Data Analysis in Studies Combining
Multiple High-Throughput
Measurement Technologies

Israel Steinfeld
Data Analysis in Studies Combining Multiple High-Throughput Measurement Technologies

Research Thesis

In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Israel Steinfeld

Submitted to the Senate of the Technion - Israel Institute of Technology

Av 5773, Haifa, August 2013
The research thesis was done under the supervision of Dr. Zohar Yakhini in the Computer Science department.

The generous financial help of the Technion- Israel Institute of Technology and of an Agilent Technologies University Relations Fellowship are gratefully acknowledged.
# Table of content

Abstract .............................................................................................................................................. 1

1 Introduction .................................................................................................................................... 4
  1.1 Breast cancer .......................................................................................................................... 4
  1.2 microRNAs ............................................................................................................................. 4
  1.3 Functional enrichment .......................................................................................................... 5
    1.3.1 Hypergeometric ($HG$) .................................................................................................... 6
    1.3.2 Minimal-hypergeometric ($mHG$) .................................................................................. 7
    1.3.3 Rank-sum ($RS$) test ....................................................................................................... 9
  1.4 Data analysis of multidimensional datasets ........................................................................ 11
    1.4.1 3HDUVRQ¶VFRUUHODWLRQ ......................................................................................... 12
    1.4.2 6SHDUPDQ¶VUDQNFRUUHODWLRQ ............................................................................ 12

2 Methods ..................................................................................................................................... 13
  2.1 Mutual enrichment in ranked lists ...................................................................................... 13
    2.1.1 Minimal-$mHG$ ($mmHG$ or $m^2HG$) ........................................................................... 14
    2.1.2 Quantized-$mHG$ ($QmHG$) ........................................................................................ 17
    2.1.3 Minimal-rank-sum ($mRS$) .......................................................................................... 20
  2.2 The integrative approach ...................................................................................................... 21
    2.2.1 Enrichment networks ..................................................................................................... 21
    2.2.2 The $mmHG$ extension .................................................................................................. 25
    2.2.3 The empirical testing correction .................................................................................... 25

3 Results ..................................................................................................................................... 28
  3.1 miRNA-mRNA Integrated Analysis Reveals Roles for miRNAs in Primary Breast Tumors ....28
  3.2 miRNA Target Enrichment Analysis Reveals Directly Active miRNAs in Health and Disease .42
  3.3 Identifying In-Trans Process Associated Genes in Breast Cancer by Integrated Analysis of Copy Number and Expression Data Integrated ........................................................................ 57

4 Discussion .................................................................................................................................. 74
  4.1 Summary of studies discussions .......................................................................................... 75
    4.1.1 miRNAs role in breast cancer [42]. .............................................................................. 75
4.1.2 Detection of miRNA activity [41] ................................................................. 77
4.1.3 Identifying driver genes in breast cancer [46] .................................................. 79

4.2 Future work ........................................................................................................ 81
4.2.1 Cooperative and complimentary miRNAs target enrichment analysis (CoCo-miTEA) ........ 82
4.2.2 Complexity of the mmHG and mRS scores computation .................................... 83
4.2.3 Identifying alternative mechanism for cancer promotion .................................... 84
4.2.4 Enrichment network visualization ..................................................................... 85

5 Appendix – Partial list of publications (during my PhD) ............................................. 87
5.1 GOrilla: a Tool for Discovery and Visualization of Enriched GO Terms in Ranked Gene Lists [58] .................................................................................................................. 87
5.2 Novel Rank-Based Statistical Methods Reveal MicroRNAs with Differential Expression in Multiple Cancer Types [34] ........................................................................................................ 87
5.3 Developmental Programming of CpG Island Methylation Profiles in the Human Genome [59] 88
5.4 EGF Decreases the Abundance of MicroRNAs That Restrain Oncogenic Transcription Factors [60] .................................................................................................................. 88

6 References ............................................................................................................. 90
List of Figures

Figure 1. The two-dimensional grid used for calculating the $mHG$ p-value. In this example $N = 30$, $B = 10$, $W = 20$ and $p = 0.1$. Light blue area describes all attainable values of $w$ and $b$. Red area describes the subset $R$: all values of $w$ and $b$ for which $HGT(N, B, n, b) \leq p$, where $n = w + b$. Two $(0, 0) \rightarrow (N,B)$ paths are depicted, representing the binary label vectors $\lambda_1 = \{1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1\}$ and $\lambda_2 = \{0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1\}$. The path $\lambda_1$ traverses $R$, demonstrating that $mHG(\lambda_1) \leq p$. The path $\lambda_2$ does not traverse $R$, demonstrating that $mHG(\lambda_2) > p$.

Taken from Eran Eden’s thesis (2007).

Figure 2. Generalization of the hypergeometric ($HG$) approach. The $HG$ approach uses two sets and estimates the significance of their intersection size. The minimal-$HG$ ($mHG$) approach uses a fixed set definition, depicted by the red circle, and generalizes the definition of the second set, depicted by the blue circle. $mHG$ statistics is computed by searching through all possible expansions of the blue set to get optimal enrichment. In the minimal-$mHG$ ($mmHG$ or $m^2HG$) we further generalize the definition of the red circle to attain all possible intersections and choosing the red circle that results in optimal enrichment of the blue one.

Figure 3. Schematic overview of the integrated analysis pipeline. Genes were ranked according to one of three options: (A) Differential expression. (B) Correlation to a pivot miRNA – for a specific miRNA, called the pivot miRNA, mRNAs were ranked according to the correlation of their expression pattern across the entire cohort (101 samples) to the pivot miRNA expression pattern. (C) miRNA target prediction – for a specific miRNA, genes were ranked according to how likely they are to be targeted by the miRNA.

Figure 4. Integrated analysis of matched high-throughput datasets. The input for this approach is the Primary Data matrix for which we have the binary Annotation Matrix and a Pivot Data Matrix that is sampled matched to the Primary Data Matrix. The aim of this approach is to uncover significant association between pivot elements (rows from the Pivot Data Matrix) and annotations (columns from the Annotation Matrix). We first generate the Correlations Matrix that holds the
correlation level between each pivot and primary element. We then for each pivot element rank the Annotation Matrix according to the pivot’s correlation vector taken from the Correlations Matrix. This step results in a collection of ranked binary vectors to which we apply the $mHG$ statistics to calculate the enrichment levels vector for each annotation (column) that will constitute a row in the $mHG$ Enrichment Matrix.

Figure 5. Significant association between miRNAs and GO terms derived from enrichment analysis in correlated/anti-correlated genes (see Figure 4). GO-terms and miRNAs are represented by rectangles and rhombi, respectively. In addition, positive and negative associations are depicted by red and green arrows, respectively. Positive/negative associations here refer to enrichment of the term in the list of genes ranked according to correlations/anti-correlations, respectively, to the pivot miRNA. Arrangement of the association network is according to cancer related biological modules: a) cell-cycle, b) glycolysis, c) metabolic process, d,e,f) development and morphogenesis, g) transport, h) DNA structure, i) cell-adhesion and j) immune response.

Figure 6. The expression profiling of HeLa cells following transfection of miR-16. Enrichment plots are depicted for different sets of miRNAs. Y-axis represent $-\log$ transformation of enrichment p-value. miR-15b (of the miR-16 family) has optimal enrichment for mono miRNA set. As can be seen this enrichment is improved more than 20 orders of magnitude in significance when considering cooperation with more miRNAs.

Figure 7. Enrichment network visualization. Visualization of enrichment network (See section 2.2.1) by ENViZ, a freely accessible Cytoscape plug-in to be released in the near future. The enrichment network built from mRNA and miRNA data from Enerly & Steinfeld et al., using WikiPathway annotation is depicted. Results are represented as a bi-partite graph: yellow to red nodes represent pathways and gray nodes represent miRNAs. Edges represent significant associations between pathways and a miRNAs where red edges represent positive correlation and blue edges represent negative correlation.
List of Tables

Table 1. QmHG statistic. The ranking of a set of elements [1,2,…,10] according to the label ranking and test ranking is depicted by columns A and P, respectively. Column P* corresponds to the ranks of the elements in A after sorting them according their ranks in P. In qP we quantize P* with 3 quanta, where ranks<3 are assigned 0, ranks >6 are assigned 2 and the rest are assigned 1. qP(1) is a binary representation of qP when only the top quantum is considered and similarly qP(2) is when only the top two quanta are considered. For each binary vector we can run the mHG and take the best score overall.
Abstract

In the studies presented here, we demonstrate the use of rigorous statistics and efficient algorithmics to analyze integrated high-throughput molecular measurement data by translating results into ranked lists and by using flexible approaches to statistically assess properties of the latter.

Over recent years, modern biology has undergone an information revolution, which is evident in a shift of thinking and practice. While typical biological studies are focused on specific pathways, like the p53 signaling pathway, the emergence of novel high-throughput technologies now enables the quantification of biological features in a genome-wide scale. The rapid development in array technology, in particular, enabled its utilization in the measurements of mRNA expression levels [1], miRNA expression levels [2, 3], DNA methylation state [4], DNA copy number aberration in cancer [5], etc. With the recent revolution in second generation sequencing technologies (reviewed in [6]) the accuracy and scope of the different high-throughput applications is constantly being improved.

The naïve approach to analyzing such rich high-throughput data is to separately cluster the samples and genes [7], usually using hierarchical clustering, and to try to characterize each cluster, commonly with the help of gene annotation repositories (e.g. Gene Ontology [8]). With an exhaustive expert examination, this methodology can yield meaningful biological results, though intricate responses are difficult to uncover. Various methodologies have been developed, in recent years, to handle integrated analysis of functional genomics data, mainly by studying the transcriptional programs and global organization of biological processes [9–11]. Still, only a few studies report the joint analysis of sample cohorts that include multiple genomic measurements [12–15].

In my thesis we explore the notion of gene set enrichment that is broadly used in analyzing genomic high-throughput results. We generalize enrichment statistical approaches to the current needs of molecular genomics, introducing the notion of mutual enrichment. In addition, we study the joint analysis of integrated high-throughput data in the context of matched samples. In such data, each sample (e.g. cell-line, subject or patient) is associated with multiple genome-wide profiles and other high-throughput information. Applying our newly developed methodologies to measurement data, we are
able to characterize biological response at the system level, capturing not only the dominant primary responses but also finer and less-easily tractable processes.
List of symbols and abbreviations

DNA – Deoxyribonucleic acid.
RNA – Ribonucleic acid.
mRNA – Messenger RNA
miRNA – microRNA
HG – Hypergeometric
HGT – Hypergeometric tail
mHG – minimal-HG
mmHG – minimal-mmHG
RS – Rank sum
mRS – minimal-RS
RCoS – Rank consistency score
miTEA – miRNA target enrichment analysis
GO – Gene ontology
iPAC – in-trans process associated and cis-correlated
1 Introduction

In this thesis we explore two main computational and statistical approaches, one is statistics in ranked lists and the other is integrated analysis of matched high-throughput measurement data (henceforth referred to as integromics). The studies presented as part of this thesis combine different aspects of these computational approaches. The biological motivation behind these studies is to understand microRNA (miRNA) role in breast cancer, to study miRNA activity in disease and other processes and to identify driver genes in breast cancer.

My introduction starts with a brief background related to cancer, breast cancer and miRNA. The main parts of my introduction provide background for the notion of gene set enrichment in ranked lists, and for the joint analysis of integrated high-throughput data in the context of matched samples.

1.1 Breast cancer

Cancer disease is a common ailment in most multi-cellular organisms in which cell growth and division lose their regulation and form the cancer tumor. In the case of the malignant tumor the uncontrollable division and growth can then invade nearby parts of the organism. Among more than 200 different known cancer types in humans, breast cancer is a type of cancer that originates in the breast tissue and is the second most common cancer (after lung cancer) and a leading cause of death in woman [16].

A range of gene signatures that predicts pathway activation, has been identified in various cancer types (reviewed in [17]). In breast cancer, mRNA profiling has been used to classify breast tumors and associate them with clinical and pathological characteristics as well as with prediction of outcome [18–21]. In particular, luminal-A and basal-like subtypes, defined using an intrinsic gene list, have distinct and reciprocal gene expression profiles as well as large differences in clinical characteristics, including survival [21–23]

1.2 MicroRNAs

MicroRNAs (miRNAs) are short (usually ~22 nt) endogenous non-coding regulatory RNA molecules. miRNAs target gene transcripts by binding to complimentary sequences mainly in the 3’ un-translated region of the transcript and by that either inhibit its translation or triggers mRNA degradation [24]. Regulation of gene
expression through mechanisms that involve miRNAs has attracted much attention during recent years. There are to date more than 900 identified human miRNAs [25], transcribed as individual units, polycistronic clusters or in concert with a protein coding host gene [26]. Many miRNAs regulate genes associated with different biological processes such as development, proliferation, apoptosis, stress response, and tumorigenesis [27–31].

Abnormal expression levels of several miRNAs have previously been shown to be associated with multiple cancer types including breast cancer [2, 32–34]. Some miRNAs correlate with specific clinical features of breast cancer, such as the levels of expression of estrogen and progesterone receptor, tumor stage, vascular invasion, and proliferation index [35–37]. In accordance with their observed abnormal expression, miRNAs were shown to have a direct role in cancer both as oncogenes and as tumor suppressors. In breast cancer, miRNAs were shown to directly regulate cancer related process (e.g. miR-10b and cell migration and invasion) as well as to have potential as prognostic biomarkers [38]. It is thus of interest to explore the power of integromics in the study of the role of miRNAs in cancer and in breast cancer in particular.

In the rest of the introduction section we provide background on literature and on methodology The introduction material is organized according to the main topics of this thesis: mutual enrichment in ranked lists and integromics.

1.3 Functional enrichment

Results of functional genomics studies can usually be represented as a ranked list of genes, as each gene is given a particular score that can serve as a basis to the ranking (e.g. differential expression, ChIP-chip experiment, sequence similarity, etc.). A common way to interpret such functional genomics results is to use curated sets of genes and test which of them can be found enriched in the top of the measurement derived ranked list. The curated sets of genes represent various biological processes and enrichment of one such set can help us understand the biological context in our original experiment. Currently there are large repositories of gene sets [8, 39], which makes this process very productive
and also attest to the fact that the enrichment approach is a common one for the community.

### 1.3.1 Hypergeometric (HG)

A typical approach used in the literature to assess the enrichment of a curated set of genes in the top of a ranked list of genes resulting from a high-throughput measurement assay (e.g. differential expression) is to set an arbitrary threshold on the ranked list to define the top genes inferred from the experiment (e.g. top 10% or 2-fold differential expression). Thereby, we get two sets of genes, the curated gene set and the results gene set. A common statistical tool for assessing mutual enrichment in the two sets is using the hypergeometric distribution (HG). Formally, let $G$ be the gene universe, usually representing the set of genes in the experiment (e.g. genes represented on the array); let $C$ be the curated subset of genes (a set of genes with a specific biological relevance), and let $R$ be the results gene subset (the set of top scored genes in our experiment). The probability of finding exactly $b$ genes in both $C$ and $R$, under a uniform distribution over all configurations of $R$, is described by the hypergeometric function:

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

(1)

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

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

(2)

As pointed out above, a common example for the use of $HG$ in functional enrichment is in gene expression analysis. Genes are given a statistical score according to their differential expression in a specific experiment. The results set of differentially expressed genes is defined by using an arbitrary threshold (e.g. top 10% or 2-fold differential expression) or using a multiple testing correction criteria (e.g. Bonferroni or 5% FDR). Using the $HG$ statistics the differentially expressed genes can be assessed
against an annotation repository, like Gene Ontology (GO [8]), and the underlying biology can be better understood.

1.3.2 Minimal-hypergeometric (mHG)

In many scenarios a fixed threshold on the ranked list of the genes is not known or is arbitrary in nature. If some ranking of the elements is given then we can consider all possible thresholds that respect the given ranking – dividing the entire set of genes into a subset of high-ranking genes and low ranking genes. We then seek a threshold that optimizes the enrichment of a given curated gene set. Formally, consider a ranked list of genes \( g_1, \ldots, g_N \) where the ranking was induced by some score function \( S \), so that \( S(g_i) > S(g_{i+1}) \), in addition consider some binary labeling of the elements \( \lambda \in \{0,1\}^N \), where \( N \) is the number of genes. The binary labels correspond to the membership of the genes in the curated set of genes tested: 1 if the gene is a member and 0 otherwise. We define the \( mHG \) score as:

\[
mHG(\lambda) = \min_{1 \leq n \leq N} \text{HGT}(N, B, n, b_n(\lambda))
\]

where,

\[
b_n(\lambda) = \sum_{i=1}^{n} \lambda_i
\]

In addition to the \( mHG \) score itself, it is also useful to note the rank \( n^* \), at which the minimal HGT was attained, as this threshold might have biological relevance.

1.3.2.1 mHG p-values

It is important to note that the multiple thresholds testing introduces a multiple testing complication and thus the \( mHG \) score cannot be considered as significance level (p-value). To enable an accurate interpretation of the \( mHG \) score significance an efficient dynamic programming procedure is available to fully characterize the distribution [40].

For completeness, a short description of the mHG p-value calculation follows. Let \( W \) and \( B \) be the number of 0’s and 1’s in the binary vector, respectively. Consider the space of all binary label vectors of size \( N = W + B \) with \( B \) 1’s and \( W \) 0’s: \( \Lambda = \{0,1\}^{(W,B)} \). Assume that we are given a vector \( \lambda_0 \in \Lambda \), for which we calculate the \( mHG \) score \( mHG(\lambda_0) = p \). We would like to determine p-value(p) by means of path counting. The
space of all label vectors $\lambda$ is represented as a two-dimensional grid ranging from (0,0) at the bottom-left to $(W,B)$ at the top-right. Each specific label vector $\lambda \in \Lambda$ is represented by a path $(0,0) \rightarrow (W,B)$ composed of $N$ distinct steps. The $i^{th}$ step in the path describing a vector $\lambda$ is $(1,0)$ if $\lambda_i = 0$ and $(0,1)$ if $\lambda_i = 1$ (see Figure 1). Each point $(w,b)$ on the grid corresponds to a threshold (on the ranked binary vector) $n = w + b$, and the respective value $b = b_n(\lambda)$. It can therefore be associated with a specific HGT score: $HGT_n(\lambda) = HGT(N,B,n,b_n(\lambda))$. A subset of the points on the grid can be characterized as those points $(w,b)$ for which $HGT(N,B,n,b) \leq p$. We denote this subset $R = R(p)$ (see Figure 1).

**Figure 1.** The two-dimensional grid used for calculating the mHG p-value. In this example $N = 30, B = 10, W = 20$ and $p = 0.1$. Light blue area describes all attainable values of $w$ and $b$. Red area describes the subset $R$: all values of $w$ and $b$ for which $HGT(N,B,n,b) \leq p$, where $n = w + b$. Two $(0,0) \rightarrow (N,B)$ paths are depicted, representing the binary label vectors $\lambda_1 = \{1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$ and $\lambda_2 = \{0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$. The path $\lambda_1$ traverses $R$, demonstrating that $mHG(\lambda_1) \leq p$. The path $\lambda_2$ does not traverse $R$, demonstrating that $mHG(\lambda_2) > p$. Taken from Eran Eden’s thesis (2007).

The $(0,0) \rightarrow (W,B)$ path represents $\lambda$ passing through $N$ distinct grid points (excluding the point $(0,0)$), which correspond to $N$ different HGT scores that are considered when calculating its mHG score, as defined in formula (3). $mHG(\lambda) \leq p$ iff the path representing $\lambda$ visits $R$. Denote by $\prod(w,b)$ the total number of paths $(0,0) \rightarrow (w,b)$ by $\prod_R(w,b)$ the number of paths $(0,0) \rightarrow (w,b)$ visiting $R$ and by $\prod_{\neg R}(w,b)$ the number of paths $(0,0) \rightarrow (w,b)$ not visiting $R$. We then have:
p-value($p$) = $\frac{\left|\{\lambda \in \Lambda: mHG(\lambda) \leq p\}\right|}{\left|\Lambda\right|}$

$$= \frac{\prod_{R}(W, B)}{\prod(W, B)} = \frac{\prod(0, 0) - \prod(0, 1)}{\prod(0, B)} = 1 - \frac{\prod(0, B)}{\prod(0, 1)}$$

(5)

We calculate $\prod_{R}(w, b)$ by means of dynamic programming. Initially, set $\prod_{R}(0, 0) = 1$, $\prod_{R}(-1, b) = 0$ for $0 \leq b \leq B$ and $\prod_{R}(w, -1) = 0$ for $0 \leq w \leq W$. Then, for each $0 \leq w \leq W$, and $0 \leq b \leq B$ calculate $\prod_{R}(w, b)$ using the formula:

$$\prod_{R}(w, b) = \begin{cases} 0 & (w, b) \in R \\ \prod_{R}(w - 1, b) + \prod_{R}(w, b - 1) & \text{o/w} \end{cases}$$

(6)

In summary, to compute the p-value of an mHG score $p$ we first calculate $\prod_{R}(W, B)$. Since the total number of possible paths is $\prod(W, B) = \binom{W + B}{B}$ the p-value($p$) may be directly computed from (4). The time complexity of the algorithm is $O(W \cdot B)$, which is $O(N^2)$.

In addition to the effective p-value calculation an effective bounds are also available which can accelerate calculations. For details on the computational aspects and for application of mHG statistics to identifying transcription factor binding sites see Eden et al. [40].

