Hart and Moffat *BMC Bioinformatics* (2016) 17:164
DOI 10.1186/s12859-016-1015-8

BMC Bioinformatics

SOFTWARE Open Access

# BAGEL: a computational framework for identifying essential genes from pooled library screens

Traver Hart[1*] and Jason Moffat[2,3]

## Abstract

**Background:** The adaptation of the CRISPR-Cas9 system to pooled library gene knockout screens in mammalian cells represents a major technological leap over RNA interference, the prior state of the art. New methods for analyzing the data and evaluating results are needed.

**Results:** We offer BAGEL (Bayesian Analysis of Gene EssentiaLity), a supervised learning method for analyzing gene knockout screens. Coupled with gold-standard reference sets of essential and nonessential genes, BAGEL offers significantly greater sensitivity than current methods, while computational optimizations reduce runtime by an order of magnitude.

**Conclusions:** Using BAGEL, we identify ~2000 fitness genes in pooled library knockout screens in human cell lines at 5 % FDR, a major advance over competing platforms. BAGEL shows high sensitivity and specificity even across screens performed by different labs using different libraries and reagents.

**Keywords:** CRISPR, Genetic screens, Cancer, Essential genes, Functional genomics

## Background

Perturbing gene activity and evaluating the resulting phenotype is a fundamental technique for identifying the biological processes in which a gene participates (i.e., "forward genetics"). Traditionally, the ability to induce complete gene knockouts on a genomic scale has been exclusively the domain of model organisms such as yeast, while experiments in higher eukaryotes, including human cell lines, have relied on RNA interference (RNAi) or gene trapping methods in the case of haploid human cells [1]. RNAi uses the endogenous RNA-induced silencing complex (RISC) machinery to target messenger RNA transcripts, which have a very large dynamic range of abundance, resulting in data that is often diluted by incomplete target knockdown and off-target effects of variable severity [2–4].

The adaptation of CRISPR-Cas9 technology to pooled library gene knockout screens in mammalian cells allows the identification of genes whose knockout contributes to gene fitness [5–9]. A pooled library screen typically contains several guide RNAs (gRNA) targeting each gene, and large numbers of cells are treated such that each cell is affected by (on average) a single gRNA clone, while each gRNA species targets hundreds of cells. Unperturbed cells, or cells with knockouts showing no growth phenotype, grow at wildtype rates, while cells harboring a guide RNA that targets a fitness gene show lower growth rates (Fig. 1a). To identify the genes whose knockout causes a fitness defect, the frequency distribution of gRNA in the population is assayed by deep sequencing and compared to the frequency distribution at an early control timepoint. Changes in the frequency distribution of gRNA are measured as log fold changes (Fig. 1a, sidebar) where severe negative fold changes reflect gRNA that cause severe fitness defects.

Aggregating individual reagent effects into an accurate estimate of gene-level effect is a major challenge in the analysis of pooled library screen data [10–14]. To analyze pooled library RNAi screens, which have similar experimental design, we previously developed a Bayesian classifier and demonstrated its superiority over contemporary approaches [3]. A key feature of this study was the establishment of reference sets of core essential and nonessential genes. Core essential genes were defined as those genes classified as hits in at half or more of the shRNA screens in [12] or [13], filtered for constitutive mRNA expression across a panel of cell lines, while nonessential

* Correspondence: traver.hart@gmail.com
[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA
Full list of author information is available at the end of the article

