*Bioinformatics*, 32, 2016, i464–i472
doi: 10.1093/bioinformatics/btw435
ECCB 2016

# Mutual enrichment in aggregated ranked lists with applications to gene expression regulation

**Dalia Cohn-Alperovich[1,2,*], Alona Rabner[3], Ilona Kifer[2], Yael Mandel-Gutfreund[3] and Zohar Yakhini[1,4]**

[1]Computer Science Department, Technion – Israel Institute of Technology, Haifa 3200003, Israel, [2]Microsoft Research and Development Center, Haifa and Herzeliya, Israel, [3]Department of Biology, Technion – Israel Institute of Technology, Haifa 3200003, Israel and [4]School of Computer Science, The Interdisciplinary Center, Herzeliya 4610101, Israel

*To whom correspondence should be addressed.

## Abstract

**Motivation:** It is often the case in biological measurement data that results are given as a ranked list of quantities—for example, differential expression (DE) of genes as inferred from microarrays or RNA-seq. Recent years brought considerable progress in statistical tools for enrichment analysis in ranked lists. Several tools are now available that allow users to break the fixed set paradigm in assessing statistical enrichment of sets of genes. Continuing with the example, these tools identify factors that may be associated with measured differential expression. A drawback of existing tools is their focus on identifying single factors associated with the observed or measured ranks, failing to address relationships between these factors. For example, a scenario in which genes targeted by multiple miRNAs play a central role in the DE signal but the effect of each single miRNA is too subtle to be detected, as shown in our results.
**Results**: We propose statistical and algorithmic approaches for selecting a sub-collection of factors that can be aggregated into one ranked list that is heuristically most associated with an input ranked list (pivot). We examine performance on simulated data and apply our approach to cancer datasets. We find small sub-collections of miRNA that are statistically associated with gene DE in several types of cancer, suggesting miRNA cooperativity in driving disease related processes. Many of our findings are consistent with known roles of miRNAs in cancer, while others suggest previously unknown roles for certain miRNAs.
**Availability and Implementation:** Code and instructions for our algorithmic framework, MULSEA, are in: https://github.com/YakhiniGroup/MULSEA.
**Contact:**dalia.cohn@gmail.com
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Biological data can often be represented as a ranked list. Examples include: genes ranked according to differential expression (DE) when comparing two conditions and using microarrays or RNA-seq (Ben-Dor *et al.*, 2001); genes ranked according to their predicted potential of being targeted by a miRNA of interest (Navon *et al.*, 2009), as provided by, say, TargetScan (Lewis *et al.*, 2005), or by experimental methods, such as CLASH (Helwak *et al.*, 2013); genomic sequences ranked according to ChIP-seq signals (Kharchenko *et al.*, 2008) or RNA sequences ranked according to CLIP (Wang *et al.*, 2015a). To gain insight into processes strongly associated with the measured data and to generate hypotheses related to driving mechanisms we often use statistical enrichment tools, building upon existing knowledge about the elements ranked. An example is the use of GO annotation in the context of genes (Ashburner *et al.*, 2000; Eden *et al.*, 2009; Gene Ontology Consortium, 2015; Subramanian *et al.*, 2005). Furthermore, ranked lists of genes can also be analyzed to assess their association to other rankings of the same genes. Such is the case, for example, when we evaluate the enrichment of the targets of the miRNA let-7 at the top of a list of DE genes. In this case, two ranked lists of genes are considered—one is the DE ranking and the second is obtained from TargetScan scores for let-7. In previous work (Leibovich and Yakhini, 2014; Steinfeld *et al.*, 2013), we described a statistical framework for this type of analysis and introduced the web-based application miTEA.

It is often the case, however, that a combination of factors can yield much stronger interpretation of the observed DE than that afforded by a single factor (Wang *et al.*, 2005). This type of combinatorial regulation is the main topic of the current paper. miRNAs regulate the expression levels of the majority of the genes in mammalians by targeting regions of the sequence and either facilitating degradation or blocking translation. A given gene is often the potential target of more than one miRNA. For example, p21, an important tumor suppressor gene in humans, is known to be targeted by several miRNAs. The role of this redundant targeting was experimentally confirmed and investigated, suggesting that it is the combination of all these activities that determines the expression of miRNA target genes (Peter, 2010; Schmitz *et al.*, 2014). Synergies of miRNA pairs were also specifically studied and several computational tools that address prediction of cooperation were suggested (Xu *et al.*, 2011). Synergies of miRNA and other regulators of biological processes give rise to interest in analyzing potential associations between sub-collections of effectors and a measured biological phenomenon.

The search for combinatorial regulation by multi-factor effectors is a difficult task, both computationally and experimentally (Wise and Bar-Joseph 2015). Here, we suggest a novel algorithm that seeks to identify combinatorial regulation, consisting of the following components. We start with a pivot ranked list and a collection of other ranked lists over the same set of elements (e.g. the pivot is a list of genes ordered by DE and the collection consists of the same genes sorted by their TargetScan scores with respect to different miRNAs). We seek a sub-collection of lists, and a model by which to aggregate them within the sub-collection, such that when combined by this model into a single list they are optimally associated to the pivot. A naïve solution—exhaustively searching over all possible sub-collections—is, of course, exponential in the number of lists considered and is generally not practical. We present efficient and effective heuristic algorithmic approaches to performing the selection task described above.

Formal definitions and descriptions are provided in Section 2. In Section 3, we describe our results in two parts. First, the results of extensive simulation studies designed to assess performance and to compare variants. In the second part, we report sets of miRNAs that form aggregates that are strongly associated with cancer DE in five different cancer types. We discuss the results as well as future directions in Section 4.

## 2 Materials and methods

### 2.1 The general algorithmic set-up

Let $\{\sigma_i | \sigma_i = (\sigma_i(1), \ldots, \sigma_i(N))\}_{i=1}^{L}$, be a collection of permutations in $S_N$, the group of permutations over N elements (that is: lists of the N elements ranked by some measured or otherwise derived property).

Let $f_{\text{agg}}$ be an aggregation function that acts on sub-collections from $\sigma_1 \ \ldots \sigma_L$. That is, $f_{\text{agg}}$ takes $Q = \{\sigma'_1, \ldots, \sigma'_k\} \subseteq \{\sigma_1, \ldots, \sigma_L\}$ and computes a permutation $f_{\text{agg}}(Q) = \pi = (\pi(1), \ldots, \pi(N))$, which is meant to jointly represent the $k$ input permutations. For example, $f_{\text{agg}}$ can compute the average rank of every element, in the members of $Q$, and then produce $\pi$ by accordingly re-sorting. Formal definitions are given in Section 2.3. (Boulesteix and Slawski, 2009) also address aggregation of ranked lists, in the context of finding a stable representation for them. This approach can also be used in our framework.

