Ehsani and Drabløs *BMC Bioinformatics* (2020) 21:134
https://doi.org/10.1186/s12859-020-3468-z

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Enhanced identification of significant regulators of gene expression

Rezvan Ehsani[1,2*] and Finn Drabløs[3*]

* Correspondence: rezvanehsani74@gmail.com; finn.drablos@ntnu.no
[1]Department of Mathematics, University of Zabol, Zabol, Iran
[3]Department of Cancer Research and Molecular Medicine, NTNU - Norwegian University of Science and Technology, NO-7491 Trondheim, Norway
Full list of author information is available at the end of the article

Abstract

Background: Diseases like cancer will lead to changes in gene expression, and it is relevant to identify key regulatory genes that can be linked directly to these changes. This can be done by computing a Regulatory Impact Factor (RIF) score for relevant regulators. However, this computation is based on estimating correlated patterns of gene expression, often Pearson correlation, and an assumption about a set of specific regulators, normally transcription factors. This study explores alternative measures of correlation, using the Fisher and Sobolev metrics, and an extended set of regulators, including epigenetic regulators and long non-coding RNAs (lncRNAs). Data on prostate cancer have been used to explore the effect of these modifications.

Results: A tool for computation of RIF scores with alternative correlation measures and extended sets of regulators was developed and tested on gene expression data for prostate cancer. The study showed that the Fisher and Sobolev metrics lead to improved identification of well-documented regulators of gene expression in prostate cancer, and the sets of identified key regulators showed improved overlap with previously defined gene sets of relevance to cancer. The extended set of regulators lead to identification of several interesting candidates for further studies, including lncRNAs. Several key processes were identified as important, including spindle assembly and the epithelial-mesenchymal transition (EMT).

Conclusions: The study has shown that using alternative metrics of correlation can improve the performance of tools based on correlation of gene expression in genomic data. The Fisher and Sobolev metrics should be considered also in other correlation-based applications.

Keywords: Regulatory impact factor, Gene regulation, Correlated gene expression, Fisher metric, Sobolev metric, Prostate cancer

Background

The transcription of genes is controlled through regulatory processes that are shared between subsets of genes, and this can lead to correlated levels of gene expression within such gene sets [1–3]. For example, increased expression of a gene producing a positive regulator can initiate higher expression values of the downstream genes that are controlled by the regulator, and this leads to correlation in expression values for

these genes across a relevant biological process, like a pathway of cellular differentiation. However, this correlation does not by itself show causality. Correlated expression pattern between two genes x and y may be a consequence of interaction between a regulator and a gene being regulated ($x \rightarrow y$), but it may also be a consequence of regulatory cascades ($x \rightarrow z \rightarrow y$), or of separate genes being regulated by the same regulator ($z \rightarrow x$; $z \rightarrow y$). This means that a general correlation network will show all possible interactions between genes, including all the indirect ones, as well as random correlations, making it challenging to identify the most important regulatory interactions in a given system.

The simplest approach for doing a more focused analysis is to specify a potential causality by postulating a specific regulator for a process and hypothesizing that correlated expression levels between the regulator and other genes indicates that this regulator has a significant regulatory impact on downstream targets in the process. This approach has in particular been implemented as regulatory impact factor (RIF) scores [4, 5]. Here transcription factors (TFs) are defined as regulators, and potential downstream targets are identified based on correlated significant differential expression (DE) using two different score values, RIF1 and RIF2. RIF1 identifies factors that are consistently co-expressed with highly abundant DE genes, whereas RIF2 identifies TFs with the ability to act as predictors of the abundance of DE genes.

The RIF approach has been used successfully in several studies. It has been used extensively for analysing data on livestock animals, like cattle and pigs. In such studies the RIF scores have been used to identify key regulators for feed efficiency [6], puberty [7, 8], and intramuscular fat content [9] in cattle, or growth and metabolism [10], high-altitude adaptation [11], and muscle characteristics [12] in pigs, to give a few examples. However, the RIF approach has also been used to analyse human data, in cases as diverse as colorectal cancer [13], effects of melphalan treatment [14], and biomarker candidates for total sleep deprivation [15]. The RIF approach was initially developed for TFs, but there have been a few examples of extension to other classes of regulators, like micro-RNAs (miRNAs) [16] and long non-coding RNAs (lncRNAs) [17], but mainly on data for livestock animals, and with limited testing. There have also been a few examples of integration of RIF scores into other relevant software tools, like RMaNI [18], INsPeCT [19], DCGL [20], RegulatorTrail [21], and REGGAE [22]. There have been some testing and comparisons of RIF to other approaches, as for example in [23, 24]. However, such comparisons can be challenging, at least partly because methods may have different requirements with respect to input data, e.g., gene lists vs. networks [24].

In this project we wanted to explore if the traditional RIF approach could be successfully extended and tested with focus on two important aspects. We have recently shown that correlations in gene expression can be identified more robustly [25] by using alternative correlation metrics like Fisher [26] or Sobolev [27], rather than the standard Pearson or Spearman correlation used in most studies, and we wanted to test if this could be applied to RIF scores. Also, since the initial implementation of RIF scores the interest in other regulators of gene expression has increased, and we wanted to test if specification of genes involved in other regulatory processes could give more insight into the role of these specific regulators. The lncRNAs [28] and genes involved in epigenetic processes (epigenetic factors, EFs) [29] are relevant examples. We therefore made a general implementation for computation of RIF scores that could be combined

with alternative correlation metrics and regulators, and tested this implementation on data for prostate cancer, generated by the TCGA Research Network [30]. The study shows that using either the Fisher or the Sobolev metric gives improved identification of relevant genes and gene sets. Specific lncRNAs and epigenetic factors are identified as important for cancer development, in addition to transcription factors.