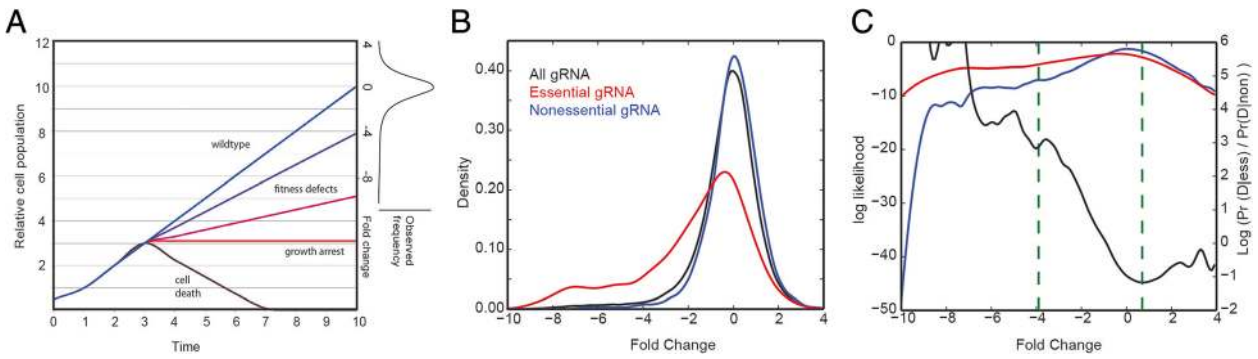


**Fig. 1** BAGEL overview. **a** Simulated growth curves of wildtype cells (*blue*), which double at every time increment. When genetic perturbations are induced (T = 3), moderate (*purple*) to severe (*magenta*) fitness defects, growth arrest (*red*), and cell death (*black*) result in different relative growth rates. At sampled timepoints, fold change relative to wildtype growth is the readout from a sequencing assay. **b** Representative data from one replicate. The fold change distribution of all gRNA targeting essential genes (*red*) is shifted relative to the fold change distribution of all gRNA targeting nonessential genes (*blue*). The fold change distribution for all gRNA (*black*) is shown for reference. **c** The log likelihood functions of the red and blue curves from (**b**), left Y axis. The BAGEL method calculates the log likelihood ratio (*black, right Y axis*) of these two curves, within empirical boundaries (*green dashes*), for each bootstrap iteration; see Methods for details

genes were defined as those which are rarely expressed across those cell lines [3]. Together, these reference sets can be used as gold standards to evaluate other algorithms in analyzing fitness screens. Here we describe BAGEL, the Bayesian Analysis of Gene EssentiaLity, an adaptation of the previously described Bayesian classifier. BAGEL features a more robust statistical model, major performance enhancements, and an improved user interface. BAGEL source code, documentation, and reference files are available at http://bagel-for-knockout-screens.sourceforge.net/.

## Implementation

The BAGEL method is implemented as a Python script and requires the freely available python modules numpy and scipy. Likelihood functions for fold change distributions of gRNA targeting reference essential and nonessential genes are estimated using kernel density estimates, implemented using the scipy.stats.gaussian_kde() function in the scipy module. BAGEL input and output are tab-delimited text files described on the BAGEL website (see Availablility and requirements for details).

## Methods

A pooled library CRISPR-Cas9 fitness screen in human cells involves having multiple gRNA reagents targeting each gene and is often evaluated at several timepoints, ideally with multiple replicates at each timepoint. BAGEL first estimates the distribution of fold changes of all gRNA targeting all genes in either the essential or nonessential training sets (Fig. 1b). Then, for each withheld gene, it evaluates the likelihood that the observed fold changes for gRNA targeting the gene were drawn from either the essential or the nonessential training distributions. The result is a Bayes Factor:

$$BF = \frac{\Pr(D \mid essential)}{\Pr(D \mid nonessential)} = \frac{\int \Pr(D|k,\ essential)\Pr(k \mid essential)\,dk}{\int \Pr(D|k,\ nonessential)\Pr(k \mid nonessential)\,dk}$$

where the data, D, is the set of observed fold changes for a given gene and $k$ is the fold change distribution of the training set, empirically estimated using a kernel density estimate function (Fig. 1b, red and blue curves).