Let $f_{\text{eval}}$ be a given (statistical evaluation) function that takes two permutations, say $(\pi_1, \ \pi_2)$, and returns a value that represents the significance of their agreement with each other. Examples of such functions, which we address in this article, are Pearson correlation and mmHG (Steinfeld *et al.*, 2013).

Given a collection of permutations, $\{\sigma_i\}_{i=1}^{L}$, we seek a sub-collection $Q = \{\sigma'_1, \ldots, \sigma'_k\} \subseteq \{\sigma_1, \ldots, \sigma_L\}$, such that $f_{\text{eval}}$ $(I, \pi = f_{\text{agg}}(Q))$ is optimal, where $I$ is the identity permutation (the pivot): $I = 1 \ \ldots \ N$.

### 2.2 Statistical evaluation of the association between two ranked lists

We examine Pearson correlation and mmHG, described below, as statistical evaluation methods.

**mmHG**

Consider a universe G and let C and R be subsets of elements within G. The probability of finding exactly b elements in $C \cap R$, under a uniform distribution over all configurations, is given by the hypergeometric function:

$$HG(N, B, n, b) = \frac{\binom{n}{b}\binom{N-n}{B-b}}{\binom{N}{B}} \tag{1}$$

where N=|G|, n=|R|, and B=|C|. The tail probability of finding b or more genes in the intersection is:

$$HGT(N, B, n, b) = \sum_{i=b}^{\min(n,B)} HG(N, B, n, i). \tag{2}$$

An example for the use of HG in functional enrichment arises in the context of gene expression analysis. Genes are assigned a statistical score according to their DE in a specific experiment (DeRisi *et al.*, 1997). The foreground set of DE genes is defined by using an arbitrary threshold or using a multiple testing correction criteria (e.g. Bonferonni or 5% FDR). Using the HG statistic, the DE genes can be tested against an annotation repository, such as gene ontology (GO), and a better understanding of the underlying biology can be deduced. In many scenarios involving statistical enrichment analysis in ranked lists of genes a defined threshold is not known, and it is often not reasonable to set one arbitrarily. We can then consider all possible thresholds with respect to the given ranking – dividing the entire set of genes into subsets of high-ranking genes and low-ranking genes. We then seek a threshold that optimizes the enrichment of an annotation at the top of the list.

Formally, given measurement values M for genes (or other elements), we number the genes as $1 \ \ldots \ N$ and then obtain the permutation $\pi$ that represents ranking them according the measurement values. Namely, $M(\pi(i)) < M(\pi(i+1)) \forall \ i$. Also consider some binary labeling of the genes $\lambda \in \{0, 1\}^N$. The binary labels correspond, for example, to the membership of the genes in the curated set of genes tested: $\lambda(i) = 1$ if the gene is a member and 0 otherwise. The mHG statistic for $\lambda$'s enrichment in $\pi$ is (Eden *et al.*, 2007, 2009; Leibovich and Yakhini, 2014):

$$mHG(\lambda) = \min_{1 \le n \le N} HGT(N, B, n, b_n(\lambda)) \text{ where } b_n(\lambda) = \sum_{i=1}^{n} \lambda_{\pi(i)} \tag{3}$$

It is important to note that the testing of many thresholds introduces a multiple testing complication and thus the mHG statistic is not a $P$-value. To enable an accurate statistical interpretation an efficient dynamic programming procedure fully characterizes the distribution of mHG in a uniform null model (Eden *et al.*, 2007).

The mHG statistics addresses enrichment of binary attributes in a ranked list. There are cases where one would like to test whether the same set of genes is active in two independent experiments. In this case, we can associate two different scores to the genes obtain two different rank orders, and assess mutual enrichment in the two ranked lists.

For this purpose, the mmHG statistic (Steinfeld *et al.*, 2013) considers all possible thresholds for one of the lists to define the top of the list. For each possible threshold, we test the enrichment at the top of the other list, finally choosing the optimal pair of thresholds. Formally, consider two permutations $\pi_1, \pi_2 \in S_N$. The mmHG statistic is:

$$\mathrm{mmHG}(\pi_1,\pi_2) = \min_{1 \le n_1,n_2 \le N} \mathrm{HGT}(N,n_1,n_2,b(n_1,n_2)) \text{ where } b(n_1,n_2) = |\pi_1[n_1] \cap \pi_2[n_2]| \textit{ and } \pi_{k=1,2}[i] = \{j | \pi_k^{-1}(j) \le i\} \tag{4}$$

We note again, that the mmHG statistic is not a *P*-value, due to multiple testing. It is, however, monotone with respect to the *P*-value that would be attained if $\pi_1$ and $\pi_2$ were to be independently and uniformly drawn. Therefore, we use mmHG as an optimization criterion. In (Leibovich and Yakhini, 2014), our group describes tight bounds on actual mmHG *P*-values.

Another issue to note is that ranking might lead to many *ex aequos*. To address it, while sorting we shuffle elements with identical value to avoid bias.

## 2.3 Aggregating ranked lists

The second component in our approach to addressing our general optimization task is the aggregation method. The different aggregation approaches represent possible models by which biological factors may cooperate. Boulesteix and Slawski (2009) and Schimek *et al.* (2015) also describe possible aggregation approaches for ranked lists.

Consider a sub-collection of $k$ lists of ranked elements: $Q = \{\sigma_1{}', \ \ldots \ \sigma_k{}'\} \subseteq \{\sigma_1 \ \ldots \ \sigma_L\}$. We aggregate these $k$ lists into one permutation $\pi \in S_\mathrm{N}$, that represents them by setting the value of each element in the aggregate list to the average of its ranks in each of the k lists. Formally, for $1 \le i \le N$ let $A(i) = \mathrm{Avge}\left(\sigma_1'^{-1}(i), \ldots, \sigma_k'^{-1}(i)\right)$, and $\pi$ is obtained by sorting $A$, where $\pi(1)$ is the index $i$ at which the minimal number in ascending order. A gene is ranked high in the permutation if it has high ranking on average in all lists. Although the set of genes at the top of $\pi$ does not have to be at the very top of any of the lists in $Q$, the sum of their ranks must be high. This approach represents a biological model of combinatorial relationship between several lists with respect to the same elements, as each of the lists in $Q$ adds value to the elements at the top of $\pi$.

Two additional alternatives are described in the Supplementary Material. We evaluate all three and compare their results (see Section 2.5).

## 2.4 Sub-collection selection

The third component of our algorithm is sub-collection selection. For this task, we propose two different approaches, both of which can be combined with any variant of the previous components.

### 2.4.1 Singleton selection

Select the best lists, considered individually in terms of their association to the pivot, as described below. We calculate the association ($f_\mathrm{eval}$) of each $\sigma_i$, $i = 1 \ldots L$, with the pivot $I$. We then sort the permutations $\{\sigma_i\}_{i=1}^{L}$ accordingly. We set a cutoff by the biggest drop in score, and select all lists above the cutoff for aggregation (Fig. 1).