Methods

General workflow

The schematic workflow is shown in Fig. 1, and uses the general strategy of Reverter et al. [4]. The starting point is a gene expression dataset involving two separate conditions, e.g., microarray data for normal and cancer tissue. The data are normalized, and a standard statistical analysis is used to find target genes with significant difference in expression level between the two conditions, known as differentially expressed (DE) genes. The set of regulators (e.g., TFs, EFs, and lncRNAs) included in the experimental data are extracted, using gene lists from published databases for identification. The correlation of co-expression between regulators and DE genes is computed for the two conditions, and the difference in correlation is used to estimate the differential wiring (DW), as shown below. This can be used to assign RIF scores to the regulators that show consistent differential co-expression with genes that are both highly abundant and differentially expressed (RIF1 score), and to the regulators with the best ability to predict the abundance of DE genes (RIF2 score). In the current project PCIT analysis is subsequently used to find interactions between regulators.

Datasets

For definition of causality (see Background) we identified three partly overlapping classes of genes representing potential regulators; transcription factors (TFs), epigenetic factors (EFs), and long non-coding RNAs (lncRNAs). We used a list of 1978 TFs from our previous work [31]. We used 815 EFs from the Epifactor database of gene products involved in epigenetics, including 98 TFs [29]. This list also includes some additional

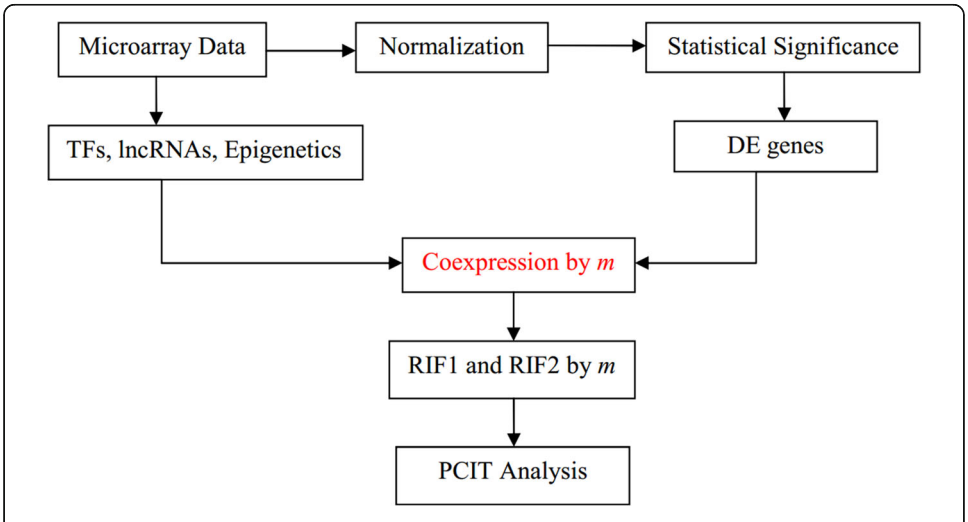

Fig. 1 General workflow of the analysis

well-known regulators, mainly various kinases, here listed together with epigenetic factors. We used 12,980 lncRNAs from Jiang et al. [32].

As a test case we identified key regulators of prostate cancer. We used data generated by the TCGA Research Network [30], using 21 paired samples on prostate adenocarcinoma (PRAD), with count data on gene expression. In each case two samples had been taken from the same donor; one sample to represent Primary Tumor, the other sample representing Solid Tissue Normal.

Normalization and differential expression

We used the edgeR package [33] to identify differentially expressed (DE) genes. The edgeR is based on a negative binomial model where a weighted fixed mean of the log expression ratios is used to normalize the sequencing depth and gene length between samples. The negative binomial model is made by using expression data where the relation between mean μ and variance is given by $= \mu + \alpha\mu^2$. The dispersion factor α is estimated using a combination of common dispersion on all the genes (estimated by a likelihood function), and a gene-specific dispersion (estimated by Bayes method). Finally, an exact test with false discovery rate (FDR) is used to identify DE genes.

Measures of RIF

For each regulator, the regulatory impact factor (RIF) tries to estimate the change in co-expression between the regulator and the DE genes. The RIF scores have been introduced and described by Reverter et al. [4], where full mathematical definitions can be found. RIF1 identifies regulators that are consistently co-expressed with highly abundant DE genes. It is computed from multiplying phenotype impact factor (PIF) with differential wiring (DW), where PIF is the difference in squared expression for a given DE gene for two conditions, whereas DW is computed from the difference in co-expression correlation between a regulator and the DE gene for the two conditions. Thus, RIF1 captures those regulators that have a large differential wiring to highly abundant highly DE genes. RIF2 identifies regulators with the ability to act as predictors of the abundance of DE genes. It is estimated from the difference of squared expression weighted by the squared co-expression correlation between the regulator and the DE genes in two conditions. The computation of RIF scores was implemented as an R program (see Availability of data and materials).

Metrics for correlation of co-expression

To identify correlation in co-expression for a typical gene g standard statistical metrics like Pearson and Spearman are normally used. However, in a recent study we used the geometrical metrics Sobolev and Fisher information to annotate lncRNAs based on co-expression [25], and could show that these geometrical metrics had better performance than the more commonly used statistical metrics. A detailed description of each of these novel metrics can be found in [25], which builds on definitions and notations as given by Villman [27] for the Sobolev metric, and definitions and notations given by Lebanon [26] for the Fisher metric.

Interactions between regulators

We used the PCIT algorithm [34] to find significant interactions between regulators. The gene list used for PCIT included the top 10 regulators with best $RIF1_m$ and $RIF2_m$ scores for each metric m separately, but only those with a direct and partial correlation ≥0.90 were used for the analysis.

Analysis of gene sets

For identification of genes previously associated with cancer we used reference data from PubMed, accessed on January 2019. Overlap with existing predefined gene sets was done with the online tool for computing overlap with MSigDB [35, 36]. The tools DAVID [37, 38] and Enrichr [39, 40] were used for enrichment analysis. We used g:Profiler [41] to convert HGNC IDs to UniProt/SwissProt IDs before DAVID analysis. The listed *p*-values for DAVID and Enrichr are after Benjamini correction as done by each of the tools.