### 1.3.3 Rank-sum (RS) test

Another threshold free method that measures the enrichment of a set of elements in a ranked list is that of rank-sum test (also referred to as the Wilcoxon test). In this test, we assign each gene a rank according to the ranked list and sum up the ranks of the curated set of genes. Thus, using the same notation as before, let $\lambda \in \{0, 1\}^N$, where 1 represents membership of the gene in the curated set and 0 otherwise. We define the rank-sum (RS) score as the following random variable:

$$RS(\lambda) = \sum_{i=1}^{N} i \cdot \lambda_i$$

(7)

We can calculate the level of significance of a particular RS score using a dynamic programming approach. Continuing with the notation from the previous section, consider the space of all binary label vectors of size $N = W + B$ with $B$ 1’s and $W$ 0’s: $A = \{0, 1\}^{(W, B)}$. Assume that we are given a vector $\lambda_0 \in A$, for which we calculate the $RS(\lambda_0) = s$. We would like to determine the significance level of $s$, p-value($s$). Let $S(k) = \left|\{\lambda \in A: RS(\lambda) = k\}\right|$ be the number of binary vectors with rank-sum of exactly $k$, then:
p-value \( (s) = \frac{|\{\lambda \in A : RS(\lambda) \geq s\}|}{|A|} = \frac{\sum_{i=1}^{s} |\{\lambda \in A : RS(\lambda) = i\}|}{|A|} = \frac{\sum_{i=1}^{s} S(i)}{|A|} \) \hspace{1cm} (8)

As in the case of the \( mHG \) statistics we can use dynamic programming to calculate \( S(k) \). To help us in the computation we will now define \( S(n,b,k) \) to be the number of binary vectors of size \( N \) with \( b \) 1’s and rank-sum of \( k \), where the last rank of a 1 bit is at most \( n \). Formally,

\[
S(n, b, k) = |\{\lambda \in \{0,1\}^{(N-b,b)} : RS(\lambda) = k, \max_i\{\lambda_i = 1\} \leq n\}|
\]

Notice that according to this definition

\[
S(k) = S(N, B, k)
\]

(9)

The way we calculate \( S(n,b,k) \) is using a dynamic programming where in each step we add all possible assignments for the last 1 bit, remove it from the current sum and continue:

\[
S(n, b, k) = \sum_{i=b}^{\min(n,k)} S(i-1, b-1, k-i)
\]

(11)

If we initialize by,

\[
S(n, 1, k) = \begin{cases} 1 & \text{if } n \geq k \\ 0 & \text{otherwise} \end{cases}, \text{ for } 1 \leq n \leq N \text{ and } 1 \leq k \leq \sum_{i=N-B}^{N} i
\]

(12)

\[
S(n, 0, k) = 0, \text{ for } 1 \leq n \leq N \text{ and } 1 \leq k \leq \sum_{i=N-B}^{N} i
\]

(13)

we get,

\[
p-value(s) = \frac{|\{\lambda \in A : RS(\lambda) \geq s\}|}{|A|} = \frac{\sum_{i=1}^{s} S(N, B, i)}{|A|}
\]

(14)

In summary, to compute the p-value of a rank-sum score \( s \) we first calculate \( \sum_{i=1}^{s} S(N, B, i) \). The number of possible binary vectors is \( \Lambda = \binom{W+B}{B} \) and thus the p-value(s) can be directly computed from (13). The time complexity of the algorithm is \( O(N^2 \cdot B^2) \), which is \( O(N^3) \) in a typical case where \( B \leq \sqrt{N} \).

In addition to the effective p-value calculation in cases where the \( W,B \geq 10 \) there is a good approximation to a normal distribution, where

\[
\mu = B(N+1)/2 \text{ and } \sigma = \sqrt{B(N-B)(N+1)/12}
\]

(15)

1.3.3.1 Comparison of mHG to RS

Although both the \( mHG \) and the \( RS \) tests tackle the same question, that is, enrichment of a set of elements in a ranked list of elements, there are few differences that can be pointed out. The main difference is in the level of robustness, while the \( mHG \) test
is more robust to noise by disregarding elements positioned in the bottom of the ranked list the RS test take into account all elements. Another important difference is in the efficiency of computation, while the RS statistic is simpler to calculate the dynamic programming for the p-value calculation is more complex than that of the mHG p-value calculation (O(N^3) for the RS test vs. O(N^2) for the mHG test). In that respect it is important to note that both tests have closed formulas for p-value approximation in the case of the RS test and for p-value tight upper bounds in the case of the mHG test.

1.4 Data analysis of multidimensional datasets

A high-throughput dataset can be represented by a matrix M, for which M(g,e) corresponds to the signal obtained for gene g in experiment e. In a multi-dimensional high-throughput matched dataset (to which I will henceforth refer to as multi-dataset) we have K datasets M_1,...,M_K all having one dimension in common. The multi-datasets used in this thesis consist of datasets with a common set of samples. That is, given m biological samples of interest we will have K different dataset M_1,...,M_K with dimensions (n_1 x m),...,(n_K x m), respectively. Note that each dataset can represent profiling results addressing a different biological aspect. Furthermore, for any datasets M_i and M_j the set of elements profiled are not necessary the same, and in particular n_i does not necessary equal n_j.

The naïve approach to analyzing such rich high-throughput data is to separately cluster the samples and genes/genomic elements [7], usually using hierarchical clustering. With an exhaustive expert examination, this methodology can yield meaningful biological results, though intricate responses are difficult to uncover. Various methodologies have been developed, in recent years, to handle integrated analysis of functional genomics data, mainly by studying the transcriptional programs and global organization of biological processes [9–11]. Still, only a few studies report the joint analysis of sample cohorts that include multiple genomic measurements [12, 15]. These studies mostly analyze only two datasets in an integrated manner. In the two-dataset scenario, the common methodology is to use correlation functions and try to find elements from both datasets that have positive or negative correlation.
I will next give a brief background on statistical correlation coefficients to assess the level of relatedness between measured elements in multi-datasets.

1.4.1 Pearson’s correlation

The most common measure for dependencies between two vectors \( u \) and \( v \) is the Pearson correlation coefficient:

\[
 r (u, v) = \frac{\sum (u - \bar{u})(v - \bar{v})}{\sqrt{\sum (u - \bar{u})^2 \sum (v - \bar{v})^2}}
\] (16)

In this case, \( r \) ranges from -1 to 1 and measures the degree to which the two vectors maintain a linear relationship, 1 corresponding to full linear relationship. It may therefore be less suitable when the data values follow some nonlinear relationship. In addition, it is sensitive to extreme outliers that are very frequent in genomic datasets.

1.4.2 Spearman’s rank correlation

A nonparametric rank statistic that measures the correlation between the ranks of the two given vectors \( u \) and \( v \) (with no ties) is Spearman’s rank correlation coefficient. Let \( \Pi_u \) denote the ranks of the samples with respect to the vector \( u \). That is, \( \Pi_u(i) \) denotes the rank of element \( u(i) \) in \( u \). In particular, \( u_{\Pi_u^{-1}(1)} < u_{\Pi_u^{-1}(2)} < \cdots < u_{\Pi_u^{-1}(N)} \), where \( N \) corresponds to the vector’s length. Similarly, let \( \Pi_v \) denote the permutation induced by the vector \( v \). Then Spearman’s rank correlation coefficient is therefore:

\[
s (u, v) = r (\Pi_u, \Pi_v)
\] (17)

This can also be formulated as

\[
s (u, v) = 1 - 6\sum_{i=1}^{N} \frac{(\Pi_u(i) - \Pi_v(i))^2}{N^3 - N}
\] (18)

Since it is distribution-free, the significance of the Spearman rank correlation coefficient may be approximated by the normal distribution. Thus, correlated pairs of elements can be extracted with higher confidence. Although the Spearman’s coefficient is more robust to noise it is also more complex to compute, with complexity of \( O(N \log N) \) as oppose to \( O(N) \) in the case of the Pearson’s coefficient, as we need to sort the vectors to obtain the ranks.
2 Methods

In the introduction we provided background on state of the art methods. In this section we provide details on methods developed during my PhD research work and refer to relevant sections in my published studies.

2.1 Mutual enrichment in ranked lists

So far we have dealt with enrichment of binary attributes, in which a 1 or 0 indicated whether or not the gene has a particular attribute and the binary vector was then sorted according to a particular scoring. 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 and obtain two different gene ranks. One ranking will correspond to the test ranked list, which we want to investigate, and the second we will call the labeling ranked list. We will use the labeling ranked list to label element in the test ranked list. We revisit our example from the previous section in which we tried to determine whether a particular set of genes are enriched in a ranked list of genes. A practical example can be viewed as testing whether genes that belong to a GO term are enriched in the top of a ranked list of genes ranked according to differential expression. The test ranked list of genes was determined according to the genes’ differential expression and the corresponding binary occurrence vector was generated according to their membership in the GO term. In the mutual enrichment scenario instead of the GO gene labeling, we will have a new scoring mechanism that will reflect the likelihood of the genes to possess the attribute in question, hence the labeling ranked list. Clearly, this information is valuable, as any threshold is arbitrary and thus should be incorporated in the enrichment analysis. A specific example, involving one ranking obtained from gene expression and the other from miRNA target prediction is discussed in Steinfeld et al. (specifically in Figure 1 of the manuscript [41]).

As described above functional genomics results are often represented as ranked lists of genes. Therefore, it is often the case that we would be interested in searching for mutual enrichment in the top of two ranked lists when comparing two functional genomics results. Note that as opposed to correlation measures, which may also be applicable to this situation, we are not interested in agreement in the bottom or even the
center of the ranked lists, but rather only in the top of the two ranked lists. Hence, mutual enrichment.

In the next section I will introduce the minimal-\(mHG\) (\(mmHG\)) approach, developed in Steinfeld et al. [41], that addresses the application of the mutual enrichment approach to identify miRNA activity. I will further describe various ways to test mutual enrichment in two ranked lists.

2.1.1 Minimal-\(mHG\) (\(mmHG\) or \(m2HG\))

The \(mmHG\) statistics is a generalization of the \(mHG\) statistics [40] presented in section 1.3.2 (Figure 2). While the \(mHG\) statistics quantifies the enrichment level of a set of elements in the top of a larger ranked list of elements, the \(mmHG\) statistics quantifies the mutual enrichment level for two ranked lists over the same set of elements. As our main application will be for genes we will, from here on, focus on genes as the ranked elements. Let two ranked lists of genes be represented as permutations, \(\pi_1 = (\pi_1(1), ..., \pi_1(N))\) and \(\pi_2 = (\pi_2(1), ..., \pi_2(N))\), over a universe of \(N\) genes where \(\pi_i(j)\) refers to the index of the gene ranked \(j\) in the \(i^{th}\) permutation. The \(mmHG\) approach seeks the top \(B\) genes in \(\pi_1\) that will have optimal enrichment in the top of the \(\pi_2\) ranking. Both \(B\) and the definition of top, above, are optimized in the process. A formal definition of the \(mmHG\) statistics follows.

Given a single permutation \(\pi \in S_N\), and for every \(i=1...N\) we define a binary vector \(\lambda_i\) in which exactly \(i\) entries are 1 and \(N-i\) are 0, as follows:

\[
\lambda_i(j) = \begin{cases} 
1 & \pi(j) \leq i \\
0 & \text{otherwise} 
\end{cases} 
\]

(1)

We define the \(mmHG\) score of a permutation \(\pi\) as:

\[
mmHG(\pi) = \min_i \text{p-value}(mHG(\lambda_i))
\]

(2)

\(mHG\) p-values used above, denoted p-value\((mHG(\lambda))\), assume a uniform null distribution of binary vectors with weight \(B\). These p-values are exact and do not require correction for the multiple thresholds tested. See details of the underlying methodology in Eden et al. [40].

For two permutations \(\pi_1\) and \(\pi_2\) the relative permutation \(\pi\), of \(\pi_2\) w.r.t \(\pi_1\), is defined by
\[ \pi(\pi(j)) = \pi_2(j), \text{ for every } j = 1 \ldots N \] (3)

or simply, using operations in the permutation group \( S_N \),

\[ \pi = \pi_2 \cdot \pi_1^{-1} \] (4)

We are now ready to define \( mmHG(\pi_1, \pi_2) \) representing the mutual enrichment of two ranked lists of genes as:

\[ mmHG(\pi_1, \pi_2) = mmHG(\pi) \] (5)

where \( \pi \) is the relative permutation of \( \pi_2 \) w.r.t \( \pi_1 \) as defined above.

Although the \( mmHG \) score is a minimum taken over a space of \( mHG \) p-values it cannot be directly considered as a significance measure, due to the additional multiple testing.

To assess the statistical significance of an \( mmHG \) result obtained in actual analysis we work over a null model that consists of a uniform measure on the group of all permutations \( S_N \). This is equivalent to a uniform measure over \( S_N \times S_N \) by the conversion of any \( \pi_1, \pi_2 \) to their relative permutation \( \pi = \pi_2 \cdot \pi_1^{-1} \) as above.

For \( mHG \) there is a dynamic programming process that provides a full characterization of the distribution of the statistics under the null model (see section 1.3.2.1). For \( mmHG \) we do not currently have an efficient process that allows for calculating exact p-values. We can however use a union bound approach as follows:

\[ p-value(mmHG(\pi)) < mmHG(\pi) \times N \] (6)

2.1.1.1 Practical implementation

Some modifications to the abstract definition of \( mmHG \) were used in Steinfeld et al. [41], to accommodate practical considerations. These are described in this section. First, instead of computing the p-value of an \( mHG \) result as required in Equation (2) we used the following bound which follows from Eden et al.[40]:

\[ p-value(mHG(\lambda_i)) \leq mHG(\lambda_i) \cdot i \] (7)

Thus, instead of computing the \( mmHG \) score we define the \( mmHG^* \) score to be

\[ mmHG^*(\pi) = \min_{1 \leq i \leq N} mHG(\lambda_i) \cdot i \] (8)

Notice that from Equation (8) it follows that

\[ mmHG(\pi) \leq mmHG^*(\pi) \] (9)

and thus from Equation (6) we get
\[ p\text{-value}(mmHG(\pi)) \leq mmHG(\pi) \cdot N \leq mmHG^*(\pi) \cdot N \] (10)

As a second practical modification, the optimization processes described in Equation (2) is not carried out in its full exhaustive scope. We only run the outer loop up to a reasonable threshold \(- n_{\text{max}}\). The altered equations are therefore:

\[ mmHG_{n_{\text{max}}}(\pi) = \min_{1 \leq i \leq n_{\text{max}}} p\text{-value}(mHG_{n_{\text{max}}}(\lambda_i)) \] (11)

where

\[ mHG_{n_{\text{max}}}(\lambda) = \min_{1 \leq n \leq n_{\text{max}}} HGT(N, B, n, b_\nu(\lambda)) \] (12)

\[ p\text{-value}(mmHG_{n_{\text{max}}}(\pi)) \leq mmHG_{n_{\text{max}}}(\pi) \cdot n_{\text{max}} \] (13)

Defining the bound-based version:

\[ mmHG^*_{n_{\text{max}}}(\pi) = \min_{1 \leq i \leq n_{\text{max}}} mHG_{n_{\text{max}}}(\lambda_i) \cdot i \] (14)

we now get:

\[ p\text{-value}(mmHG_{n_{\text{max}}}(\pi)) \leq mmHG^*_{n_{\text{max}}}(\pi) \cdot n_{\text{max}} \] (15)

Since the complexity of computing the \( mHG \) p-value is \( O(N^2) \) we get that the complexity of computing \( mmHG(\pi) \) is \( O(N^3) \). Using the bounds presented above we get a reduced complexity of \( O(N^2) \) for computing \( mmHG^*(\pi) \), or \( O(n_{\text{max}}^2) \) if we consider Equation (15).
Figure 2. Generalization of the hypergeometric (HG) approach. The HG approach uses two sets and estimates the significance of their intersection size. The minimal-HG (mHG) approach uses a fixed set definition, depicted by the red circle, and generalizes the definition of the second set, depicted by the blue circle. mHG statistics is computed by searching through all possible expansions of the blue set to get optimal enrichment. In the minimal-mHG(mmHG or m²HG) we further generalize the definition of the red circle to attain all possible intersections and choosing the red circle that results in optimal enrichment of the blue one.

2.1.2 Quantized-mHG (QmHG)

Another approach to capture the notion of mutual enrichment in two ranked lists is the Quantized-mHG model (QmHG), which extends the previously defined mHG
framework for enrichment analysis to non-binary label vectors. The basic idea of this methodology is to use $k$ thresholds on the labeled ranked list to define $k+1$ quanta in the test ranked list. Each threshold will define a set of genes (defined to be the top of the ranked list) that will be tested for enrichment in the test ranked list using the $mHG$ statistic. Formally, consider two ranked lists of genes: $P = (P(1), ..., P(N))$ being the test ranking, and $\Lambda = (\Lambda(1), ..., \Lambda(N))$ being the label ranking. We divide $\Lambda$ to $k+1$ subsets $S_0, S_1, ..., S_k$ of respective sizes $B_0, B_1, ..., B_k$. Where, $S_i = \{\Lambda(r_i+1), \Lambda(r_i+2), ..., \Lambda(r_i+B_i)\}$, $r_0=0$ and $r_i=\sum_{j=0}^{i-1} B_j$, $i=1,..,k$. We now define the integer vector $\lambda$ mapping each element from $\Lambda$ to $P$ and we label the elements in $P$ according to their origin set $S_i$ in $\Lambda$. That is:

$$\lambda_i = j \text{ iff } P(i) \in S_j$$

Thus the vector $\lambda$ has $B_0$ 0’s, $B_1$ 1’s, ..., and $B_k$ k’s. Notice that now 0’s correspond to elements that originated from the top quantum of the label vector while the k’s correspond to elements that originated from the bottom quantum of the label vector (an outline of $QmHG$ is illustrated in Table 1).

We define the $QmHG$ of such a vector $\lambda$ to be:

$$QmHG(\lambda) = \min_{0 \leq i < k} \left\{ \min_{1 \leq n \leq N} HGT\left(N, \sum_{j=0}^{i} B_j, n, b_n(\lambda, i)\right) \right\}$$

where,

$$b_n(\lambda, r) = \sum_{j=1}^{n} f(\lambda_j, r)$$

and,

$$f(t, l) = \begin{cases} 1 & t \leq l \\ 0 & o/w \end{cases}$$

Or if we define $\lambda^l$ to be the binary vector for which

$$\lambda^l_j = \begin{cases} 1 & \lambda_j \leq l \\ 0 & o/w \end{cases}$$

then

$$QmHG(\lambda) = \min_{0 \leq i < k} \{mHG(\lambda^l)\}$$

That is – when working with $QmHG$ we are trying to find, in a data driven manner, what definition (threshold) of top elements in $\Lambda$ (traversing all quanta) maximizes their enrichment in the top of $P$. 

18
The exact p-value of the $QmHG$ can be computed using a path enumeration strategy, similar to the one described in section 1.3.2.1, in a $k+1$ dimensional space. The details on how to compute this p-value in a 3-dimensional space are explained in the next section.

<table>
<thead>
<tr>
<th>$A$</th>
<th>$P$</th>
<th>$P^*$</th>
<th>$qP$</th>
<th>$qP(1)$</th>
<th>$qP(2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>8</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1. $QmHG$ statistic. The ranking of a set of elements $\{1,2,\ldots,10\}$ according to the label ranking and test ranking is depicted by columns $A$ and $P$, respectively. Column $P^*$ corresponds to the ranks of the elements in $A$ after sorting them according their ranks in $P$. In $qP$ we quantize $P^*$ with 3 quanta, where ranks $<3$ are assigned 0, ranks $>6$ are assigned 2 and the rest are assigned 1. $qP(1)$ is a binary representation of $qP$ when only the top quantum is considered and similarly $qP(2)$ is when only the top two quanta are considered. For each binary vector we can run the $mHG$ and take the best score overall.

### 2.1.2.1 p-value of the Quantized-mHG score

In previous section 1.3.2.1 we explained how to compute the exact p-values of $mHG$ scores by means of path counting. Here we demonstrate how to extend the idea to $QmHG$ scores to derive the exact p-values. Consider a 3-dimensional case for which we have fixed values of $W$, $B$, and $Y$, where $N = W + B + Y$. The p-value of a given 3-dimensional $QmHG$ score $s$ is calculated using a 3-dimensional grid ranging from $(0, 0, 0)$ to $(W, B, Y)$. A specific trinary label vector $\lambda \in \{0, 1, 2\}^N$ is represented by a path $(0, 0, 0) \rightarrow (W, B, Y)$ where the $i^{th}$ step in the path is $(1, 0, 0)$ if $\lambda_i = 0$, $(0, 1, 0)$ if $\lambda_i = 1$ or $(0, 0, 1)$ if $\lambda_i = 2$. Each point $(w, b, y)$ on the grid corresponds to a specific Quantized-HGT score.
Quantized-HGT(w, b, y) = min{ HGT(N, W, n, w), HGT(N, W + B, n, w + b) } \tag{22}

As in the binary case, we define $R$ to be the subset of points on the grid for which the $\text{Quantized-HGT}_n(\lambda) \leq s$, where $\text{Quantized-HGT}_n(\lambda)$ corresponds to the $\text{Quantized-HGT}$ score of the vector $\lambda$ after $n$ moves. We then count the number of paths $(0, 0, 0) \rightarrow (W, B, Y)$ not traversing $R$ by dynamic programming, denoted $\prod_R(W, B, Y)$. The $QmHG$ p-value is thus

$$p\text{-value}(s) = 1 - \frac{\prod_R(W, B, Y)}{B, Y} \tag{23}$$

This can be computed in $O(W \cdot B \cdot Y) = O(N^3)$. For a generic vector $\lambda \in \{0, 1, \ldots, k\}^N$, containing $B_0$ 0’s, $B_1$ 1’s, ..., and $B_k$ $k$’s the p-value of $QmHG(\lambda)$ can be calculated thus in $O(N^{k+1})$.

### 2.1.3 Minimal-rank-sum (mRS)

The same line of generalization can be applied to the rank-sum score. In this case instead of testing the rank-sum of a set of elements of interest we are given two ranked lists and through exhaustive threshold testing we optimize enrichment. Similar to the $mmHG$ approach, the threshold testing is carried out in one of the ranked lists, where each threshold assignment defines the set of top elements, with which we test for the optimal enrichment in the second ranked list. Formally, let two ranked lists of genes be represented as permutations $\pi_1 = (\pi_1(1), \ldots, \pi_1(N))$ and $\pi_2 = (\pi_2(1), \ldots, \pi_2(N))$, over a universe of $N$ genes. We define $mRS(\pi_1, \pi_2)$ to be

$$mRS(\pi_1, \pi_2) = \min_i RS(\lambda_i) \tag{24}$$

Where $RS$ is defined in section 1.3.3 and $\lambda_i$ defined as in section 2.1.1 Equation (1). We note that $RS$ has a normal approximation that is simple to calculate (using a closed formula) and thus the calculation of the mRS statistic is only $O(N)$ using this approximation. A full characterization of the random variable $mRS(\pi_1, \pi_2)$, when permutation are uniformly drawn, remains an open question.
2.2 The integrative approach

2.2.1 Enrichment networks

Our approach to integrated data analysis in high-throughput molecular biology tries to harness the power of enrichment statistics and knowledge of genomic annotation databases to assign relevant function annotations to profiled elements. This approach is a ‘guilty by association’ one - we use information available on one profiled dataset to assign relevant information on elements from another profiled dataset. The basic idea is to first rank the informative elements according to a particular aspect of the non-informative dataset and then search for enrichment of interesting annotations in the ranked list of elements. The enrichment results are therefore directly assigned to the un-annotated profiled elements.