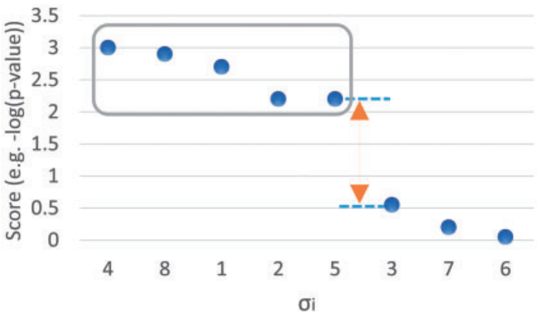

**Fig. 1.** Illustration of the Singletons selection approach

### 2.4.2 MUlti-list SElection algorithm (MULSEA)

Our main algorithmic approach takes into account relationships between lists and greedily constructs a heuristically optimal aggregate of lists.

Our algorithm accepts as input a pivot and a collection of ranked lists and produces an aggregate of a subset of these lists that is most statistically similar to the pivot. The algorithm, which we name MULSEA, also depicted in Figure 2 and in the Supplementary Material (pseudo-code), goes as follows. We start with aggregate

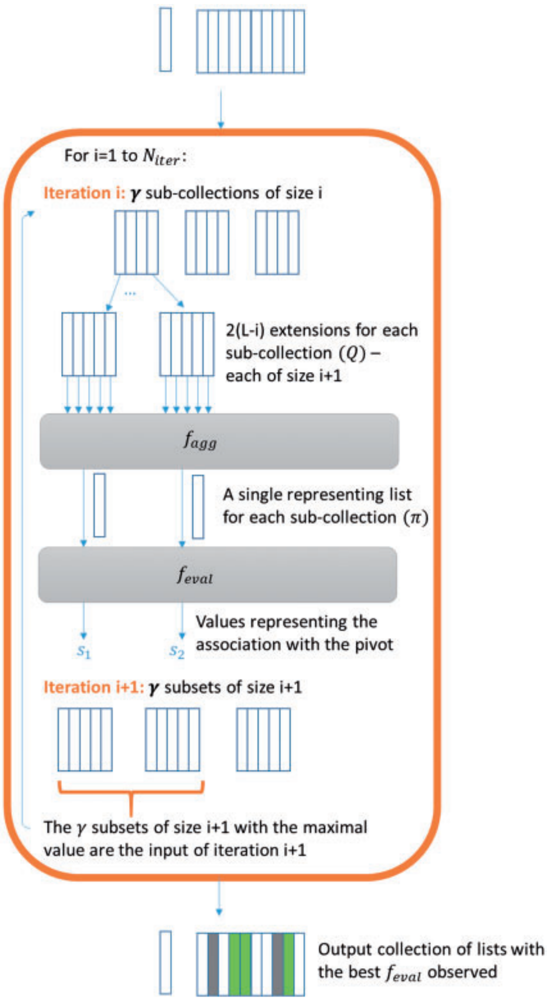
For i=1 to $N_{iter}$:
Iteration i: γ sub-collections of size i
...
2(L-i) extensions for each sub-collection (Q) – each of size i+1
$f_{agg}$
A single representing list for each sub-collection (π)
$f_{eval}$
$s_1$
$s_2$
Values representing the association with the pivot
Iteration i+1: γ subsets of size i+1
The γ subsets of size i+1 with the maximal value are the input of iteration i+1
Output collection of lists with the best $f_{eval}$ observed

**Fig. 2.** A flowchart describing the MULSEA algorithm. The output consists of the green and grey lists. The grey lists represent the reversed members of the output. The white ones are not included in the output collection

collections of size 1, where each single list comprises its own collection. If the number of effectors is L then the number of collections we start with is 2L, since every list is considered in both directions (e.g. over-expressed and under-expressed). In each iteration, a list is added to every sub-collection that 'survived' the previous round (namely, it appeared at the top-scoring $\gamma$ collections of the current iteration) and an aggregated list is computed according to the method $f_{\mathrm{agg}}$. Next, the score of the new aggregation with respect to the pivot list is computed. Here, the score is either the corrected *P*-value of the Pearson correlation or the mmHG as described in the Section 3.5.

The sub-collections that continue to the next round are the $\gamma$ (a parameter—the level of greed in this approach) top scoring sub-collections. The number of iteration is also a parameter—$N_{\mathrm{iter}}$.

The parameters that drive the algorithm and need to be set *a priori* are as follows: $f_{\mathrm{eval}}$ (variants are referred to in the paper as MULSEA-m with mmHG and Pearson-MULSEA with Pearson) described above; greed level ($\gamma$)—the number of top scoring sub-collections to be considered in the next iteration; maximal sub-collection size, or number of iterations ($N_{\mathrm{iter}}$)—the largest size of $Q$ to consider; An indicator which indicated whether to flip the lists to examine collections containing reversed lists (negatively associated).

If $f_{\mathrm{eval}}$ is mmHG then two more parameters needs to be set: max $n_1$ —the maximal index in the identity permutation to consider and max $n_2$ —the maximal index in $\pi$ to consider for mmHG.

MULSEA's running time is proportional to $\gamma \cdot N_{iter} \cdot L$.

## 2.5 Corrected *P*-value

As the number of sub-collections of size $k$ depends on $k$, the statistical evaluation of sub-collections of different sizes has to be separately corrected for multiple testing. We use a Bonferroni correction. Namely:

$$\begin{aligned} &P\begin{pmatrix}\text{one of the } \binom{L}{k} \text{ sub}-\text{collections of size } k \text{ has an}\\ \text{aggregate representation } \pi \text{ with mmHG}(I,\pi) < s\end{pmatrix}\\ &\quad\le \binom{L}{k} \cdot P(\text{the mmHG score of two independent lists} \le s)\\ &\quad\le \binom{L}{k} \cdot s \cdot (\max\, n_1) \cdot (\max\, n_2) \qquad (5)\end{aligned}$$

and similarly for the Pearson-MULSEA variant, as defined in Section 2.2.

For mmHG, we score every sub-collection Q by –log of the Stirling approximation of the above, namely:

$$\begin{aligned}\text{score}(Q) = &-\log((\text{mmHG score of } Q) \cdot (\max\, n_1) \cdot (\max\, n_2))\\ &- k \cdot (\log L - \log k). \qquad (6)\end{aligned}$$

## 2.6 Reporting robust sub-collections

The essence of our methodology, despite the Bonferroni–Stirling correction, inherently gives advantage to larger collections. This may result in infiltration of sporadic, unrelated members into the final collection. To avoid this artifact, we base our final report on filtering the set of $L_{\mathrm{top}}$ top scoring collections (a parameter; set to 20 in the implementation reported and provide her). The final collection is constructed as follows. We start with the collection that obtained the top score calculated using Equation (6) (if it is a singleton, we use the second one, etc.). A list is removed from this collection if it appears in two or more of the $L_{\mathrm{top}}$ collections with opposite direction—i.e. ascending and descending or if it appears in less than 25% of the $L_{\mathrm{top}}$ sub-collections.

