figure 8. The Daisy model of gene essentiality. (a) The Daisy model, where each petal represents a tissue or context in which a gene is essential. Petals overlap to varying degrees but all share a core set of essential housekeeping genes that should be detectable in any cell-based assay. Whole-organism studies will sample from the whole flower, not specific petals. (b) Human orthologs of mouse essential genes were divided into core and non-core ("peripheral") essentials. Peripheral essentials show strong enrichment for disease genes while core essentials do not. (c) Frequency of putative deleterious mutation by gene class, normalized for transcript length, derived from population exome studies (Tennessen *et al.*, 2012). Inset, fraction of genes by class in which no variant was observed. Little variation is tolerated among core essentials, probably explaining the infrequency with which they are associated with disease.