Genomics Data 4 (2015) 123–126

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Using shRNA experiments to validate gene regulatory networks

Catharina Olsen [a,b], Kathleen Fleming [c], Niall Prendergast [c], Renee Rubio [c], Frank Emmert-Streib [d,g], Gianluca Bontempi [a,b], John Quackenbush [c], Benjamin Haibe-Kains [e,f,*]

[a] Machine Learning Group, Université Libre de Bruxelles, Brussels, Belgium
[b] Interuniversity Institute of Bioinformatics in Brussels (IB)^2, Belgium
[c] Computational Biology and Functional Genomics Laboratory, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, MA, USA
[d] Computational Medicine and Statistical Learning Laboratory, Department of Signal Processing, Tampere University of Technology, Korkeakoulunkatu 1, 33720 Tampere, Finland
[e] Bioinformatics and Computational Genomics Laboratory, Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada
[f] Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada
[g] Institute of Biosciences and Medical Technology, Biokatu 10, 33520 Tampere, Finland

ARTICLE INFO

Article history:
Received 20 March 2015
Received in revised form 23 March 2015
Accepted 23 March 2015
Available online 1 April 2015

Keywords:
Knock-down
Gene expression
Microarray
Colon cancer
shRNA

ABSTRACT

Quantitative validation of gene regulatory networks (GRNs) inferred from observational expression data is a difficult task usually involving time intensive and costly laboratory experiments. We were able to show that gene knock-down experiments can be used to quantitatively assess the quality of large-scale GRNs via a purely data-driven approach (Olsen et al. 2014). Our new validation framework also enables the statistical comparison of multiple network inference techniques, which was a long-standing challenge in the field.
In this Data in Brief we detail the contents and quality controls for the gene expression data (available from NCBI Gene Expression Omnibus repository with accession number GSE53091) associated with our study published in Genomics (Olsen et al. 2014). We also provide R code to access the data and reproduce the analysis presented in this article.

Specifications	
Organism/cell line/tissue	*Human/SW480, SW620/colorectal tumor tissue*
Sex	*Male*
Sequencer or array type	*GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array*
Data format	*Raw and fRMA normalized*
Experimental factors	*shRNA*
Experimental features	*RNAi-mediated gene knock-down experiments in two colorectal cell lines, targeting eight key genes in the RAS pathway. The experiments were done in six biological replicates of each knockdown and controls in both cell lines. From each sample, we profiled gene expression using the Affymetrix GeneChip HGU133PLUS2 platform.*
Consent	*None necessary, data are publicly available.*
Sample source location	*ATCC*

Direct link to deposited data

Deposited data can be found here: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53091

* Corresponding author at: Princess Margaret Cancer Centre, University Health Network, Toronto Medical Discovery Tower, 11th floor, Room 310, 101 College Street, Toronto, ON M5G 1L7, Canada.
E-mail address: bhaibeka@uhnresearch.ca (B. Haibe-Kains).

Experimental design, materials and methods

Short hairpin RNA experiments

We selected eight genes known to be involved in the RAS pathway, namely CDK5, HRAS, MAPK1, MAPK3, MAP2K1, MAP2K2, NGFR and RAF1. They are hereafter referred to as 'core genes' due to their relevance in the RAS pathway [2] and their consequential importance in colorectal cancer [8].

The knock-down experiments were performed on the eight core genes using short hairpin RNA (shRNA) in two colorectal cancer cell lines SW480 and SW620 [4]. For each knock-down there are six replicates (except CDK5 in SW620 with five replicates) with three different types of controls (empty vector, nontarget and non transduced), totaling 125 samples. We used the Affymetrix GeneChip HG-U133PLUS2 platform to profile the gene expression of each sample.

Quality control

We used the simpleaffy Bioconductor package [7] to check the quality of each individual CEL file. Fig. 1 shows that a majority of files contains a sufficiently large percentage of present calls (>40%, except for biological replicates one and two and HRAS biological replicate three with 39.58%) and all scale factors lie within a 3-fold range which complies with the good quality guidelines from Affymetrix [1]. In more