The final output is reported as the filtered top scoring collection, and for the MULSEA variants—also a matrix with all examined sub-collections, ranked by their score (not relevant for the singleton selection method).

## 2.7 Simulation methodology—'fishing out' planted sub-collections of lists

In our simulations, we aim to mimic a biological scenario in which there is a causal relationship between a set of biological entities of a certain type and an entity of a different type. For example, a set of miRNAs that jointly regulate a change in an mRNA program. Our simulation starts by generating L lists, each corresponding in this example to the measured effect of a single miRNA on expression levels of all known genes, from a distribution typical of biological data. We continue by randomly selecting a sub-collection of the L lists, representing a set of miRNAs which have a causal relationship to an observed mRNA differential expression. We then use these selected miRNAs to construct a pivot list. This pivot therefore represents the combinatorial effect of the selected miRNAs on gene expression levels. Finally, we add noise to all L lists, as is typical of biological measurements.

To evaluate the performance of our approach we provide MULSEA with the pivot list and the additional (noisy) L lists, without indicating which of them were part of the pivot construction. MULSEA then attempts to identify this randomly selected sub-collection of lists (Fig. 3).

Below is a detailed outline of our procedure for generating a single instance of simulated data.

*Underlying simulated data*. L lists of $N = 1000$ elements are generated. We set the value of element i, in list j, $V_j(i)$, by randomly drawing from $e^X$ with $X \sim N(\ln(0.25), VAR)$. Several different values of VAR were studied. The log-normal distribution was selected as it represents a non-negative distribution that adequately represents many biological quantities, including gene expression. The expected value of ln(0.25) was set to fit the distribution of DE datasets.

*Pivot generation*:

Select $k = k_{\mathrm{planted}}$ lists out of the L generated above: $V_1{}', \ \ldots \ V_k{}'$.

Set the pivot values for the current instance, $V_P(i)$, based on $V_j{}'(i), \ j = 1 \ldots k$ according to each one of the following pivot

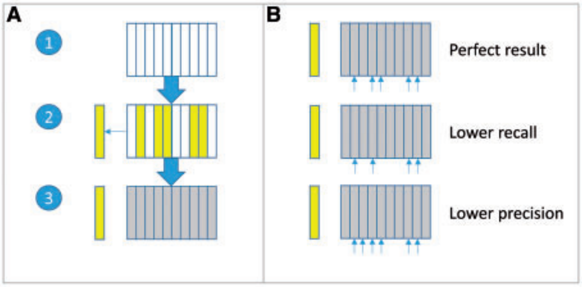

**Fig. 3.** (A) Simulations methodology. In stage (1) L lists are generated by randomly drawing values from $e^X$ with $X \sim N(\ln(0.25), VAR)$. In stage (2), k out of the L lists are randomly selected and the pivot is generated based on these. Finally in stage (3), a noise drawn from independent instances of Y ~ Norm(0, NF/1000) is added to every entry of each of the L lists, in order to "hide" the k lists. (B) Algorithm's output. The perfect result would be exactly the k lists which were used to generate the pivot (yellow in A). A missing yellow list reduces recall (how many of the yellow lists were retrieved), and an additional non-yellow list reduces precision (how many of the retrieved lists were yellow)

generation models (see more extensive biological reasoning in the corresponding $f_{\mathrm{agg}}$ models in Section 2.3 and in the Supplementary Material):

*Cumulative*. $V_P(i) = \sum_{j=1}^{k} V_j'(i)$. Models combinatorial relationship with respect to individual elements.

*Max*. $V_P(i) = \max_{j=1\ldots k} V_j'(i)$. Models redundancy.

*Min*. $V_P(i) = \min_{j=1\ldots k} V_j'(i)$. Models combinatorial relationship with respect to different elements.

Generate a permutation by sorting by $V_P(i)$, $i = 1 \ldots N$. Consider the resulting permutation as the id permutation $I$ (the pivot) for this instance.

Add noise drawn from independent instances of $Y \sim$ *N(0, NF/1000)* to each entry of each of the L lists from A.

Sort the resulting lists to obtain L permutations.

Numerous instances of simulated data were generated, with varying values of VAR (the variance of $X$ defined in A), and of NF (level of noise added to the list values, ranging from 0 to 2000). For each simulation instance, MULSEA attempts to identify the lists that were used to construct the pivot (including their number), as well as the type of aggregation approach used to create the simulated data, both of these unknown to the algorithm at the input stage. For this purpose, MULSEA is invoked several times per simulation instance, once for every aggregation approach ($f_{\mathrm{agg}}$) described above. The final output is the aggregation method and indices of lists that give the best statistical score (see Section 2.5, Equation 6).

## 2.8 Software and implementation

The algorithmic framework described herein is implemented in C#, and runs on Azure cloud environment. It is implemented with two levels of parallelization—internal and external. Internal parallelization is obtained by simultaneously evaluating $f_{\mathrm{eval}}$ for subcollections of the same size in a single iteration, and running MULSEA for different $f_{\mathrm{eval}}$ assignments in parallel. External parallelization is obtained using multiple Azure machines for runs with different set of parameters as detailed in Section 2.4. Azure allows a very high scale and a seamless integration with no necessary additional set-up. It is the natural environment for our algorithm, implemented in C#. An executable (for Windows, any CPU) and instructions are available at: https://github.com/YakhiniGroup/MULSEA. The source code is available upon request (and can run on other platforms).

A single run with a simulated $L = 20$ lists of size $N = 1000$ each takes approximately 6 min on a standard PC. A single run of the DE and miRNAs datasets as described in Section 3.2 takes approximately 1.5 h on a standard PC, but much faster in Azure (actual speed depends on configuration).

## 2.9 Micro-RNA targets

Following progress in molecular profiling and in the experimental validation of miRNA activity it is now possible to analyze data to infer interactions between different miRNAs and mechanisms of cooperation. Over the years several approaches have been developed to identify combinatorial regulation by miRNAs, for review see (Friedman *et al.*, 2013). We take a step in this direction by analyzing DE data from five different types of cancer (see below). The collection of ranked lists examined as factors was generated by sorting the genes according to Target Scan context ++ scores (Agarwal *et al.*, 2015), that represent the predicted likelihood to be targeted by the relevant miRNA. We therefore work with an input of $L = 401$ ranked lists.

## 2.10 Differential expression datasets

### 2.10.1 Micro-array based dataset

We analyzed DE data of breast invasive carcinoma (BRCA) based on patients' data from (Haakensen *et al.*, 2010), containing breast biopsies from 143 women: 79 women with no malignancy (healthy women) and 64 newly diagnosed breast cancer patients. We calculated t-test to obtain a DE value for each gene comparing cancer and healthy samples.