Results

Important regulators in prostate cancer

Using the data set from TCGA with 21 paired samples on normal and cancerous prostate tissue, the RIF score was computed for all regulators, using external lists of transcription factors, epigenetic factors, and long non-coding RNAs for defining regulators, and four different metrics for correlation of expression values. Significant interactions between regulators were identified by using the PCIT algorithm [34]. The number of significant regulators of each type and metric is shown in Table 1, and the overlap between the results is shown as a Venn diagram in Fig. 2. There is a clear overlap between the different methods, and the largest group of overlaps is for a set of 9 genes found by all correlation metrics. However, in total most entries are unique to each method. This highlights the importance of identifying the most robust correlation measure for subsequent analysis.

Overlap with known key regulators

It seems reasonable to assume that a high-quality prediction should include several important regulators. It may be difficult to define exactly what we mean by "important" in this context, but one approach is to assume that these regulators already have been identified as important, and therefore have been studied (and described) in many publications. We therefore used a simple approach where we counted the number of publications on each gene, as listed in PubMed. This was based on standard PubMed searches using either the gene name itself, or the gene name in combination with "cancer" or "prostate cancer", and we did this for both transcription factors and epigenetic factors, based on the assumption that these

Table 1 Number of significant regulators for each correlation metric and in total

	Pearson	Spearman	Fisher	Sobolev	Total
TFs	18	12	18	15	32
EFs	8	9	14	15	22
lncRNAs	16	8	15	8	33
Sum	42	29	47	38	87

Total is the number of unique entries across all methods

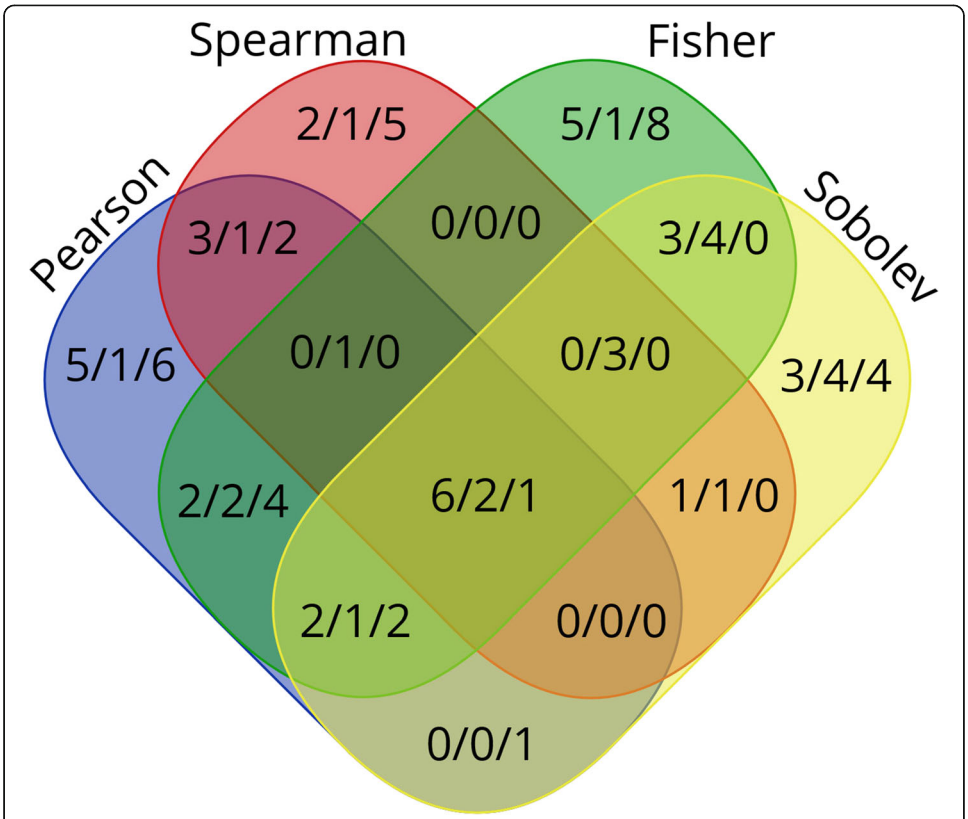

Fig. 2 Venn diagram of the significant regulators according to correlation metric. The figure shows numbers for TFs / EFs / lncRNAs, respectively. The Venn diagram tool at [42] was used

two groups of genes represent the most extensively studied regulators in our study, compared to lncRNAs. The gene for HR (HR lysine demethylase and nuclear receptor corepressor) was not included in this analysis, due to the high number of non-relevant hits during the PubMed searches. The result is shown in Table 2. An extended table with gene names and statistics is available (Additional file 1), as well as tables of gene sets used for the analysis (Additional file 2). The PubMed analysis is a simplistic approach, and it assumes a relatively standardized use of gene names in publications, but it still shows a clear trend, where the alternative metrics Sobolev and Fisher consistently have retrieved the gene sets with most publications, with the Fisher metric doing slightly better than Sobolev.

It is possible that this result could be dominated by a small number of highly described genes, where any method retrieving these specific genes (possibly by chance) would get a high score. We therefore identified the genes that were most frequently described in combination with prostate cancer, for TFs (8 genes) and EFs (7 genes), i.e., 15 genes in total. We then checked how many of these genes were found by each method, which was 5 genes for Pearson, 3 genes for Spearman, 11 genes for Fisher, and 8 genes for Sobolev.

Table 2 Average number of publications in PubMed

	Pearson	Spearman	Fisher	Sobolev
TFs	10.53	4.00	24.82	22.60
EFs	3.75	5.33	21.29	15.20

Numbers are for searching with gene name and "prostate cancer" over all significant regulators

This again shows that the Fisher and Sobolev metrics gave the best performance with respect to identification of well-known regulators.

Overlap with predefined gene sets

It also seems reasonable to assume that a good set of regulators should be informative by showing significant overlap with predefined gene sets from literature. Several such gene sets have been collected in the MSigDB database, which also offers a web-based tool for computing overlap between input gene sets and gene sets in the MSigDB database. This was used to estimate the overlap between the gene set for each of the methods against all gene sets in MSigDB, except gene sets based on GO terms. GO terms were excluded because the initial set of genes is based on their role in gene regulation (like transcription factors), therefore all GO terms related to gene regulation will be highly significant, which will then dominate the analysis, whereas in this case we want to focus on what we can learn *in addition* to what we already know. For each gene set (Pearson, Spearman, Fisher, Sobolev) we retrieved the FDR q-score for the overlap between the sets and transformed it to its -log value for all significant overlaps (limited to the top 100). The results are shown in Fig. 3. Again, the results show that Fisher and Sobolev is better than Pearson and Spearman, as those metrics define gene sets that show a more significant overlap with existing gene sets from literature.