The integral is estimated by bootstrap resampling of genes in the training sets. At each iteration the $k$ distributions are calculated and, for each withheld gene, a log BF is calculated:

$$BF_g = \frac{\Pr(D_g|k_{ess})}{\Pr(D_g|k_{non})}$$

$$\log(BF_g) = \log(Pr(D_g|k_{ess})) - \log(\Pr(D_g|k_{non}))$$

$$\log(BF_g) = \sum_i (\log(\Pr(\mathrm{fc}_i|k_{ess})) - \log(\Pr(\mathrm{fc}_i|k_{non})))$$

where $\mathrm{fc}_i$ are the observed fold changes for gRNA targeting gene $g$. One thousand bootstrapping iterations are conducted; Bayes Factors for withheld genes are calculated for each iteration (resulting in ~360 posterior BFs for each gene) and the mean and standard deviation of the resulting posterior distribution of BFs is reported.

Two factors inherent in the data require that empirical boundaries be applied to the calculations. First, when taking the ratio of two curves, the ratio can take on extreme values when the denominator approaches zero. Second, kernel density estimates become unstable in regions of sparse data. For these reasons, we identify the lowest fold change (x-coordinate) at which the $k_{non}$

density estimate, the denominator above, exceeds $2^{-7}$, and set this as a lower bound (Fig. 1c). This boundary is, in our experience, a conservative threshold that captures the smooth region of the $k_{non}$ kernel density estimate across all data sets we examined. All observed changes below this boundary are set to the boundary value. Similarly, we calculate the fold change at which the log ratio of the curves is a minimum and set this as an upper bound (Fig. 1c). These boundaries ensure that individual observations do not dominate the final BF score while, in our experience, making no material change to gene estimates: observed fold changes outside these boundaries are not stronger evidence that a gene does or doesn't induce a fitness defect, given the normal constraints of the experiment (number of cells, sequencing depth, etc.). Note that this approach makes no statement about whether a gene knockout can increase cell fitness, only whether perturbation causes a growth defect.

For very large CRISPR libraries, the calculation as described can be computationally expensive. To speed up the calculations, we include two optimizations. First, we round all calculated fold changes to the nearest 0.01. Second, for each bootstrap iteration, we calculate the value of the log ratio function (Fig. 1c) at each 0.01 within the empirical boundaries described above and store the values in a lookup table. Then, instead of recalculating the values for each gRNA, we pull the value of the log ratio function from the lookup table. These optimizations decrease processing time by over an order of magnitude, with no impact on final results (Pearson's $r \sim 0.999$ for final BFs; data not shown).

For knockout screens with multiple timepoints, the BF is calculated at each timepoint, and a final BF is the sum of the timepoint BFs. Since the posterior BF distributions are approximately normal (by KS tests, not shown), the variance of the final BF is estimated as the sum of the variances at the timepoints.

Screen performance is evaluated by calculating precision-recall (PR) curves, using the reference essential and nonessential gene lists as the test set. As noted above, during the bootstrap process, BFs are only calculated for genes not selected in the bootstrap resampling of the fold change density estimates; therefore no circularity is introduced. We confirmed this by comparing BF-bootstrap results to BFs calculating using 5-fold cross-validation; the resulting BFs virtually identical ($R^2 > 0.99$). False discovery rate is (FP/FP + TP), precision is 1- FDR, and recall = TP/(TP + FN), where positives and negatives are defined in the reference sets.

## Results and discussion

We demonstrate this approach with screens from the Toronto KnockOut (TKO) library in four cell lines: a patient-derived glioblastoma cell line (GBM, Fig. 2a), HCT116 colorectal carcinoma cell line (Fig. 2b), HeLa cervical carcinoma cell line (Fig. 2c), and RPE1 retinal pigmented epithelial cells (Fig. 2d) [15]. All the screens were sampled at multiple timepoints. Using the gold-standard reference sets from [3], BFs were calculated for each timepoint and precision-recall (PR) curves were plotted. In all cases, later timepoints showed improved recall over the earliest timepoint. The "integrated" sample is the sum of the timepoint BFs and can be considered a summary result for the entire screen; the PR curve for the integrated sample is in every case as good or better than the timepoint curves. In all cases screens yielded a very large number of fitness genes: on average, ~2000 genes at 5 % false discovery rate (FDR) using the integrated results, and these genes show very high functional coherence (see [15] for a more complete evaluation).

