Bioinformatics, 32(2), 2016, 260–267
doi: 10.1093/bioinformatics/btv556
Advance Access Publication Date: 28 September 2015
Original Paper

Systems biology

ScreenBEAM: a novel meta-analysis algorithm for functional genomics screens via Bayesian hierarchical modeling

Jiyang Yu[1,†,*], Jose Silva[2,*] and Andrea Califano[1,*]

[1]Department of Biomedical Informatics, Department of Systems Biology, Center for Computational Biology and Bioinformatics, Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY 10032, USA and [2]Department of Pathology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

*To whom correspondence should be addressed.
[†]Present address: Department of Precision Medicine, Oncology Research Unit, Pfizer Inc., Pearl River, NY 10965, USA
Associate Editor: Janet Kelso

Received on March 12, 2015; revised on September 9, 2015; accepted on September 21, 2015

Abstract

Motivation: Functional genomics (FG) screens, using RNAi or CRISPR technology, have become a standard tool for systematic, genome-wide loss-of-function studies for therapeutic target discovery. As in many large-scale assays, however, off-target effects, variable reagents' potency and experimental noise must be accounted for appropriately control for false positives. Indeed, rigorous statistical analysis of high-throughput FG screening data remains challenging, particularly when integrative analyses are used to combine multiple sh/sgRNAs targeting the same gene in the library.
Method: We use large RNAi and CRISPR repositories that are publicly available to evaluate a novel meta-analysis approach for FG screens via Bayesian hierarchical modeling, Screening Bayesian Evaluation and Analysis Method (ScreenBEAM).
Results: Results from our analysis show that the proposed strategy, which seamlessly combines all available data, robustly outperforms classical algorithms developed for microarray data sets as well as recent approaches designed for next generation sequencing technologies. Remarkably, the ScreenBEAM algorithm works well even when the quality of FG screens is relatively low, which accounts for about 80–95% of the public datasets.
Availability and implementation: R package and source code are available at: https://github.com/jyyu/ScreenBEAM.
Contact: ac2248@columbia.edu, jose.silva@mssm.edu, yujiyang@gmail.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recent technological advances have significantly improved our ability to perform systematic and informative functional genomics (FG) studies in mammalian cells. Specifically, during the past decade, RNA interference (RNAi) has become a standard technique for studying phenotype-specific gene function via suppression of gene specific mRNA expression or translation (Cheung *et al.*, 2011; Hammond *et al.*, 2000; Silva *et al.*, 2008, 2009). More recently, the CRISPR/Cas9 system (Cong *et al.*, 2013; Mali *et al.*, 2013) has emerged as an even more effective tool to implement complete gene knock-out. As a result, loss of function genome-scale screens have become popular, especially in pooled shRNA library format (Moffat *et al.*, 2006; Paddison *et al.*, 2004; Silva *et al.*, 2004, 2005). These have been widely used not only to identify essential genes in a specific context (Konig *et al.*, 2007; Luo *et al.*, 2008; Marcotte *et al.*, 2012; Zhou *et al.*, 2014), but also genes that are differentially essential in different contexts (Deng *et al.*, 2014; Hahn *et al.*, 2009; Zuber *et al.*, 2011) or synthetic-lethal genes (Hoffman *et al.*, 2014;

Luo *et al.*, 2009). Results from these screens may ultimately inform discovery of novel therapeutic targets in tissue-specific, subtype-specific, or mutation-specific contexts, not only in cancer but in many other diseases and physiologic contexts. FG screen design has also been extended to identify tumor suppressors that increase cell growth upon repression (Silva *et al.*, 2009) or genes that modulate drug sensitivity (Prahallad *et al.*, 2012). Complementing direct knock out methodologies, the CRISPR/Cas9 gene-editing technology has also been scaled up for high-throughput FG screens to study the effect of specific mutations on cellular phenotypes (Shalem *et al.*, 2014; Wang *et al.*, 2014; Zhou *et al.*, 2014).

However, progress in large-scale FG screens has been hampered by a nontrivial modelling of false positives (Kaelin, 2012). For example, many candidate therapeutic targets identified from shRNA screens have failed validation in independent assays (Babij *et al.*, 2011; Begley and Ellis, 2012; Prinz *et al.*, 2011). Among the possible reasons for lack of robustness, one must consider RNAi off-target effects, variable potency and knock-down/out efficiency of RNAi and CRISPR reagents, biological and technical noise in high-throughput screens (Echeverri *et al.*, 2006; Hart *et al.*, 2014; Shao *et al.*, 2013), cell line-specific efficiency of viral pool infection and toxicity from additional viral vector expression cassettes (e.g. fluorescence reporter proteins). This suggests that sophisticated statistical methods may be required to model the complexity of these experiments, thus reducing false positive and negative rates and leading to improved, more robust interpretation of FG screen results.

A key computational challenge, in genome-scale FG screens, is to score gene-level activity from individual reagents. Whole genome FG libraries normally include multiple optimally designed shRNAs hairpins or CRISPR sgRNAs targeting the same gene to increase the likelihood of an effective gene knock-down/out. For instance, on average 2–3 or 5 hairpins per gene were used with GIPZ (Paddison *et al.*, 2004; Silva *et al.*, 2004, 2005) and TRC (Moffat *et al.*, 2006) shRNA libraries, respectively. Similarly 4 and 10 sgRNAs per gene were used in the Zhang (Shalem *et al.*, 2014) or Sabatini_Lander (Wang *et al.*, 2014) CRISPR libraries, respectively. It is thus critical that such diverse evidence from multiple sh/sgRNAs targeting a gene is integrated when assessing its contribution to a specific endpoint phenotype. In addition, both microarray (Silva *et al.*, 2004) and next generation sequencing technologies (NGS) (Sims *et al.*, 2011) technologies have been used to quantitatively assess sh/sgRNAs representation or abundance in pooled FG screens, each one introducing bias and measurement noise.