We will now discuss the results in more detail. Based on the results regarding metrics we will focus on regulators identified by the Fisher and Sobolev metrics, and in particular clusters of regulators that could be found in both analyses. Figure 4 show networks of significant regulators and interactions in prostate cancer as estimated with these two metrics.

Discussion

Overlap with specific MSigDB gene sets

The overlap of sets of regulators with specific MSigDB gene sets, including overlap with the GO sets, can provide interesting information on relevant processes that were identified by the analysis. As expected, both sets (Fisher and Sobolev) showed significant

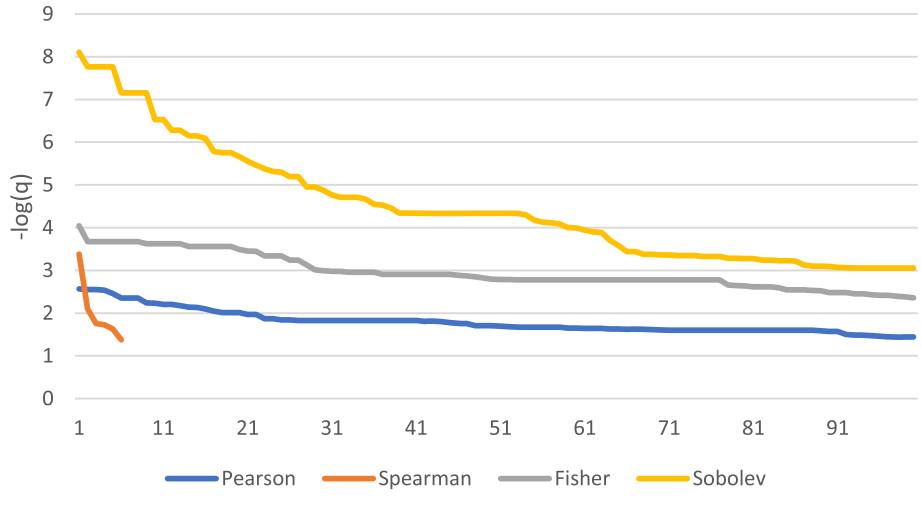

Fig. 3 Significance of overlap with MSigDB gene sets. The graph shows ranked FDR q-scores (as -log) of overlap between each gene set and the MSigDB reference sets

overlap with gene sets related to gene regulation, like GO_POSITIVE_REGULATION_OF_GENE_EXPRESSION. However, the most significant terms for Sobolev were GO_CHROMOSOME_ORGANIZATION and GO_CHROMATIN_ORGANIZATION, indicating a strong involvement of chromatin-specific processes. The original analysis (without including the GO-specific gene sets) had best significance for Sobolev against GNF2_BUB1B (Neighborhood of BUB1B, which is an important kinase involved in the mitotic spindle checkpoint [43], ROSTY_CERVICAL_CANCER_PROLIFERATION_CLUSTER, and MODULE_54 (Cell cycle expression cluster). The overlap with the cell cycle expression cluster consisted of the genes TOP2A, BUB1, AURKA, AURKB, HJURP, ASF1B, ANP32E and RAD54L. Most of these genes were also important for the other overlaps with high significance, and several of the genes will be discussed below. The overlap for the Fisher gene set also included several interesting gene sets, like BURTON_ADIPOGENESIS_3 (Differentiation into adipocytes), and involved in particular the genes CDK1, TOP2A, and AURKA.

Enrichment analysis

Properties represented by the sets of regulators can also be described by enrichment analysis. We used DAVID and Enrichr as two complementary approaches. Both do standard enrichment analysis, but DAVID can also do clustering on enrichment results, which can highlight interesting trends across the individual enrichments, whereas Enrichr has a very large set of reference libraries, which can highlight additional enriched properties.

In DAVID the gene sets from both Fisher and Sobolev gave clusters with strong enrichment for processes associated with transcription regulation, as expected, and clusters with strong enrichment for zinc fingers. In addition, the Sobolev gene set gave clusters with enrichment for terms associated with cell cycle, centromere and chromosome, including cell proliferation, mitosis and cell division. This trend was reinforced by enrichment results for GOTERM_CC_FAT, where significant enrichments included terms associated with centromeric region ($p = 1.3e\text{-}2$), centrosome ($p = 8.8e\text{-}2$), and spindle (8.8e-2). The gene set retrieved by using the Sobolev metric therefore highlights the importance of fundamental processes directly associated with cell division.

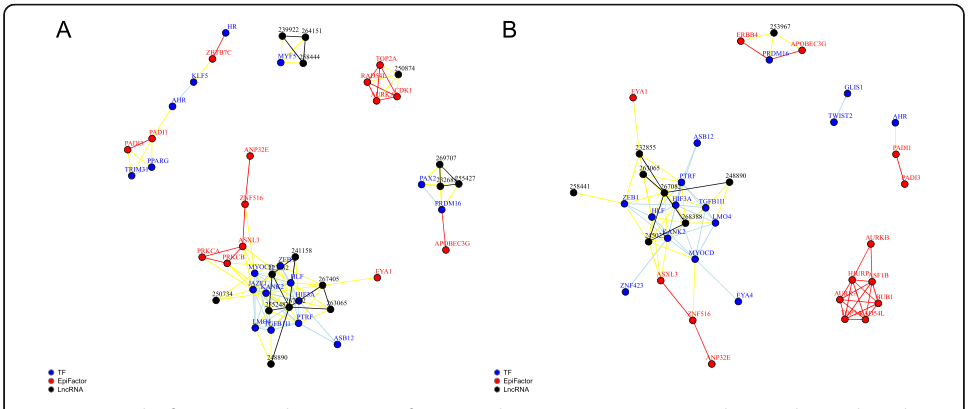

Fig. 4 Network of interactions between significant regulators in prostate cancer. The correlations have been computed according to the Fisher metric (**a**) and the Sobolev metric (**b**), and the network nodes of significant regulators are TFs (blue), EFs (red), and lncRNAs (black, with ENSG numbers as ENSG00000xxxxxx)