http://dx.doi.org/10.1016/j.gdata.2015.03.011

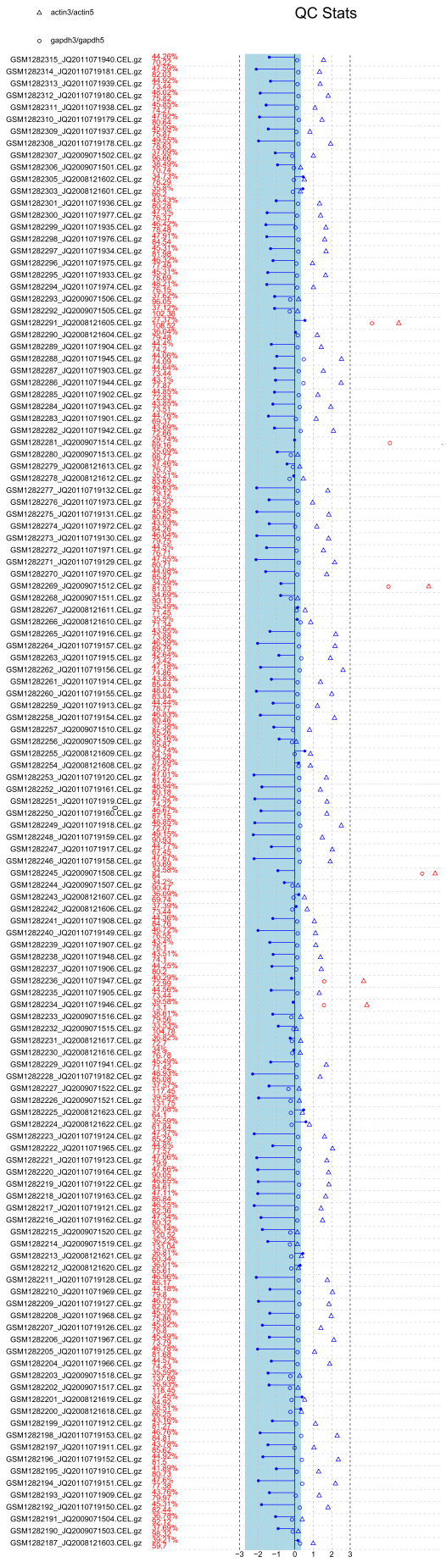

Fig. 1. Quality controls for the Affymetrix Raw data generated in [5]. CEL file names for each experiment is provided on the left side, followed by the percentage of present and absent calls (in red) following the Affymetrix guidelines. The blue region in the middle of the plot represents the 3-fold region for scale factor as this region is considered as acceptable according to Affymetrix guidelines; any scale factor outside this region is drawn in red as it is considered an indicator of poor quality. Beta-actin and GAPDH 3′–5′ ratios are also represented on the right side by triangles and circles, respectively; ratio higher than 1.25 are drawn in red as they are considered indicators of poor quality.

detail, we can observe that those CEL files that were generated in the early stages of the data generation are of lower quality than the rest of the files, namely the biological replicates one and two (Fig. 2 and Table 1).

Normalization

The raw and normalized data are available from NCBI Gene Expression Omnibus repository [3] with accession number GSE53091.

Basic analysis

A successful knock-down experiment should result in significantly lower expression compared to the unperturbed samples. Here, we assess the quality of a knock-down experiments by testing the difference between matched knock-down samples versus nontarget control samples using a Wilcoxon signed rank test [6].

In Fig. 3, we show the difference between knock-down and control sample expression for each of the eight knock-downs for both cell-lines together. In each plot, the knocked-down gene is highlighted in blue and the obtained p-values are represented by symbols in the bottom of each plot. The significance levels are represented as follows: '***' for $p < 0.001$, '**' for $p < 0.01$, '*' for $p < 0.05$ and '-' for $p < 0.1$.

From the eight plots in Fig. 3, we can observe that the difference in expression between knock-down and control samples is significant for all of the eight knock-downs (the differential expression of the blue boxes is significantly lower than zero). In our study [5], we determined the set of significantly affected genes for each of the eight knock-downs. For example, the knock-down of RAF1 significantly changes the expression of MAP2K2 and NGFR with p-values < 0.001 (only considering the eight core genes). We then used the set of significantly affected genes to quantitatively validate inferred gene regulatory interactions.

Discussion

In this article we described a unique data set containing RAS pathway-related gene knock-down experiments in two different colon cancer cell-lines. It contains the expression values from the knock-down of eight genes as well as three different controls in six biological replicates. The genome-wide gene expression was measured using the Human Genome U133 Plus 2.0 Array. This data was recently used in a published study on the validation of regulatory gene networks [5].

Acknowledgments

Funding. CO and BHK were supported by the Natural Sciences and Engineering Research Council of Canada (NSERC Grant File Number: RGPIN-2015-03654). JQ, KF, NP, and RR were supported by a grant from the US National Library of Medicine of the National Institutes of Health (1R01LM010129-01).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2015.03.011.