Traditional methods to summarize gene-level activity usually rely on single-probe level analysis. Specifically, shRNA are first individually scored and then the scores of representative (e.g. high-scoring) shRNAs or of all shRNAs targeting a specific gene are combined. Several algorithms have been proposed to select or combine shRNA-level evidence, including choosing the second best or most depleted shRNA (Luo *et al.*, 2008) (RIGER_SB), averaging the two shRNAs that produced the largest scores (Luo *et al.*, 2008) (RIGER_WS), performing enrichment analysis of all shRNAs targeting one gene against all shRNAs in the library (Luo *et al.*, 2008) (RIGER_KS), comparing rank distributions of effective size of all shRNAs per gene (Konig *et al.*, 2007) (RSA) and more recent model-based MAGeCK (Li *et al.*, 2014) and HitSelect (Diaz *et al.*, 2015). An intrinsic limitation with all of these approaches is that they rely on the accurate assessment of an individual sh/sgRNA activity, which is difficult to achieve in large-scale screens that typically have a relatively small number of replicate samples. Moreover, off-target effects, variable silencing efficiency, differences among sh/sgRNAs targeting the same gene, and experimental/technical noise make heuristic selection of representative sh/sgRNAs problematic, causing significant false discovery rates.

To overcome these limitations, we propose a novel ScreenBEAM (Screening Bayesian Evaluation and Analysis Method) algorithm via Bayesian hierarchical modeling to directly assess gene-level activity from all relevant measurements. Due to its robustness, hierarchical modeling (Gelman *et al.*, 2004; Gelman and Hill, 2007), also known as multilevel modeling, has been increasingly valuable in large-scaled 'omics studies (Ji and Liu, 2010). In this context, ScreenBEAM algorithm analyses all sh/sgRNAs targeting the same gene as a set, instead of one at the time, and then fits a linear mixture model that directly models the potential activity variability of different hairpins, as a random effect. This 'multi-probe' analysis strategy improves parameter estimation, by increasing sample size, and reduces prediction error and false positive rate, by integrating information from multiple shRNAs. Use of Bayesian inference with Markov chain Monte Carlo (MCMC) techniques, in this analysis, further improves accuracy and robustness of scoring metrics.

Systematic benchmark assays, using large-scale, publicly available shRNA (RNAi) and sgRNA (CRISPR) screens designed to profile gene essentiality by microarray (Marcotte *et al.*, 2012) or NGS (Cowley *et al.*, 2014), suggest that the ScreenBEAM method robustly outperforms existing single-probe analysis algorithms. The ScreenBEAM algorithm improvements are especially significant with assays with lower data quality, which accounts for about 80-95% of the FG screens considered in this manuscript.

2 Methods

2.1 Benchmark microarray-based datasets

To benchmark different algorithms for meta-analysis of shRNAs targeting the same gene, we selected three representative datasets (MCF7, HPAFII and OVCAR5) from a panel of 72 microarray-based RNAi screens designed for profiling essential genes in breast, pancreatic and ovarian cancer cell lines (Marcotte *et al.*, 2012). These screens use 80K TRC hairpins combined in a pooled library to target 16K human genes, with an average of 5 shRNAs per gene. In the original study (Marcotte *et al.*, 2012), cells representative of each cell line were collected at three time points, including T0, to study shRNA silencing dynamics. However, to detect depleted hairpins, representing genes that are essential for cell viability, we consider only T0 and the final time point in this analysis. Thus, depleted genes that are essential for cell viability are identified independent of whether they represent early or late events. Such a two-time-point design is also generally applied in literature for cost consideration, relying on long selection times to capture late-dropped-out genes.

2.2 Benchmark NGS-based Achilles datasets

Similarly, to evaluate performance of different algorithms on NGS-based FG screens, we selected three representative datasets (OVCAR8, HPAFII and FUOV1) and seven additional randomly-selected cell lines (Supplementary Fig. S5) from the Achilles data (Cowley *et al.*, 2014), the largest repository of NGS-based shRNA dropout screens across 216 cancer cell lines using 80K TRC library.

2.3 Additional NGS-based CRISPR datasets

We selected a NGS-based CRISPR dropout screen on HL60 cells (Wang *et al.*, 2014) which also had NGS-based RNAi screening data available (Cowley *et al.*, 2014) and used the overlap of top hits between CRISPR and RNAi as a metric to evaluate difference algorithms.

2.4 Data quality metrics

Importantly, the three selected cell lines in microarray or NGS-based dataset also represent three different levels of screen data quality, indicated by a metric MRC (minimum replicate consistency). MRC is defined as the minimal pair-wise Pearson correlation among biological replicates for each group of a dataset, which represents the consistency of biological replicates in a dataset. For example, there are three replicates of each T0 or TX group for MCF7, so MRC for MCF7 dataset will be the minimal Pearson correlation of three pair-wise T0 and three pair-wise TX comparisons. Surprisingly, the emerging NGS-based genome-wide shRNA screens have worse data quality than traditional microarray-based data as the percentage of high-quality data (MRC > 0.9) is 6% versus 22% (Supplementary Fig. S1).

In the microarray dataset, MCF7, HPAFII and OVCAR5 represent high (MRC > 0.9), medium (MRC between 0.8 and 0.9) or low (MRC < 0.8) data quality categories respectively (Supplementary Fig. S2), each of which accounts for 22, 50 and 28% of the total 72 screens. Similarly, in the NGS dataset, OVCAR8, HPAFII and FUOV1 represent high (MRC > 0.9), medium (MRC between 0.75 and 0.9) or low (MRC < 0.75) data quality categories respectively (Supplementary Fig. S4), each of which accounts for 6%, 62% and 32% of the total 255 screens.