The Enrichr analysis showed a strong enrichment for downregulated genes from GEO data on prostate cancer ($p = 3.6e\text{-}5$), as expected. Enrichr also highlighted knockdown data from LINCS L1000 Kinase Perturbations for WEE1 (p = 8.8e-4) and ERBB3 ($p = 4.3e\text{-}4$). Especially WEE1 (WEE1 G2 checkpoint kinase) is known to be an important cell cycle checkpoint kinase [44], but also ERBB3 (Erb-B2 receptor tyrosine kinase 3) is important in cancer development and has been identified as a potential target for treatment (e.g. [45],). There was also a significant enrichment for FOXM1 ChEA 2016 ChIP-Seq data ($p = 2.9e\text{-}2$) and E2F4 ENCODE data ($p = 7.3e\text{-}4$), and both FOXM1 (Forkhead box M1) and E2F4 (E2F transcription factor 4) are known to be important for cell cycle regulation, see for example [46]. This highlights again the central role of cell cycle regulation in cancer development.

Specific gene clusters and individual genes

The computational analysis also identified potential interactions between genes, leading to a network-like representation of clusters of interacting genes, indicating genes with shared activities. These clusters can be seen in Fig. 4. Specifically, there is a highly interacting set of genes consisting of TOP2A, RAD54L and AURKA, possibly with the inclusion of CDK1, AURKB, HJURP, ASF1B, and BUB1 (here discussed as gene set A). There is also a more diverse set, but with a very high overlap between Fisher and Sobolev consisting of EYA1, ZEB1, ASB12, PTRF, HIF3A, HLF, TGFB1I1, LMO4, MYOCD, ASXL3, ZNF516, and ANP32E, possibly with the inclusion of ZNF423, EYA4, PRKCA, PREKCB, JAZF1 and KANK2 (gene set B). Many of these genes represent well-known cases of regulators involved in cancer development, including prostate cancer, as can be seen from Additional file 1, although at different levels of experimental verification. The more novel predictions obviously must be experimentally verified, but the relevance of individual cases (based on the PubMed data) indicates that many of the predictions are likely to be true positive. This is particularly relevant for the lncRNAs, where in most cases very little is known about their function.

Gene set A overlaps with cell cycle signatures [47, 48], including the MSigDB cell cycle signature mentioned above, confirming that this gene set represents cell cycle regulation. The individual genes are known to be involved in several processes, including processes related to DNA topology, centrosomes and mitotic spindle formation. This includes AURKA (Aurora kinase A), which is associated with centrosome maturation and separation, and regulates spindle assembly and stability. It has been shown that AURKA is important for development of prostate cancer, and that it represents a possible target for treatment [49]. TOP2A (DNA Topoisomerase IIα) controls and alters the topological states of DNA and is important for proper segregation of daughter chromosomes [50]. RAD54L (RAD54 Like) is a helicase which seems to influence DNA topology in different ways, e.g., in chromatin remodelling, homologous recombination and interaction with Holliday junctions [51]. The Fisher gene set also includes CDK1 (Cyclin dependent kinas), which is known to be essential for cell division, and is known to be involved in regulation of cell cycle through centrosomes / spindle assembly [52]. The Sobolev gene set also includes ASF1B (Anti-silencing function 1B histone chaperone), which is essential for the mitotic spindle checkpoint during the cell cycle [53], AURKB (Aurora kinase B), which participates in the regulation of alignment and

segregation of chromosomes [54], HJURP (Holliday junction recognition protein), which mediates the centromere-specific assembly of CENP-A nucleosomes important for chromosome segregation [55, 56], and BUB1 (Mitotic checkpoint serine/threonine kinase), which is essential for spindle-assembly checkpoint signalling [57]. This is a strong indication that cell cycle processes associated with chromosome handling and the mitotic spindle are important in prostate cancer.

Gene set B is more diverse and does not show very strong overlap with any specific gene set in MSigDB. However, many of the genes in gene set B are known to be involved in the endothelial to mesenchymal transition (EMT), which is a key process in prostate cancer. ZEB1 (zinc finger E-box binding homeobox 1) has frequently been associated with cancer in general, and prostate cancer in particular. It has been shown that up-regulation of ZEB1 drives EMT in human prostate cancer cells [58]. HIF3A (hypoxia inducible factor 3 subunit alpha) is another gene that has been linked to prostate cancer. It is a negative regulator of HIF1A [59], and it has been shown that destabilisation of HIF1A by ERβ can induce EMT [60]. EYA1 (EYA Transcriptional coactivator and phosphatase 1) is a transcriptional coactivator of the SIX homeobox genes and is a coregulator of TGF-β signalling during EMT [61]. LMO4 (LIM domain only 4) has been shown to be an essential cofactor in EMT at least in neuroblastoma and neural crest cells [62]. ANP32E (Acidic nuclear phosphoprotein 32 family member E) is a histone chaperone that mediates removal of histone H2A.Z from the nucleosome [63], and it has been shown that H2A.Z is a master regulator of EMT [64]. TGFB1I1 (Transforming growth factor beta 1 induced transcript 1) is a coactivator of the androgen receptor, it is known to be associated with prostate cancer, and it can induce EMT, at least in astrocytomes. From the Fisher gene set, JAZF1 (JAZF zinc finger 1) has been shown to promote prostate cancer progression through JNK/Slug, leading to enhanced EMT [65]. Although these genes are just examples of genes identified by our analysis and the processes they may be involved in, this list seems to be a clear confirmation of the importance of EMT in prostate cancer [66].