In Enerly & Steinfeld et al. [42] we studied 101 breast cancer patients and the profiled datasets were mRNA and miRNA. In this study the aim was to assign miRNAs to cancer related biological processes. In this paper the annotation database used were GO term sets [8], MSigDB sets [39] and miRNA target sets [43]. These different databases give comprehensive information on coding genes that were also profiled in the same cohort using mRNA microarrays. The enrichment analysis was carried using the mHG statistics as described above (see section 1.3.2). The following ranking schemes were used in the analyses performed in this study (see Figure 3): (A) Differential expression – using TNoM [44] or other differential expression score. (B) Correlation to pivot miRNA – for a specific miRNA, called the pivot miRNA, mRNAs were ranked according to the correlation (Spearman’s coefficient) of their expression pattern across the entire cohort (101 samples) to the pivot miRNA expression pattern. (C) miRNA target prediction – for a specific miRNA, genes were ranked according to how likely they are to be targeted by the miRNA. Context score values, taken from Targetscan V5.1 target prediction tool [43], were used as prediction scores.
Figure 3. Schematic overview of the integrated analysis pipeline. Genes were ranked according to one of three options: (A) Differential expression. (B) Correlation to a pivot miRNA – for a specific miRNA, called the pivot miRNA, mRNAs were ranked according to the correlation of their expression pattern across the entire cohort (101 samples) to the pivot miRNA expression pattern. (C) miRNA target prediction – for a specific miRNA, genes were ranked according to how likely they are to be targeted by the miRNA.

A generic description of this approach is described in Figure 4. The Primary Data and Pivot Data are the matched profiled datasets on the same set of samples. The Annotation Data is a binary matrix that represents a collection of gene sets – a column for each gene set – and describes for each element in the Primary Data whether it belongs to each of the gene sets. The process described above is carried out by first generating the Correlations Matrix that holds for each pair of Primary Data element and Pivot Data element, their correlation coefficient levels. We then take each element in the Pivot Data separately and use its relevant vector of correlation levels to rank the Annotation Binary Matrix rows. This step results in a collection of ranked binary vectors to which we apply the mHG statistics to calculate the enrichment levels for each annotation (column). Thus, for each pivot element a vector of enrichment significance levels is generated (one for each annotation term). Carrying this process out for all pivots, the mHG Enrichments Matrix is generated.
Figure 4. Integrated analysis of matched high-throughput datasets. The input for this approach is the Primary Data matrix for which we have the binary Annotation Matrix and a Pivot Data Matrix that is sampled matched to the Primary Data Matrix. The aim of this approach is to uncover significant association between pivot elements (rows from the Pivot Data Matrix) and annotations (columns from the Annotation Matrix). We first generate the Correlations Matrix that holds the correlation level between each pivot and primary element. We then for each pivot element rank the Annotation Matrix according to the pivot’s correlation vector taken from the Correlations Matrix. This step results in a collection of ranked binary vectors to which we apply the mHG statistics to calculate the enrichment levels vector for each annotation (column) that will constitute a row in the mHG Enrichment Matrix.

The final mHG Enrichments Matrix is the result of multiple statistical tests. In order to focus on significant results the entries of this matrix need to be corrected for the multiple tests carried out in the process. After the multiple testing corrections, the mHG Enrichments matrix can be viewed as a binary matrix where each positive entry
represents a significant association between the relevant pivot and annotation. This binary matrix is best visualized using a bipartite graph where we have the pivot elements in one side and the annotations on the other and edges present significant associations between the two. In Figure 5 the miR-GO association network was plotted using Graphviz software [45] for the dataset used in Enerly & Steinfeld et al. [42].

Figure 5. Significant association between miRNAs and GO terms derived from enrichment analysis in correlated/anti-correlated genes (see Figure 4). GO-terms and miRNAs are represented by rectangles and rhombi, respectively. In addition, positive and negative associations are depicted by red and green arrows, respectively. Positive/negative associations here refer to enrichment of the term in the list of genes ranked according to correlations/anti-correlations, respectively, to the pivot miRNA. Arrangement of the association network is according to cancer related biological modules: a) cell-cycle, b) glycolysis, c) metabolic process, d,e,f) development and morphogenesis, g) transport, h) DNA structure, i) cell-adhesion and j) immune response.

A formal description of the enrichment network calculation follows:

Let $D_{N\times S}$ be the primary data matrix and let $P_{M\times S}$ be the pivot data matrix and let $T_{N\times A}$ be the annotation binary matrix. We define the correlation data matrix $C_{M\times N}$ to hold the
correlation coefficient between each row from \( D \) and \( P \). That is, let \( u \) be the \( m \)th row in \( P \) and \( v \) be the \( n \)th row in \( D \), in that case if Pearson’s correlation is used it follows that

\[
C[m, n] = \frac{\Sigma(u-\bar{u})(v-\bar{v})}{\sqrt{\Sigma(u-\bar{u})^2\Sigma(v-\bar{v})^2}}
\]  

(25)

From \( C \) and \( T \) we now calculate the \( mHG \) enrichment matrix \( En_{MA} \), where each entry \( En[m, a] = -\log\text{-p-value} \) of the enrichment level calculated by ranking the \( a \)th column in \( T \) according to the correlation score vector defined by the \( m \)th row in \( P \), and running the \( mHG \) statistics (as defined in section 1.3.2) on the ranked binary vector. The significance level is then corrected by the Bonferroni multiple testing correction for the \( M\times A \) tests and thus only p-values \(<0.05/M\times A\) are considered as significant. We can now define the bipartite graph \( G(U, V, E) \), where each element in \( U \) represents a row in \( M \) and each element in \( V \) represents a column in \( T \) and an edge \((u, v)\) is in \( E \) iff the entry in \( En[u, v] \) \( > -\log(0.05/M\times A) \).

### 2.2.2 The \textit{mmHG} extension

In Steinfeld \textit{et al.} [41] we employed an integrative approach which required the analysis of miRNA target information that is quantitative in nature instead of the binary one used in the section above. When assessing association between a pivot and an annotation, instead of a ranked binary vector we now get two ranked list of genes: one from the pivot correlated genes and the other from the quantitative annotation. Therefore, the \textit{mmHG} statistics for mutual enrichment (see section 2.1.1) was used and instead of the \textit{mHG} Enrichment Matrix an \textit{mmHG} Enrichment Matrix is generated. A similar multiple testing correction follows.

### 2.2.3 The empirical testing correction

In Aure & Steinfeld \textit{et al.} [46] we describe a workflow of selection steps to focus on genes that are likely to drive the cancer process. In the final step of the workflow we use the integrated approach described above where the Pivot Data is an expression matrix of a set of genes that are likely to participate in the cancer process. That is, these genes follow two important criteria: a) copy number data shows that they are commonly
aberrant (at least 10% of the samples) and b) their expression is highly correlated with their aberration profile over the 100 breast cancer samples profiled (\textit{in-cis} correlation is observed). The Primary Data holds the expression levels of all other genes after compensation for their observed aberration profile, so when a gene is observed to be over expressed it will not be as a consequence of high copy number in its locus. For the Binary Annotation Matrix the GO database was used, as in Enerly & Steinfeld \textit{et al.} [42]. The idea of this process was to associate genes represented in the Pivot Data with biological processes. The final output of this workflow is a set of genes that are commonly aberrant in many tumors, have high \textit{in-cis} correlation between expression and aberration calling, and also seem to drive a particular biological process in the examined cohort of breast cancers.

In this study we encountered high level of false positive results that was a result of high levels of dependency between profiled genes that pertain to the same annotation term. The uniform null model assumption of \textit{mHG} is less valid because of this dependency. We therefore extended the statistical framework of the integrated approach to also include empirical testing. To obtain empirical p-values for the associations selected above, 100 random simulations were performed. In each simulation instance, we shuffled the order of the samples in the Pivot Data Matrix only and recalculated all associations. This approach preserves existing expression dependencies between genes. So, let $E_{js}$ be the enrichment score (\textit{mHG} p-value) obtained for the association between a pivot gene $j$ and gene set (GO term) $s$, and let $E_{js}^{*k}$ be the enrichment score of this association in the $k^{th}$ simulation instance. We now consider a given pivot gene $j$ to be significantly associated with a gene set $s$ if

$$E_{js} < \min_{t,k} (E_{ts}^{*k}),$$

(26)

for every $k$ (shuffling simulation) and $t$ (pivot).

That is, a pivot gene was associated with a gene set if their association level was higher than the association levels obtained for that gene set in all simulations and for all pivot genes randomly simulated in the process.

To obtain the binary matrix that represent the significant associations we assign a positive entry only to entries that follow the criteria in Equation (26). Note that this is used instead of using the Bonferroni multiple testing correction used in section 2.2.1. We
can now evaluate the false discovery rate of our results by comparing the amount of our observed significant results to the expected amount of significant results.
3 Results

3.1 miRNA-mRNA Integrated Analysis Reveals Roles for miRNAs in Primary Breast Tumors

miRNA-mRNA Integrated Analysis Reveals Roles for miRNAs in Primary Breast Tumors

Espen Enerly1,4, Israel Steinfeld2,9, Kristine Kleivi1,2, Suvi-Katri Leivonen3, Miriam R. Aure1, Hege G. Russnes1,10, Jo Anders Rønneberg1,4, Hilde Johnsen1, Roy Navon5, Einar Rodland6, Rami Mäkelä3, Bjørn Naume7, Merja Perälä3, Olli Kallioniemi3,8, Vessela N. Kristensen1,4,9, Zohar Yakhini2,5,*, Anne-Lise Børresen-Dale1,4*

1 Division of Surgery and Cancer, Department of Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, Oslo, Norway, 2 Department of Computer Science, Technion, Haifa, Israel, 3 Medical Biotechnology, VTT Technical Research Centre of Finland, Turku, Finland, 4 Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway, 5 Agilent Laboratories, Tel Aviv, Israel, 6 Department of Informatics, University of Oslo, Oslo, Norway, 7 Division of Surgery and Cancer, Department of Oncology, Oslo University Hospital, Oslo, Norway, 8 Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland, 9 Institute for Clinical Epidemiology and Molecular Biology (EPIGEN), Nordbyhagen, Norway, 10 Division of Pathology, Oslo University Hospital, Oslo, Norway

**Abstract**

**Introduction:** Few studies have performed expression profiling of both miRNA and mRNA from the same primary breast carcinomas. In this study we present and analyze data derived from expression profiling of 799 miRNAs in 101 primary human breast tumors, along with genome-wide mRNA profiles and extensive clinical information.

**Methods:** We investigate the relationship between these molecular components, in terms of their correlation with each other and with clinical characteristics. We use a systems biology approach to examine the correlative relationship between miRNA and mRNAs using statistical enrichment methods.

**Results:** We identify statistically significant differential expression of miRNAs between molecular intrinsic subtypes, and between samples with different levels of proliferation. Specifically, we point to miRNAs significantly associated with TP53 and ER status. We also show that several cellular processes, such as proliferation, cell adhesion and immune response, are strongly associated with certain miRNAs. We validate the role of miRNAs in regulating proliferation using high-throughput lysate-microarrays on cell lines and point to potential drivers of this process.

**Conclusion:** This study provides a comprehensive dataset as well as methods and system-level results that jointly form a basis for further work on understanding the role of miRNA in primary breast cancer.

---

**Introduction**

Expression profiling of mRNA has been used to molecularly characterize various tissues and tumors. A range of gene signatures that predicts pathway activation, has been identified in various cancer types [reviewed in [1]]. In breast cancer, mRNA profiling has been used to classify breast tumors and associate them with clinical and pathological characteristics as well as with prediction of outcome [2,3,4,5]. In particular, luminal-A and basal-like subtypes, defined using an intrinsic gene list, have distinct and reciprocal gene expression profiles as well as large differences in clinical characteristics, including survival [5,6,7,8].

Gene expression regulation through mechanisms that involve microRNAs (miRNAs) has attracted much attention during recent years. miRNAs are a class of endogenous small regulatory RNA molecules that target mRNAs and trigger either translation repression or mRNA degradation [9]. There are to date more than 900 identified human miRNAs [10], transcribed as individual units, polycistronic clusters or in concert with a protein coding host gene [11]. Many miRNAs regulate genes associated with different biological processes such as development, proliferation, apoptosis, stress response, and tumourigenesis [12,13,14,15,16].

Abnormal expression levels of several miRNAs have previously been shown to be associated with multiple cancer types including breast cancer [17,18,19,20]. Some miRNAs correlate with specific clinical features of breast cancer, such as estrogen and progesterone receptor expression, tumor stage, vascular invasion, and proliferation index [19,21,22,23]. In a study by Blenkiron et al. a set of 309 miRNAs were profiled in 93 human primary breast tumors, 5 normal breast samples and 21 cell lines, identifying 31
miRNAs associated with molecular subtype, estrogen receptor status or grade [24]. In addition, this study reported a strong co-regulation of miRNA genomic clusters and showed that for the majority of miRNAs differential expression cannot be attributed to chromosomal loss or gain in their genomic region. The study reports some findings that pertain to jointly analyzing the miRNA data with its matching mRNA data. Continuing this direction and taking a systematic approach to joint analysis will further enhance our understanding of the role of miRNA in breast cancer pathogenesis and progression.

In this work we present expression profiling of 799 miRNAs in 101 human primary breast tumor samples, along with genome-wide matched mRNA profiling and extensive clinical information. We applied several approaches to statistically analyze the resulting data. We identified statistically significant differential expression of miRNAs that distinguishes the reciprocal basal-like and luminal-A breast cancer subtypes. Our analysis confirmed some observations from previous studies including Blenkiron et al. [24], but also revealed subtype specific expression of previously uncharacterized miRNAs. We put emphasis on the joint analysis of miRNA and mRNA data, and analyzed correlations between miRNA and mRNA expression data. We show that particular cellular processes such as proliferation, cell adhesion, and immune response are significantly enriched in the co-regulated clusters, suggesting a central role for miRNAs in regulating these pivotal pathways. We performed functional assays using direct measurement techniques to validate the influence of miRNA on proliferation.

Results

miRNA differential expression in molecular breast cancer subtypes

miRNA expression profiling was carried out for 101 human primary breast cancer samples (Table S1) using microarrays covering 799 miRNAs, from Agilent Technologies. After filtering miRNAs that were not expressed in most of the cohort (see Materials and Methods), 489 miRNAs were considered for further analyses. Applying hierarchical clustering based on the 100 most variably expressed miRNAs, we observed a cluster consisting of tumors with mainly basal-like subtype characterization, that was distinguished by a higher expression of the miR-17-92 cluster/ family from 7q22.1, 13q31.3 and Xq26.2 (Figure 1A).

Since the basal-like breast cancer subtype is defined by specific morphological and pathological characteristics as well as by a distinct miRNA expression profile we explored the full list of miRNAs related to the basal-like tumors. Comparing miRNA expression in 15 basal-like and 41 luminal-A samples, the subtypes with strongest reciprocal miRNA expression profiles, 111 differentially expressed miRNAs were identified at an FDR (False Discovery Rate [25]) of 5%. The top 26 miRNAs separated almost perfectly the basal-like and luminal-A samples. The threshold number of misclassification (TNoM) ≤ 6, see Materials and Methods). These results confirm that the subtype difference is manifested also at the level of miRNA expression (Figure 1B, Table S2). Only five miRNAs were intronic in genes that play a role in the miRNA subtype classification (hsa-mir-324 in AGADVL, hsa-mir-153 in PTPRN2, hsa-mir-934 in VGLL1, hsa-mir-593 in PTPRN2, hsa-mir-744 in MIA2P4), demonstrating that the miRNA differential expression is not merely a recapitulation of the mRNA classification.

The top miRNAs with elevated expression levels in basal-like samples were miR-18a/b (TNoM p<2E-10) and other members of the miR-17-92 cluster (miR-17/17*, miR-18a/b, miR-19a, miR-20a and miR-106a). In addition we found miR-9/9* (TNoM p-value < 4E-9), which had no detectable expression in most of the non-basal-like samples (Figure S1A). Among the prominently down regulated miRNAs in basal-like tumors were representatives of the miR-29 family (TNoM p<7E-12) along with miR-190b (TNoM p<2E-10) (Figure S1B-C). miR-29 family members were moderately expressed in subtypes other than luminal-A and basal-like. miR-190b had an almost discrete binary expression mode with higher expression in the luminal-A/B subtypes than in the basal-like/ERBB2-enriched subtypes (TNoM p<4E-15). We further discuss the role of miR-29 in our cohort in later sections.

miRNA expression and TP53 mutational status

Mutations in the TP53 gene are well studied and have been associated with cancer progression and worse prognosis [26]. In the present cohort we explored the miRNAs that were differentially expressed between the 64 wild-type samples (WT) and 36 TP53 mutant samples (Figure 2). Since the breast cancer subtype classification is not independent of the TP53 status, with most of the basal-like and ERBB2-enriched tumors having TP53 mutations, many of the same miRNAs were found to be differentially expressed in both partitions. At top among the 81 differentially expressed miRNAs (at 5% FDR) we identified miR-342-3p (TNoM p<2E-08) to have significantly lower expression in the TP53 mutant tumors. We note that miR-34a, previously shown to regulate TP53 [14], is not observed to be differentially expressed in comparing TP53 mutational status in this cohort (Table S3).

As previously reported [26] there is a strong association between estrogen receptor (ER) and TP53 status (Figure 2). Tumor samples with wild type TP53 are mostly ER-positive and samples with mutated TP53 are mostly ER-negative. In line with this co-occurrence we observed a substantial overlap between miRNAs differentially expressed between ER+ and ER- samples and between WT and mutant TP53 samples. Removing this confounding factor, by comparing 50 TP53 WT/ER+ samples vs. 11 TP53 mutant/ER+ samples or 12 TP53 WT/ER- samples vs. 26 TP53 mutant/ER- samples, we observed different repertoires of differentially expressed miRNAs (Table S3).

Joint enrichment analysis associates miRNAs to distinct biological processes

To better understand the role of miRNAs in different biological modules, as evidenced in our cohort of primary breast cancer samples, we further take a systems biology approach to examine the correlation of miRNA expression pattern to the expression pattern of the pivot miRNA. Using GO enrichment analysis carried out on the ranked list of miRNAs (see Materials and Methods) we were able to elucidate the biological modules that are correlated or anti-correlated to the expression level of the pivot miRNA (Table S4). We note that this association does not imply a direct regulation by the miRNA but rather indicates the biological process in which the pivot miRNA plays a role. For several cases we are able to find enrichment of the pivot miRNA targets in its anti-correlated miRNAs. In these cases we point to a potential direct regulation effect by the miRNA.

Several pivot miRNAs showed strong association to cell-cycle genes. Specifically, we observed an enrichment of cell-cycle genes when considering positive correlation to members of the miR-17-92 cluster, with miR-93 and miR-18b yielding the strongest enrichment (minimum-hypergeometric (mHG) p<4E-74 and p<4E-73, respectively). In accordance with previous results [27] we also observed a significant enrichment of genes regulated by E2F (mHG p<4E-24 and p<4E-27 for the above two miRNAs,
We note that miR-19b targets are enriched amongst its anti-correlatees (mHG \( p \leq 2 \times 10^{-12} \)). For miR-493 and the miR-214 cluster we observed cell-cycle genes to be enriched in the negatively correlated genes.

Another biological module with many significantly associated miRNAs is that of the immune response (Table S4). In particular, miR-150 was found to have the strongest enrichment of the immune response term amongst its positive correlated mRNAs (mHG \( p \leq 1.47 \times 10^{-17} \), Figure 3). In addition we identified other miRNAs with a strong positive correlation to the immune response module (e.g. miR-146 at mHG \( p \leq 1.32 \times 10^{-12} \), miR-142 at mHG \( p \leq 1.08 \times 10^{-12} \), miR-155 at mHG \( p \leq 1.23 \times 10^{-17} \), and miR-223 with mHG \( p \leq 1.11 \times 10^{-17} \)). Interestingly, their association was more related to T-cell activation genes while miR-150 was strongly associated with genes related to the inflammatory response.

### miRNA expression associated with proliferation

Our systematic approach led us to pay special attention to the cell-cycle module in the context of our cohort of early stage breast cancer patients. As activation of cell-cycle genes is closely related to proliferation we further examined the expression of miRNAs in samples with different proliferative states. To do so, we used Ki67 immunohistochemistry staining (IHC) and scoring of mitotic count from tumor sections to partition the samples into a High-Proliferative class (HP), consisting of 24 samples, and a Low-Proliferative class (LP) consisting of 35 samples (Figure 4A, see Materials and Methods). We identified 123 differentially expressed miRNAs (at 5% FDR, Table S5) when comparing HP and LP classes. This high number reflects the significant difference between the miRNA expression signatures of the two classes (Figure 4B, Table S5).

No basal-like nor luminal-B samples were assigned to the LP class, while no luminal-A and no normal-like samples, except one, were grouped to the HP class. However, ERBB2-enriched samples were equally assigned to the different proliferation classes (6 in HP and 5 in LP). To test for subtype independent miRNA differential expression we examined differential expression using only the ERBB2-enriched samples. We identified 21 miRNAs (at TNoM
A

![Graph showing hsa-miR-150 expression levels](image)

**Subtype legend**
- Basal
- Lum A
- Lum B
- ERBB2
- Normal

B

![Diagram illustrating P-value color scale](image)

**Cell activation**
- Lymphocyte activation during immune response
  - Adaptive immune response
    - Inflammatory response to stress
      - Defense response to wounding
        - Cell activation during immune response
          - Adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains
Figure 3. Expression of miR-150 and its mRNA correlates. (A) The samples are ordered according to the expression levels of miR-150 for that sample. The absolute signal intensities of miR-150 are presented in the top panel. The top 50 correlated genes are sorted from top to bottom. The color in the bar beneath the heatmap indicates the different subtypes. As can be seen many luminal-A samples have low levels of miR-150 expression though no clear cut can be deduced to separate the luminal-A samples from the rest of the subtype samples. (B) Graphical representation of GO-term enrichment of genes positively correlated to miR-150. The strongest enrichment is seen for the “immune response” (p<1.2E-147) term. The graph is color coded according to degree of enrichment. Figure obtained using the GOrilla web tool [62].

miR-29c is associated with cell adhesion/extra cellular matrix

The tumors studied here have previously been classified into four classes based on a signature of extracellular matrix (ECM) genes [29]. Strong degree of differential expression is seen at the miRNA level between ECM1 and ECM3 as well as between ECM1 and ECM4. We focused on the ECM1 and ECM4 classes (23 and 16 samples, respectively) and identified 47 differentially expressed miRNAs (at 5% FDR, Figure S3A). miR-29c, shown in previous sections to be down regulated in basal-like samples as compared to luminal-A samples, was found to be the most differentially expressed miRNA between the two ECM classes (TNOM p<3E-5, lower in ECM1, Table S8). We used the TargetScan prediction tool [30] to rank all genes according to their miR-29c target prediction scores. The top predicted targets of miR-29c were found to be enriched with genes related to the cell-adhesion GO-term (Figure S4). Further supporting its direct role in our cohort, we found miR-29c targets as well as cell adhesion genes to be significantly anti-correlated with miR-29c expression (mHG p<2E-11 and mHG p<2E-13, respectively; see Figure S3B and C and Materials and Methods).

miRNAs related to survival

For all 489 miRNAs, we ran univariate Cox regressions to predict survival. The top 9 miRNAs resulted in 37% FDR (Table S9).

High expression levels of the immune module have previously been associated with better survival in ER-/HER2- patients [31]. In agreement with the strong association of miR-150 with genes related to the immune response, we found high expression of miR-150 alone to be predictive of better prognosis (log-rank p<0.085) within the corresponding set of patients in our cohort, namely the ER-/non-ERBB2 enriched patients (Figure S5).