2.5 Reference essential genes

Without a gold standard in place for selecting human essential genes, housekeeping and evolutionary-conserved genes (most likely critical for cell viability) were used as positive controls (Marcotte *et al.*, 2012). We thus use the same criterion to compare the ScreenBEAM approach to existing algorithms for essential gene inference. In this study, we collected four independent gene sets into a positive control 'gold standard' (Supplementary Table S1). This includes two sets reported in a previous study (Marcotte *et al.*, 2012) and two from more recent studies on human housekeeping genes (Chang *et al.*, 2011; Tu *et al.*, 2006). Common genes ($N = 222$) across four independent reference sets, which are most likely to be essential across different cell lines, were used as the gold standard set. Housekeeping or orthologous genes that were not present in the shRNA library were filtered out. In additional to the above knowledge-based house-keeping genes, we also used a recent essential gene list identified from genome-wide RNAi screens of 48 cancer cell lines (Hart *et al.*, 2014). We selected genes ($N = 272$) that were identified as essential in over 24 (50%) out of 48 cell lines as an independent gold standard reference of essential genes (Supplementary Table S1). We then assessed the percent overlap between the gold standard set and the top k genes predicted as essential by each method. To avoid selection bias on k, we sampled k from 0 up to 2000, with a sliding widow of 5. Algorithms were then compared, based on the consistency of the overlap of their predictions with the gold standard set.

3 Results

3.1 Profiling cell essential genes by microarray or NGS-based RNAi screens

As described in the previous section, negative genome-wide RNAi screening is commonly used to identify essential genes for proliferation and viability in cancer cells. A typical procedure to identify essential genes using microarray or NGS-based pooled sh/sgRNA screens is shown in Figure 1A. The pool of sh/sgRNA plasmid vectors is transduced into a target cell population at a multiplicity of

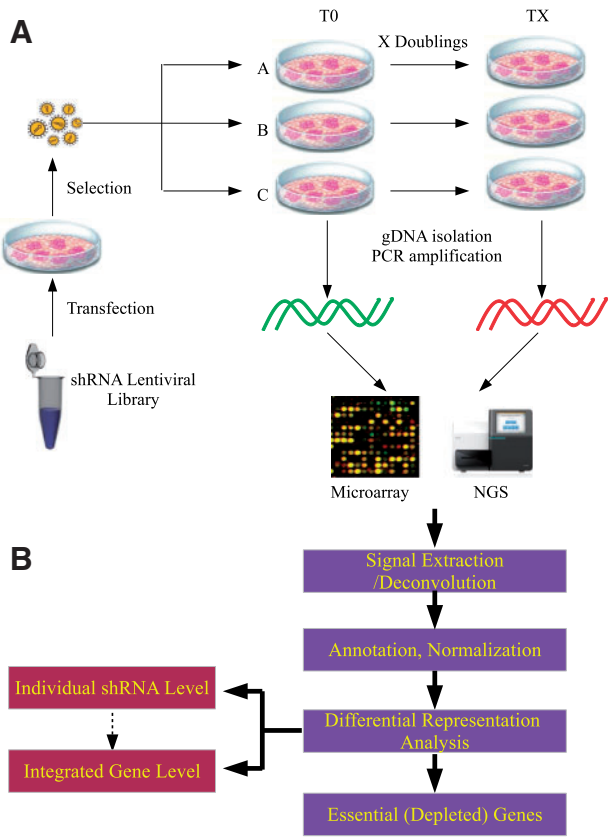

Fig. 1. Outline of a typical negative sh/sgRNA screen by microarray or NGS to profile cell essential genes showing (**A**) experimental procedures and (**B**) analysis pipelines

0.3 to achieve no more than one shRNA per cell. Transduced cells are then harvested at a time point *TX*, generally representing a predefined multiple of the cell specific doubling time, in triplicates. Genomic DNAs are then extracted from cells collected at T0 and TX, PCR-amplified, labeled and hybridized to a multiplexed microarray or indexed for multiplexing deep sequencing.

Analysis of microarray or NGS readout involves several steps (Fig. 1B). Intensity signals or sh/sgRNA sequence counts representing relative abundance are extracted or deconvoluted, and processed with background correction and normalization. Differential representation analysis (DRA) on normalized data at TX and T0 time points are performed to identify sh/sgRNAs that are underrepresented at time TX, whose targets are candidate essential genes. Since each gene is targeted by multiple sequences, DRA can be conducted at the individual sh/sgRNA level and then integrated at the gene level. Classical single-probe approaches produce a statistical confidence measure (score) for each sh/sgRNAs separately and then combine the scores of the sequences targeting the same gene to derive a gene-level score; in contrast, the proposed multi-probe approach fits a hierarchical model to the data representing all of the sequences targeting a specific gene at once, to directly estimate a global gene-level activity.

3.2 Classical single-probe approach on microarray-based data

All previous methods to estimate activity of genes targeted by multiple sequences in large-scale screening microarray-based data are based on a single-probe analysis strategy as illustrated in an example

from benchmark datasets that a gene is targeted by three shRNA clones (Fig. 2A). There are two well-developed algorithms for this type of approach: RNAi gene enrichment ranking (Luo *et al.*, 2008) (RIGER) and redundant siRNA activity (Konig *et al.*, 2007) (RSA). RIGER has three sub-algorithms to integrate multiple shRNA scores including Kolmogorov-Smirnov statistic-based enrichment analysis (RIGER_KS), weighted sum of the best two hairpins (RIGER_WS) and the second best hairpin (RIGER_SB). RSA employs a hypergeometric distribution or Fisher's exact test-based statistical method to rank gene activities. These algorithms can be reclassified into summarizing all shRNAs targeting the same gene (RIGER_KS and RSA) and heuristic selection of representative shRNAs (RIGER_SB and RIGER_WS).