### 2.10.2 RNA-seq based datasets

RNAseq data from cancer patients (level 3) of four different cancer types was downloaded from TCGA GDAC Firehouse (Broad Institute TCGA Genome Data Analysis Center, 2016). For each patient paired samples were analyzed, taken from the cancer and normal tissues of the same patient. The cancer types analyzed are kidney renal clear cell carcinoma (KIRC) – (15 paired samples), lung squamous cell carcinoma (LUSC) – (17 paired samples), head and neck squamous cell carcinoma (HNSC)-(6 paired samples) and uterine corpus endometrial carcinoma (UCEC)-(2 paired samples). DE profiles were generated by DESeq2 Love *et al.*, 2014 R package. For patients' barcodes see the Supplementary Material. The gene DE data was ranked once by high expression in cancer and once by high expression in the normal samples.

## 2.11 MicroRNA expression profile

We analyzed miRNASeq data from cancer and normal tissues of 30 KIRC patients. The miRNASeq data was downloaded from TCGA GDAC Firehouse (Broad Institute TCGA Genome Data Analysis Center, 2016). The data was normalized by DESeq2 (Love *et al.*, 2014) and then log transformed. T heat-map in Figure 8 was generated for the z-score scaled data using the made4 (Culhane *et al.*, 2005) script from the R package. For patients' barcodes see the Supplementary Material.

# 3 Results

We test our approach both on simulated data as well as on real biological data. Our incentive for performing simulations is four-fold. First, extensive and complete biological data, measuring the effect of multiple biological factors (e.g. miRNAs translated into lists in our setting), on a quantitative property of a set of other biological moieties (e.g. genes) is relatively rare to come by. Second, simulations allow us to thoroughly test various scenarios (e.g. aggregation approaches and noise levels), as well as the effect of different algorithm related parameters (such as number of iterations and greed level) on our method's performance. Third, while on biological data we can speculate and provide supporting evidence to the relevance of our results, simulations provide a means of measuring actual performance levels as the ground truth is known *a priori*. And forth, MULSEA is applicable in non-biological contexts and simulated data is an unbiased test case for broader usage.

## 3.1 Evaluating algorithm performance on simulated data

We examined numerous configurations using the simulation methodology described in Section 2.7. We present the results for the cumulative model, which translates to a biological scenario in which several factors with a fairly modest contribution to some event can have a large effect when working together. Results of other models and configurations led to similar conclusions. More of the results are presented in the Supplementary Material.

### 3.1.1 MULSEA-m resistance to noise

Figure 4 presents simulation runs results for different levels of noise. Evidently, MULSEA-m (mmHG version of MULSEA) performs well, even when challenged with high values of noise. An accepted estimation of cellular noise in the context of gene expression is $CV = 0.2$ (Efron and Tibshirani, 2007), which corresponds to $\frac{50}{1000}$ ($NF = 50$) in our setting. Therefore, our results indicate that our approach is robust for reasonable NF values. MULSEA-m manages to 'fish out' the planted lists with a recall and precision of 1 for NF = 1000 with VAR = 1 and for NF = 100 with VAR= 0.5 (Fig. 4).

### 3.1.2 MULSEA-m versus Pearson-MULSEA and Pearson based singleton selection

We tested two different evaluation functions, $f_{\text{eval}}$, for scoring the association between the identity permutation $I$ and the permutation $\pi$ of the aggregated lists in a sub-collection—mmHG and Pearson correlation. For each of the two $f_{\text{eval}}$ functions we tested MULSEA, the multi-list algorithm, as well as the singletons algorithm in which the lists are not aggregated but selected according to their direct association with the pivot. Results are depicted in Figure 5. All methods detect the lists that were used to construct the pivot, as is indicated by the high recall rates. The main differences lie in the precision. While MULSEA-m is robust for high NF values, the Pearson-MULSEA approach consistently adds more irrelevant lists to the final output collection for higher values of noise. MULSEA-m's score against the pivot considers only the top of the lists, namely, the targeted list elements which provide evidence for the connection to the pivot. Pearson correlation scores the permutations based on the entire list, mostly consisting of elements which are irrelevant to the relationship to the pivot. Since the pivot values for the targeted list elements are often at the right tail of the distribution, they are less affected by additional noise as is indicated by MULSEA-m's consistently high precision. The Pearson singletons obtain poor precision values even for very low value of noise.

### 3.1.3 MULSEA-m versus mmHG singletons

We compared the two sub-collection selection approaches described in Section 2.4: singletons and multi-list. Comparison was performed using mmHG as it performed better than Pearson (see 0). MULSEA-m's recall and overall performance is significantly more robust, although its precision is slightly less robust (Fig. 6). This may be due to MULSEA's scoring function (Equation 6), which favors larger collections. In the context of hypothesis generation, it is more important not to miss key factors than to avoid redundant ones, as false results will transpire by experimental validation.

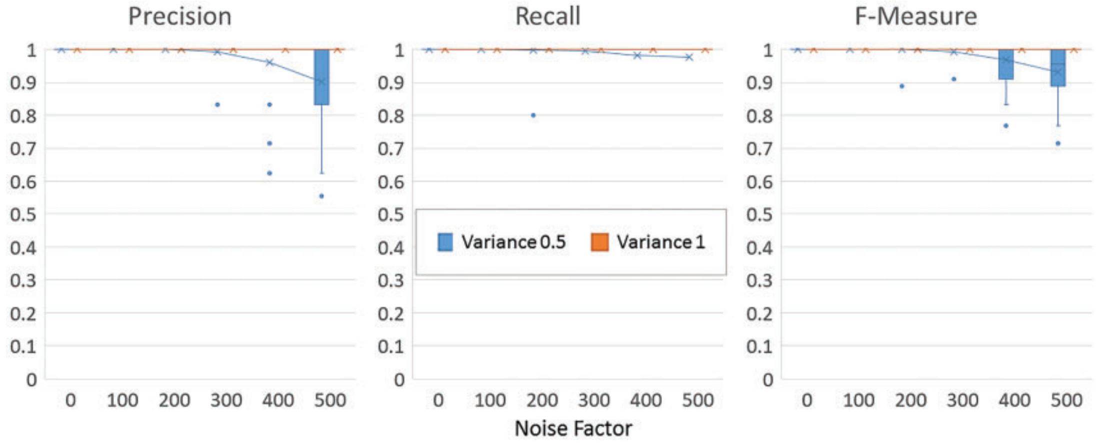

**Fig. 4.** MULSEA-m noise resistance. Performance results for 100 random instances with 20 lists total of 1000 elements each, 5 planted lists, VAR = 0.5 or 1. The pivot is generated by the cumulative model. The noise levels depicted are 0 to $\frac{500}{1000} = 0.5$