One question that arises from these results is whether the lower performance at the early timepoint, relative to the later ones, reflects the screening technology or the biology of the systems being perturbed. We address this question by looking at functional enrichment in genes unique to the early hits, genes unique to the late hits, or genes in the intersection, using the GORILLA web service [16]. We find that most (75-89 %) early hits are also observed at the last timepoint (Fig. 3a), and that genes exclusively in the early hit set are not meaningfully enriched for annotated biological processes. Looking specifically at the GBM cell line as an example, genes in the intersection are highly enriched for core biological processes one could reasonably expect to cause fitness defects (Fig. 3b). Genes in the intersection comprise only 53-65 % of the total number of hits at the last timepoint; however, genes exclusive to the last timepoint typically extend coverage of the biological processes identified in the intersection, as shown for GBM cells in Fig. 3b, and identify few novel processes.

Though late fitness genes typically reflect the processes observed in early fitness genes, genes which encode proteins involved in mitochondrial function offer an interesting contrast. Genes in both the early and late timepoints are enriched for some mitochondrial processes, including protein transport to the mitochondrion and mitochondrial translation. However, the late-only genes are enriched for a small number of GO BP terms that are centered around functions related to oxidative phosphorylation, including "respiratory chain complex I assembly" (7 hits of 18 annotated genes, 7.4-fold enrichment), "respiratory chain complex IV assembly" (4/8 genes, 9.4-fold), and "mitochondrial electron transport, NADH to ubiquinone" (12/36 genes, 6.3-fold). This difference may reflect a more subtle phenotype (i.e., lower fitness defect) among oxphos genes that only becomes detectable at the later timepoint (Fig. 1a).

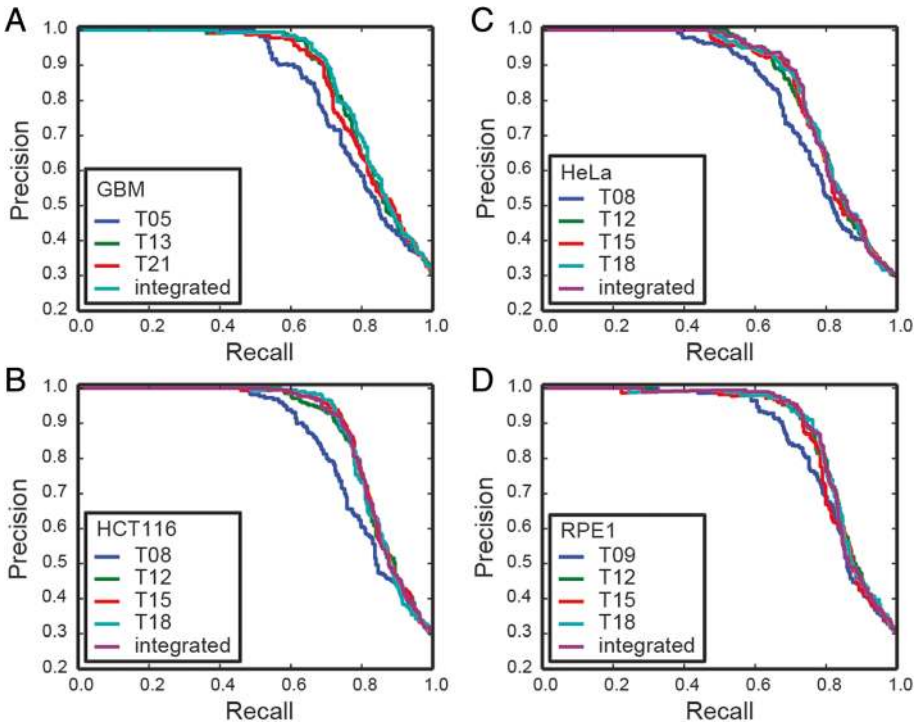


**Fig. 2** Precision-recall curves for BAGEL results for GBM (**a**), HCT116 (**b**), HeLa (**c**), and RPE1 (**d**) screens using the TKO library. Where indicated, a single timepoint is plotted. "Integrated" = Bayes Factors summed across all timepoints in the experiment