Various metrics have been proposed to score individual shRNA behavior (Fig. 2A) at TX and T0 time points including Student's *t*-statistic, *z*-statistic, coefficient of linear regression model, signal to noise ratio, logarithm of fold change and difference of mean. Student's *t*-statistic or *z*-statistic of coefficient in linear model is commonly used due to their statistical integration of replicate variance. Student's *t*-test is equivalent to a linear model with Gaussian noise. However, with the fact that the sample size in this context is usually small, a Bayesian linear model with a Gaussian prior for coefficients (Fig. 2B) is suggested for its robustness.

3.3 ScreenBEAM: multi-probe analysis using a Bayesian hierarchical model

Instead of a single-probe, two-step analysis, we propose a single multi-probe analysis, ScreenBEAM, to fit a complex hierarchical or multilevel model using Bayesian approach into data of all shRNAs targeting the same gene. To address the risk of inaccurate estimation due to small sample size and microarray noise, we formulate this approach as a Bayesian hierarchical model (BHM). A BHM introduces an additional level to classical linear fitting model, where the parameters representing the silencing effect of each shRNA group (indexed by j) are drawn from a distribution (Fig. 3). In terms of the prior distributions of the parameters in BHM, we employ conjugate priors, i.e. Gaussian model for the coefficients and an Inverse-Gamma distribution for the variance for computational convenience (Fig. 3). To get robust estimation of the parameters, we use Markov chain Monte Carlo (MCMC) simulations. The multilevel model can be rewritten as a linear mixture model in which 'fixed effects' correspond to gene-level activity and 'random effects' reflect silencing efficiency differences by multiple shRNAs targeting the same gene (Fig. 3).

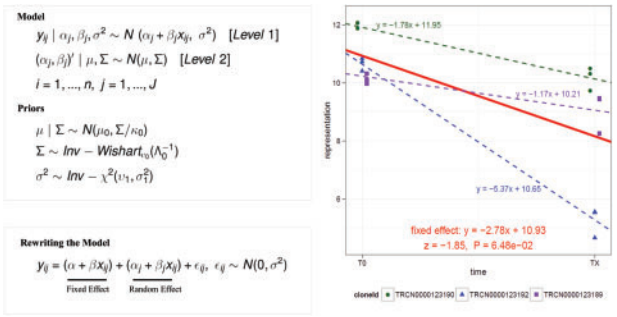

Fig. 3. Global approach ScreenBEAM via Bayesian hierarchical modeling. Model: the data of all shRNAs targeting one gene can be fit by a hierarchical model, in which the extra level is indexed by j, indicating the shRNA group the sample belongs to. Sample index i is up to n, the total number of samples for one gene; j is up to J, the number of shRNA classes. Parameter μ, a vector of slope and intercept, reflects the gene-level activity and allows variation for each shRNA class. Conjugate priors are set for parameters. Rewriting the model: the model can be rewritten to a two-component mixture model in which 'fixed effect' corresponds to gene-level behavior and 'random effect' indicates the noise of each shRNA group. A practical application of the Bayesian hierarchical model to the example in Figure 2 is summarized in the plot. Red solid line indicates fitted gene-level/fixed effects in the model. Estimated parameters and summary statistics including *z*-statistic and *P*-value are displayed on bottom middle. Each colored dashed line reflects individual activity of each shRNA class by adding random effect to fixed gene-level effect

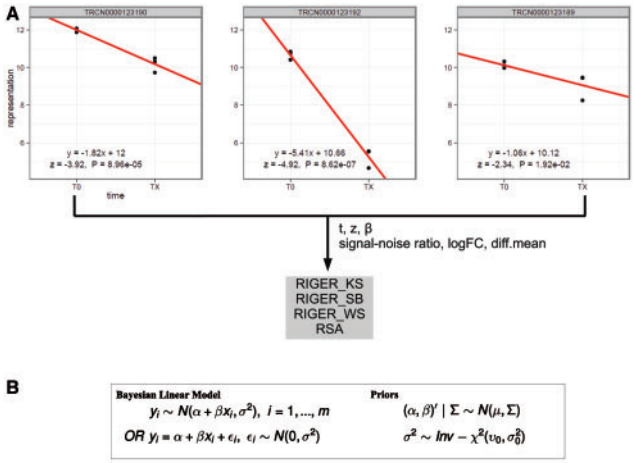

Fig. 2. Single-Probe analysis approach (**A**) An example of three shRNAs targeting KPNB1 gene from MCF7 dataset is selected to illustrate this approach. A Bayesian linear model is fit into data of each shRNA respectively. Estimated parameters (fitted lines in red) and summary statistical metrics are displayed on bottom left of each shRNA plot. Z scores and p-values are calculated by Wald test using a standard Gaussian as null distribution. Individual shRNA scores as input for algorithms to combine them can be calculated by *t* (Student's *t*-statistic), *z* (*z*-statistic of β in linear model), β (the coefficient in the linear model), signal-noise ratio (mean difference of TX versus T0 over sample standard deviation), logFC (logarithm of fold change of TX versus T0) and diff.mean (mean difference of TX versus T0). (**B**) In the linear regression model under Bayesian framework, yi indicates time point, TX or T0, and xi represents shRNA abundance for sample i; m is the sample size of the corresponding shRNA; noise follows a Gaussian distribution with mean 0 and variance σ^2; β is the parameter of interest, indicating the silencing effects on cell viability by the shRNA in consideration. As for priors, a two-variable multi-Gaussian is set for coefficients and an Inverse-Gamma is for variance. For NGS-based count, we employed log-normal distribution and performed log2-transformation of count values before fitting the above model

3.4 Microarray-based data: comparison of ScreenBEAM with RIGER and RSA

To evaluate the performance of this ScreenBEAM algorithm, compared with classical RIGER (RIGER_KS, RIGER_SB, RIGER_WS) and RSA methods we tested the overlap of their predictions with the gold standard housekeeping or orthologous genes set (see description in methods section), using three benchmark shRNA screens from the microarray-based dataset (Marcotte *et al.*, 2012). For these studies we did not use the essential gene list identified from genome-wide RNAi screens (Hart *et al.*, 2014) to prevent over-fitting due to the use of the same dataset. For the RIGER and RSA methods, we defaulted to the *t*-statistics to score individual shRNA hairpins (as implemented in their corresponding software packages). We plotted the percent overlap of the gold standard set against the top 0 up to 1000 inferred essential genes by each algorithm, using each of the three test datasets (Fig. 4). The area under the curve (AUC), based on the sensitivity or recall versus precision or prediction rate, for each algorithm, provides an unbiased metric to assess algorithm performance in terms of both false positives and prediction rates.