Interestingly, gene set B also includes 3 lncRNAs that were found by both Fisher and Sobolev, consisting of ENSG00000263065, which is antisense to exon 21 and 22 of MYH11, ENSG00000248890, which is antisense to HHIP, close to TSS, and ENSG00000267082, which is antisense to DOCK6. MYH11 (Myosin heavy chain 11) has been used as a marker of mesenchymal and endothelial differentiation [67], and HHIP (Hedgehog interacting protein) can be linked to EMT through the hedgehog pathway [68]. Although these associations between lncRNAs and cancer are more circumstantial, they are consistent with the general view that EMT is important in prostate cancer.

We have compared the gene sets from this study to other sets of "cancer genes", using the library of 299 cancer driver genes of Bailey et al. [69], the COSMIC library of 723 cancer genes [70], and the knockout library of 684 genes of Hart et al. [71]. In all cases the overlap was very low, in the range of just 1–2 genes. However, this is hardly surprising, as they represent quite different classes of cancer genes. The cancer genes are genes where a mutation may drive cancer, possibly without changing the expression level of the gene itself. The knockout experiments, on the other hand will identify genes that are essential to survival of the cancer cell, but not necessarily linked to changes in expression level. The RIF approach focuses on regulators where it is possible to observe

a change in expression level, linked to cancer, and where this change is reflected in a downstream process, possibly regulated by the regulator. This means that we identify parts of the regulatory system that are affected by the changes introduced through cancer, rather than the cancer genes themselves, and the overlap between these different approaches will therefore be low.

Conclusions

We have developed an extended approach for identification of important regulators in biological processes, based on correlations in gene expression. We have implemented alternative metrics for correlation and used this on additional groups of regulators; epigenetic factors and lncRNAs. There is no benchmark dataset for this type of analysis, which makes it difficult to compare our extensions to previous approaches. However, we have shown that using alternative metrics as correlation measures identifies more genes that previously have been associated with cancer and identifies gene sets with better overlap with known gene sets, using prostate cancer as a test case. A case-by-case study of the genes and gene sets shows that the identified genes are relevant for understanding the processes involved in development of prostate cancer, and that they point at mitotic spindle formation and the endothelial to mesenchymal transition as important processes. The study also identifies specific lncRNAs that are likely to be important in these processes.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-3468-z.

Additional file 1. Gene lists with PubMed statistics.

Additional file 2. Gene sets used for enrichment analysis.

Additional file 3. Identifiers for TCGA datasets on prostate cancer as downloaded from the GDC data portal.

Abbreviations

DE: Differential expression; DW: Differential wiring; EF: Epigenetic factor; EMT: Epithelial-mesenchymal transition; FDR: False discovery rate; GO: Gene ontology; lncRNA: Long non-coding RNA; PIF: Phenotype impact factor; RIF: Regulatory impact factor; TF: Transcription factor

Acknowledgements

The results shown here are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Authors' contributions

RE initiated the study, developed the software, and did the computational analysis. FD did the functional interpretation of the data. Both authors have contributed to the manuscript and have read and approved the final version.

Funding

The project has been partly funded by University of Zabol to RE. The funding body played no roles in the design of the study, or in collection, analysis, and interpretation of data, or in writing the manuscript.

Availability of data and materials

The Cancer Genome Atlas (TCGA) datasets for prostate cancer supporting the conclusions of this article are available in the repository in Genomic Data Commons Data Portal at https://portal.gdc.cancer.gov/. The list of datasets is available in Additional file 3.
The software for computing RIF scores with alternative correlation measures is available via GitHub at [72].
Project name: RIF scores with alternative correlation measures.
Project home page: https://github.com/RezvanEhsani
Archived version: Please see the GitHub repository.
Operating system(s): Platform independent.
Programming language: R.
Other requirements: Please see the GitHub repository.
License: GNU GPL.
Any restrictions to use by non-academics: None.

Ethics approval and consent to participate
Not applicable. The study uses open data, and ethics approval and consent is not needed.

Consent for publication
Not applicable.

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

Author details
[1]Department of Mathematics, University of Zabol, Zabol, Iran. [2]Department of Bioinformatics, University of Zabol, Zabol, Iran. [3]Department of Cancer Research and Molecular Medicine, NTNU - Norwegian University of Science and Technology, NO-7491 Trondheim, Norway.

Received: 12 September 2019 Accepted: 24 March 2020
Published online: 06 April 2020