We compared BAGEL to MAGeCK, a contemporary method for analyzing CRISPR knockout screens [11]. MAGeCK ranks gRNAs by P-value derived from a negative binomial model comparing control to experimental timepoints, then calculates gene-level P-values using modified Robust Ranking Aggregation. To facilitate a more equal comparison, we compared MAGeCK results to BAGEL results using only the final timepoint from the TKO screens described above, and plotted PR curves using the same reference sets (Fig. 4a-d). In all cases, BAGEL outperformed MAGeCK, yielding more recall and more overall hits in a reasonable range of empirically-calculated FDR (5–15 %). Most striking, however, was the severe lack of sensitivity using the theoretical model of MAGeCK. Although gene rankings for the two methods were generally similar (Spearman correlations 0.76–0.81 for the top 3000 genes in each set), the MAGeCK algorithm yielded only 674 (mean; range 489–905) genes at 10 % FDR, using its own FDR estimates (Fig. 4). We are confident that the higher numbers of fitness genes detected by BAGEL are in fact real: we analyze their expression level, biological function, and other functional genomic data in detail in [15].

We also compared the two algorithms using a newly published data set from Wang et al. [17], where four leukemia and lymphoma cell lines were screened for essential genes using a large gRNA library. As with the TKO screens, the BAGEL algorithm yields equal or superior precision-recall curves and greater sensitivity, though with a smaller margin of improvement (Fig. 4e-h). MAGeCK identifies 1571 (mean, range 1241–1800) hits at 10 % FDR while BAGEL identifies on average 2272 (range 1963–2482) essential genes at 5 % FDR.

The reason behind the difference in sensitivity between BAGEL and MAGeCK likely lies in the variable effectiveness of CRISPR reagents. Examining the fold change distribution of all guides targeting genes in the reference set of high-confidence essentials (Fig. 1b), it is evident that many gRNAs targeting essential genes do not show significant dropout. The BAGEL algorithm chooses between the essential and nonessential distributions, and is able to detect even a slight shift in overall effect, whereas a statistical test based solely on excluding the null hypothesis – generally speaking, that the observed fold changes are not likely to be drawn from the blue curve in Fig. 1a—requires either deeper sampling (i.e., more replicates and/or more guides targeting each gene) or a more severe phenotype. In fact, this is reflected in the MAGeCK results for the four TKO cell lines tested: the GBM and RPE1 cell lines were screened with a 90 k library and MAGeCK