Discussion

In this work we introduce an extensive analysis of miRNA expression in 101 tumor samples from breast cancer patients. We show that miRNA expression alone is sufficient to distinguish luminal-A from basal-like samples, two types that represent different degrees of aggressiveness of the disease. Specifically, the oncogenic miR-17-92 module is shown to be distinctly over-expressed in highly proliferative samples including the basal-like samples, which are also characterized with high frequency of TP53 mutations. Recently we have shown that the miR-17-92 module is repressed by wild-type TP53 in an E2F1-mediated manner [27]. Our integrated analysis supports the E2F association with the miRNA cluster. Other over-expressed members of the cluster, such as miR-17-5p/miR-20, have been linked to the regulation of...
cell proliferation through a Cyclin D1 regulatory feedback loop [32] and through the inhibition of AIB1 translation [33] in breast cancer. miR-18a directly targets ESR1 [28] and has been shown to promote estrogen receptor alpha (ESR1) dependent proliferation in hepatocellular carcinoma cells [34]. In addition we have recently observed that the miR-18 cluster is over-expressed in a panel of samples from various cancers types [20]. miR-18a/b may therefore be an important contributor to the different overall gene expression profiles that distinguish between malignant to non malignant tumors, specifically ER positive tumors.

We further explore and characterize the miRNA expression signature that distinguishes between basal-like and luminal-A samples, known to be reciprocal with respect to miRNA expression and clinical properties. We show that miR-9/9* are markers for aggressive tumors, being expressed specifically in the basal-like tumors. miR-9/9* are over-expressed in c-Myc induced mouse mammary tumors [35], as well as in brain [36] and ovarian primary cancers [37] which points to a more general role in cancer progression for this miRNA. In our cohort we see a positive correlation between miR-9 and c-Myc (Pearson’s r = 0.22) and

---

**Figure 4. Proliferation associated miRNAs.** The panels show miRNAs that are both positively and negatively associated with proliferation in in vivo profiling from tumors. (A) Immunohistochemistry staining of Ki67 of tumors scored as highly proliferative (HP, sample 267) and low proliferative (LP, sample 570). The right panel shows the signal distribution in the 101 samples for three selected miRNAs, miR-142-3p, miR-19a, and 449a, with signal intensities for sample 570 (red dot) and 627 (green dot) highlighted. (B) Volcano plot of all miRNAs with TNoM p-value of differential expression against fold change differences in low (24 samples) versus high (35 samples) proliferation groups. Pink dots represent significant miRNAs (p<0.001). Left part contains miRNAs that are up-regulated in highly proliferative samples and right part down-regulated miRNAs. (C) The plot shows the scores for each miRNA, where each miRNA is represented by a dot. On the Y-axis differential expression score is –log(p-value) if miRNA is up-regulated and log(p-value) if miRNA is down regulated, thus assigning positive and negative scores according to differential expression between the high and low proliferative groups. Significance of differential expression is calculated using TNoM as described in Materials and Methods. “Cell-Cycle” (CC) score is –log(p-value) if the CC genes are enriched within the genes positively correlated to the miRNA, and log(p-value) if the CC genes are enriched within the genes negatively correlated to the miRNA, p-value for CC enrichment is calculated using the mHG statistic as described in Materials and Methods. doi:10.1371/journal.pone.0016915.g004
We introduce several new approaches to analyze miRNA and mRNA expression signature with respect to other molecular characteristics of the tumors like TP53 mutations. TP53 mutant samples largely overlap the basal-like samples and the ER-negative samples. Therefore the majority of the miRNAs found to be over-expressed in basal-like samples were also over-expressed in TP53 mutated samples and ER-negative samples and vice versa. Among the differentially expressed miRNAs we find miR-34c which is a direct transactivation target of TP53 and miR-18a/b which targets TP53. Although TP53 mutations have previously been linked to tumorigenesis and further characterization is needed to understand their relationship to TP53 and ER. Although miR-34a was previously shown to regulate TP53, we do not see it differentially expressed in the context of TP53 mutational status which might indicate a lack of feedback loop in this regulation.

We introduce several new approaches to analyze miRNA and mRNA expression data in an integrated manner using a systems biology approach. In particular we assess the enrichment of various gene sets amongst the genes correlated/anti-correlated to expression of a pivot miRNA. Our straightforward approach allows to comprehensively associate miRNAs to biological processes in a statistically sound and functionally relevant manner. Thus, our approach leads us to additional insight into the role of miRNAs in breast cancer and enables the identification of key players. Other approaches of data integration [41,42] are aimed at finding a set of miRNAs and a set of mRNAs that are expressed in a concerted manner either in a subset of the samples or in the entire cohort. We note that while these approaches are computationally sophisticated they are not driven by functional relevance.

In particular we point to a significant association of many miRNAs to the cell-cycle module. Our methodology also shows that miRNAs over-expressed in proliferative samples, are correlated to positive regulators of cell-cycle. Similarly, miRNAs under-expressed in proliferative samples are correlated with negative regulators of cell-cycle. The relationship between enrichment of these gene sets and proliferation related differential expression is, in general, monotone, as depicted in Figure 4. This monotonicity further demonstrates the sensitivity of our approach in terms of detecting miRNA association to biological processes.

The integrated analysis also revealed the association of several miRNAs to the immune response module, a major biological process closely associated with cancer progression and development [43]. We found miR-150, as well as miR-155 and miR-142, to have strong positive correlation to the immune response module. As miRNAs are considered to be negative regulators of expression the positive association of miR-150 to the immune response indicates that miR-150 is not a direct regulator of this process but rather a part of the immune response transcriptional program. Several studies have identified miR-150 to be involved in controlling B-cell differentiation by targeting the transcription factor c-Myb [44,45,46]. In our dataset (data not shown) we see an inverse correlation (Pearson’s $r = -0.18$) between Myb and miR-150 expression levels which might point to a similar regulation pathway of immune response in breast cancer. Immune response was previously linked to clinical outcome in ER-/HER2- samples [31], and in accordance with its strong association to the immune response module we see that miR-150 can predict clinical outcome in our ER-/HER2- samples. This makes miR-150 a good marker for the activity of the immune response in breast cancer samples, as high expression of miR-150 is associated with active immune response and better prognosis. Since we could not find any association of miR-150 to any of the known breast cancer subtypes we propose that its expression can act as a mean for classifying breast cancer samples based on immune response. However, more

### Table 4.1

<table>
<thead>
<tr>
<th>miRNA</th>
<th>Proliferative effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-34c-5p</td>
<td>-</td>
</tr>
<tr>
<td>miR-145</td>
<td>-</td>
</tr>
<tr>
<td>miR-154</td>
<td>- (†)</td>
</tr>
<tr>
<td>miR-299-3p</td>
<td>-</td>
</tr>
<tr>
<td>miR-331-3p</td>
<td>-</td>
</tr>
<tr>
<td>miR-342-3p</td>
<td>-</td>
</tr>
<tr>
<td>miR-449a</td>
<td>-</td>
</tr>
<tr>
<td>miR-19a</td>
<td>+</td>
</tr>
<tr>
<td>miR-130b</td>
<td>+</td>
</tr>
<tr>
<td>miR-142-3p</td>
<td>+</td>
</tr>
<tr>
<td>miR-146b</td>
<td>+</td>
</tr>
<tr>
<td>miR-150</td>
<td>+</td>
</tr>
<tr>
<td>miR-362-5p</td>
<td>+</td>
</tr>
</tbody>
</table>

Figure 5. Proliferation assay in miRNA transfected cell-lines. The panels show miRNAs that are both positively and negatively associated with proliferation in transfected cell-lines. (A) List of miRNAs that showed a significant effect on proliferation in cell-lines with a corresponding differential expression in tumors. († indicates opposite effect in MCF-7 and BT-474). (B) Lysate microarray (LMA) screening of MCF-7 and BT-474 cells transfected with 20 nM human Pre-miRTM miRNA Precursor library v2. Ki67 readout after 48 and 72 hours. The two miRNAs from panel A with strongest positive and negative effect on proliferation for each cell line are shown (see Table S7 for complete list).

doi:10.1371/journal.pone.0016915.g005
studies are needed to better elucidate causal relationships, if any, between miR-150 and the immune response as well as between miR-150 and prognosis.

We further examined the involvement of miRNAs in regulating cell proliferation. Comparing the highly proliferative (HP) samples to the low proliferative ones (LP), we discovered a distinct miRNA expression signature. Specifically, we validate in vivo the association of the miR-17-92 module to proliferation in breast cancer, in agreement with the same association observed in tumor cell lines [47]. In addition, we find miR-19a and miR-214 to be significantly down regulated in HP samples, specifically in ER-negative samples. Both miRNAs reside on 1q24.3 and were previously shown to be downregulated in ER+ samples [24] and to induce cell survival by targeting PTEN and subsequently activating the Akt pathway in ovarian cancer [48]. Overall, we observed a higher number of miRNAs with lower expression in HP than in LP samples. This is consistent with the higher number of lower expressed miRNAs in ER-negative than in ER-positive tumor samples observed by others [24,49].

As differential expression does not imply causality, we carried out functional assays to validate and characterize the effect of individual miRNAs on proliferation. In the cases of miR-130b and miR-19a, up regulated in HP (with positive correlation with the cell-cycle genes), and miR-449a, miR-29c, miR-154 and miR-145, downregulated in HP (with negative correlation with the cell-cycle genes), the effect of miRNA over-expression on proliferation was confirmed in cell lines. We therefore propose that for these miRNAs the association with proliferation is not only manifested in the transcriptomics level but rather they are likely to be drivers of the process. The strongest effect was seen for miR-449a. In prostate cancer cell lines miR-449a was shown to have growth suppressing activity partly through inhibition of HDAC-1 expression [50]. The expression patterns of miR-449a and of HDAC-1, in our cohort, are anti-correlated (Pearson’s r = -0.26) and it might be that the mechanism is similar in breast cancer. It is therefore of interest to characterize its relationship to proliferation and assess its therapeutic potential. Consistent with our findings, this recent study [50] also reports miR-145 to inhibit proliferation. We note that Blenkiron et al. [24] observed higher expression of miR-145 in luminal-A samples.

There is an apparent disagreement between the functional proliferation assay results for several other miRNAs (e.g. miR-18a/b). Since association does not imply causality this apparent disagreement is, in fact, expected. In addition, this disagreement can be a result of the complexity of transferring observations from individual cell-lines to clinical tumor cohorts. For the case of miR-18a/b we note that [27] showed that this miRNA and the miRNA cluster it resides in (miR-17-92 cluster) are activated by E2F. Therefore, it is reasonable to expect its expression level to be driven by the cell-cycle process rather than to be a determinant of that process.

Analyzing miRNA expression with respect to extracellular matrix component signature of the studied cohort, we found miR-29c to be the most prominently differentially expressed. It is under-expressed in the ECM1 class and over-expressed in the luminal-A subtype. Predicted targets of the miR-29 family are enriched with genes associated with cell-adhesion and show significant anti-correlation to the expression of miR-29c which points to a direct involvement of miR-29c in regulating cell-adhesion. Over-expression of the miR-29 family was shown to revert aberrant methylation patterns in lung cancer [51], and recently it was shown that miR-29c can induce apoptosis in a TP53 dependent manner [52]. In our cohort we also found miR-29c to be significantly under-expressed in proliferative samples (Table S5), which may suggest breast tumor suppressive activity mediated by the regulation of the ECM related genes.

We have run univariate Cox analysis to assess the association of miRNAs and survival in the entire cohort. We have not found any significant association of any of the tested miRNAs, after correcting for the multiple testing. We do find miR-150 to be associated to survival in part of the cohort as described above.

Conclusion

We introduce a dataset of mRNA and miRNA expression profiles measured in a well studied patient cohort. We show that miRNAs can distinctly differentiate between tumor subtypes and various clinical sub-classifications. In addition, we present experimental support linking some miRNAs to proliferation. Finally, we show that miRNAs can act as reliable proxies to the activity of known biological processes related to breast cancer progression such as cell-cycle, immune response and cell adhesion.

Materials and Methods

Patient characteristics and classifications

The 101 breast cancer patients in this study are part of a cohort previously described [53]. The study was approved by the Norwegian Regional committee for medical research ethics, Health region II (reference number S-97105). All patients have given written consent for the use of material to research purposes. Total RNA isolation was performed using TRIzol (Invitrogen) as described previously [54]. The mRNA expression derived subtype classification has previously been performed and presented [8]. The same is true for the extracellular matrix (ECM) based classification [29].

Expression profiling

miRNA microarray hybridization. miRNA profiling from total RNA was performed using Agilent Technologies “Human miRNA Microarray Kit (V2)” according to manufacturer’s protocol. Scanning on Agilent Scanner G2565A and Feature Extraction (FE) v9.5 was used to extract signals. Excluding two samples, experiments were performed using duplicate hybridizations (99 samples) on different arrays and time points. miRNA signal intensities for replicate samples were averaged and log2 transformed. The expression levels were normalized to the 75th percentile. That is, the expression levels in each sample, i, were multiplied by a constant, ci such that the 75th percentile of the expression levels in that sample will equal to a constant e – the 75th percentile in the entire dataset. miRNA expression status was scored as present or absent for each gene in each sample by default settings in FE v9.5. miRNAs in samples that were run in replicates were considered present if scored in one of the two arrays. The microarray contains probes for 76 viral and 723 human miRNAs (based on miRBase v10.1). We filtered out all miRNAs that were detected in less than 10% of the samples. This filtering resulted in 489 miRNAs considered to be expressed in this set of human breast tumors and used in further analysis steps. For these miRNAs, all expression values were used for further analysis. Thus, no missing values were used. The miRNA expression data is MIAME compliant and have been submitted to the Gene Expression Omnibus (GEO) with accession number GSE19536.

miRNA expression using RT-PCR. Quantification of nine selected mature miRNAs was performed on 20 samples with TaqMan® MicroRNA Assays (Applied Biosystems). The miRNAs selected for this validation were miR-17-5p, miR-18a, miR-18b, miR-19a, miR-29c, miR-34c-5p, miR-142-3p, miR-150 and miR-449a, and the endogenous control used was RNU6B. RT-PCR
mHG p-values presented herein are exact and do not require correction for multiple thresholds tested [59].

In this paper we use GO terms [60], MSigDB sets [61] and miRNA target sets [30] to define different sets of genes (playing the role of H above). GO enrichment analysis was performed using the GOrilla web tool [62], which takes as input a ranked list of genes.

The following ranking schemes are used in the analyses performed in this study (see Figure S7): (A) Differential expression – using TNoM or other scores, as described above. (B) Correlation to pivot miRNA – for a specific miRNA, called the pivot miRNA, miRNAs were ranked according to the correlation of their expression pattern across the entire cohort (101 samples) to the pivot miRNA expression pattern. To obtain the results presented here we used Spearman’s correlation. (C) miRNA target prediction – for a specific miRNA, genes were ranked according to how likely they are to be targeted by the miRNA. Context score values, taken from TargetScan V5.1 target prediction tool [30], were used as prediction scores.

In the case of miRNA target enrichment, the genes were ranked according to their anti-correlation to a pivot miRNA. The top 2000 targets of the pivot miRNA were then tested for enrichment amongst the top anti-correlates. Enrichment in this case indicates a significant anti-correlation between the expressions of the miRNA and the expression of its targets.

Survival analysis. The univariate Cox scores for each of the 489 miRNAs was calculated using SAS. The log-rank calculation for miR-150 was performed using Matlab version 2008b (Mathworks, Inc).

Assessment of proliferation

Formalin-fixed paraffin-embedded tissue was available from 93 of the 101 patients as part of a Tissue Micro Array (TMA). Briefly, the TMA was composed of three 0.6 mm cores from each of the tumor specimen assembled in a recipient paraffin block using a manual device from Beecher Instruments, Silver Spring, USA. Immunohistochemistry (IHC) was performed with antibody directed towards Ki67 (MB1 from DAKO, diluted 1:100) using the Envision+ detection system (DAKO). The slides were scored visually using a conventional microscope and grouped by level of positivity into three groups; negative (<1% stained cells), moderate (1-10%) and high (>10%) regardless of intensity. Samples with no interpretable cores were excluded (three patients). The highest value was used if the three cores from a patient showed discrepant scores. In total, 90 patients had interpretable Ki67 IHC staining; 29 were classified as negative, 39 classified as moderate and 22 classified as high Ki67 (summary of scoring results presented in Table S10) [33].

Whole tissue sections were available from 92 patients, and the mitotic count was assessed by microscopy as part of estimating the histological grade. Briefly, the number of mitoses in 10 high power fields were counted and the samples were classified as having low (0–5 mitoses), moderate (6–11 mitoses) or high (>11 mitoses) mitotic index (MI). Out of 95 samples, 39 had low MI, 21 had moderate MI and 32 had high MI.

Proliferation groups were created by dividing the samples in two groups. Samples with a high score on both Ki67 and mitotic index or high and moderate were considered highly proliferative (24 samples). Samples that scored low/negative on both Ki67 and mitotic index or low and moderate were considered weakly proliferative samples (35 samples) (Table S10).
Pre-miR transfections and scoring

MCF-7 cells were obtained from Interlab Cell Line Collection (ICLC, Genova, Italy) and BT-474 from American Type Culture Collection (ATCC, Manassas, VA, USA). For lysate microarray (LMA) screening, the MCF-7 and BT-474 cells were transfeced with 20 nM human Pre-miRTM miRNA Precursor library v2 (Ambion Inc., Austin, TX) as previously described [28]. Thereafter, the cells were lysed and printed on nitrocellulose-coated microarray FASTTM slides (Whatman Inc., Florham Park, NJ). KI67 was detected by staining the slides with Ki67 antibody (#M7240, Dako, Glostrup, Denmark) followed by exposure to Alexa Fluor 680 -tagged secondary antibody (Invitrogen Inc., Carlsbad, CA). For total protein measurement, the arrays were stained with Sypro Ruby Blot solution (Invitrogen Inc.). The slides were scanned with Tecan LS400 (Tecan Inc., Durham, NC) microarray scanner and Odyssey Licor IR-scanner (LI-COR Biosciences, Lincoln, NE) to detect the Sypro and Ki67 signals, respectively. Array-Pro Analyzer microarray analysis software (Median Cybernetics Inc., Bethesda, MD) was used for analyzing the data. For each miRNA the signal intensity was normalized to a negative control miRNA (z-score). miRNAs that gave at least one hit with a z-score > 2 or <-2 as well as both z-scores > 1 or <-1 at both 48 h and 72 h were considered as having an effect on proliferation.

Supporting Information

Figure S1 Expression profiles of miR-9*, miR-29c and miR-190b in subtypes. Expression profiles (signal intensities) of miR-9*, miR-29c and miR-190b ordered by subtypes. Note that Y-axes in different panels are in different scales. (EPS)

Figure S2 Proliferative miRNAs versus cell-cycle related enrichments. The plot shows the scores for each miRNA, where each miRNA is represented by a dot. On the Y-axis the differential expression score is –log(p-value) if the miRNA is upregulated and log(p-value) if the miRNA is down regulated, yielding positive and negative scores according to differential expression between the high and low proliferative groups. Significance of differential expression is calculated using TNoM as described in Materials and Methods. (A) “Positive regulation of mitotic cell cycle” score is –log(p-value) if CC genes are enriched in the miRNA positively correlated genes, and log(p-value) if CC genes are enriched in the miRNA negatively correlated genes. P-value for CC enrichment is calculated using the mHG statistic as described in Materials and Methods. Here CC genes are those annotated in GO as ‘positively regulation of mitotic cell cycle’. (B) “Negative regulation of S phase of mitotic cell cycle” scores are calculated in the same manner as “Positive regulation of mitotic cell cycle” scores. The difference is in the definition of CC genes. Here we use genes annotated in GO as “Negative regulation of S phase of mitotic cell cycle”. (EPS)

Figure S3 Association of miR-29c with extracellular matrix. (A) Extracellular matrix miRNA differential expression. miRNAs ordered by significance of differential expression between two reciprocal extracellular matrix classes ECM1 and ECM4. For visualization expression values of each miRNA were linearly stretched. miR-29c shows the highest significance of differential expression between the two classes (TNoM p<4E-5, see Table S8 for full list). (B) miR-29c is anti-correlated to its miRNA targets. All miRNAs were ranked according to their anti-correlation to miR-29c expression profile. The absolute signal intensities of miR-29c are presented in the top bar. The top 50 anti-correlated genes are ordered from top to bottom. We find a significant enrichment of miR-29c targets, as derived from TargetScan V5.1, in the anti-correlated gene ranking (mHG p<3E-11). The color bar beneath the heatmap illustrates the different subtypes. (C) GO enrichment in miR-29c anti-correlated genes. The figure depicts the GO enrichment result, as carried out by GOriila [62]. We observed an enrichment of terms related to extracellular matrix (e.g. cell-adhesion). (EPS)

Figure S4 GO enrichment in miR-29c targets. Target prediction context scores of miR-29c were taken from TargetScan V5.1. We find enrichment of several GO terms in the high scoring genes, with respect to miR-29c targets, using GOriila web tool [62]. The graph is color coded according to degree of enrichment. (EPS)

Figure S5 miR-150 and survival. ER-/non-ERBB2 enriched patients from our cohort were divided to two groups: patients with high expression of miR-150 (above the average in the entire cohort which was 1041) and patients with low expression of miR-150 (below average). Using a log-rank test we found that high expression of miR-150 is predictive of better prognosis (log-rank p<0.085) in the ER-/non-ERBB2 enriched patients. (EPS)

Figure S6 RT-PCR analysis of miRNA expression. Each panel displays Agilent expression (vertical axis) versus negative TaqMan expression (horizontal axis) for a miRNA on all 20 samples. Two lines are shown in each panel: the least squares fit to the data (green) and a robust regression line found by iteratively reweighted least squares with a bisquare weighting function (magenta). Pearson’s correlation coefficients and corresponding p-values are shown above each panel. (EPS)

Figure S7 Workflow. Schematic overview of the data analysis methods applied. (A) Differential expression – using TNoM or other scores. (B) Correlation to pivot miRNA – for a specific miRNA, called the pivot miRNA, miRNAs were ranked according to the correlation of their expression pattern across the entire cohort (101 samples) to the pivot miRNA expression pattern. (C) miRNA target prediction – for a specific miRNA, genes were ranked according to how likely they are to be targeted by the miRNA. A, B and C all lead to ranked lists of genes which are analyzed using the mHG method. (EPS)

Table S1 Characterization of the samples. (XLS)

Table S2 miRNAs differentially expressed between basal-like and luminal-A like samples. (XLS)

Table S3 miRNAs differentially expressed between TP53 WT and mutant and between ER-positive and ER-negative. (XLS)

Table S4 Enrichment levels of miRNA-GO associations. (XLS)

Table S5 miRNAs significantly associated with proliferation. (XLS)
Table S6  Proliferation associated miRNAs in ER-positive and ER-negative samples.  
(XLS)

Table S7  LMA screen of Ki67.  
(XLS)

Table S8  miRNAs differentially expressed between ECM1 and ECM4.  
(XLS)

Table S9  miRNAs associated with survival.  
(XLS)

Table S10  Proliferation scoring scheme.  
(XLS)

References


Acknowledgments

We are grateful to Elin Borgen, MD for the histological grading including mitotic index of all tumors and for training HGR in visual scoring of Ki67 IHC. We also acknowledge Tove Anita Slyngstad for the construction of the TMA and for performing the IHC procedure for Ki67. Thanks to Ivan Potapenko for computational assistance. We thank Anke van den Berg for discussion and very useful suggestions.