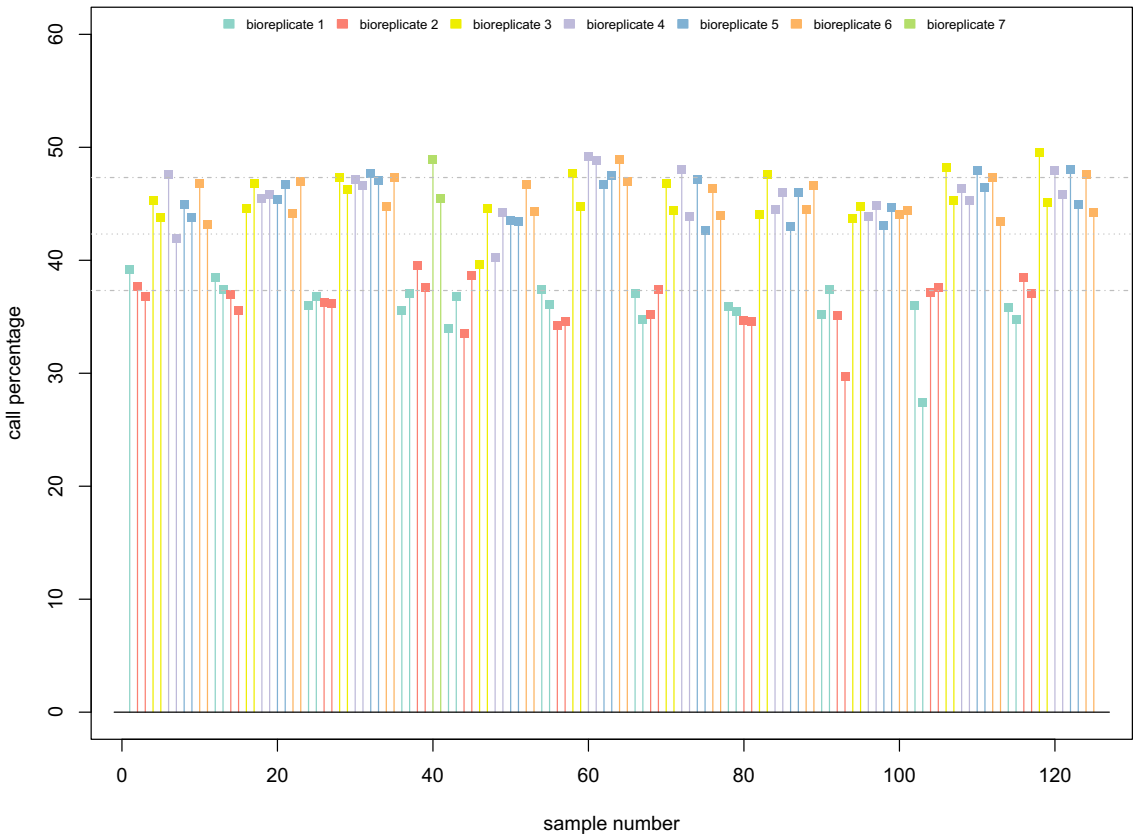

Fig. 2. Call percentage for each CEL file. The colors correspond to the biological replicate number. The quality of the first two replicates is lower than for the remaining five replicates.

Table 1
For each biological replicate, the time of data generation is specified. There are three main batches: 2008 (biological replicates 1), 2009 (biological replicates 2) and 2011 (biological replicates 3–7).

		Date				
		2008-12-16	2008-12-17	2009-07-15	2011-07-19	2011-07-20
Biological replicate	1	11	10	0	0	0
	2	0	0	22	0	0
	3	0	0	0	15	5
	4	0	0	0	19	1
	5	0	0	0	17	3
	6	0	0	0	18	2
	7	0	0	0	2	0

References

[1] Affymetrix, Quality Control Assessment in Genotyping Console. 2008.
[2] Gökmen Altay, Nejla Altay, David Neal, Global assessment of network inference algorithms based on available literature of gene/protein interactions. Turk. J. Biol. Turk. Biyol. Derg./Sci. Tech. Res. Counc. Turkey 37 (5) (2013) 547–555 (The Scientific and Technological Research Council of Turkey).
[3] Tanya Barrett, Tugba O. Suzek, Dennis B. Troup, Stephen E. Wilhite, Wing-Chi Ngau, Pierre Ledoux, Dmitry Rudnev, Alex E. Lash, Wataru Fujibuchi, Ron Edgar, NCBI GEO: mining millions of expression profiles—database and tools. Nucleic Acids Res. 33 (Suppl. 1) (2005) D562–D566.
[4] A. Leibovitz, J.C. Stinson, W.B. McCombs III, C.E. McCoy, K.C. Mazur, N.D. Mabry, Classification of human colorectal adenocarcinoma cell lines. Cancer Res. 36 (12) (1976) 4562–4569.
[5] C. Olsen, A. Djebbari, K. Fleming, N. Prendergast, R. Rubio, F. Emmert-Streib, G. Bontempi, B. Haibe-Kains, J. Quackenbush, Inference and validation of predictive gene networks from biomedical literature and gene expression data. Genomics 103 (5-6) (2014) 329–336.
[6] F. Wilcoxon, Individual comparisons by ranking methods. Biom. Bull. 1 (1945) 80–83.
[7] Claire L. Wilson, Crispin J. Miller, Simpleaffy: a bioconductor package for affymetrix quality control and data analysis. Bioinformatics 21 (18) (2005) 3683–3685.
[8] Kypros Zenonos, Katy Kyprianou, RAS signaling pathways, mutations and their role in colorectal cancer. World J. Gastrointest. Oncol. 5 (5) (2013) 97–101.