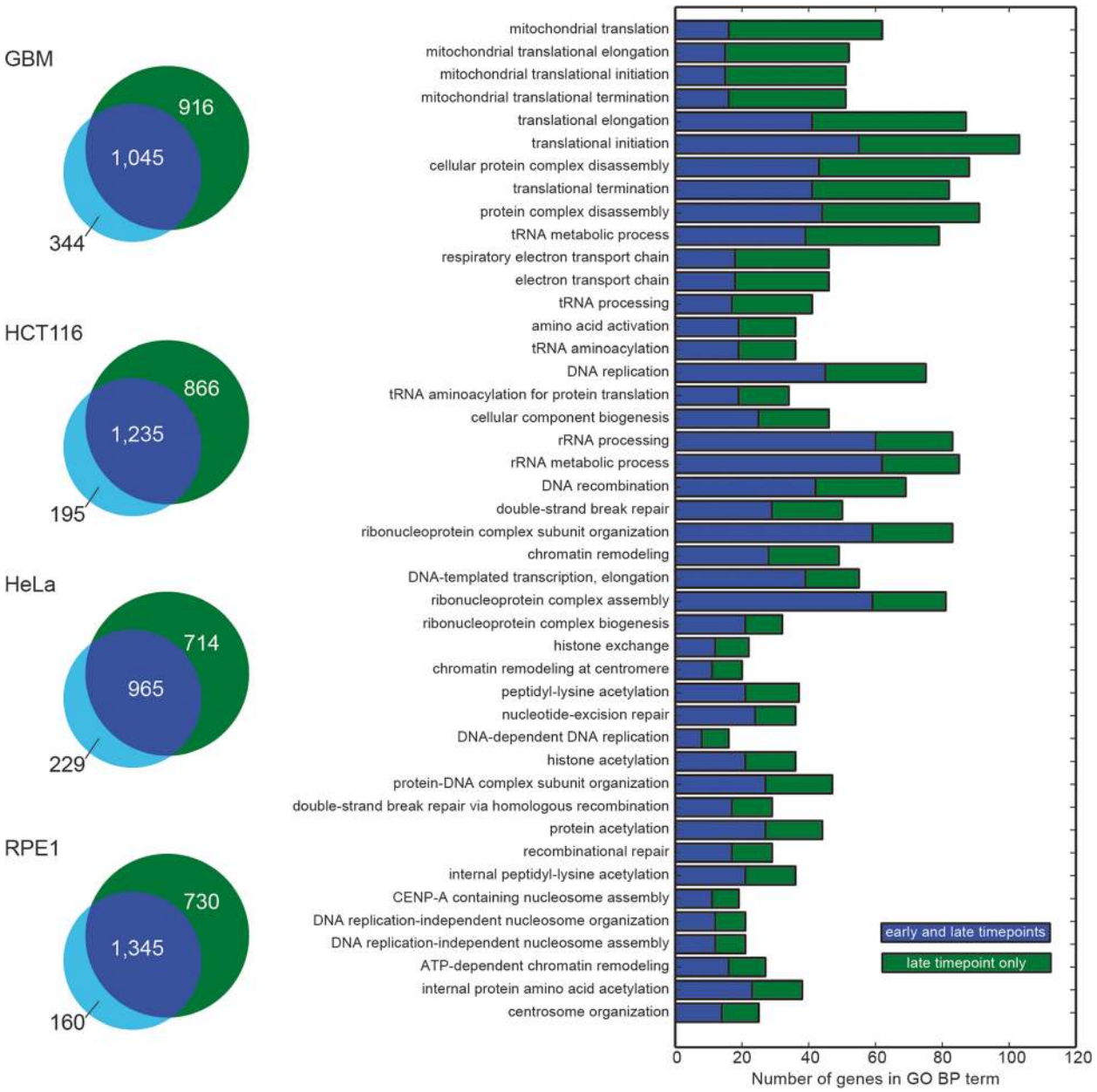


**Fig. 3** Comparing early and late hits. **a** Number of fitness genes detected at early timepoint (*cyan*), late timepoint, (*green*), or both (*blue*) in each TKO screen. **b** Representative data from GBM screen. Most GO_BP terms enriched in late-only genes (*green*) extend observations of terms enriched in genes found in both early and late timepoints

yielded 586 and 489 hits, respectively, while the HeLa and HCT116 lines were screened with a 177 k library and MAGeCK yielded 718 and 905 hits – on average, ~50 % more hits using the larger library. The screens described in Wang et al. used a sequence-optimized 180 k gRNA library and used a more conservative experimental design, resulting in a lower proportion of non-performing guides and contributing to substantially improved sensitivity for both BAGEL and MAGeCK, though BAGEL still identifies ~50 % more hits in each screen.