Analysis of these results (Fig. 4 and Supplementary Fig. S3) shows that, as expected, all methods perform consistently and significantly better than random selection. The AUC of RIGER_KS, RIGER_WS, RIGER_SB and RSA were among the lowest, across all tested datasets, suggesting little difference in performance among these three classical single-probe analysis methods. Comparing with classical RIGER and RSA, the ScreenBEAM method significantly and consistently outperformed RIGER and RSA methods in the HPAFII and OVCAR5 studies. Performance improvements were more mixed in the MCF7 cell screens, suggesting that methods have more uniformly similar behavior in these experiments. This is further explained by the fact that ScreenBEAM method's improvements are inversely proportional to data quality, i.e. the greater improvements are observed with the lower quality data. Indeed, we observed that ScreenBEAM's performance improvement over RIGER or RSA, by up to 1/2 increase of sensitivity without loss of precision, was monotonically increasing from MCF7 to HPAFII to OVCAR5 studies. This matched the decreasing pattern in noise level or replicate consistency, from high in MCF7, to medium in HPAFII, to low in OVCAR5. This confirms our expectation that a multi-probe approach implemented as a ScreenBEAM algorithm would outperform classical single-probe analysis especially when data quality was low. Critically, ~80% of the shRNA screens in the panel of 72 cancer cell lines had data quality that was consistent with that of HPAFII and OVCAR5 cells (i.e. medium to low data quality), where ScreenBEAM's contribucontribution is most significant.

For high quality screens, such as those with MCF7 cells, method selection is less relevant as all of the tested methods had similar performance. Yet, for some of the lower quality hairpins, within a high quality dataset, ScreenBEAM would still produce significant improvements. As a result, there would be potential advantages even in this kind of datasets. In addition, high quality data in high-throughput RNAi screens is rarely achieved, especially when based on microarray data, due to a variety of error and variability sources, suggesting that the ScreenBEAM method may have broad applicability and value.

3.5 NGS-based data: ScreenBEAM robustly outperforms MAGeCK and HitSelect

Next we evaluated the performance of ScreenBEAM analyzing NGS-based screens. Recently, two novel algorithms, MAGeCK (Li *et al.*, 2014) and HitSelect (Diaz *et al.*, 2015), have been developed to analyze these type of data and they have been shown to outperform RIGER and RSA. As shown in the original studies, we independently confirmed that MAGeCK and HitSelect outperform RIGER and RSA methods when analyzing NGS data (data not shown). Therefore, for simplicity, we only show here the comparison of ScreenBEAM with these two algorithms.

As surrogate measurement of performance we tested the overlap of their predictions with uboth the gold standard housekeeping and orthologous genes set and the essential gene list identified from genome-wide RNAi screens. These studies revealed that ScreenBEAM still performs the best to identify essential genes (Fig. 5A, Supplementary Fig. S5) when low-quality data sets such as FUOV1 (MRC=0.39, Supplementary Fig. S4), need to be evaluated. In high-quality OVCAR8 (MRC=0.92) and relative good-quality HPAFII

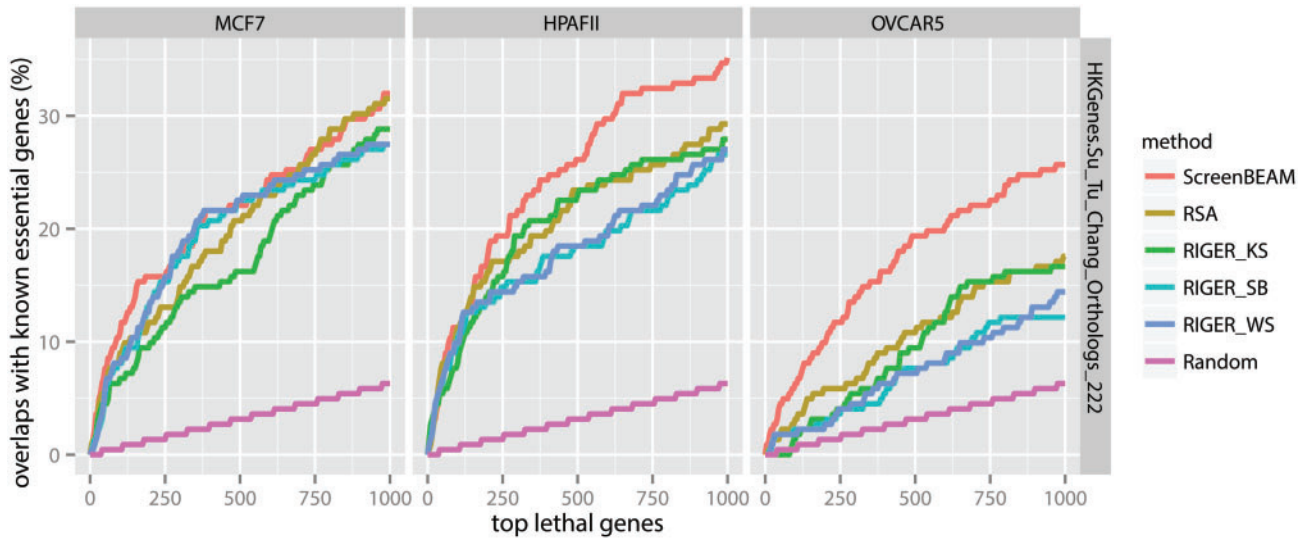

Fig. 4. Evaluation results of ScreenBEAM and existing methods on microarray-based RNAi screens. Each colored curve shows the percentage (Y axis) of gold standard essential gene set (intersection of four independent reference sets) intersected by top 0 to 1000 hits (*X* axis) predicted as essential genes by the corresponding algorithm (ScreenBEAM, RSA, RIGER_KS, RIGER_SB, RIGER_WS and Random) in each cell line dataset. The slope of 'Random' method line (in purple) is proportional to the frequency of the reference set out of all genes in the library. The greater the area under the curve, the more powerful the algorithm is