Author Contributions

Conceived and designed the experiments: EE IS OK VNK ZY A-LB-D. Performed the experiments: EE KK S-KL MRA HGR JAR HJ RM BN MP. Analyzed the data: IS EE RN ER. Wrote the manuscript: EE IS OK VNK ZY A-LB-D. Designed computational tools used in analysis: IS ZY.

Integrated RNA Analysis of Human Breast Tumors
3.2 miRNA Target Enrichment Analysis Reveals Directly Active miRNAs in Health and Disease

miRNA target enrichment analysis reveals directly active miRNAs in health and disease

Israel Steinfeld¹*, Roy Navon², Robert Ach³ and Zohar Yakhini¹,²

¹Computer Science Department, Technion—Israel Institute of Technology, Haifa 32000, ²Agilent Laboratories, Agilent Technologies, Tel Aviv 49527, Israel and ³Agilent Laboratories, Agilent Technologies, Santa Clara, CA 95051, USA

Received April 18, 2012; Revised October 23, 2012; Accepted October 24, 2012

ABSTRACT

microRNAs (miRNAs) are short non-coding regulatory RNA molecules. The activity of a miRNA in a biological process can often be reflected in the expression program that characterizes the outcome of the activity. We introduce a computational approach that infers such activity from high-throughput data using a novel statistical methodology, called minimum-mHG (mmHG), that examines mutual enrichment in two ranked lists. Based on this methodology, we provide a user-friendly web application that supports the statistical assessment of miRNA target enrichment analysis (miTEA) in the top of a ranked list of genes or proteins. Using miTEA, we analyze several target prediction tools by examining performance on public miRNA constitutive expression data. We also apply miTEA to analyze several integrative biology data sets, including a novel matched miRNA/mRNA data set covering nine human tissue types. Our novel findings include proposed direct activity of miR-519 in placenta, a direct activity of the poorly characterized miR-768 in both healthy tissue types and cancer cell lines. The miTEA web application is available at http://cbl-gorilla.cs.technion.ac.il/miTEA/.

INTRODUCTION

microRNAs (miRNAs) are short (usually ~22 nt) non-coding regulatory RNA molecules. Hundreds of miRNAs have been discovered in recent years and several have been functionally characterized (1). In mammals, miRNAs are well known to take part in regulating tissue differentiation (2) and for several miRNAs a well-defined tissue specific signature is known (3). As such, miRNAs are known to regulate major biological processes such as development, cancer (4,5) and heart function (6).

Many studies attempted to elucidate the mechanism by which miRNAs act to regulate target genes. With recent experimental studies, many of the major factors that partake in the recognition mechanism of miRNA targeting have been revealed (7,8). Better understanding of the miRNA mechanism of regulation led to the development of a large variety of computational tools designed to predict which genes are targeted by any miRNA of interest (9). We herein refer to these tools as miRNA target prediction algorithms (miTPAs). The current study uses the publicly available predictions of several such miTPAs.

The refined characterization of miRNA targets enables better understanding of the role of miRNAs in various biological processes by combining measurement in relevant samples with analysis that takes information about targets into account. Several groups have developed computational tools to infer miRNA activity by analyzing their targets in mRNA transcription profiles (10–15). Sood et al. (10) used the Pictar miTPA (16) and defined cell-type specific signatures of miRNAs by searching for enriched miRNA targets in expression profiles using the Wilcoxon rank sum test. Cheng and Li (11) used the miRanda miTPA (17) to identify miRNA activity in miRNA transfected HeLa cells. The statistical approach of Cheng and Li (11) employs a generalization of the enrichment score used by GSEA (18) which requires a permutation step to infer the significance level. Liang et al. (14) have used TargetsCan (19) to report miRNA activity in a breast cancer data set and in miRNA transfected HeLa cells. Their method assesses miRNA activity in every sample in the cohort. These per-sample activity scores, based on t-test variants, are then compared for different classes of samples in the cohort, using variants of the Kruskal–Wallis test. In general, current approaches, including the methods mentioned earlier, work with a fixed set of genes as the target set of a specific miRNA, either by using miTPA predictions or simply by considering the miRNA
seed sequence (20, 21). Subsequently, the expression distribution of this set is tested for divergence compared to the overall transcription profile. When significant divergence in observed the miRNA itself can be deduced to be active. A few groups have also provided online services to infer miRNA activity (14, 15). While miRNA targets behave in a coherent manner in synthetic and controlled environments, where the miRNA is artificially induced or repressed, their coordinated activity is obscured in most biological conditions. Therefore, there is a need for more sensitive analysis approaches that can capture more subtle trends in the data.

In accordance with the notion that transcription networks are not discrete and should be modelled as quantitative relations (22) and to enable the performance of more accurate and sensitive analysis of miRNA target enrichment, we avoid the use of a predefined set of mRNAs to describe the targets of a specific miRNA. We leverage the additional information provided by the miTPA, in the form of a prediction score or significance level. We thereby produce better predictions of miRNA activity, including subtle trends as earlier. In this article, we address mutual enrichment in two ranked lists of elements and develop a statistical framework and software for this purpose, including miRNA target enrichment analysis (miTEA)—a publically available web-based application. We apply our methods to synthetic expression profiles to show the strength of our approach as well as to assess the robustness level of different miTPAs, in the context of these specific experiments. We further demonstrate an analysis of an integrative biology high-throughput data set of mRNA and miRNA expression profiling and shed light on several aspects of miRNA mechanisms of regulation in healthy human tissue data sets as well as in cancer data sets.

**MATERIALS AND METHODS**

**Expression profiling**

Protein and mRNA expression profiles for miRNA transfected HeLa cells were taken from (23, 24). Matched mRNA and miRNA expression profiles for cancer cell lines and primary tumours were taken from (25, 26), respectively. mRNA profiles from nine human tissue types were described in Ach et al. (27). For mRNA profiles from these tissue types, the same total RNA preps used for miRNA profiling in Ach et al. (27) were labelled in duplicates and profiled on human whole genome gene expression DNA microarrays from Agilent Technologies, according to the manufacturer’s protocols (www.agilent.com). Data were deposited in GEO with the accession number GSE31904.

Repeat measurements of the same tissue type were averaged resulting in one profile for each tissue type. To obtain a tissue type-specific ranked list of genes, the expression of each gene was standardized (mean = 0, SD = 1) across samples and for each tissue type genes were ranked according to their standardized expression signal.

**Minimum-mHG**

The mmHG statistics is a generalization of the mHG statistics (28–32). While the mHG statistics quantifies the enrichment level of a set of elements in the top of a ranked list of elements, the mmHG statistics quantifies the mutual enrichment level for two ranked list of elements. As such, the mmHG statistics is applicable to any two ranked lists of common elements. While any parametric or non-parametric correlation statistics (e.g. Spearman’s correlation coefficient), that takes the same input, calculates the overall agreement between the two ranked lists, the mmHG statistic focuses only on agreement at the top of the two ranked lists. The mmHG calculates how many elements are common in the top of both lists, without predefining what the top is. Its output is simply the chance for getting the obtained size of intersection at the top of the two ranked lists of elements at random (the enrichment P-value). In this article, we assess the mutual enrichment in two ranked lists of genes—one ranked according to an expression-based measurement and one according to a miTPA score. Another example of an application is to assess mutual enrichment when genes are ranked according to differential expression in two types of disease or other biological condition.

As our main application will be for genes we will, from here on, focus on genes as the ranked elements. A formal definition of the mmHG statistics follows. Given a single permutation \(\pi \epsilon S_N\) and for every \(i = 1 \ldots N\) we define a binary vector \(\lambda_i\) in which exactly \(i\) entries are 1 and \(N - i\) are 0, as follows:

\[
\lambda_i(j) = 1 \text{ iff } \pi(j) \leq i
\]  

We define the mmHG score of a permutation \(\pi\) as:

\[
\text{mmHG}(\pi) = \min_{1 \leq \lambda \leq N} P\text{-value}(\text{mHG}(\lambda, \pi))
\]

We have previously introduced the mHG statistics to evaluate the enrichment of a fixed gene set within a ranked list of genes (28, 29). For completeness, we define the mHG score of a ranked binary vector \(\lambda\) as:

\[
\text{mHG}(\lambda) = \min_{1 \leq B \leq N} HGT(N, B, n, b_n(\lambda))
\]

Where,

\[
N = |\lambda|, b_n(\lambda) = \sum_{i=1}^{n} \lambda_i, B = b_N(\lambda)
\]

and

\[
HGT(N, B, n, b) = \Pr(X \geq b) = \sum_{i=b}^{\min(n, b)} \binom{n}{i} \left( \frac{N - n}{B - i} \right)
\]

is the tail probability for a random variable \(X\), having an appropriate hypergeometric distribution. mHG P-values used earlier, denoted \(P\text{-value}(\text{mHG}(\lambda))\), assume a uniform null distribution of binary vectors with weight \(B\). These P-values are exact and do not require correction...
for the multiple thresholds tested. See details of the underlying methodology in Eden et al. (28).

For two permutations, \( \pi_1 = (\pi_1(1), \ldots, \pi_1(N)) \) and \( \pi_2 = (\pi_2(1), \ldots, \pi_2(N)) \), over a universe of \( N \) genes, the relative permutation \( \pi \), of \( \pi_2 \) w.r.t. \( \pi_1 \), is defined by
\[
\pi(\pi_2(j)) = \pi_1(j), \quad \text{for } j = 1 \ldots N \text{ or simply, using operations in the permutation group } S_N:
\]
\[
\pi = \pi_2 \cdot \pi_1^{-1}
\]
(6)

We are now ready to define \( \text{mmHG}(\pi_1, \pi_2) \) representing the mutual enrichment of two ranked list of genes as:
\[
\text{mmHG}(\pi_1, \pi_2) = \text{mmHG}(\pi)
\]
(7)

where \( \pi \) is the relative permutation of \( \pi_2 \) w.r.t. \( \pi_1 \) as defined earlier.

Although the \( \text{mmHG} \) score is a minimum taken over a space of \( \text{mHG} \) \( P \)-values, it cannot be considered as a significance measure, due to the additional multiple testing. To assess the statistical significance of an \( \text{mmHG} \) result obtained in actual analysis, we work over a null model that consists of a uniform measure on the group \( S_N \). This is equivalent to a uniform measure over \( S_N \times S_N \) by the conversion of any \( \pi_1, \pi_2 \) to their relative permutation \( \pi = \pi_2 \cdot \pi_1^{-1} \) as earlier.

For \( \text{mHG} \), there is a dynamic programming process that provides a full characterization of the distribution of the statistics under the null model. For \( \text{mmHG} \), we do not currently have an efficient process that allows for calculating exact \( P \)-values. We can however use a union bound approach as follows:
\[
P-value(\text{mmHG}(\pi)) \leq \text{mmHG}(\pi) \times N
\]
(8)

### mmHG for miRNA target enrichment in this study

In this article, we use gene scores given by any miTPA to define gene ranking \( \pi_1 \). For a specific miRNA, genes are therefore ranked according to how likely they are to be targeted by that miRNA. The following ranking schemes are used in the analyses performed in this study to define the gene ranking \( \pi_2 \):

(A) Protein expression values and related changes, derived from SILAC/pSILAC experiments.

(B) mRNA expression values derived from high-throughput transcription profiling and

(C) In integrated miRNA–mRNA data—for a specific miRNA, called the pivot miRNA, mRNAs were ranked according to the anti-correlation level (Pearson’s \( r \)) of their expression patterns, across the entire cohort of samples, to the pivot miRNA expression pattern. See Enerly et al. (26) for an example of a similar analysis.

When using any specific miTPA we reduce the gene universe to the genes for which the miTPA reports results (genes targeted by any miRNA). In addition, for each expression experiment the gene universe was further reduced to the genes that were actually measured in the experiment and reported to have produced an interpretable signal. The same is true for the protein expression experiments.

### Practical implementation

Some modifications to the abstract definition of mmHG were used in this study, to accommodate practical considerations. These are described in this section. First, instead of computing the \( P \)-value of an \( \text{mHG} \) result as required in Equation (2), we used the following bound which follows from Eden et al. (28):
\[
P-value(\text{mmHG}(\lambda_i)) \leq \text{mHG}(\lambda_i) \cdot i
\]
(9)

Thus, instead of computing the \( \text{mmHG} \) score we define the \( \text{mmHG} \) score to be:
\[
\text{mmHG}^*(\pi) = \min_{1 \leq i \leq N} \text{mHG}(\lambda_i) \cdot i
\]
(10)

Notice that from Equation (9) it follows that:
\[
\text{mmHG}(\pi) \leq \text{mmHG}^*(\pi)
\]
(11)

and thus from Equation (8) we get:
\[
P-value(\text{mmHG}(\pi)) \leq \text{mmHG}(\pi) \cdot N \leq \text{mmHG}^*(\pi) \cdot N
\]
(12)

As a second practical modification, the optimization processes described in Equations (2) and (3) are not carried out in their full exhaustive scope. We only run the outer loop up to a reasonable threshold—\( n_{\text{max}} \).

The altered equations are therefore:
\[
\text{mmHG}_{n_{\text{max}}}(\pi) = \min_{1 \leq i \leq n_{\text{max}}} P-value(\text{mmHG}_{n_{\text{max}}}(\lambda_i))
\]
(13)

\[
\text{mmHG}_{n_{\text{max}}}(\lambda) = \min_{1 \leq n \leq n_{\text{max}}} \text{HGT}(N, B, n, b_0(\lambda))
\]
(14)

\[
P-value(\text{mmHG}_{n_{\text{max}}}(\pi)) \leq \text{mmHG}_{n_{\text{max}}}(\pi) \cdot n_{\text{max}}
\]
(15)

Defining the bound-based version:
\[
\text{mmHG}^*_{n_{\text{max}}}(\pi) = \min_{1 \leq i \leq n_{\text{max}}} \text{mmHG}_{n_{\text{max}}}(\lambda_i) \cdot i
\]
(16)

we now get:
\[
P-value(\text{mmHG}_{n_{\text{max}}}(\pi)) \leq \text{mmHG}^*_{n_{\text{max}}}(\pi) \cdot n_{\text{max}}
\]
(17)

For all parts of this study, we used the bound presented in the right hand side of Equation (17) to represent the significance level of an \( \text{mmHG} \) result. The bound used in the analysis was \( n_{\text{max}} = 2000 \) excluding the integrative analysis part which used \( n_{\text{max}} = 3000 \).

Since the complexity of computing the \( \text{mHG} \) \( P \)-value is \( O(N^2) \), we get that the complexity of computing \( \text{mmHG}(\pi) \) is \( O(N^3) \). Using the bounds presented earlier, we get a reduced complexity of \( O(N^2) \) for computing \( \text{mmHG}^*(\pi) \) or \( O(n_{\text{max}}^2) \) if we consider Equation (17).

### miRNA mutual exclusion

A scored list of miRNAs were mutually excluded according to miRNA family or similar 6-bp seed, as obtained from miRBase (1), giving precedence to high-scoring miRNAs.
Tool description (website)

A web-based application miTEA is publically available, free for non-commercial use, at http://cbl-gorilla.cs.technion.ac.il/miTEA/

miTEA is designed to statistically deduce miRNA activity from high-throughput measurement results. The input for miTEA is a ranked list of genes for which the mmHG statistics is used. The miTPA ranking is embedded into miTEA and is updated periodically. The web interface for miTEA is shown in Supplementary Figure S1. miTEA supports the following formats of gene nomenclature: gene symbol, protein/gene RefSeq, Uniprot, Unigene or Ensembl. miTEA automatically removes gene redundancies keeping the highest ranking occurrence (in each of the two lists, when applicable). This includes dealing with duplicates that hide behind different nomenclatures.

Simulated data

Each simulated data set, $E$, was composed of 100 samples, where in each sample we generated a simulated expression profile spanning all 17 297 genes that are available in Targetscan (V5.2). The expression level of each gene in each sample was independently randomly drawn from a standard normal distribution, $E(g,s) \sim N(0,1)$ for gene $g$ and sample $s$. In each simulated data set 10 random miRNAs were simulated to be active at different levels of influence: the first repressing its top 100 targets, the second repressing its top 200 targets, . . . , the last repressing its top 1000 targets. In addition, for each simulated data set, the level of activity, $\alpha$, of the miRNAs is changed. The miRNA repression was simulated for only 50 of the 100 samples. The expression levels of the miRNA targets in the affected samples were reduced by $(\alpha + \varepsilon)$, where $\varepsilon$ is drawn from a standard normal distribution. That is, $E(g,s) \sim (N(0,1) - (\alpha + N(0,1)))$ for a target $g$ of an active miRNA, in one of the affected samples $s$. For each level of $\alpha$, three simulated data sets were generated.

While mirAct uses the entire data set as input, DIANA-mirExTra takes only a target set and background set of genes as input and miTEA uses a ranked list of genes as input. Therefore, when testing the simulated data set using miTEA and DIANA-mirExTra genes were ranked according to their $t$-test $P$-value down-regulation in the affected samples. For DIANA-mirExTra, genes with $P < 0.05$ were assigned to the target set and the rest to the background set. The use of miTEA, mirAct and DIANA-mirExTra was carried out with default parameters as provided in the online versions.

RESULTS

Web-based miTEA tool

In this study, we introduce miTEA, an online web application designed to detect and measure miRNA regulatory activity in high-throughput measurement results. miTEA takes as input a ranked list of genes that represents the results of a high-throughput measurement experiment. It employs the output of any selected miTPA to define, in a robust manner, a ranked list of genes representing the targets of each miRNA in the relevant organism. Using mmHG, a novel statistical enrichment method (see ‘Materials and Methods’ section), miTEA finds mutual enrichment in the two ranked lists of genes and thus statistically infers miRNA activity (see Figure 1 and ‘Materials and Methods’ section for more details). The main advantages of miTEA, compared to other similar tools (10–15) are:

- The use of all prediction scores given by the miTPA rather than setting a threshold and predefining the set of genes targeted by a specific miRNA.
- A novel statistical approach that provides an assessment of the statistical significance of the obtained results without using simulations.
- Its availability as an efficient web tool easily accessible to the community.

Evaluation through a comparison of miTPAs

As a benchmark for miRNA activity, we first tested miTEA on two seminal studies that measured global miRNA as well as protein expression profiles in response to a perturbation of a single miRNA, in a controlled environment (24, 23). The data span 8 protein expression profiling and 13 mRNA expression profiling results following constitutive miRNA over-expression experiments. To better assess the sensitivity of our approach, we applied miTEA using eight different miTPAs: Targetscan (v5), microCosm (v5), PITA, PicTar, MicroT (v3), RNA22, targetRank and EIMMo (v4) (1, 16, 19, 33–37). In addition, we also applied miTEA using a Targetscan version of only conserved sites (Figure 2). In each of the experiments, all genes were ranked according to their down-regulation and these ranked lists of genes were used as input to miTEA (Figure 1). In the majority of the experiments and for many of the miTPAs, miTEA detects a significant enrichment of the targets of the perturbed miRNA in the down-regulated genes—as expected (Figure 2). Furthermore, miTEA finds the perturbed miRNA targets to be most enriched compared to targets of all other non-perturbed
miRNAs in most of the experiments (Figure 2). In particular, when the miTPA used was Targetscan miTEA detected the correct perturbed miRNA in 20 out of the 21 experiments tested.

The high range of enrichment scores obtained for different miTPAs could be a result of the different level of accuracy in the target predictions provided by the miTPAs. Similarly, different levels of noise and robustness can explain the range of enrichment scores obtained for the different experiments. In this respect, two additional interesting insights arise. First, Targetscan shows the best specificity and sensitivity for miRNA detection by yielding the highest enrichment scores in most experiments. The proteomics work was done using pSILAC (23) and SILAC (24) technologies. A second insight is that the enrichment scores in the pSILAC data are more significant than the enrichment scores in the SILAC data. In particular, the two studies measure response to miR-1 over-expression. The targets (as predicted by Targetscan) of miR-1 are enriched in the pSILAC data with mmHG $P < 10^{-71}$ while in the SILAC experiment they are enriched with mmHG $P < 10^{-9}$.

Comparing miTEA to existing online tools

To evaluate miTEA performance in comparison to other existing online tools we compared miTEA to mirAct (14) and mirExTra (15) using a simulated data set as well as using the miRNA over-expression experiments used earlier. The simulated data sets were designed to model the activity of a small set of miRNAs in a subset of the samples in different levels of activity (see ‘Materials and Methods’ section for complete description of data simulation process). Our results show that miTEA is able to detect the active miRNAs with higher level of sensitivity and specificity even in cases of low activity (Figure 3). For the over-expression experiments miTEA is able to find the activity of the perturbed miRNA with higher level of significance—finding the perturbed miRNA to be most active (Supplementary Table S1). For example, for the mRNA profiling data measured 32 h after let-7 transfection (23) mirExTra finds miR-608 to yield better results ($P < 3.1 \times 10^{-9}$, data not shown) than those of let-7. Similarly, in the same experiment, mirAct finds miR-491 to be more active (data not shown). Both are inconsistent
with the experimental set-up, miTEA clearly identifies let-7 as the most active miRNA ($P < 9 \times 10^{-13}$).

### Tissue-specific miRNA activity

The strength of miTEA is very well established in the highly controlled miRNA over-expression experiments described earlier. In the more practical contexts of less controlled data sets, such as clinical ones, we should consider a milder effect of miRNAs on their targets. Therefore, to better capture miTEA sensitivity in detecting miRNA activity, we applied it to mRNA profiling data set of human tissue types measured for the purpose of this study. In this analysis, we applied miTEA in combination with Targetscan (V5.2), since among the different miTPAs the latter demonstrates the highest sensitivity and specificity in detecting miRNA activity compared to the other miTPAs.