References

1. Gu Q, Nagaraj SH, Hudson NJ, Dalrymple BP, Reverter A. Genome-wide patterns of promoter sharing and co-expression in bovine skeletal muscle. BMC Genomics. 2011;12:23.
2. Marco A, Konikoff C, Karr TL, Kumar S. Relationship between gene co-expression and sharing of transcription factor binding sites in Drosophila melanogaster. Bioinformatics. 2009;25(19):2473–7.
3. Michalak P. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics. 2008;91(3):243–8.
4. Reverter A, Hudson NJ, Nagaraj SH, Perez-Enciso M, Dalrymple BP. Regulatory impact factors: unraveling the transcriptional regulation of complex traits from expression data. Bioinformatics. 2010;26(7):896–904.
5. Hudson NJ, Dalrymple BP, Reverter A. Beyond differential expression: the quest for causal mutations and effector molecules. BMC Genomics. 2012;13:356.
6. Alexandre PA, Naval-Sanchez M, Porto-Neto LR, Ferraz JBS, Reverter A, Fukumasu H. Systems biology reveals NR2F6 and TGFB1 as key regulators of feed efficiency in beef cattle. Front Genet. 2019;10:230.
7. Nguyen LT, Reverter A, Canovas A, Venus B, Anderson ST, Islas-Trejo A, et al. STAT6, PBX2, and PBRM1 emerge as predicted regulators of 452 differentially expressed genes associated with puberty in Brahman heifers. Front Genet. 2018;9:87.
8. Canovas A, Reverter A, DeAtley KL, Ashley RL, Colgrave ML, Fortes MR, et al. Multi-tissue omics analyses reveal molecular regulatory networks for puberty in composite beef cattle. PLoS One. 2014;9(7):e102551.
9. Cesar AS, Regitano LC, Koltes JE, Fritz-Waters ER, Lanna DP, Gasparin G, et al. Putative regulatory factors associated with intramuscular fat content. PLoS One. 2015;10(6):e0128350.
10. Ayuso M, Fernandez A, Nunez Y, Benitez R, Isabel B, Fernandez AI, et al. Developmental stage, muscle and genetic type modify muscle transcriptome in pigs: effects on gene expression and regulatory factors involved in growth and metabolism. PLoS One. 2016;11(12):e0167858.
11. Jia C, Kong X, Koltes JE, Gou X, Yang S, Yan D, et al. Gene co-expression network analysis unraveling transcriptional regulation of high-altitude adaptation of Tibetan pig. PLoS One. 2016;11(12):e0168161.
12. Ovilo C, Benitez R, Fernandez A, Nunez Y, Ayuso M, Fernandez AI, et al. Longissimus dorsi transcriptome analysis of purebred and crossbred Iberian pigs differing in muscle characteristics. BMC Genomics. 2014;15:413.
13. Nagaraj SH, Reverter A. A Boolean-based systems biology approach to predict novel genes associated with cancer: application to colorectal cancer. BMC Syst Biol. 2011;5:35.
14. Yang Y, Xing Y, Liang C, Hu L, Xu F, Mei Q. In search of underlying mechanisms and potential drugs of melphalan-induced vascular toxicity through retinal endothelial cells using bioinformatics approach. Tumour Biol. 2016;37(5):6709–18.
15. Uyhelji HA, Kupfer DM, White VL, Jackson ML, Van Dongen HPA, Burian DM. Exploring gene expression biomarker candidates for neurobehavioral impairment from total sleep deprivation. BMC Genomics. 2018;19(1):341.
16. Marmol-Sanchez E, Ramayo-Caldas Y, Quintanilla R, Cardoso TF, Gonzalez-Prendes R, Tibau J, et al. Co-expression network analysis predicts a key role of microRNAs in the adaptation of the porcine skeletal muscle to nutrient supply. J Anim Sci Biotechnol. 2020;11:10.
17. Nolte W, Weikard R, Brunner RM, Albrecht E, Hammon HM, Reverter A, et al. Biological network approach for the identification of regulatory long non-coding RNAs associated with metabolic efficiency in cattle. Front Genet. 2019;10:1130.
18. Madhamshettiwar PB, Maetschke SR, Davis MJ, Ragan MA. RMaNI: regulatory module network inference framework. BMC Bioinformatics. 2013;14 Suppl 16:S14.
19. Madhamshettiwar PB, Maetschke SR, Davis MJ, Reverter A, Ragan MA. INsPeCT: INtegrative platform for cancer transcriptomics. Cancer Inform. 2014;13:59–66.
20. Yang J, Yu H, Liu BH, Zhao Z, Liu L, Ma LX, et al. DCGL v2.0: an R package for unveiling differential regulation from differential co-expression. PLoS One. 2013;8(11):e79729.
21. Kehl T, Schneider L, Schmidt F, Stockel D, Gerstner N, Backes C, et al. RegulatorTrail: a web service for the identification of key transcriptional regulators. Nucleic Acids Res. 2017;45(W1):W146–53.
22. Kehl T, Schneider L, Kattler K, Stockel D, Wegert J, Gerstner N, et al. REGGAE: a novel approach for the identification of key transcriptional regulators. Bioinformatics. 2018;34(20):3503–10.
23. Wu C, Zhu J, Zhang X. Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. BMC Bioinformatics. 2012;13:182.
24. Yu H, Mitra R, Yang J, Li Y, Zhao Z. Algorithms for network-based identification of differential regulators from transcriptome data: a systematic evaluation. Sci China Life Sci. 2014;57(11):1090–102.