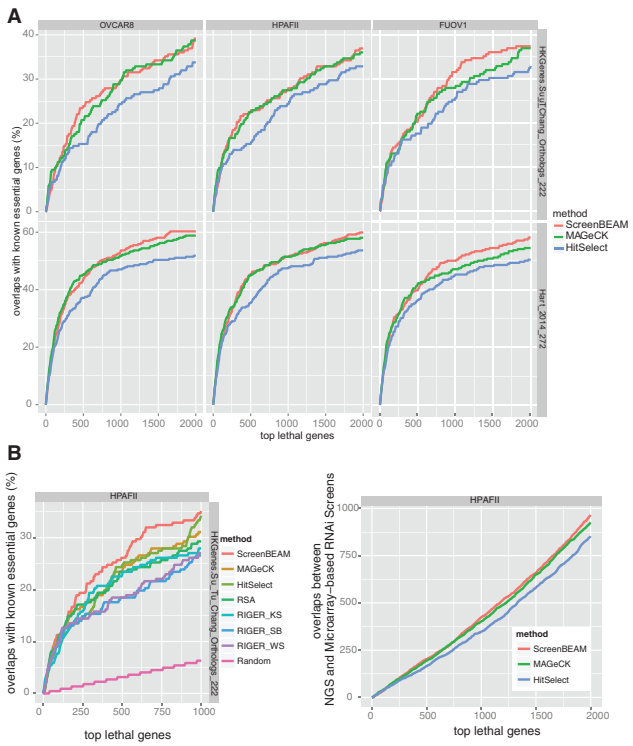

Fig. 5. Evaluation results of ScreenBEAM with the other top two algorithms on NGS-based RNAi screens using (**A**) both the housekeeping and orthologous genes set (upper panels) and the essential gene list identified from genome-wide RNAi screens (lower panel) as gold standards. (**B**) The right panel shows overlaps of top hits with microarray-based screens on the same cell line with the same shRNA library. The left panel shows the analysis of the microarray data set after this transformation

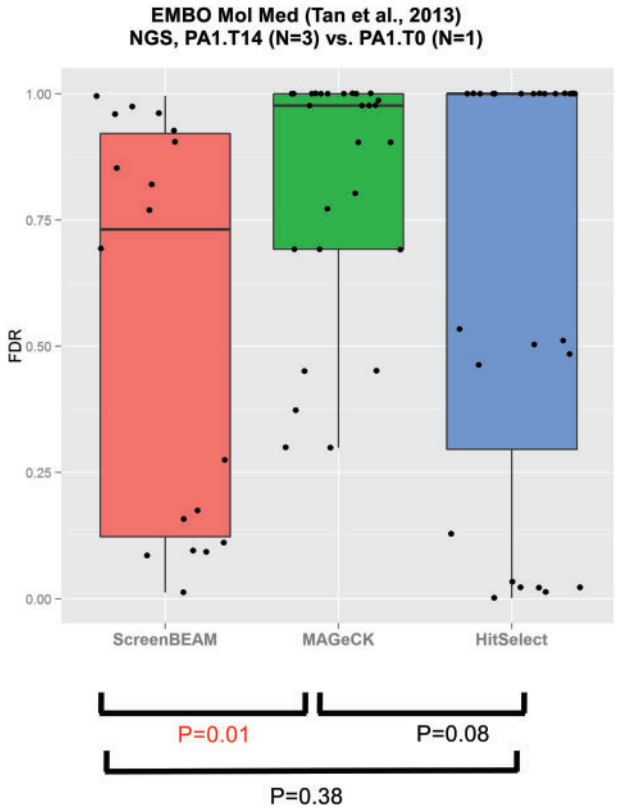

Fig. 6. FDR of experimental validated hits scored by the three top algorithms: ScreenBEAM, MAGeCK and HitSelect on Tan et al. (2013) NGS-based FG screening data (PA1 cell line). The FDRs were calculated using the Benjamini–Hochberg method. *P* values were calculated by Student's *t*-test

(MRC=0.75) ScreenBEAM and MAGeCK were indistinguishable and both consistently outperform HitSelect.

Additionally, we also compared these three algorithms by looking at the overlaps of top hits from matched NGS-based and microarray-based screens on the same cell line. In this study, it is expected that the best algorithm generate higher number of overlapping hits.

We decided to utilize the data sets from the HPAFII cell line using the TRC library as it represents the most common class (~62%) of NGS-based FG screens. To make MAGeCK and HitSelect, which only take NGS-based count data, comparable on microarray-based data, we rounded the raw microarray intensity values as count data. Consistent with essential gene comparison, both ScreenBEAM and MAGeCK signficanlty outperforms Hitselect (Fig. 5B). Interestingly, while the difference using gold-standard essential gene is not significant ScreenBEAM performed slightly better than MAGeCK using this metric. As expected, there was a great overlap among the hits identified by the different methods (Supplementary Fig. S6).