## Conclusions

The ability to perform accurate, saturating forward genetic screens in human cell lines will transform molecular genetics in the coming years. To maximize potential—and to avoid pitfalls similar to the costly false starts encountered in the RNAi field—rigorous

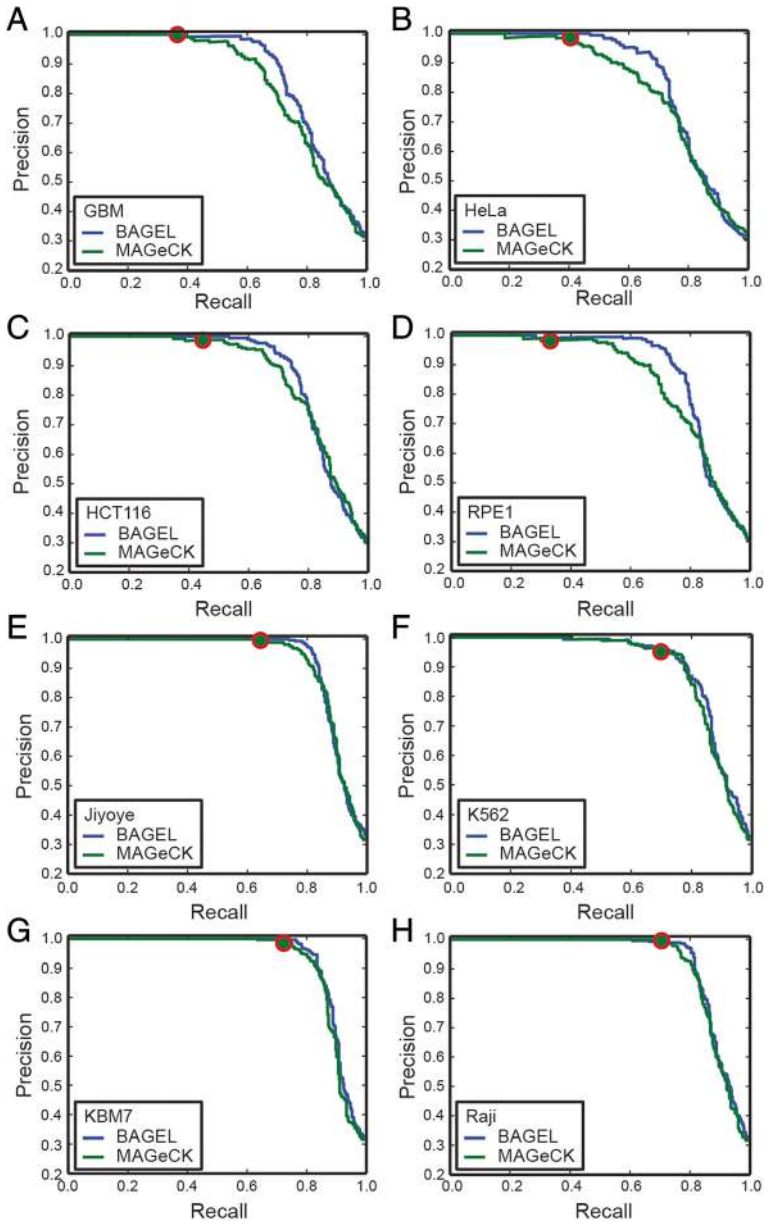


**Fig. 4** Comparing BAGEL with MAGeCK. For each cell line, precision-recall curves were plotted for BAGEL and MAGeCK results using the last timepoint of the screen. Red circle indicates results at MAGeCK-reported 10 % FDR cutoff. **a**-**d** TKO screens from Hart et al. [15] **e**-**h** Screens from Wang et al. [17]

analytical methods must be applied that are able to effectively discriminate true hits from false positives. While data suggests that off-target effects in CRISPR-Cas9 pooled library screens are much less of a concern than with RNAi, the variable effectiveness of early reagent pools makes it important that analytical methods are able to detect subtle phenotypes. BAGEL accurately models the wide variability in phenotype shown by reagents targeting known essential genes, enabling the sensitive and precise identification of fitness genes, even under conditions of suboptimal data quality.

## Availability and requirements

Project name: bagel-for-knockout-screens
Project home page: http://bagel-for-knockout-screens.sourceforge.net/
Operating system(s): platform independent
Programming language: Python

Licensing: This software is provided without restriction for commercial or academic use.
TKO screen data are available at http://tko.ccbr.utoronto.ca/

#### Competing interests
The authors declare that they have no competing interests.

#### Authors' contributions
TH developed the algorithm and wrote the software; TH and JM wrote and edited the manuscript. Both authors read and approved the final manuscript.