![Table](image)

<table>
<thead>
<tr>
<th>Selbach mRNA</th>
<th>miR-1</th>
<th>miR-155</th>
<th>miR-155_32h</th>
<th>miR-16</th>
<th>miR-16_32h</th>
<th>miR-1_8h</th>
<th>miR-1_32h</th>
<th>miR-30a_8h</th>
<th>miR-30a_32h</th>
</tr>
</thead>
<tbody>
<tr>
<td>let-7_8h</td>
<td>1.49</td>
<td>0.00</td>
<td>0.00</td>
<td>0.34</td>
<td>0.00</td>
<td>1.53</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>let-7_32h</td>
<td>12.05</td>
<td>8.01</td>
<td>3.41</td>
<td>0.87</td>
<td>6.14</td>
<td>1.53</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>miR-155_8h</td>
<td>10.11</td>
<td>2.74</td>
<td>4.61</td>
<td>0.44</td>
<td>0.00</td>
<td>7.41</td>
<td>5.59</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>miR-155_32h</td>
<td>63.33</td>
<td>12.44</td>
<td>34.71</td>
<td>7.70</td>
<td>5.51</td>
<td>64.00</td>
<td>55.02</td>
<td>1.13</td>
<td>40.61</td>
</tr>
<tr>
<td>miR-16_8h</td>
<td>10.79</td>
<td>6.82</td>
<td>3.87</td>
<td>1.03</td>
<td>0.00</td>
<td>2.66</td>
<td>6.83</td>
<td>5.55</td>
<td>0.00</td>
</tr>
<tr>
<td>miR-16_32h</td>
<td>46.48</td>
<td>30.21</td>
<td>20.15</td>
<td>4.30</td>
<td>17.75</td>
<td>48.87</td>
<td>39.04</td>
<td>0.78</td>
<td>28.31</td>
</tr>
<tr>
<td>miR-1_8h</td>
<td>39.99</td>
<td>31.04</td>
<td>15.55</td>
<td>7.53</td>
<td>23.91</td>
<td>34.35</td>
<td>28.35</td>
<td>1.88</td>
<td>28.06</td>
</tr>
<tr>
<td>miR-1_32h</td>
<td>75.76</td>
<td>60.88</td>
<td>26.86</td>
<td>8.78</td>
<td>36.05</td>
<td>73.54</td>
<td>66.35</td>
<td>8.49</td>
<td>61.10</td>
</tr>
<tr>
<td>miR-30a_8h</td>
<td>2.94</td>
<td>2.65</td>
<td>2.02</td>
<td>1.41</td>
<td>0.00</td>
<td>4.52</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>miR-30a_32h</td>
<td>47.41</td>
<td>42.13</td>
<td>11.69</td>
<td>7.26</td>
<td>21.86</td>
<td>39.59</td>
<td>40.16</td>
<td>1.64</td>
<td>39.52</td>
</tr>
</tbody>
</table>

Figure 2. Results of applying miTEA using a data set of miRNA transfection experiments. The table provides the enrichment scores [given in $-\log_{10}(\text{mmHG } P\text{-value})$] obtained for two high-throughput data sets of miRNA transfection experiments (23, 24). Each row represents a different experiment, for either transcriptomics or proteomics and each column represents an application of miTEA with a different miTPA. Each row is coloured in gradient to red according to the enrichment score compared to other enrichment scores obtained for the relevant transfected miRNA experiment. Enrichment scores are marked in blue when the targets of the transfected miRNA, in that specific experiment, were not found to be the most enriched compared to targets of other miRNAs. TS_con stands for Targetscan representation of only conserved sites. Higher scores could be due to better target predictions or due to less noisy and more robust profiling experiments. As can be observed, using Targetscan (V5.2), miTEA obtains the highest enrichment scores in 15 out of the 21 experiments and also finds the correct perturbed miRNA in 20 out of 21 experiments. Thus, Targetscan exhibits the highest sensitivity and specificity in detecting miRNA activity compared to the other miTPAs.

The data set we used consist of 20 human tissue samples spanning nine different healthy tissue types of origin (see ‘Materials and Methods’ section). To test miRNA activity in the different tissue types, we ranked all the genes, for each tissue type separately, according to the level of their down-regulation in that specific tissue. Applying miTEA on each of the resulting tissue-specific ranked list of genes, we composed a miRNA activity map for the different tissue types (see Figure 4 and Supplementary Table S2). The results confirmed the specific miRNA activity for miR-124 in brain (38) (mmHG $P < 4 \times 10^{-11}$), miR-122 in liver (39) (mmHG $P < 2 \times 10^{-6}$) and miR-1/miR-206 in skeletal muscle (40, 41) (mmHG $P < 6 \times 10^{-10}$).
Our analysis also suggests an activity of miR-519a in placenta \( \text{mmHG} \times 3 \times 10^{-5} \). Supplementary Figure S2a. miR-519 was observed to be specifically expressed in placenta (42,43). A direct effect on its targets, however, has not been reported so far, to our knowledge.

**DA-miRs in human tissues types**

miRNA regulatory contributions to differentiation are not expected to be manifested only in single tissue specificity, but rather in differences that pertain to the entire differentiation programs that lead to distinct tissue types. Using miRNA expression profiling data from the same cohort of healthy tissue samples, we next took a more exploratory approach to detect complex miRNA activity, as oppose to tissue specific miRNA activity. We selected each miRNA as a pivot and ranked all the genes according to the level of anti-correlation between each gene’s mRNA expression profile and the pivot miRNA expression profile across the cohort of samples. We then applied miTEA, using Targetscan (V5.2), to the above-ranked list to test the enrichment of pivot targets in its anti-correlated genes. We call miRNAs for which such enrichment is observed directly active miRNAs \( (DA\text{-miRs}) \). As expected, our results indicate that the list of \( DA\text{-miRs} \) forms a superset of the tissue-specific active miRNA described earlier. Specifically, we see that of the 470 profiled miRNAs that are also covered by Targetscan database, 71 are found to be \( DA\text{-miRs} \) \( P < 0.05 \) after Bonferroni multiple testing correction, see Supplementary Table S3.
After mutually excluding miRNAs from the same family we end up with 30 DA-miRs found in this data set of human tissue-type samples (Supplementary Table S3). Further examination of these DA-miRs showed that, although found in healthy human tissues, many were either shown to have elevated expression levels in cancer [e.g. miR-135b and miR-182 (44)] or directly implicated in cancer development and metastasis [e.g. miR-15b (45), miR-200 (46) and the oncogenic miR-17/92 cluster (47)]. However, to our knowledge, not all of the DA-miRs have been implicated in association with cancer (e.g. miR-512-3p, miR-768). Their direct activity in different human tissues might imply a role in differentiation, a role that when impaired or disrupted might lead to cancer development.

Study of miRNA regulation in health and disease

The link between the DA-miRs in healthy human tissues and oncogenicity has led us to investigate the miRNA activity map in cancer samples. The NCI-60 panel is composed of 60 cell lines representing nine different cancer types (48). We applied our miTEA-based analysis pipeline presented in the previous section to the NCI-60 matched miRNA–mRNA expression data set (49) to obtain the NCI-60 miRNA activity map. Out of 258 miRNAs that can be commonly mapped in both data sets, 42 and 44 are DA-miRs in the healthy and in the NCI-60 data sets, respectively, with a significant overlap of 14 DA-miRs (hypergeometric \( P \approx 0.004 \)). It is possibly more interesting to note those miRNAs that lose/gain their direct activity when comparing normal to highly proliferative cellular conditions (see Figure 5 and Supplementary Table S5). We find 30 miRNAs that are specifically active in cancer (e.g. miR-142 and miR-29b) and 28 miRNAs that are specifically active in healthy tissue types (e.g. miR-15b and miR-148a). Moreover, deviating from the widely recognized role of miRNAs as repressors, we see a small set of miRNAs (19 in the tissue type data set and 5 in the NCI-60 data set) that are significantly correlated with their targets rather than anti-correlated. For example, miR-377 yields enrichment levels of mmHG \( P < 6 \times 10^{-11} \) and mmHG \( P < 9 \times 10^{-6} \) in the tissue types and NCI-60 data sets, respectively, when considering positive correlates.

DA-miRs in breast cancer

Thus far, we describe miTEA results for heterogeneous data sets composed of several different tissue types. To further develop our understanding of miRNA direct activity and test miTEA in a more homogenous context,
we ran the above workflow on a matched integrative mRNA–miRNA data set of 100 primary breast cancer samples (26). Generating the miRNA activity map for this cohort, we found eight miRNAs from five miRNA families to be DA-miRs, expanding the findings reported in Enerly et al. (26) (Supplementary Table S6). We compared these results to miRNA high-throughput transfection assays that measured cell proliferation in MCF7 breast cancer cell lines (26). We found that all transfected DA-miRs play a direct role in cell proliferation in the cell lines (Supplementary Figure S3).

DISCUSSION

In this study, we introduce miTEA—a framework for miRNA target enrichment analysis. miTEA takes as input a ranked list of genes and then finds miRNAs of which targets are enriched in the top of the input list. miTEA uses a novel statistical analysis that takes into account the rich information available from high-throughput experiments as well as from the different miTPAs. We show how miTEA can be applied to detect miRNA activity in different experiments and shed light into mRNA activity map in healthy and cancer samples. miTEA is a web-based tool allowing the community an easy and efficient free access.

Most results obtained from recent high-throughput measurement technologies are naturally given as ranked lists of genes rather than as fixed sets of genes. The question of mutual enrichment in two ranked lists of genes is thus highly relevant to the analysis of such data sets. The mmHG approach focuses on commonalities in the top ends of the two analyzed lists. Statistical properties of such commonalities are not adequately addressed by other models.

To assess the accuracy and robustness of the miTEA approach, we first applied it to study well-designed and controlled experiments where mRNA and protein expression profiling were performed following over-expression of specific miRNAs. Using Targetscan as the underlying miTPA, miTEA specifically detected the activity of the over-expressed miRNA in 20 out of the 21 tested cases with high significance levels. Among the various miTPAs used in this comparison, Targetscan yielded the most accurate and robust result, even when compared to the version of Targetscan with only conserved sites and thus was used for further analysis in this study. We note that this comparison is limited to the particular data sets analyzed in our investigation. The strong target enrichment results obtained in the mRNA expression experiments shed light on the mode of mRNA regulation by miRNAs. It points to mRNA degradation rather than translation inhibition, which is consistent with several recent studies (24,50), but also justifies the use of miTEA to detect miRNA activity using mRNA expression profiles, as we describe in this article.

We next utilized miTEA to study a data set of healthy human tissue samples and confirmed a direct activity for known tissue-specific miRNAs (e.g. miR-122 in liver). We note that not all miRNAs with tissue specific expression are found to be active (e.g. miR-215 and liver, Supplementary Figure S2b), supporting the need for a robust enrichment analysis. We also observed a direct activity for miR-519a in placenta (mmHG $P < 3 \times 10^{-5}$, Supplementary Figure S2a). miR-519a was previously shown to be specifically expressed in placenta alongside with numerous other miRNAs (42,43). Our finding is, to
functions of the miRNAs that are not related to cancer or to tissue differentiation. Tumour suppression is one such attribute that can explain the cancer-specific direct activity of miRNAs. Indeed many of the miRNAs which we found to be specifically directly active in the NCI-60 data set were shown to be tumour suppressors in various cancer types [e.g. miR-29b/c (52,53), miR-142 (54,55) and miR-101 (56,57)]. It would be interesting to further examine other such miRNAs for their potential role as tumour suppressors (e.g. miR-30e, miR-203). Specific direct activity of miRNAs in the healthy data set can be attributed to specific tissue type regulation. Our results also show that there is a subset of miRNAs that have a significant positive correlation with their targets (19 and 5 in the healthy and cancer cohorts, respectively). It was previously shown that miRNAs can up-regulate translation of their targets in quiescent cells (58) which is in line with our observation in the healthy cohort as compared to the cancer cohort. Positive correlation between miRNAs and their targets and even more so, miRNA roles as activators, has weak support in existing literature. It is also possible that these results represent false positive results in these cohorts. Therefore, more work is required to put our positive correlation results in broader perspective.

miR-768 is located within the sequence of a known snoRNA—HBII-239. Since the snoRNA has a better evolutionary support, miR-768 annotation was discarded from miRBase (1). It is thus not clear whether miR-768 functions as a miRNA. Recent studies have shown that snoRNA-like miRNAs should be examined with care (59). Specifically, an indication for miR-768 expression was observed in HeLa cells (60). In our study, we used miR-768 predicted targets, taken from Targetscan (V5.2) and compared their expression profiles to the measured expression profile of miR-768. Our results clearly show its predicted targets to be significantly anti-correlated to its measured expression profile across the cohort of samples in both data sets of healthy human tissues and cancer cell lines. Thus, we provide evidence not only for its existence but also for its regulatory activity in the two independent sample cohorts.

We also analyzed a homogenous integrated mRNA/mRNA data sets of 100 breast cancer samples. Interestingly, we detected eight DA-miRs in spite of the homogeneity of the data set. To validate direct activity, we compared these results to miRNA transfection assays that measured cell proliferation in MCF7 breast cancer cell lines. The cell lines showed elevated or reduced levels of proliferation when transfected with these breast cancer DA-miRs, in agreement with their observed direct activity in the breast cancer cohort (26). This homogenous data context demonstrates miTEA’s high sensitivity in detecting miRNA activity.

miTEA’s novel statistical approach, called mmHG, enables the detection of miRNA activity without predefining the set of genes targeted by the miRNA nor the set of genes that are over/under expressed in the expression experiment. In this respect, miTEA expands and extends the mHG approach (26,28,30) and uses two ranked lists of genes. The statistical approach used by miTEA finds the two thresholds in the top of the two
ranked lists of genes that maximize the mutual enrichment and thereby allows the data to define the set of affected genes in the experiment and the set of relevant miRNA targets. This approach is bound to be more sensitive than previous approaches that fix the threshold on the prediction score provided by the miTPA or define the miRNA targets according to a 6/7/8-seed match (20,21), as these approaches use a fixed set of genes to characterize the miRNA targets. Similarly, using regression or a simple correlation approach to compare two ranked lists of genes may result in such mutual enrichment being obscured by the majority of the genes further down the lists (see Supplementary Table S1). This notion is exemplified in Figure 1, where fixing a threshold to either define the down-regulated genes in the experiment or to define the genes targeted by the perturbed miRNA in the experiment might result in loss of observed significance. Moreover, a pair of thresholds that optimizes enrichment in one configuration will not be optimal in a different miRNA and expression configuration. Thus, the different thresholds selected in each of the experiments tested much improve the enrichment analysis results, supporting the need for a statistical tool that enables flexible threshold selection. Indeed, when compared to alternative online tools miTEA was able to obtain results with higher specificity and sensitivity in both simulated data sets as well as in a miRNA perturbation data set (Figure 3).

The threshold optimization process obviously introduces statistical multiple testing. We address multiple testing correction through a combination of an exact 


duces statistical multiple testing. We address multiple 
enrichment. Further characterization of the distribution of the 
simulated data sets as well in a miRNA perturbation data 

obtain results with higher specificity and sensitivity in both 

within the mmHG statistics under the null model (a uniform 

enrichment in one configuration will not be optimal in a 

cance. Moreover, a pair of thresholds that optimizes en-

the experiment might result in loss of observed signifi-

by a MicroRNA. Science, 316, 575–579.


ACKNOWLEDGEMENTS

We thank Limor Leibovich (Technion) for useful discussion and anonymous reviewers for their constructive comments.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–6 and Supplementary Figures 1–5.

FUNDING

EU FP7 project GlycoHIT (to I.S., in part). Funding for open access charge: Agilent Technologies.

Conflict of interest statement. None declared.

REFERENCES


12 Nucleic Acids Research, 2012

pairing, often flanked by adenosines, indicates that thousands of human genes are MicroRNA targets. Cell, 120, 15–20.


3.3 Identifying In-Trans Process Associated Genes in Breast Cancer by Integrated Analysis of Copy Number and Expression Data

Identifying In-Trans Process Associated Genes in Breast Cancer by Integrated Analysis of Copy Number and Expression Data

Miriam Ragle Aure1,2,*, Israel Steinfeld3,*, Lars Oliver Baumbusch1,9, Knut Liestøl4,5, Doron Lipson3, Sandra Nyberg1,2, Bjørn Naume6, Kristine Kleivi Sahlberg1,2, Vessela N. Kristensen1,2,7, Anne-Lise Borresen-Dale1,2, Ole Christian Lingjærde2,4,5,*, Zohar Yakhini3,8*

1 Department of Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, Oslo, Norway, 2 K. G. Jebsen Centre for Breast Cancer Research, Institute for Clinical Medicine, University of Oslo, Oslo, Norway, 3 Laboratory of Computational Biology, Computer Science Department, Israel Institute of Technology, Haifa, Israel, 4 BiomedicalInformatics Lab, Department of Computer Science, University of Oslo, Oslo, Norway, 5 Centre for Cancer Biomedicine, University of Oslo, Oslo, Norway, 6 Division of Cancer Medicine and Radiotherapy, Department of Oncology, Oslo University Hospital Radiumhospitalet, Oslo, Norway, 7 Institute for Clinical Epidemiology and Molecular Biology (EpiGen) Akershus University Hospital, Akershus, Norway, 8 Agilent Laboratories, Tel Aviv, Israel

Abstract

Genomic copy number alterations are common in cancer. Finding the genes causally implicated in oncogenesis is challenging because the gain or loss of a chromosomal region may affect a few key driver genes and many passengers. Integrative analyses have opened new vistas for addressing this issue. One approach is to identify genes with frequent copy number alterations and corresponding changes in expression. Several methods also analyse effects of transcriptional changes on known pathways. Here, we propose a method that analyses in-cis correlated genes for evidence of in-trans association to biological processes, with no bias towards processes of a particular type or function. The method aims to identify cis-regulated genes for which the expression correlation to other genes provides further evidence of a network-perturbing role in cancer. The proposed unsupervised approach involves a sequence of statistical tests to systematically narrow down the list of relevant genes, based on integrative analysis of copy number and gene expression data. A novel adjustment method handles confounding effects of co-occurring copy number aberrations, potentially a large source of false positives in such studies. Applying the method to whole-genome copy number and expression data from 100 primary breast carcinomas, 1373 genes were identified as commonly aberrant, 578 were highly in-cis correlated, and 56 were in addition associated in-trans to biological processes. Among these in-trans process associated and cis-correlated (iPAC) genes, 28% have previously been reported as breast cancer associated, and 64% as cancer associated. By combining statistical evidence from three separate subanalyses that focus respectively on copy number, gene expression and the combination of the two, the proposed method identifies several known and novel cancer driver candidates. Validation in an independent data set supports the conclusion that the method identifies genes implicated in cancer.


Editor: Paolo Provero, University of Turin, Italy

Received: June 11, 2012; Accepted: November 22, 2012; Published: January 30, 2013

Copyright: © 2013 Aure et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: (1) Norwegian Research Council (Grant no 193387/50, Understanding breast cancer genomics); (2) Norwegian Cancer Society (Grant no 138296 – PR-2008-0108, Exploring The Systems Biology of Breast Cancer); (3) South-Eastern Norway Regional Health Authority (Project no 2011042, OSBREAC: Towards personalized therapy for breast cancer); (4) K.G. Jebsen Center for Breast Cancer Research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: One author (Zohar Yakhini) is affiliated with Agilent Laboratories, Tel Aviv, Israel. This does not alter the authors’ adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: ole@ifi.uio.no (OCL); zohar_yakhini@agilent.com (ZY)
† These authors contributed equally to this work.

Introduction

Genomic copy number alterations resulting from genomic instability are commonly observed in cancer [1,2]. Substantial effort has been invested in identifying aberration events playing a critical role in the disease development. In breast carcinomas, the genomic architectural changes are diverse and involve various events such as loss and gain of whole chromosome arms, inversions, translocations, and more focal gains and losses [3,4]. Several array comparative genomic hybridization (aCGH) studies of breast tumors and breast cancer cell lines point to commonly observed gains and losses on regions of chromosome 8, 13 and 17 – regions known to contain breast cancer associated genes such as BRCA2, ERBB2 and AHC [5,6,7,8,9].

Recurring aberrations in tumors may be indications of selection driven by changes in the expression of key genes in the affected regions. Since recurrent segmental gains and losses frequently involve several genes, their relative contribution to increased or decreased cell viability and proliferation cannot be inferred from copy number alone. This problem, often portrayed as distinguishing between ‘drivers’ and ‘passengers’, is a key challenge in the task of linking copy number alterations to genes and processes involved in cancer development and progression. One way to proceed would be to focus on genes for which copy number variation
substantially affects gene expression. Integrated analyses of copy number and gene expression data have revealed that the strength of the in-cis correlation between copy number and expression varies extensively between genes [10], and subsets of genes with high correlation have been identified and proposed as candidate driver genes [10,11,12,13,14,15].

It has been suggested that the oncogenic effect of molecular alterations is to cause perturbations at the network level, leading cells to malignant phenotypic states (see, e.g. [16]). Several studies have aimed at identifying pathways and networks perturbed by copy number aberrations, thus establishing associations between genomic profiles and aberrant pathways in cancer [17,18,19], clinical outcome and survival [13,20,21,22,23]. One may ask whether a particular gene through its genomic aberrations has an effect on higher-order phenotypes such as processes, pathways and networks. A natural way to approach this would be to first investigate how other genes are affected by the aberration, and second to study whether any biological processes are overrepresented in the list of affected genes. Following this idea, we propose a workflow for integration of copy number and gene expression data based on the stepwise application of a series of gene selection criteria. The method combines correlation analysis, regression analysis, and gene set enrichment, and to avoid confounding effects, the method adjusts for co-occurring copy number aberrations. A key element of the approach is the direct integration of a statistical enrichment step enabling the assignment of statistical confidence to in-trans associations between genes and biological processes. The resulting genes are referred to as in-trans process associated and cis-correlated (iPAC) genes.

The purpose of combining in-cis and in-trans analyses is here to identify genes that are cis-regulated and for which the correlation structure in the gene expression data provides further support for a role in the alteration of cell phenotype in cancer. The method was applied to a matched data set of aCGH and mRNA expression from 100 well-characterized human primary breast tumors [24,25,26,27], and subsequent application to a second, independent breast cancer cohort showed consistent behavior of the iPAC genes found in the first data set. A small selection of iPAC genes were further studied using siRNA knockdown experiments.

Materials and Methods

Ethics statement

The study was approved by the Norwegian regional committee for medical research ethics, Health region II (reference number S-97103), and patients have given written consent for the use of material for research purposes.

Patient samples and array experiments

Primary breast carcinoma samples from 100 patients previously described as part of the MicMa cohort were used [24]. All samples were fresh frozen and contained at least 40% tumor cells. The majority of the tumor specimens represent tumor size T1/T2, node status N0/N1 (9/11), and histological grade 2 or 3. Tumor DNA was extracted using an ABI 341 Nucleic Acid Purification System (Applied Biosystems, CA, USA) according to the manufacturer’s protocol. Tumor RNA was isolated using TRIzol reagent (Invitrogen, CA, USA) as previously described [28]. The subtype classification deriving from mRNA expression has previously been presented [29]. The aCGH and the mRNA expression data sets have previously been published [26,30]. The expression data (measured using Agilent 4 by 44K one-color oligonucleotide arrays) are available in Gene Expression Omnibus (GEO) with accession number GSE19783 [26], and the copy number data (measured using Illumina Human-1 109K BeadChip SNP arrays) are available on request to OCL. A breast cancer data set from the University of North Carolina, Chapel Hill (UNC), including 73 samples profiled on the same array platforms as described above, was used for validation (see [25,31] for details). Genomic locus annotation for all analyzed data is based on the human genome build hg17.