Finally, we also compared ScreenBEAM with MAGeCK and HitSelect on two other NGS-based FG screens (Qin *et al.*, 2014; Tan *et al.*, 2013), which had massive individual validations of identified hits. We compared the false discovery rates of validated positive hits scored by the three algorithms. In Tan dataset, both ScreenBEAM and HitSelect scored better than MAGeCK, but only ScreenBEAM reached the statistical significance level ($P = 0.01$ by *t*-test) compared to MAGeCK (Fig. 6). In Qin dataset (Supplementary Fig. S7), both ScreenBEAM and HitSelect again scored better than MAGeCK though the significance levels are at 0.1–0.2 level. The lower performance of MAGeCK in both cases might be explained by the small sample size in both cases (3 versus 1 and 1 versus 1). As we demonstrated in the benchmark analysis, MAGeCK tends to favor high-quality data with relative large-sized samples while HitSelect favors noisy or small sample-sized data. Overall, ScreenBEAM demonstrated the most robust and consistent performance identifying true hits in all scenarios.

Complete gene knock out through the CRISPR/Cas9 gene-editing technology has also been scaled up for high-throughput FG screens. (Shalem *et al.*, 2014; Wang *et al.*, 2014; Zhou *et al.*, 2014). Thus, we also evaluated the performance of ScreenBEAM analysing this type of data. The available CRISPR genome-wide data sets are coupled to NGS-based data. From the analytical point of view these data do not differ from the above evaluated RNAi screens and MAGeCK and HitSelect have been already shown to perform better than classical RIGER and RSA (Diaz *et al.*, 2015; Li *et al.*, 2014). Thus, we compared ScreenBEAM with MAGeCK and HitSelect only. We analyzed the screen performed in HL60 cells (Wang *et al.*, 2014) as matched RNAi genome-wide studies (Cowley *et al.*, 2014) are also available providing us additional comparisons. Unfortunately, the absence of biological replicates in the limited number of CRISPR screens that are publically available greatly reduces our power to identify differences between methods. Thus, not surprisingly, this study revealed no differences between the three methods (Supplementary Fig. S8).

4 Discussion

Meta-analysis of shRNA screening data increase gene-level inference robustness remains difficult. We proposed a novel multi-probe analysis strategy, ScreenBEAM (Screening Bayesian Evaluation and

Analysis Method), implemented via a Bayesian hierarchical modeling, to address this problem. The evaluation results demonstrated that the ScreenBEAM method outperformed traditional single-probe analysis approaches (RIGER and RSA) and recent model-based methods (MAGeCK and HitSelect). This was especially relevant when the screen data was in relatively low quality which account for about 80–95% cases.

Hierarchical modelling, also known as partial pooling, can be viewed as a compromise between two extremes. One extreme, complete pooling, assumes the equal knock-down effect across all shRNA classes targeting the same gene. The other extreme, no pooling, ignores the similarity of the replicates within one shRNA group and treat each hairpin replicate separately. The assumptions of these two extreme methods are too strong for shRNA screening design to be considered for integration of multiple shRNA evidences because different shRNAs targeting the same gene in the library might have significantly different silencing efficiencies. Hierarchical modeling comprises two extremes by allowing between-group variance and considering within-group effects, thus making an appropriate solution to this question.

The problem of multiple comparisons can also disappear in Bayesian hierarchical models (Gelman *et al.*, 2011). Partial pooling in hierarchical models shifts estimates toward each other whereas classical procedures for multiple comparison correction typically adjust p-values corresponding to intervals of fixed width. Thus ScreenBEAM fitting results in reliable and conservative estimates for main effects or gene-level effects in this context.

For single-probe analysis strategy, a few other possible algorithms might be considered to integrate shRNA-level scores for the same gene, for example, Fisher's method (Fisher, 1948) to combine signed p-values, or Stouffer's method to combine *z*-statistics (Stouffer *et al.*, 1949). However, these integrating *P*-values or *z*-scores methods easily over-estimate the significance of gene-level activity and generate a long list of significant candidates. Also, they ignore the magnitude of knock-down effects for each hairpin by only considering the statistical significance of how the effect is away from zero, and require strong assumptions. Thereby, these methods might not be comparable to this ScreenBEAM algorithm, or could be even worse than the other single-probe analysis methods.

Additionally, other enrichment analysis algorithms such as GSA (Efron and Tibshirani, 2007) have been used in this context (Sims *et al.*, 2011) and might perform better than KS-based GSEA method; however, these algorithms still bear the drawbacks of single-probe analysis strategy, making them less powerful than ScreenBEAM. The valuable point from enrichment-type methods that might improve ScreenBEAM is to borrow information from all shRNAs or genes in the library because current ScreenBEAM algorithm only considers shRNAs corresponding to one gene. Looking at entire list of candidates might produce more robust statistics for cut-off based hits selection, but probably would not change the rank of a gene as a potential candidate.

NGS has dominated as a cost-effective technology for quantitatively measuring the abundance of short-length DNA or RNA in a short time, and this multiplexing parallel technology has been used in genome-wide FG shRNA and CRISPR sgRNA screens. Compared to microarray-based approaches, NGS offers several potential advantages in terms of coverage of targeting genes, flexibility of input library, scalability and dynamic range, which will possibly replace microarray for FG screens in the near future, however, the data quality of NGS-based genome-wide FG screens still has a big room to improve as only 6% of over 250 Achilles screens are in good category, compared to 22% of microarray-based data (Supplementary Fig. S1). The two existing best algorithms (MAGeCK and HitSelect) were specifically designed for NGS-based count data; however, ScreenBEAM can handle both microarray-based intensity data and NGS-based count data.

Using conserved housekeeping genes or RNAi screen-identified essential genes as the gold-standard of essential genes identified by loss-of-function screens, we demonstrated that the ScreenBEAM algorithm improves the sensitivity by up to 50% of that by classical approaches without loss of precision.