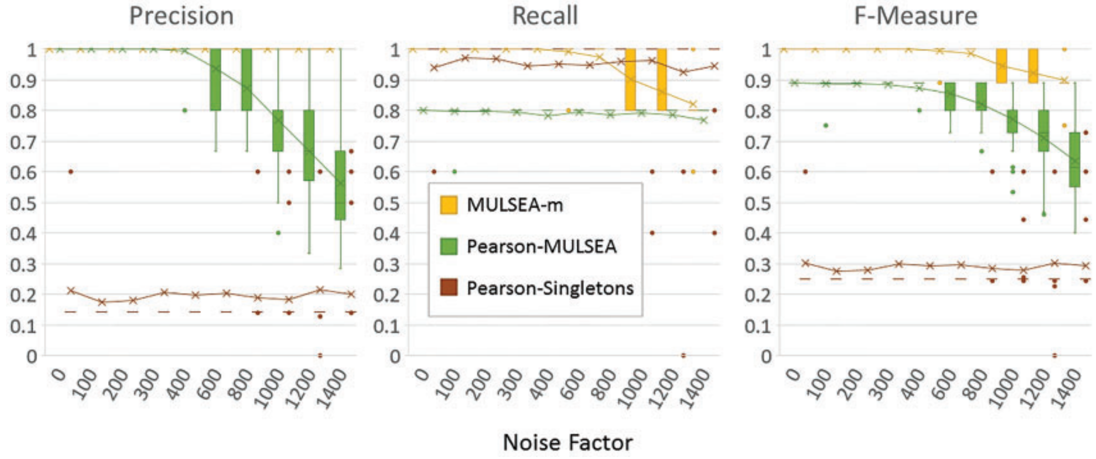

**Fig. 5.** Comparison of MULSEA-m, Pearson-MULSEA and Pearson Singletons. Performance results for 100 random instances with 20 lists total of 1000 elements each, 5 planted lists, VAR = 1. The pivot is generated by the cumulative model. Noise levels depicted: 0 to $\frac{1400}{1000} = 1.4$

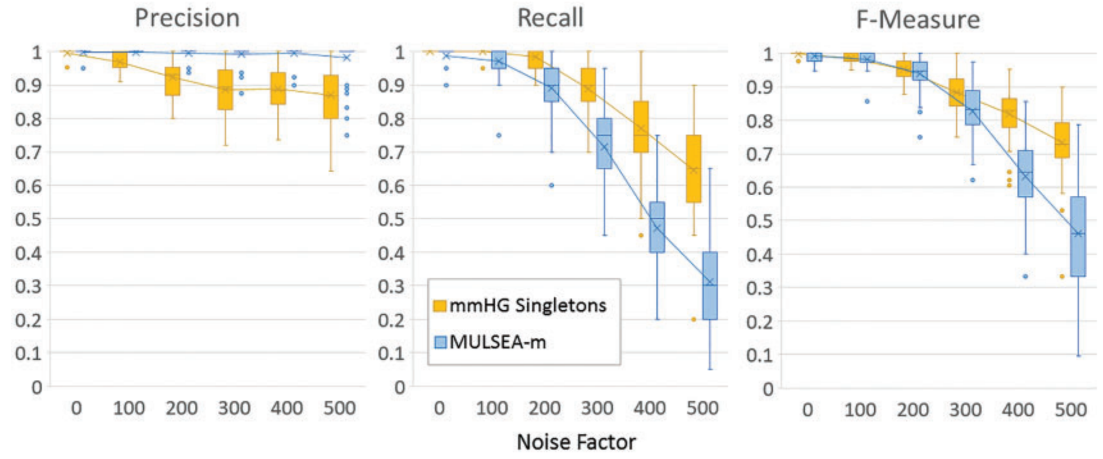

**Fig. 6.** Comparison of MULSEA-m and mmHG Singletons. Performance results for 100 random instances with 50 lists total of 1000 elements each, 20 planted lists, VAR = 0.5. The pivot is generated by the cumulative model. The noise levels depicted are 0 to $\frac{500}{1000} = 0.5$

**Table 1.** miRNA sub-collections that are jointly associated with cancer differential expression

| | -log(p-value) | max singe-list -log(p-value) | let-7d-5p | let-7f-5p | let-7g-5p | let-7i-5p | miR-100-5p | miR-10a-5p | miR-129-2-3p | miR-129-5p | miR-1298-5p | miR-133a-3p.2 | miR-133b | miR-134-5p | miR-137 | miR-1-3p | miR-142-5p | miR-144-3p | miR-151a-3p | miR-151b |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| KIRC | 16.67 | 0 | | | | | | | | | X | | | | | | | | | |
| LUSC | 11.44 | 5.54 | X | X | | | X | X | | | | | | | X | | | X | | |
| UCEC | 19.64 | 0 | | | | X | | | X | X | | | | | | X | X | | X | |
| HNSC | 13.77 | 1.97 | | | | | | | | | | | X | X | | X | | | | X |
| BRCA | 12.91 | 4.69 | | | X | | | | | | | X | | | | | | | | |

| | miR-185-5p | miR-18a-3p | miR-192-5p | miR-200a-3p | miR-208a-3p | miR-210-3p | miR-216a-5p | miR-216b-5p | miR-218-5p | miR-296-3p | miR-346 | miR-374b-5p | miR-378b | miR-382-5p | miR-410-3p | miR-450a-5p | miR-455-3p.2 | miR-532-3p | miR-543 | miR-551b-5p | miR-665 | miR-758-3p | miR-9-5p |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| KIRC | X | X | | | | | | | | | | X | X | X | | | | | X | X | X | | |
| LUSC | | | | X | | | | | | | | | | | | | | | | | | | |
| UCEC | | | | | | | X | | | X | | | | | X | | | | | | | | |
| HNSC | | | | | | | | | X | | X | | | | | X | X | | | | | | X |
| BRCA | | | X | | X | X | | X | | | | | | | | | | X | | | | X | |

Sub-collections presented per DE pivot, ranking genes over-expressed in cancer at the top. The –log(*P*-value) of the sub-collection and of the top scored single list are indicated (left). For all pivots, the top sub-collection association was attained using average aggregation.

## 3.2 Applying MULSEA to detect aggregate association between miRNA targets and gene differential expression in cancer

We applied MULSEA-m (details of the parameters can be found in the Supplementary Material) to five differential expression datasets described in Section 2.10 as pivots and miRNA targets as the factor lists (see Section 2.9). Table 1 presents the robust output sub-collections (see Section 0)—a single sub-collection per DE dataset, ranking genes by over-expression in cancer: genes that are highly over-expressed in cancer are at the top of the DE pivot. More results in the Supplementary Material.

Our first important observation from this analysis is that for all cancer DE pivots either no significant score or a much less significant score was obtained in examining single miRNAs and their target lists. Namely, the statistical effect of a collection is dramatically more significant, even after correction, than that of any single list (see score in Table 1, Supplementary Material and Fig. 7).