Figure 1. Workflow of the proposed method to identify iPAC genes. (1) Starting with all genes, the commonly aberrant genes are selected as those that have more than 10% gains or losses; (2) Next, those genes which in addition have an in-cis Pearson correlation above 0.6 are selected and referred to as in-cis genes; (3) Finally, statistical enrichment analysis is performed to assess in-trans functionality, leading to identification of the 56 iPAC genes.

doi:10.1371/journal.pone.0053014.g001
Overview of analysis

The proposed method is based on the stepwise application of a series of gene selection criteria, and a core element is the use of a rigorous statistical enrichment technique to reveal significant associations in trans between the selected genes and biological processes (see Figure 1). This enrichment step is combined with a novel correction method designed to alleviate the problem of co-occurring copy number alterations across the genome.

Segmentation

Copy number data were log2-transformed, and each sample was segmented by fitting a piecewise constant regression function to the data using the piecewise constant fitting (PCF) algorithm in the R package copynumber. A fitted value (“PCF value”) was then obtained for each segment (and was inherited by each probe in the segment) by averaging the log-transformed copy number values for all probes located in that segment. The user controls the trade-off between sensitivity and specificity with a penalty parameter ($c$) and the minimal number of probes per segment ($k_{min}$). We chose $c \approx 70$ which is fairly conservative and thus provides robustness against the presence of potential local (spurious) trends which are common in aCGH data due to varying GC-content and other reasons (see [34] for details), and use the default value $k_{min} = 3$.

Matching copy number and expression values

In order to obtain matching copy number and expression data sets, we first identified all expression probes annotated with a gene symbol in the data set. For each such probe, the copy number probe mapping to the nearest location in the genome was identified. Copy number and gene expression data were then averaged over the corresponding probe values for each gene symbol, resulting in a unique copy number value and expression value for each patient and each gene. The corresponding pair of values was assigned a genomic position by averaging over the locations of the expression probes associated with the gene symbol. Analogous methods are also used in other studies [12,35,36]. This procedure yielded two 25,688 x 100 matrices of copy numbers and corresponding gene expressions, where each row represents a gene and each column a patient sample.

Aberration calling

To call aberrations, a parameter $\theta > 0$ determining the sensitivity of the aberration calling (and hence what is considered a significant aberration) was introduced. Probes with a PCF value larger than $\theta$ were called as gains, and probes with a PCF value less than $-\theta$ were called as losses. Following the recommended practice for threshold selection in the R package copynumber [34], we concluded that $\theta = 0.2$ was an appropriate threshold, which is very similar to the threshold used in [33] where a subset of the copy number data considered in this paper was analyzed.

Identification of common gains and losses

To identify genomic loci where the copy number events are skewed towards either gain or loss, a sign test was applied. Let $n$ denote the total number of samples with an aberration in a particular locus, and suppose $G$ of these aberrations are gains and $L$ are losses (so that $n = G + L$). Modeling the number of gains as a binomial distribution with $n$ draws and success probability $p$, $\text{Bin}(n,p)$, we may formally infer whether gains are overrepresented by testing the null hypothesis $H_0 : p = 0.5$ against the alternative $H_1 : p > 0.5$. Using the difference $D = G - L$ as the test

![Figure 2. Copy number aberrations and in-cis correlations.](image-url)

The frequency of samples with gains (red) and losses (green) is shown at the top. Each gray point shows the level of in-cis correlation between copy number and expression for a particular gene. The chromosomal positions of the genes selected in our workflow are shown at the bottom. This includes commonly aberrant genes (n = 6373; upper band), in-cis genes (n = 578; middle band), and the iPAC genes (n = 56; lower band). Colors indicate whether the gene is most frequently amplified (red) or deleted (green). doi:10.1371/journal.pone.0053014.g002
statistic, we want to determine the rejection region \( D \geq d \) where \( d > 0 \) is a given threshold. Since \( D = G - (n - G) = 2G - n \), we have that

\[
\Pr(D \geq d) = \Pr\left( G \geq \frac{n + d}{2} \right) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}
\]

where \( r(n) = \lceil (n + d) / 2 \rceil \). Assuming a significance level of \( \alpha = 0.05 \), we seek the least integer \( d > 0 \) for which we have \( \Pr(D \geq d) \leq 0.05 \) for all \( n \) under the null hypothesis of \( p = 0.5 \). In practice (see Results) the number of aberrations never exceeds 38 in any given locus, and \( n \) may be restrained correspondingly above. A simple calculation then shows that the appropriate threshold is \( D \geq 11 \), and this value was used in the analyses. Thus, all genes for which \( G - L > 10 \) were defined as being commonly gained.

By an analogous argument, all genes for which \( L - G > 10 \) were defined as being commonly lost. Whenever \( |G - L| > 10 \), the gene was referred to as being commonly aberrant. Note that the purpose of this step was to filter out the bulk of aberrant genes with no indication of skewness towards either gain or loss, and hence the above significance criterion was designed to be very mild and did not take into account multiple comparisons.

Identification of in-cis correlated genes

To seek the genes for which the expression is significantly influenced by the copy number, we identified in-cis correlated genes. To identify significant in-cis correlations between log copy number and log gene expression, the in-cis correlations of all the commonly aberrant genes were compared to a background distribution of in-cis correlations. The background distribution was generated by performing 2000 shuffling simulations where in each, only the gene order in the aCGH data set was shuffled and the in-cis correlations were recalculated. By selecting the genes with (Pearson) in-cis correlation \( r > 0.6 \) we achieved a false discovery rate (FDR) of less than 2%. This cut-off corresponds to a coefficient of determination of \((0.6)^2 = 0.36\), meaning that at least 36% of the variation in log-expression is accounted for by the in-cis variation of log-copy number.

The above procedure corresponds to keeping only the genes \( g \) for which the following log-linear model provides a good fit to the observed copy number and gene expression levels:

\[
\text{Figure 3. Association between expression and copy number.} \text{ Linear regression of log-expression as a function of log-copy number for four selected iPAC genes.}
\]
avoid this problem we calculated for each gene potential expression-mediated effects of one gene on another. To factor, and this should be taken into account when assessing occurring copy number aberrations can act as a confounding expression levels of the two genes as a possible result. Thus, co-

considered the correlation between the in-
cis

expression values

g

and

log copy number (and noise). Suppressing the gene subscript and ignoring the noise term, expression is a linear function of log copy number (and noise).

The purpose of this step was to quantify the level of association between the in-
cis
genes and other genes. To do this, we considered the correlation between the in-
cis
gene and all other genes. A potential problem in this context is that genes close to each other on the same chromosome may be affected by the same copy number alterations, with inflated correlation between the expression levels of the two genes as a possible result. Thus, co-

occurring copy number aberrations can act as a confounding factor, and this should be taken into account when assessing potential expression-mediated effects of one gene on another. To avoid this problem we calculated for each gene g the residual log expression values

\[
R_{g,i} = \log E_{g,i} - \bar{a}_g - b_g \log C_{g,i}
\]

over all the samples, where the coefficients \( \bar{a}_g \) and \( b_g \) were found by fitting the model in (1), and quantified the in-
trans
effect of an in-
cis
gene \( G \) on gene \( g \) by the Pearson correlation between the observed log expressions \( \log E_{G,i} \) of the in-
cis
gene and the residual log expressions \( R_{g,i} \) of the gene in trans.

Identification of in-
cis
genes associated in-
trans

with processes

In order to identify in-
cis
genes that were associated with processes in trans, we considered 8284 gene sets defined by Gene Ontology (GO) biological process terms [37]. Using each in-
cis
gene \( G \) in turn as a pivot, all other genes \( g \) were ranked according to the correlation between \( \log E_{G,i} \) and \( R_{g,i} \) (from high positive correlation to high negative correlation), and an enrichment score was calculated for each GO term in the ranked list of genes. This was done separately for the genes in the top and the bottom of the ranked list. The enrichment score was defined as the p-value from the minimum hypergeometric (mHG) test [38,39]. Such scores were calculated for each in-
cis
gene and each GO biological process term. For further analysis, we only considered associations between in-
cis
genes and GO terms with a p-value score p<0.05 (after Bonferroni correction).

To obtain empirical p-values for the associations selected above, 100 random simulations were performed. In each simulation instance, we shuffled the order of the samples in the residual expression data set only and recalculated all enrichment scores. This approach preserves existing expression dependencies between genes. Let \( P_{g,s} \) be the enrichment score (mHG p-value) of the association between in-
cis
gene \( g \) and gene set (GO term) \( s \), and let \( P_{s,k} \) be the enrichment score of this association in the \( k^{th} \) simulation instance \( (k=1,\ldots,100) \). We considered an in-
cis
gene \( g \) to be significantly associated with a gene set \( s \) if \( P_{s,k} < \min_{s,k}(P_{s,k}) \), where \( k=1,\ldots,100 \) and \( t \) ranges over all in-
cis
genes. That is, a relation between an in-
cis
gene and a gene set was called significant if the observed enrichment score (mHG p-value) was less than the enrichment scores obtained for that gene set for all in-
cis
genes in the simulations. This step alleviates differences in attainable p-
values due to correlated null hypotheses (the GO gene sets have strong overlaps and genes within a set may be strongly dependent).

Enrichment analysis using GOrilla
GOrilla (http://cbl-gorilla.cs.technion.ac.il/) [38,39] was used with default parameters to investigate and visualize the enrichment of GO biological process in ranked lists of selected gene sets.

Results
We have presented a computational framework for identification of aberrant genes potentially leading to a substantial shift in transcriptional programs. The proposed method was applied to matched copy number and expression data from a cohort of 100 breast carcinomas. The resulting iPAC genes were further validated in a data set from another breast cancer cohort. The workflow of our approach is depicted in Figure 1 and details are provided in Materials and Methods.

Common aberrations
The first step of our workflow was the identification of genes that were commonly aberrant between the patient samples. Among the 25,688 genes profiled, a total of 6373 genes were found to be commonly aberrant, of which 3499 were commonly amplified and 2874 commonly deleted (notice that by the definition of commonly aberrant genes given in Materials and Methods, a gene cannot be both commonly amplified and commonly deleted). These genes are scattered throughout the genome with highest frequency on chromosomes 1, 8, 11, 13, 16, and 17 (Figure 2; Figure S1A). For all genes combined, 7.5% of the variance of the expression values was explained by copy number alterations in cis. Considering only commonly aberrant genes, this fraction increased to 11.5%.

In-trans associations
The in-trans correlation is shown for all genes in Figure 2 and Figure S1B. Ranking the 6373 commonly aberrant genes according to their in-trans correlation reveals that the genes with highest correlation are enriched with the GO terms of DNA repair, cell cycle, DNA recombination, and chromatin modification and organization (see Figure S2 and Table S1 for a full list of results). Genes with high in-trans correlation (Pearson’s r > 0.6) were selected among the commonly aberrant genes, resulting in 578 in-trans genes (see Figure 2 and 3, and Table S2). These genes were predominantly found on chromosomes 1, 8, 16, and 17. Of these, 423 genes were commonly amplified and 155 commonly deleted (Figure 2).

The in-trans genes included known cancer-associated genes such as EBB2, MAP3K7, MDM4, FGFR1, CCND1 and FADD. Further annotation of the 578 genes showed that 19% code for enzymes, 8% regulators of transcription, 7% transporters, 4% kinases, 2% peptidases, and 2% phosphatases (Figure S3A). The remaining genes encode various sorts of proteins, e.g. zinc finger proteins, ribosomal proteins, RNA binding proteins, and mitochondrial proteins (see Table S2 for description). The fraction of the variance in expression explained by copy number alterations increased to 46.6% when considering only the in-trans genes. Although the in-trans genes exhibit strong correlation between copy number and expression, a substantial proportion of the variability in these genes across samples is also related to other influences. Thus, their expression reflects copy number as well as various other factors.

In-trans associations to biological processes
The final step of the workflow led to the identification of in-trans genes significantly associated with at least one biological process in trans. For this purpose, the copy number-adjusted residual expression was calculated for all 25,688 genes. Each in-trans gene was taken separately as a pivot and all 25,688 genes were ranked according to the in-trans correlation between their copy number-adjusted residual expression and the non-adjusted expression of the pivot gene. The importance of adjusting for copy number is
### Table 1. Description and properties of the 56 iPAC genes.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Full gene name</th>
<th>Cytoband</th>
<th>Highest associated GO term (trait)</th>
<th>Score</th>
<th>Annot</th>
</tr>
</thead>
<tbody>
<tr>
<td>DARS2</td>
<td>aspartyl-tRNA synthetase 2, mitochondrial</td>
<td>1q25.1</td>
<td>nucleic acid metabolic proc.</td>
<td>94.31</td>
<td></td>
</tr>
<tr>
<td>ATAD2</td>
<td>ATPase family, AAA domain containing 2</td>
<td>8q24.13</td>
<td>cell cycle</td>
<td>91.53</td>
<td>BC</td>
</tr>
<tr>
<td>SMC4</td>
<td>structural maintenance of chromosomes 4</td>
<td>3q25.33</td>
<td>cell cycle</td>
<td>90.42</td>
<td></td>
</tr>
<tr>
<td>ACTL6A</td>
<td>actin-like 6A</td>
<td>3q26.33</td>
<td>cell cycle</td>
<td>87.78</td>
<td></td>
</tr>
<tr>
<td>RECL4</td>
<td>RecQ protein-like 4</td>
<td>8q24.3</td>
<td>cell cycle</td>
<td>86.78</td>
<td>BC</td>
</tr>
<tr>
<td>ECT2</td>
<td>epithelial cell transforming sequence 2 oncogene</td>
<td>3q26.31</td>
<td>cell cycle</td>
<td>82.84</td>
<td></td>
</tr>
<tr>
<td>POGK</td>
<td>pogo transposable element with KRAB domain</td>
<td>1q24.1</td>
<td>nucleic acid metabolic proc.</td>
<td>82.62</td>
<td></td>
</tr>
<tr>
<td>MTBP</td>
<td>Mdm2, transformed 3T3 cell double minute 2, p53 binding protein</td>
<td>8q24.12</td>
<td>cell cycle</td>
<td>80.67</td>
<td>C</td>
</tr>
<tr>
<td>YPS2</td>
<td>vacuolar protein sorting 72 homolog (S. cerevisiae)</td>
<td>1q21.2</td>
<td>nucleic acid metabolic proc.</td>
<td>80.29</td>
<td></td>
</tr>
<tr>
<td>NUDCD1</td>
<td>NudC domain containing 1</td>
<td>8q23.1</td>
<td>nucleic acid metabolic proc.</td>
<td>80.27</td>
<td></td>
</tr>
<tr>
<td>MTERFD1</td>
<td>MTERF domain containing 1</td>
<td>8q22.1</td>
<td>nucleic acid metabolic proc.</td>
<td>78.77</td>
<td></td>
</tr>
<tr>
<td>WDOSF1</td>
<td>DDB1 and CUL4 associated factor 13</td>
<td>8q22.3</td>
<td>nucleic acid metabolic proc.</td>
<td>78.14</td>
<td>BC</td>
</tr>
<tr>
<td>NUP85</td>
<td>nucleoparin 85kDa</td>
<td>17q25.1</td>
<td>cell cycle</td>
<td>75.28</td>
<td></td>
</tr>
<tr>
<td>RAD21</td>
<td>RAD21 homolog (S. pombe)</td>
<td>8q24.11</td>
<td>cell cycle</td>
<td>74.77</td>
<td>BC</td>
</tr>
<tr>
<td>KPNA2</td>
<td>karyopherin alpha 2 (RAG cohort 1, importin alpha 1)</td>
<td>17q24.2</td>
<td>cell cycle proc.</td>
<td>74.09</td>
<td></td>
</tr>
<tr>
<td>C8orf76</td>
<td>chromosome 8 open reading frame 76</td>
<td>8q24.13</td>
<td>cell cycle proc.</td>
<td>69.94</td>
<td></td>
</tr>
<tr>
<td>POP1</td>
<td>processing of precursor 1, ribonuclease P/MRP subunit</td>
<td>8q22.2</td>
<td>cell division</td>
<td>64.40</td>
<td>C</td>
</tr>
<tr>
<td>TATDN1</td>
<td>Tad D Nase domain containing 1</td>
<td>8q24.13</td>
<td>cellular macromol. metabolic proc.</td>
<td>61.06</td>
<td>BC</td>
</tr>
<tr>
<td>PDCD10</td>
<td>programmed cell death 10</td>
<td>3q26.1</td>
<td>cellular macromol. metabolic proc.</td>
<td>82.81</td>
<td></td>
</tr>
<tr>
<td>THRAP6</td>
<td>mediator complex subunit 30</td>
<td>8q24.11</td>
<td>cellular nitrogen compound metab. proc.</td>
<td>55.46</td>
<td>BC</td>
</tr>
<tr>
<td>RPL30</td>
<td>ribosomal protein L30</td>
<td>8q22.2</td>
<td>cellular macromolecule biosynth. proc.</td>
<td>46.14</td>
<td></td>
</tr>
<tr>
<td>PRCC</td>
<td>papillary renal cell carcinoma (translocation-associated)</td>
<td>1q23.1</td>
<td>organelation</td>
<td>38.66</td>
<td></td>
</tr>
<tr>
<td>C1orf35</td>
<td>chromosome 1 open reading frame 35</td>
<td>1q42.13</td>
<td>chromosome organization</td>
<td>34.69</td>
<td></td>
</tr>
<tr>
<td>PARP1</td>
<td>poly (ADP-ribose) polymerase 1</td>
<td>1q42.12</td>
<td>chromosome organization</td>
<td>34.10</td>
<td>BC</td>
</tr>
<tr>
<td>MRPS23</td>
<td>mitochondrial ribosomal protein S23</td>
<td>1q23.2</td>
<td>positive regulation of ligase activity</td>
<td>28.47</td>
<td></td>
</tr>
<tr>
<td>PSMD4</td>
<td>proteasome (prosome, macropain) 26S subunit, non-ATPase, 4</td>
<td>1q21.2</td>
<td>response to DNA damage stimulus</td>
<td>27.44</td>
<td></td>
</tr>
<tr>
<td>SETDB1</td>
<td>SET domain, bifurcated 1</td>
<td>1q21.2</td>
<td>chromatin modification</td>
<td>23.51</td>
<td>C</td>
</tr>
<tr>
<td>HNRPU</td>
<td>heterogeneous nuclear ribonucleoprotein U</td>
<td>1q44</td>
<td>chromatin modification</td>
<td>20.52</td>
<td>C</td>
</tr>
<tr>
<td>BOP1</td>
<td>block of proliferation 1</td>
<td>8q24.3</td>
<td>DNA conformation change</td>
<td>17.27</td>
<td>C</td>
</tr>
<tr>
<td>SIAHBP1</td>
<td>poly-U binding splicing factor 60kDa</td>
<td>8q24.3</td>
<td>mitotic sister chromatid segregation</td>
<td>16.14</td>
<td>C</td>
</tr>
<tr>
<td>PRPF3</td>
<td>PRP3 pre-mRNA processing factor 3 homolog (S. cerevisiae)</td>
<td>1q21.2</td>
<td>mRNA transport</td>
<td>15.77</td>
<td>C</td>
</tr>
<tr>
<td>PM1D1</td>
<td>protein phosphatase, Mg2+/Mn2+ dependent, 1D</td>
<td>1q23.2</td>
<td>mitotic cell cycle checkpoint</td>
<td>14.92</td>
<td>BC</td>
</tr>
<tr>
<td>FAM33A</td>
<td>spindle and kinetochore associated complex subunit 2</td>
<td>1q23.2</td>
<td>mitotic cell cycle checkpoint</td>
<td>14.12</td>
<td>BC</td>
</tr>
<tr>
<td>MRPL9</td>
<td>mitochondrial ribosomal protein L9</td>
<td>1q21.3</td>
<td>establishment of organelle localization</td>
<td>13.74</td>
<td></td>
</tr>
<tr>
<td>C22orf28</td>
<td>chromosome 22 open reading frame 28</td>
<td>22q12.3</td>
<td>cellular protein metabolic proc.</td>
<td>13.41</td>
<td></td>
</tr>
<tr>
<td>SLMO2</td>
<td>slowmo homolog 2 (Drosophila)</td>
<td>2q13.32</td>
<td>spindle checkpoint</td>
<td>11.92</td>
<td></td>
</tr>
<tr>
<td>CHRAC1</td>
<td>chromatin accessibility complex 1</td>
<td>8q24.3</td>
<td>mitotic metaphase plate congestion</td>
<td>11.91</td>
<td>C</td>
</tr>
<tr>
<td>C16orf61</td>
<td>chromosome 16 open reading frame 61</td>
<td>16q23.2</td>
<td>spindle checkpoint</td>
<td>11.28</td>
<td>BC</td>
</tr>
<tr>
<td>ISG20L2</td>
<td>interferon stimulated exonuclease gene 20kDa-like 2</td>
<td>1q23.1</td>
<td>DNA-dependent DNA replication init.</td>
<td>11.07</td>
<td></td>
</tr>
<tr>
<td>CSNK1E</td>
<td>casein kinase 1, epsilon</td>
<td>2q13.1</td>
<td>neural tube development</td>
<td>10.25</td>
<td>C</td>
</tr>
<tr>
<td>FAM91A1</td>
<td>family with sequence similarity 91, member A1</td>
<td>8q24.13</td>
<td>establishment of mitotic spindle loc.</td>
<td>10.08</td>
<td></td>
</tr>
<tr>
<td>TOMM20</td>
<td>translocase of outer mitochondrial membrane 20 homolog (yeast)</td>
<td>1q42.3</td>
<td>transcription</td>
<td>10.83</td>
<td></td>
</tr>
<tr>
<td>C20orf20</td>
<td>chromosome 20 open reading frame 20</td>
<td>20q13.33</td>
<td>mitotic cell cycle spindle checkpoint</td>
<td>9.67</td>
<td>C</td>
</tr>
<tr>
<td>GALNS</td>
<td>galactosamine (N-acetyl)-6-sulfate sulfatase</td>
<td>16q24.3</td>
<td>carbohydrate catabolic proc.</td>
<td>9.05</td>
<td>BC</td>
</tr>
<tr>
<td>AZIM1</td>
<td>antizyme inhibitor 1</td>
<td>8q22.3</td>
<td>histone mRNA metabolic proc.</td>
<td>8.19</td>
<td>C</td>
</tr>
<tr>
<td>MTL5</td>
<td>metallothionein-like 5, testis-specific (tesmin)</td>
<td>11q13.2</td>
<td>water-soluble vitamin biosynthetic proc.</td>
<td>–8.37</td>
<td></td>
</tr>
<tr>
<td>TP532</td>
<td>tumor protein DS2</td>
<td>8q21.13</td>
<td>regeneration</td>
<td>–8.57</td>
<td>BC</td>
</tr>
<tr>
<td>ARID4B</td>
<td>AT rich interactive domain 4B (RBP1-like)</td>
<td>1q42.3</td>
<td>organ regeneration</td>
<td>–8.80</td>
<td>C</td>
</tr>
<tr>
<td>THC2340878</td>
<td>NA</td>
<td>8q13.2</td>
<td>programmed cell death</td>
<td>–9.06</td>
<td></td>
</tr>
</tbody>
</table>
most pronounced for genes in close proximity (see Figure 4A), and the effect of using copy number-adjusted residual expression increases with the in-\textit{cis} correlation (see Figure 4B and Figure 5). Overrepresentation of Gene Ontology (GO) biological process terms in the above ranked list of genes was statistically assessed (Figure 6). Out of the $578 \times 8284 = 4.7 \times 10^6$ potential associations (for every in-\textit{cis} gene and every GO term tested), we first selected those with enrichment score $p < 1.04 \times 10^{-8}$ (corresponding to $p < 0.05$ after Bonferroni correction). This resulted in 19,606 associations covering 467 GO terms and all 578 in-\textit{cis} genes. Finally, simulations were used to call significant gene-process associations (see Materials and Methods). This yielded a total of