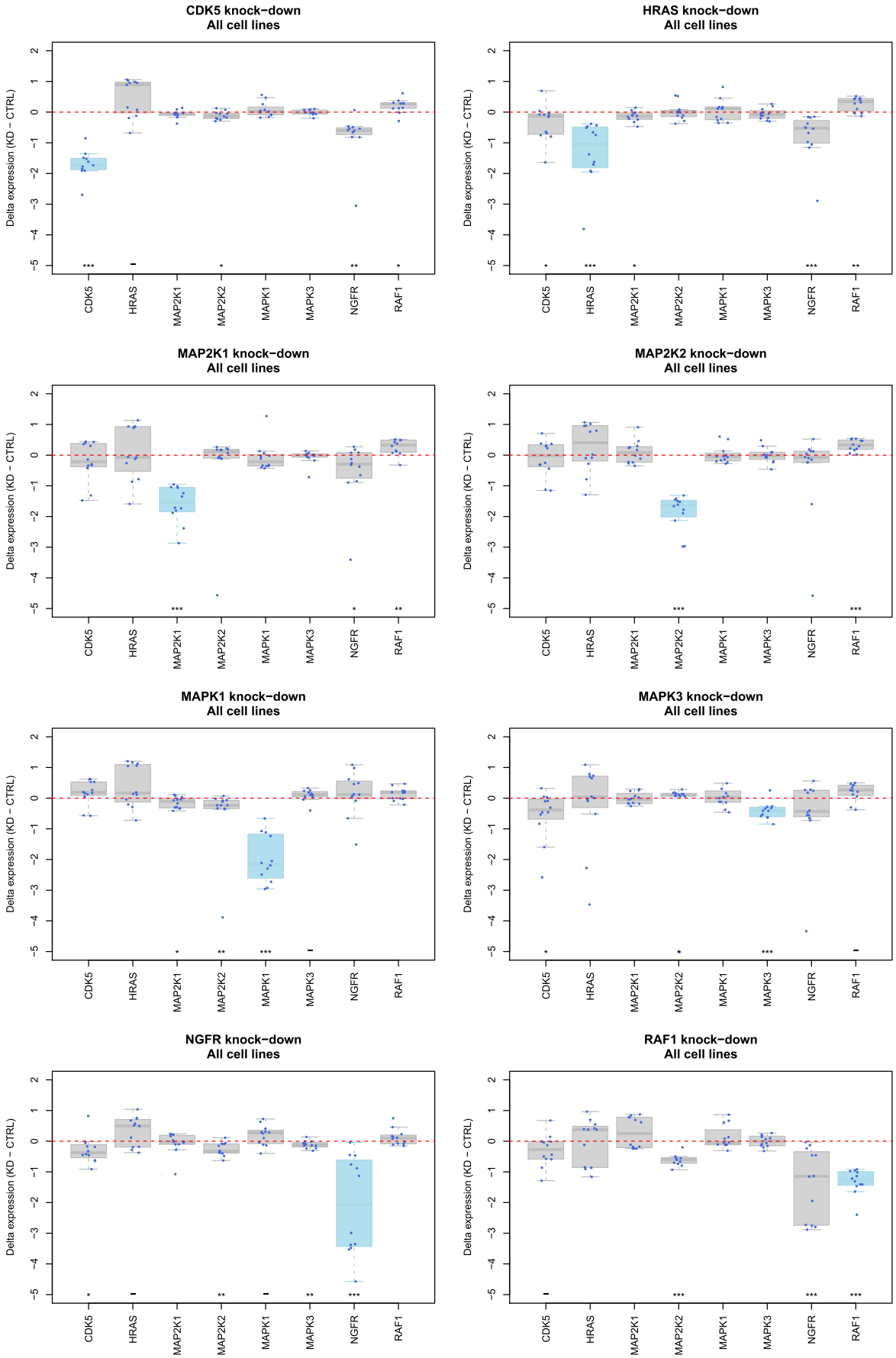

Fig. 3. Each plot shows the difference of expression for the eight core genes. The knocked down gene highlighted in light blue. The significance level is indicated by '-' for $p < 0.1$, '*' for $p < 0.05$, '**' for $p < 0.01$ and '****' for $p < 0.001$ using a Wilcoxon signed rank test.