25. Ehsani R, Drablos F. Measures of co-expression for improved function prediction of long non-coding RNAs. BMC Bioinformatics. 2018;19(1):533.
26. Lebanon G. Learning riemannian metrics. In: *Proceedings of the nineteenth conference on uncertainty in artificial intelligence. Acapulco*: Morgan Kaufmann Publishers Inc; 2003. p. 362–9.
27. Villmann T. Sobolev metrics for learning of functional data - mathematical and theoretical aspects. In: Villmann T, Schleif F-M, editors. Machine learning reports, vol. 1. Leipzig: Medical Department, University of Leipzig; 2007. p. 1–13.
28. Jarroux J, Morillon A, Pinskaya M. History, discovery, and classification of lncRNAs. Adv Exp Med Biol. 2017;1008:1–46.
29. Medvedeva YA, Lennartsson A, Ehsani R, Kulakovskiy IV, Vorontsov IE, Panahandeh P, et al. EpiFactors: a comprehensive database of human epigenetic factors and complexes. Database (Oxford). 2015;2015:bav067.
30. TCGA Research Network. https://www.cancer.gov/tcga. Accessed 1 Apr 2020.
31. Ehsani R, Bahrami S, Drablos F. Feature-based classification of human transcription factors into hypothetical sub-classes related to regulatory function. BMC Bioinformatics. 2016;17(1):459.
32. Jiang Q, Ma R, Wang J, Wu X, Jin S, Peng J, et al. LncRNA2Function: a comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data. BMC Genomics. 2015;16 Suppl 3:S2.
33. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
34. Watson-Haigh NS, Kadarmideen HN, Reverter A. PCIT: an R package for weighted gene co-expression networks based on partial correlation and information theory approaches. Bioinformatics. 2010;26(3):411–3.
35. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739–40.
36. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
37. da Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.
38. da Huang W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1–13.
39. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14:128.
40. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):W90–7.
41. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. G:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47(W1):W191–8.
42. Venn diagrams. http://bioinformatics.psb.ugent.be/webtools/Venn/. Accessed 1 Apr 2020.
43. Hahn MM, Vreede L, Bemelmans SA, van der Looij E, van Kessel AG, Schackert HK, et al. Prevalence of germline mutations in the spindle assembly checkpoint gene BUB1B in individuals with early-onset colorectal cancer. Genes Chromosomes Cancer. 2016;55(11):855–63.
44. Matheson CJ, Backos DS, Reigan P. Targeting WEE1 kinase in cancer. Trends Pharmacol Sci. 2016;37(10):872–81.
45. Liu X, Liu S, Lyu H, Riker AI, Zhang Y, Liu B. Development of effective therapeutics targeting HER3 for cancer treatment. Biol Proced Online. 2019;21:5.
46. Grant GD, Brooks L 3rd, Zhang X, Mahoney JM, Martyanov V, Wood TA, et al. Identification of cell cycle-regulated genes periodically expressed in U2OS cells and their regulation by FOXM1 and E2F transcription factors. Mol Biol Cell. 2013;24(23):3634–50.
47. Gobin M, Nazarov PV, Warta R, Timmer M, Reifenberger G, Felsberg J, et al. A DNA repair and cell-cycle gene expression signature in primary and recurrent glioblastoma: prognostic value and clinical implications. Cancer Res. 2019;79(6):1226–38.
48. Engeland K. Cell cycle arrest through indirect transcriptional repression by p53: I have a DREAM. Cell Death Differ. 2018;25(1):114–32.
49. Lee EC, Frolov A, Li R, Ayala G, Greenberg NM. Targeting Aurora kinases for the treatment of prostate cancer. Cancer Res. 2006;66(10):4996–5002.
50. Broderick R, Niedzwiedz W. Sister chromatid decatenation: bridging the gaps in our knowledge. Cell Cycle. 2015;14(19):3040–4.
51. Mazin AV, Mazina OM, Bugreev DV, Rossi MJ. Rad54, the motor of homologous recombination. DNA Repair (Amst). 2010;9(3):286–302.
52. Crasta K, Lim HH, Zhang T, Nirantar S, Surana U. Consorting kinases, end of destruction and birth of a spindle. Cell Cycle. 2008;7(19):2960–6.
53. D'Angiolella V, Mari C, Nocera D, Rametti L, Grieco D. The spindle checkpoint requires cyclin-dependent kinase activity. Genes Dev. 2003;17(20):2520–5.
54. Shuda K, Schindler K, Ma J, Schultz RM, Donovan PJ. Aurora kinase B modulates chromosome alignment in mouse oocytes. Mol Reprod Dev. 2009;76(11):1094–105.
55. Dunleavy EM, Roche D, Tagami H, Lacoste N, Ray-Gallet D, Nakamura Y, et al. HJURP is a cell-cycle-dependent maintenance and deposition factor of CENP-A at centromeres. Cell. 2009;137(3):485–97.
56. Foltz DR, Jansen LE, Bailey AO, Yates JR 3rd, Bassett EA, Wood S, et al. Centromere-specific assembly of CENP-a nucleosomes is mediated by HJURP. Cell. 2009;137(3):472–84.
57. Jia L, Li B, Yu H. The Bub1-Plk1 kinase complex promotes spindle checkpoint signalling through Cdc20 phosphorylation. Nat Commun. 2016;7:10818.
58. Graham TR, Zhau HE, Odero-Marah VA, Osunkoya AO, Kimbro KS, Tighiouart M, et al. Insulin-like growth factor-I-dependent up-regulation of ZEB1 drives epithelial-to-mesenchymal transition in human prostate cancer cells. Cancer Res. 2008;68(7):2479–88.
59. Heikkila M, Pasanen A, Kivirikko KI, Myllyharju J. Roles of the human hypoxia-inducible factor (HIF)-3alpha variants in the hypoxia response. Cell Mol Life Sci. 2011;68(23):3885–901.
60. Mak P, Leav I, Pursell B, Bae D, Yang X, Taglienti CA, et al. ERbeta impedes prostate cancer EMT by destabilizing HIF-1alpha and inhibiting VEGF-mediated snail nuclear localization: implications for Gleason grading. Cancer Cell. 2010;17(4):319–32.
61. Liu Y, Kong D, Wu H, Yuan X, Xu H, Zhang C, et al. Interplay of retinal determination gene network with TGF-beta signaling pathway in epithelial-mesenchymal transition. Stem Cell Investig. 2015;2:12.

62. Ferronha T, Rabadan MA, Gil-Guinon E, Le Dreau G, de Torres C, Marti E. LMO4 is an essential cofactor in the Snail2-mediated epithelial-to-mesenchymal transition of neuroblastoma and neural crest cells. J Neurosci. 2013;33(7):2773–83.
63. Obri A, Ouararhni K, Papin C, Diebold ML, Padmanabhan K, Marek M, et al. ANP32E is a histone chaperone that removes H2A.Z from chromatin. Nature. 2014;505(7485):648–53.
64. Domaschenz R, Kurscheid S, Nekrasov M, Han S, Tremethick DJ. The histone variant H2A.Z is a master regulator of the epithelial-mesenchymal transition. Cell Rep. 2017;21(4):943–52.
65. Sung Y, Park S, Park SJ, Jeong J, Choi M, Lee J, et al. Jazf1 promotes prostate cancer progression by activating JNK/slug. Oncotarget. 2018;9(1):755–65.
66. Montanari M, Rossetti S, Cavaliere C, D'Aniello C, Malzone MG, Vanacore D, et al. Epithelial-mesenchymal transition in prostate cancer: an overview. Oncotarget. 2017;8(21):35376–89.
67. Lu CC, Liu MM, Clinton M, Culshaw G, Argyle DJ, Corcoran BM. Developmental pathways and endothelial to mesenchymal transition in canine myxomatous mitral valve disease. Vet J (London, England : 1997). 2015;206(3):377–84.
68. Choi SS, Omenetti A, Witek RP, Moylan CA, Syn WK, Jung Y, et al. Hedgehog pathway activation and epithelial-to-mesenchymal transitions during myofibroblastic transformation of rat hepatic cells in culture and cirrhosis. Am J Physiol Gastrointest Liver Physiol. 2009;297(6):G1093–106.
69. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive characterization of cancer driver genes and mutations. Cell. 2018;173(2):371–85 e318.
70. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2019;47(D1):D941–d947.
71. Hart T, Tong AHY, Chan K, Van Leeuwen J, Seetharaman A, Aregger M, et al. Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. G3 (Bethesda, Md). 2017;7(8):2719–27.
72. RIF Scores. https://github.com/RezvanEhsani/RIF-Scores-with-Alternative-Correlation-Measures. Accessed 1 Apr 2020.

Publisher's Note