To improve our statistical confidence in the results we also generate 20 randomized pivots from each DE dataset, by randomly shuffling the genes appearing in the original pivot, and applying MULSEA to these pivots with the original miRNA targets as factor lists. The results are presented in Figure 7. In all datasets but one (LUSC), the statistical significance of the output collection was higher than the one observed for all randomized pivots. In 3 of the 5 datasets, BRCA, KIRC and UCEC, it is notably higher.

In general, miRNAs have suppressing effects on their targets. Therefore, for a target gene, higher expression potentially reflects lower miRNA activity and vice versa. This highly simplified thinking leads us to expect tumor suppressor miRNAs to be associated with pivots ranking high cancer over-expression at the top, and oncomiRs for the opposite pivots. The miRNAs observed in more than one cancer type, i.e. *let-7* (LUSC (Takamizawa *et al.*, 2004), UCEC (Yanokura *et al.*, 2010) and BRCA (Hu *et al.*, 2013)), *miR-133* (HNSC (Kinoshita *et al.*, 2012), BRCA (Cui *et al.*, 2013)) and *miR-216* (UCEC, BRCA (Zheng *et al.*, 2014)), are all known tumor suppressors in multiple cancer types. Many of the other miRNAs were shown to be tumor suppressors in the specific pivot cancer type (e.g. *miR-10a* (Yu *et al.*, 2015), miR-137 (Shen *et al.*, 2016) and *miR-200* (Zhen *et al.*, 2015) in LUSC, *miR-9* (Minor *et al.*, 2012) and *miR-1* (Nohata *et al.*, 2011) in HNSC). In

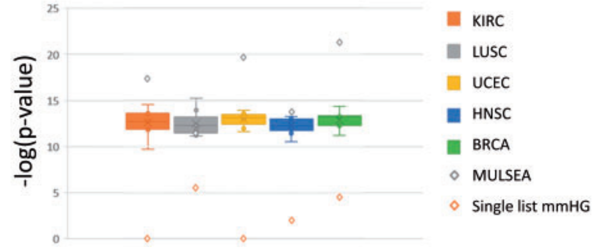

**Fig. 7.** Top –log(*P*-value) compared to randomized pivots. For each DE dataset, the genes were randomly shuffled 20 times to obtain random pivots that were used as input to MULSEA, against the original miRNA target lists. Box plots—scores of the 20 randomized pivots runs. Gray diamonds—the original pivot's score. Orange diamonds—best *P*-value attained by computing mmHG for the original DE pivot against any single miRNA list of targets

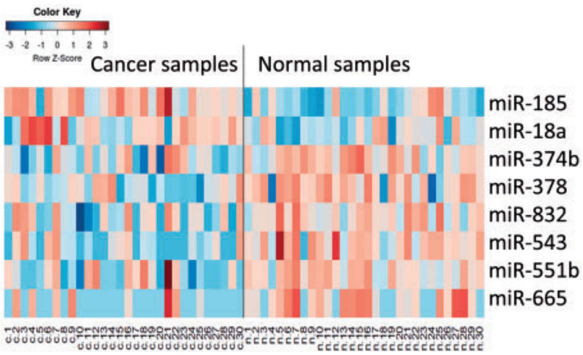

**Fig. 8.** Heat-map representing miRNA expression in tumor (c) and normal (n) samples from 30 KIRC patients. The miRNAs form the sub-collection corresponding to the KIRC pivot. Only miRNAs with miRNAseq data in the TGCA database are shown

examining the genes at the top of the mmHG evaluation of the pivot and the aggregated list in BRCA, for example, we found expected genes such as BRCA1. We also found NIPBL, which is known to be regulated by *miR-208*, and MAD2L1 and TFRC which are strong targets of *miR-758*. These genes were found by several studies to be overexpressed in breast cancer (Majidzadeh-A *et al.*, 2011; Wang *et al.*, 2015). Both *miR-208* and *miR-758* were found in our BRCA sub-collection. They were not, to the best of our knowledge, previously shown to be associated with breast cancer.

## 4 Discussion

We presented several alternative approaches to selecting a sub-collection of factors that are jointly associated with a pivot ranked list of interest. Our approach differs from other approaches to sub-collection selection, such as clustering and co-expression, in many aspects. The most important one is the fact that we center our analysis around the pivot—a list that defines the association that we are looking for. It also differs in our use of mmHG statistical assessment to identify mutual enrichment—an approach that focuses significant correlation only in part of the set of ranked elements (for example, only for genes that are highly over-expressed). MULSEA provides multiple alternatives as an output. The latter is also an advantage of MULSEA compared to the singletons algorithm presented in this article. The use of ranking is also a key feature of MULSEA. Ranking is more robust and can tolerate a much higher rate (see Pearson and mmHG comparison in 0). Moreover, it allows comparison of values from completely different scales and distribution with no normalization required, as TargetScan's score ++ and differential expression measurements.

Our simulations study, while spanning several parameter and generation model configurations, can only address the tested cases. It may be useful to better understand the data characteristics that might influence the expected results and the selection of algorithm components to be applied to given input data. We intend to invest more work in such characterization, including the consideration of models derived from other potential application domains. Also, we intend to explore more accurate statistical correction. The current approach tends towards larger collections and sometimes includes unrelated factors in the top-scoring collection. A more accurate multiple testing correction will likely further improve upon our method's precision. We also presented an analysis of several cancer datasets. We further analyzed miRNA expression in patients for some of the cohorts analyzed to produce Table 1.

Figure 8 depicts the expression heat-map for the miRNA sub-collection found by MULSEA using the KIRC pivot. MULSEA results are addressing the potential interaction of the miRNAs in the sub-collection and not necessarily direct interactions between them. We therefore do not expect MULSEA results to tightly reflect the miRNA expression levels. However, these profiles are interesting to investigate in this context. For 6 out of the 8 miRNAs analyzed, we observe an activity that is aligned with this paradigm—a tumor-suppressor-like activity. The remaining two are not aligned with the direction of their target DE. *miR-185* was previously shown to be a tumor suppressor in renal cancer (Imam *et al.*, 2010). *miR-18a*, specifically, is a known oncomiR (Komatsu *et al.*, 2014), therefore making the explanation of the contribution of its target DE to the MULSEA output an even more interesting question. DIRAS2 is highly overexpressed in the KIRC dataset, ranked as number 9 in the list. It is also likely targeted by miR18 but not by any of the other miRNAs include in the output sub-collection. High expression of *miR-18a* is not likely to drive high DIRAS2 expression but this observation should be taken into account in jointly analyzing this miRNA and mRNA datasets.