Overall, ScreenBEAM demonstrated the most robust and consistent performance identifying true hits in all scenarios, even from small-sized noisy high-throughput screens which accounts for about 80–95% of the public datasets. High quality data in high-throughput loss-of-function screens is rarely achieved (22% in microarray-based and 6% in NGS-based data), due to a variety of error and variability sources. For high quality screens, method selection is less relevant as all of the tested methods had small-difference performance. Yet, for the lower quality sh/sgRNAs, within a high quality dataset, ScreenBEAM would still produce significant improvements. As a result, there would be potential advantages even in this kind of datasets.

In summary, we developed a novel hierarchical modelling algorithm within Bayesian framework for meta-analysis of large-scale FG screens. This novel multi-probe approach performs more robustly than previously established analysis methods, especially with noisy high-throughput data.

Acknowledgements

The authors thank all the members of the Califano and Silva labs for helpful discussions, Dr. Aaron Diaz for sharing data, Dr. Barbara Weir for Achilles data access.

Funding

This work was partially supported by the NIH grants R01/EUREKA R01CA153233 and for the U01/5U01CA168426-04.

Conflict of Interest: none declared.

References

Babij,C. *et al.* (2011) STK33 Kinase activity is nonessential in KRAS-dependent cancer cells. *Cancer Res.*, **71**, 5818–5826.

Begley,C.G. and Ellis,L.M. (2012) Drug development: raise standards for preclinical cancer research. *Nature*, **483**, 531–533.

Chang,C.W. *et al.* (2011) Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PLoS One*, **6**, e22859.

Cheung,H.W. *et al.* (2011) Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc. Natl. Acad. Sci. USA.*, **108**, 12372–12377.

Cong,L. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.

Cowley,G.S. *et al.* (2014) Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci. Data*, **1**, 140035.

Deng,T. *et al.* (2014) shRNA kinome screen identifies TBK1 as a therapeutic target for HER2+ breast cancer. *Cancer Res.*, **74**, 2119–2130.

Diaz,A.A. *et al.* (2015) HiTSelect: a comprehensive tool for high-complexity-pooled screen analysis. *Nucleic Acids Res.*, **43**, e16.

Echeverri,C.J. *et al.* (2006) Minimizing the risk of reporting false positives in large-scale RNAi screens. *Nat. Methods*, **3**, 777–779.

Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107–129.

Fisher,R.A. (1948) Questions and answers #14. *The American Statistician*, **2**, 30–31.

Gelman,A. *et al*. (2004) *Bayesian Data Analysis*. Chapman & Hall, Boca Raton, FL, USA.

Gelman,A. and Hill,J. (2007) *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY, USA.

Gelman,A. *et al*. (2011) Why we (usually) don't have to worry about multiple comparisons. *Technical Report*.

Hahn,W.C. *et al*. (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, **462**, U108–U122.

Hammond,S.M. *et al*. (2000) An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature*, **404**, 293–296.

Hart,T. *et al*. (2014) Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol. Syst. Biol.*, **10**, 733.

Hoffman,G.R. *et al*. (2014) Functional epigenetics approach identifies BRM/SMARCA2 as a critical synthetic lethal target in BRG1-deficient cancers. *Proc. Natl. Acad. Sci. USA*, **111**, 3128–3133.

Ji,H.K. and Liu,X.S. (2010) Analyzing 'omics data using hierarchical models. *Nat. Biotechnol.*, **28**, 337–340.

Kaelin,W.G. (2012) Use and abuse of RNAi to study mammalian gene function. *Science*, **337**, 421–422.

Konig,R. *et al*. (2007) A probability-based approach for the analysis of large-scale RNAi screens. *Nat. Methods*, **4**, 847–849.

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

Luo,B. *et al*. (2008) Highly parallel identification of essential genes in cancer cells. *Proc. Natl. Acad. Sci. USA*, **105**, 20380–20385.

Luo,J. *et al*. (2009) A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. *Cell*, **137**, 835–848.

Mali,P. *et al*. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823–826.

Marcotte,R. *et al*. (2012) Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Disc.*, **2**, 172–189.

Moffat,J. *et al*. (2006) A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell*, **124**, 1283–1298.

Paddison,P.J. *et al*. (2004) A resource for large-scale RNA-interference-based screens in mammals. *Nature*, **428**, 427–431.

Prahallad,A. *et al*. (2012) Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature*, **483**, 100–103.

Prinz,F. *et al*. Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.*, **10**, 712–712.

Qin,H. *et al*. (2014) Systematic identification of barriers to human iPSC generation. *Cell*, **158**, 449–461.

Shalem,O. *et al*. (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, **343**, 84–87.

Shao,D.D. *et al*. (2013) ATARiS: computational quantification of gene suppression phenotypes from multisample RNAi screens. *Genome Res.*, **23**, 665–678.

Silva,J.M. *et al*. (2009) Cyfip1 is a putative invasion suppressor in epithelial cancers. *Cell*, **137**, 1047–1061.

Silva,J.M. *et al*. (2005) Second-generation shRNA libraries covering the mouse and human genomes. *Nature Genet*., **37**, 1281–1288.

Silva,J.M. *et al*. (2008) Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science*, **319**, 617–620.

Silva,J.M. *et al*. (2004) RNA interference microarrays: high-throughput loss-of-function genetics in mammalian cells. *Proc. Natl. Acad. Sci. USA*, **101**, 6548–6552.

Sims,D. *et al*. (2011) High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing. *Genome Biol.*, **12**, R104.

Stouffer,S.A. *et al*. (1949) *Adjustment During Army Life*. Princeton: Princeton University Press.

Tan,T.Z. *et al*. (2013) Functional genomics identifies five distinct molecular subtypes with clinical relevance and pathways for growth control in epithelial ovarian cancer. *EMBO Mol. Med.*, **5**, 983–998.

Tu,Z.D. *et al*. (2006) Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics*, **7**, 31.

Wang,T. *et al*. (2014) Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, **343**, 80–84.

Zhou,Y. *et al*. (2014) High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature*, **509**, 487–491.

Zuber,J. *et al*. (2011) RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. *Nature*, **478**, 524–528.