#### Acknowledgements
We would like to thank Megha Chandrashekhar, Michael Aregger, Graham MacLeod, and Stephane Angers for acquiring the data for this study.

#### Author details
[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [2]Donnelly Centre, University of Toronto, Toronto, Canada. [3]Department of Molecular Genetics, University of Toronto, Toronto, Canada.

Received: 26 November 2015 Accepted: 6 April 2016
Published online: 16 April 2016

#### References
1. Carette JE, Guimaraes CP, Wuethrich I, Blomen VA, Varadarajan M, Sun C, Bell G, Yuan B, Muellner MK, Nijman SM, et al. Global gene disruption in human cells to assign genes to phenotypes by deep sequencing. Nat Biotechnol. 2011;29(6):542–6.
2. Echeverri CJ, Beachy PA, Baum B, Boutros M, Buchholz F, Chanda SK, Downward J, Ellenberg J, Fraser AG, Hacohen N, et al. Minimizing the risk of reporting false positives in large-scale RNAi screens. Nat Methods. 2006;3(10):777–9.
3. Hart T, Brown KR, Sircoulomb F, Rottapel R, Moffat J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. Mol Syst Biol. 2014;10:733.
4. Kaelin Jr WG. Molecular biology. Use and abuse of RNAi to study mammalian gene function. Science. 2012;337(6093):421–2.
5. Chen S, Sanjana NE, Zheng K, Shalem O, Lee K, Shi X, Scott DA, Song J, Pan JQ, Weissleder R, et al. Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. Cell. 2015;160(6):1246–60.
6. Koike-Yusa H, Li Y, Tan EP, Velasco-Herrera Mdel C, Yusa K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nat Biotechnol. 2014;32(3):267–73.
7. Parnas O JM, Eisenhaure TM, Herbst RH, Dixit A YC, Przybylski D, Platt RJ, Tirosh I, Sanjana NE SO, Satija R, Raychowdhury, R MP, Carr SA, Zhang F, Hacohen N, A R. A genome-wide CRISPR screen in primary immune cells to dissect regulatory networks. Cell. 2015;162(3):675–86.
8. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelson T, Heckl D, Ebert BL, Root DE, Doench JG et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science. 2013;343(6166):84–7.
9. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR/Cas9 system. Science. 2013.
10. Konig R, Chiang CY, Tu BP, Yan SF, DeJesus PD, Romero A, Bergauer T, Orth A, Krueger U, Zhou Y, et al. A probability-based approach for the analysis of large-scale RNAi screens. Nat Methods. 2007;4(10):847–9.
11. Li W, Xu H, Xiao T, Cong L, Love MI, Zhang F, Irizarry RA, Liu JS, Brown M, Liu XS. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol. 2014;15(12):554.
12. Luo B, Cheung HW, Subramanian A, Sharifnia T, Okamoto M, Yang X, Hinkle G, Boehm JS, Beroukhim R, Weir BA, et al. Highly parallel identification of essential genes in cancer cells. Proc Natl Acad Sci U S A. 2008;105(51):20380–5.
13. Marcotte R, Brown KR, Suarez F, Sayad A, Karamboulas K, Krzyzanowski PM, Sircoulomb F, Medrano M, Fedyshyn Y, Koh JL, et al. Essential gene profiles in breast, pancreatic, and ovarian cancer cells. Cancer Discov. 2012;2(2):172–89.
14. Yu J, Silva J, Califano A. ScreenBEAM: a novel meta-analysis algorithm for functional genomics screens via Bayesian hierarchical modeling. Bioinformatics. 2015;32(2):260–7.
15. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmerman M, Fradet-Turcotte A, Sun S et al. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. Cell. 2015; doi:10.1016/j.cell.2015.11.015.
16. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10:48.
17. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM. Identification and characterization of essential genes in the human genome. Science. 2015;350(6264):1096–101.