A specific use case scenario for the task of selecting a sub-collection of factors is as a first step in a hypothesis generation and testing process. We seek a hypothesis that maximizes chance of success. Under given constraints on the number of factors that can be manipulated together, we will select to study the aggregate that is found to be most associated or anti-associated with the phenomenon of interest, represented by the pivot. One notable observation, from the simulation study above, is that the simple singleton approach provides reasonable performance in certain ranges. In the context of generating the most potentially successful aggregate for further testing, it is more important to opt for better recall and investing the computing time required by MULSEA is likely to be the right approach.

By applying the methods described herein we hope to investigate the role of non-coding RNA in other contexts, including better understanding co-operation in gene regulation. Aggregation models to better capture such interactions are a subject for further development.

## Acknowledgements

We thank Israel Steinfeld, from Agilent Laboratories, for useful comments and Shlomi Dvir for advice and help in the DE analysis. We thank the award program Microsoft Azure for Research for providing access to Azure resources.

*Conflict of Interest*: none declared.

## References

Agarwal,V. *et al.* (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, e05005.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Broad Institute TCGA Genome Data Analysis Center (2016). Analysis-ready standardized TCGA data from Broad GDAC Firehose stddata__2015_11_01 run.

Ben-Dor,A. *et al.* (2001) Class discovery in gene expression data. In: *Proceedings of RECOMB*, ACM Press , New York, pp. 31–38.

Boulesteix,A.L. and Slawski,M. (2009) Stability and aggregation of ranked gene lists. *Brief. Bioinform.*, **10**, 556–568.

Culhane,A.C. *et al.* (2005) MADE4: an R package for multivariate analysis of gene expression data. *Bioinformatics*, **21**, 2789–2790.

Cui,W. *et al.* (2013) microRNA-133a regulates the cell cycle and proliferation of breast cancer cells by targeting epidermal growth factor receptor through the EGFR/Akt signaling pathway. *FEBS J.*, **280**, 3962–3974.

DeRisi,J.L. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Eden,E. *et al.* (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol.*, **3**, e39.

Eden,E. *et al.* (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.

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

Friedman,Y. *et al.* (2013) Working together: combinatorial regulation by microRNAs. *Adv. Exp. Med. Biol.*, **774**, 317–337.

Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.

Haakensen,V. *et al.* (2010) Expression levels of uridine 5'-diphospho-glucuronosyltransferase genes in breast tissue from healthy women are associated with mammographic density. *Breast Cancer Res.*, **12**, R65.

Helwak,A. *et al.* (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, **153**, 654–665.

Hu,X. *et al.* (2013) The heterochronic microRNA let-7 inhibits cell motility by regulating the genes in the actin cytoskeleton pathway in breast cancer. *Mol. Cancer Res.*, **11**, 240–250.

Imam,J.S. *et al.* (2010) MicroRNA-185 suppresses tumor growth and progression by targeting the Six1 oncogene in human cancers. *Oncogene*, **29**, 4971–4979.

Kinoshita,T. *et al.* (2012) Tumor suppressive microRNA-133a regulates novel targets: moesin contributes to cancer cell proliferation and invasion in head and neck squamous cell carcinoma. *Biochem. Biophys. Res. Commun.*, **418**, 378–383.

Komatsu,S. *et al.* (2014) Circulating miR-18a: a sensitive cancer screening biomarker in human cancer. *In Vivo*, **28**, 293–297.

Leibovich,L. and Yakhini,Z. (2014) Mutual enrichment in ranked lists and the statistical assessment of position weight matrix motifs. *Algorithms Mol. Biol.*, **9**, 11.

Lewis,B.P. *et al.* (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Gen Bio*, **15**, 550.

Majidzadeh-A,K. *et al.* (2011) TFRC and ACTB as the best reference genes to quantify urokinase plasminogen activator in breast cancer. *BMC Res. Notes*, **4**, 215.

Minor,J. *et al.* (2012) Methylation of microRNA-9 is a specific and sensitive biomarker for oral and oropharyngeal squamous cell carcinomas. *Oral Oncol.*, **48**, 73–78.

Navon,R. *et al.* (2009) Novel rank-based statistical methods reveal microRNAs with differential expression in multiple cancer types. *PLoS One*, **4**, e8003.

Nohata,N. *et al.* (2011) miR-1 as a tumor suppressive microRNA targeting TAGLN2 in head and neck squamous cell carcinoma. *Oncotarget*, **2**, 29–42.

Peter,M.E. (2010) Targeting of mRNAs by multiple miRNAs: the next step. *Oncogene*, **29**, 2161–2164.

Kharchenko,P.V. *et al.* (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, **26**, 1351–1359.

Schimek,M.G. *et al.* (2015) TopKLists: a comprehensive R package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists. *Stat. Appl. Genet. Mol. Biol.*, **14**, 311–316.

Schmitz,U. *et al.* (2014) Cooperative gene regulation by microRNA pairs and their identification using a computational workflow. *Nucleic Acids Res.*, **42**, 7539–7552.

Shen,H. *et al.* (2016) MicroRNA-137 inhibits tumor growth and sensitizes chemosensitivity to paclitaxel and cisplatin in lung cancer. *Oncotarget*, **7**, 20728–20742.

Steinfeld,I. *et al.* (2013) miRNA target enrichment analysis reveals directly active miRNAs in health and disease. *Nucleic Acids Res.*, **41**, e45.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Takamizawa,J. *et al.* (2004) Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res*, **64**, 3753–3756.

Wang,T. *et al.* (2015a) Design and bioinformatics analysis of genome-wide CLIP experiments. *Nucleic Acids Res.*, **43**, 5263–5274.

Wang,W. *et al.* (2005) Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. *Proc. Natl. Acad. Sci. USA*, **102**, 1998–2003.

Wang,Z. *et al.* (2015) Biological and clinical significance of MAD2L1 and BUB1, genes frequently appearing in expression signatures for breast cancer prognosis. *PLoS One*, **10**, e0136246.

Wise,A. and Bar-Joseph,Z. (2015) cDREM: inferring dynamic combinatorial gene regulation. *J. Comput. Biol.*, **22**, 324–333.

Xu,J. *et al.* (2011) MiRNA-miRNA synergistic network: construction via coregulating functional modules and disease miRNA topological features. *Nucleic Acids Res.*, **39**, 825–836.

Yanokura,M. *et al.* (2010) MicroRNA and endometrial cancer: Roles of small RNAs in human tumors and clinical applications [Review]. *Oncol. Lett.*, **1**, 935–940.

Yu,T. *et al.* (2015) MiRNA-10a is upregulated in NSCLC and may promote cancer by targeting PTEN. *Oncotarget*, **6**, 30239–30250.

Zhen,Q. *et al.* (2015) MicroRNA-200a targets EGFR and c-Met to inhibit migration, invasion, and gefitinib resistance in non-small cell lung cancer. *Cytogenet. Genome Res.*, **146**, 1–8.

Zheng,L. *et al.* (2014) Regulation of the P2X7R by microRNA-216b in human breast cancer. *Biochem. Biophys. Res. Commun.*, **452**, 197–204.