Table 1. Cont.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Full gene name</th>
<th>Cytoband</th>
<th>Highest associated GO term (trait)</th>
<th>Score</th>
<th>Annot</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHTOP</td>
<td>chromatin target of PRMT1</td>
<td>1q21.3</td>
<td>activation of plasma proteins</td>
<td>-11.73</td>
<td></td>
</tr>
<tr>
<td>TEMEM70</td>
<td>transmembrane protein 70</td>
<td>8q21.11</td>
<td>regulation of Rho protein signal transd.</td>
<td>-13.02</td>
<td>BC</td>
</tr>
<tr>
<td>DPM1</td>
<td>dolichyl-phosphatmannosyltransferase poly peptide 1, cat. subunit</td>
<td>20q13.13</td>
<td>negative regulation of gene expression</td>
<td>-13.36</td>
<td>C</td>
</tr>
<tr>
<td>PRCRL</td>
<td>pyrroline-5-carboxylate reductase-like</td>
<td>8q24.3</td>
<td>membrane invagination</td>
<td>-15.15</td>
<td></td>
</tr>
<tr>
<td>IMPAD1</td>
<td>inositol monophosphatase domain containing 1</td>
<td>8q12.1</td>
<td>positive regulation of cell death</td>
<td>-15.88</td>
<td></td>
</tr>
<tr>
<td>STX16</td>
<td>syntaxin 16</td>
<td>20q13.32</td>
<td>cellular protein metabolic process</td>
<td>-15.88</td>
<td></td>
</tr>
<tr>
<td>PIGM</td>
<td>phosphatidylinositol glycan anchor biosynthesis, class M</td>
<td>1q23.2</td>
<td>response to external stimulus</td>
<td>-22.09</td>
<td></td>
</tr>
</tbody>
</table>

Scores in the table are the negative logarithms of the enrichment scores, the sign indicating whether the association of the trait to the genes is positively or negatively correlated with the iPAC gene. The annotation column indicates genes previously linked with breast cancer (BC) and among those that are not, genes linked to cancer in general (C), based on annotation of the genes obtained with IPA (Ingenuity® Systems, www.ingenuity.com).

doi:10.1371/journal.pone.0053014.t001

Figure 7. Associations between iPAC genes and traits (biological processes). A hierarchical clustered heatmap representation of traits associated with at least four iPAC genes. A red entry indicates a significant association between an iPAC gene and the corresponding trait (see Figure S4 for all the significant associations). The Expander suite [66] using average Euclidian distance was used to calculate and visualize the hierarchical clustering analysis.

doi:10.1371/journal.pone.0053014.g007
Figure 8. Distribution of in-cis correlation levels between copy number and expression in the MicMa and UNC cohorts. Green bins in the histogram show distribution of in-cis correlation levels of all genes in the data set, while red bins show the distribution for only the identified iPAC genes. The left-hand y-axes in each histogram show the count in each bin among all genes, and the right-hand axes show the count for iPAC genes in each bin. (A) Distribution of the in-cis correlation levels in the MicMa cohort. (B) Distribution of the in-cis correlation levels in the UNC cohort. The iPAC genes were inferred from the MicMa cohort.

doi:10.1371/journal.pone.0053014.g008

Figure 9. Association consistency of iPAC genes in the validation cohort. Blue dots represent associations between an iPAC gene and a GO term. The blue dots are plotted according to the level of association, as signed \(-\log(p\text{-value})\), in the MicMa cohort (x-axis) and in the UNC cohort (y-axis), where signed \(-\log(p\text{-value})\) refers to \(-\log(mHG\ p\text{-value})\) for positive associations and \(\log(mHG\ p\text{-value})\) for negative associations. A monotone relation is observed, supporting the iPAC behavior of the MicMa inferred iPAC genes in the validation cohort. A bar with a red dot in the center is plotted for each blue dot representing 1 standard deviation (SD) of the associations generated by associating 100 random genes from the UNC cohort to the relevant GO term.

doi:10.1371/journal.pone.0053014.g009
276 highly significant associations, covering 56 in-cis genes (henceforth called iPAC genes) and 97 unique GO terms (called traits) at a false discovery rate (FDR) of less than 1% (see Figure 2, Table 1 and Table S3). Cell cycle related processes commonly occurred as traits of the iPAC genes, consistent with an association to tumor development and progression related processes (see Figure 7 and Figure S4).

Properties of the identified iPAC genes

Four of the 56 iPAC genes were commonly deleted and 52 commonly amplified. The iPAC genes encode proteins with various biological roles, including enzymes, regulators of transcription and translation, and transporter molecules (Figure S3B). Five of them were themselves members of the biological process(es) they were found to be associated with (MTBP, RAD21, RECQL4, SETDB1, and SMC4). Comparing the 56 iPAC genes against the background of all other genes using GOvila, the 56 genes were found to be associated to four biological processes all commonly disrupted in cancer: cell cycle, cell cycle process, nucleic acid metabolic process, and chromosome organization (Table S4 and Figure S5).

Among the iPAC genes, 38 mapped to chromosomes 1 (n = 16) and 8 (n = 22), while the rest were located on chromosomes 3, 11, 16, 17, 20, and 22 (Figure 2 and Figure S6). There was a tendency for iPAC genes to reside in blocks of commonly aberrant segments. As would be expected by their mutual proximity and their high in-cis correlation, the expression levels of iPAC genes residing in the same block were highly correlated (Figure S7). Accordingly, the scope for further narrowing down the list of candidates based on copy number and expression data alone was limited. However, in several cases, iPAC genes in close proximity were found to be associated with different biological processes. For example, PRPF3 and SETDB1 are less than 1 Mb apart from each other and were associated with mRNA transport and chromatin modification, respectively.

Figure S8 shows how the patient samples clustered according to the expression of the iPAC genes. Most of the luminal samples clustered together, as did the basal-like samples, with the latter having a tendency towards higher expression of the iPAC genes. We note that the expression levels of the iPAC genes were not found to be significantly associated with survival (data not shown).

Knockdown experiment with siRNA

To investigate the effect of the selected iPAC genes on cell viability, siRNA knockdown was performed for three iPAC genes (ECT2, PSMD4 and MTBP) in two breast cancer cell lines (MCF7 and MDA-MB-231; see the Supporting Information uploaded to the file inventory: File S1.pdf). For one of the siRNAs tested against ECT2 we observed a ~30% reduction in cell viability (p < 0.05, Figure S9) in the MCF7 cell line. ECT2 is a guanine nucleotide exchange factor for Rho family GTPases and was most strongly associated with the cell cycle GO term and amplified in 15% of the samples. The reduced cell viability after knockdown emphasizes the importance of the iPAC gene ECT2 in the MCF7 cell line. A smaller reduction in cell viability was also observed for PSMD4 (data not shown).

Robust iPAC signature in a validation cohort

In order to validate the robustness of the 56 identified genes, we investigated their iPAC characteristics in an independent breast cancer cohort (UNC), consisting of 73 patients [25]. Out of the 56 iPAC genes identified in our study, 51 were among the genes measured in the UNC study. The five remaining genes (IMPAD1, FAM33A, FAM91A1, PARP1 and THIC2340878) had been removed further upstream in the analysis and data preprocessing in the UNC study.

The in-cis correlation between copy number and expression for the iPAC genes ranged from 0.16 (SETDB1) to 0.69 (PPM1D) in the validation cohort, with an average of 0.43 (Figure 8). To assess the in-trans associations for the iPAC genes in the validation cohort, all genes were ranked according to the level of correlation between their copy number-adjusted residual expression and the expression of each iPAC gene in the validation cohort. The association between each iPAC gene and each of the 97 GO terms identified in the original analysis was then assessed. The results showed high level of consistency of enrichments between the two cohorts (Figure 9). As a further confirmation of consistency, we compared these results to the association levels (negative logarithms of the enrichment scores) of 100 random genes to each of the above GO terms. It was found that 95% of the iPAC gene/trait pairs tested in the validation cohort had association levels exceeding m + SD, where m is the average and SD is the standard deviation of association levels obtained for the random genes, and 80% of the pairs had association levels exceeding m + 2SD (Figure 9). This shows that the level of association of the iPAC genes with their relevant biological process (represented by the GO term) in the validation cohort is not random.

Discussion

Copy number aberrations are common in breast cancer, but to what extent such aberrations affect cancer cell phenotype through alterations of the transcriptional program is not yet known. The methods we propose here aim to identify genes subject to selection in breast cancer by detecting commonly aberrant genes affected on the gene expression level by genomic aberrations. Furthermore, the method requires the identified genes to be correlated with genes collectively enriched with respect to GO biological processes. Thus, it is through the influence on other genes and their associated processes that the iPAC genes are identified.

Integrative analysis as a tool for inferring causality

Numerous high-throughput profiling expression studies have identified clusters of genes with expression varying in a coordinated manner over time or across disease states. However, such studies generally give no information about the directionality of gene interactions unless additional information is available. Strong association between the mRNA expression levels of two genes may result from one gene regulating the other, both being regulated by a common factor, or a combination of both. The combined use of copy number and expression data allows the distinction between a situation where the expression of one gene influences the expression of another gene and a situation where the expression levels of the two genes are merely correlated [17].

Relationship to other methods

Several strategies that aim to identify driver genes in cancer exploit the integration of matched copy number and expression data. Woo et al. [23] worked with an integrated copy number and expression data set and used the prognostic significance of genes to guide the selection process. Akavia et al. [17] also utilized this sort of integration in their CONEXIC algorithm. Their study assumed that a driver mutation would occur more often than by chance in multiple tumors, that the mutation would be correlated with the expression of a group of genes (a module), and that copy number changes often had an effect on expression of the driver that thus further influenced the expression of the module [17]. The CONEXIC approach is founded on the notion that the expression
levels of the driver, rather than the determinants of that expression level, confers a fitness advantage to the tumor. Alteration of copy number is only one way of achieving this, manifested by a high frequency of aberrations in a patient cohort.

The iPAC approach has a similar rationale as CONEXIC. However, our method differs from the approaches described above in several aspects. First, we use residual expression for the in-trans correlation analysis, thus bypassing the potential confounder effect of co-occurring copy numbers. Second, we use a robust enrichment analysis approach to identify aberrations that lead to a significant shift in cancer-related transcriptional programs. Using the enrichment framework, we assign statistical significance to gene-process associations. By taking advantage of the residual expression, the modulator properties of the iPAC genes are more robustly captured. Such modulator effects on biological processes interrupted in cancer may go beyond the direct effects on a pathway; transcriptional responses launched by the cell after physiologic alterations may result from various indirect influences and mechanisms [17], and in this respect, the iPAC genes represent a diverse set of candidates.

Characteristics of the iPAC genes

The list of iPAC genes includes 16 genes previously associated with breast cancer and 20 additional genes associated with cancer in general (Table 1). For example, ATAD2 was highly associated to the cell cycle process, indicating that the cell cycle module is activated when ATAD2 is amplified and overexpressed. ATAD2 is an ATPase and was recently reported to be a cofactor for the MYC oncogene [40]. While copy number was a predominant determinant of ATAD2 expression levels, other factors also probably influence ATAD2 expression levels and through its expression level, ATAD2 is proposed to affect its target process. Another example is TPD52 which was highly associated to regeneration processes; this gene has previously been suggested as a potential driver gene and reported amplified and overexpressed in various cancer types, including breast cancer [41,42,43,44]. Furthermore, PPMD12 was strongly associated to the mitotic cell cycle checkpoint; this gene encodes a serine/threonine phosphatase, maps to the 17q23.2 amplicon and has been shown to be involved in the regulation of several tumor suppressor pathways, including the p53 pathway [45,46]. Amplification of this gene has previously been found to be correlated with overexpression in breast cancer [47]. The iPAC gene KPNA2 was associated with the trait of nuclear division, and is a member of the importin family of proteins involved in nuclear transport. KPNA2 has been proposed to be a prognostic marker in breast cancer [48], and overexpression of this gene has been associated with poor prognosis, expression signatures of high proliferation, and tumor grade [49,50].

The iPAC genes also include several genes not previously associated with cancer. One interesting example is the gene MTL5 which was negatively correlated with the water-soluble vitamin biosynthetic process and encodes a protein with homology to the metal-binding motif of the metallothionein (MT) family [51]. MTL5 is located on chromosome 11q13.2 and was found amplified in 17% of our investigated breast cancer samples. Through their ability to bind metal, MT proteins can affect the activity of several proteins and enzymes dependent on metals as co-factors. In this respect, MT proteins play important roles in apoptosis and proliferation [52]. Furthermore, elevated expression of MT proteins has been reported in various cancer types, including breast cancer [52,33,34] and was also linked to modulation of p53 activity through zinc exchange [55,56]. Dividing our samples according to p53 mutational status, MTL5 was one of the top 2% most down regulated genes in mutated p53 (p<10^-5) (data not shown). As MTL5 was found to be amplified in a significant proportion of the samples in our cohort, and because of its iPAC properties, our results indicate that the gene may have an important role in breast cancer, similar to the homologous MT proteins. Many homologs of MTL5 exist both in animals and plants, suggesting that the function of this gene is conserved [57].

Proof-of-concept knockdown experiments

We selected three iPAC genes for siRNA knockdown experiments. Out of these, silencing of ECT2 led to significant decrease in cell viability. By using our approach, this gene was found to be most highly associated with cell cycle related traits. The protein ECT2 has been shown to regulate cytokinesis [58], which can explain the effect on cell viability after knockdown. ECT2 has been found to be up-regulated during transition to malignancy in a mouse model [59], to be amplified and overexpressed in non-small cell lung cancer [60], and to have an elevated expression in colorectal cancer [61].

In another study, siRNA-mediated knockdown of the iPAC gene RAD21 was found to decrease cell growth and enhance cytotoxicity in MCF7 and T47D breast cancer cell lines [62]. RAD21 encodes a phosphoprotein and is a component of the cohesin complex essential for chromosome segregation during mitosis/meiosis and DNA repair [63,64]. In our breast cancer cohort, RAD21 was found to be amplified in 36% of the tumor samples and to be highly associated with the cell cycle trait. Strong association to cell cycle has been shown to correlate with cell proliferation for the same patient samples [26], and enhanced expression of this protein has been associated with poor prognosis and resistance to chemotherapy in breast cancers [65].

Conclusion

Whole-genome integrative analyses of copy number and gene expression data is a useful tool in genome-wide searches for candidate driver genes in cancer. The first phase of analysis is typically to detect genes with frequent aberrations in copy number and strong in-cis correlation to gene expression. For example, in our study, the gene ERBB2 was ranked 7 out of 6373 genes with respect to the in-cis correlation level, indicating a direct link between copy number and expression. However, even among those genes that satisfy these criteria there are potentially many passengers with no direct oncogenic role. In the opposite direction, there may be genes that manifest moderate in-cis expression but still drive cancer-related processes through their expression level. Regulation of these expression levels may be selected for in the cancer through copy number changes as well as other mechanisms (e.g. altered methylation). Our aim has been to detect genes for which the gene-gene correlation structure of the expression data reveals additional evidence to support a link to a phenotype. The iPAC gene ATAD2, which was ranked only 450 in the in-cis correlated genes and hence would easily have been missed by in-cis focused methods, illustrates this point. Several similar examples are described above, indicating that the iPAC procedure does indeed capture biologically relevant genes not found on the top of the list of in-cis correlated genes.

Validation in an independent cohort of the proposed methodology and of the observation regarding the 56 iPAC genes found in our initial analysis supports method robustness and justify focus on the identified genes with respect to their tumorogenic role. In this study, we have selected GO biological process terms as they represent a comprehensive view of functional traits. It is clearly possible to select other annotation approaches for this purpose. For example, one could assess the enrichment of molecular pathways or transcription factor networks.
among the in-trans correlated genes. We provide cell line based experimental data for the effect of ECT2 on cell viability; however, further functional validation is still needed to firmly establish the role of the 56 iPAC genes in breast cancer.

The framework for the identification of in-trans regulatory mechanisms, as exemplified here in human breast cancer, is applicable to any kind of data with existing comparable aCGH, expression profiles and a collection of gene sets representing transcriptional programs. We propose this method as an unbiased and robust approach for the identification of genes of relevance to tumorigenesis.

**Supporting Information**

**Figure S1** Copy number and expression correlations. (A) Pearson correlation of copy number data for all the 25,688×25,688 genes. (B) Pearson correlation of copy number and expression of all 25,688×25,688 genes, with in-trans correlation along the diagonal. Color map represents the Pearson correlation coefficient. (TIF)

**Figure S2** GO terms enriched among the in-trans correlated genes. The GO biological process statistical enrichment analysis was performed by GOrilla. The input for GOrilla in this analysis was the list of 6373 commonly aberrant genes ranked according to their in-trans correlation. (TIF)

**Figure S3** Functional annotation of genes. (A) The 578 in-trans genes; (B) The 56 iPAC genes. The genes were annotated using IPA (Ingenuity® Systems, www.ingenuity.com). (TIF)

**Figure S4** Associations between iPAC genes and their traits (GO terms). Extension of Figure 7. A hierarchical clustered heatmap representation of all significant associations between iPAC genes and biological processes. A red entry indicates a significant association between an iPAC gene and the corresponding traits. The Expander suite [66] using average Euclidian distance was used to calculate and visualize the hierarchical clustering analysis. (TIF)

**Figure S5** Statistical enrichment analysis of the 56 iPAC genes for GO biological processes. Performed by GOrilla, on the list of 56 iPAC genes, compared to a background gene list consisting of all the remaining genes. (TIF)

**Figure S6** Sample-wise genomic copy number aberrations. Copy number aberrations are shown for chromosomes harboring at least one iPAC gene. The x-axis represents chromosomal location and the y-axis represents sample no (1–100). Green lines are regions of loss ($\theta < -0.2$), and red lines are regions of gain ($\theta > 0.2$). The vertical black lines indicate the locations of the 56 iPAC genes. (TIF)

**Figure S7** Correlation plots. (A) Pairwise correlations of log copy number of the 56 iPAC genes. (B) Pairwise correlations of log expression levels of the 56 iPAC genes. Chromosomes are indicated with numbers. (TIF)

**Figure S8** Hierarchical clustering of the expression levels of the 56 iPAC genes. Samples are color-coded according to gene expression subtype. The clustering was made with Pearson correlation using Ward linkage. Three samples could not be subtyped and were omitted from the analysis. Color map represents log expression values. (TIF)

**Figure S9** siRNA knockdown of the iPAC gene ECT2. (A) Effect of siRNA knockdown of ECT2 on cell viability in the MCF7 cell line. Four various siRNAs against ECT2 were tested in addition to controls (bars show SD from eight replicates). The ECT2_5 siRNA shows a statistically significant reduction in cell viability compared to the non-transfected cells (asterisk; Student’s t-test, $p<0.05$). (B) Relative quantification (RQ) of ECT2 mRNA after siRNA transfections (9 replicates), showing the specificity of the knockdown in the MCF7 cell line. The data were normalized to the control (cells + transfection lipid). (TIF)

**Table S1** GOrilla results from ranking the 6373 commonly aberrant genes. (XLSX)

**Table S2** The 578 in-trans genes. (XLSX)

**Table S3** Description of the 56 iPAC genes. (XLSX)

**Table S4** GOrilla results from 56 iPAC genes versus background gene set. (XLSX)

**File S1** Cell culture and siRNA transfection. (PDF)

**Acknowledgments**

We thank Silje Nord and Grethe I. G. Alnæs for valuable discussions and assistance in data normalization and technical support. We thank Roy Navon and Amir Ben-Dor for useful discussions.

**Author Contributions**

Designed the software used in the analysis: MRA IS OCL ZY. Data analysis: MRA IS LOB KL OCL ZY. Interpretation of results: MRA IS LOB KL KKS ALBD OCL ZY. Conceived and designed the experiments: MRA SN BN KKS VNK ALBD OCL ZY. Contributed reagents/materials/analysis tools: IS KL DL BN KKS VNK ALBD OCL ZY. Wrote the paper: MRA OCL IS LOB ALBD OCL ZY.

**References**


4 Discussion
In this thesis we addressed data analysis in studies combining multiple high-throughput measurement technologies and developed methods to support such analysis. Our research is focused on two main computational challenges: A) analysis of matched high-throughput measurement assays in an integrated manner (henceforth referred to as integromics), B) statistics in ranked lists, specifically the statistical question of mutual enrichment in two ranked lists (herby referred to as statistics in ranked lists). The work presented in this thesis combines both aspects of integromics and statistics in ranked lists to extract biological meaningful results from studies involving inference from high-throughput measurement data. The underlying theme common to the methods developed in this work is the emphasis on data driven analysis with sound statistical support.

In the work by Enerly & Steinfeld et al. [42] we explore the integrated analysis of mRNA and miRNA dataset collected for 101 breast cancer samples. Various methodologies have been developed, in recent years, to handle integrated analysis of functional genomics data, mainly by studying the transcriptional programs and global organization of biological processes [9–11]. Still, only a few studies report the joint analysis of sample cohorts that include multiple genomic measurements [12–15]. In Enerly & Steinfeld et al. [42] we develop a novel statistical framework that employs available rank statistics in a high-throughput manner to uncover miRNA biological relevance in a case specific manner. Specifically – a data driven approach is taken to study the association of all miRNAs to broader transcription programs. The application of our statistical framework to the breast cancer cohort enabled us to assign miRNAs to specific biological functions in breast cancer. While some assignments were previously observed, we report several novel assignments and thus improve our understanding of the involvement of miRNAs in breast cancer. Some of our results are further validated in the lab.

In the second work by Steinfeld et al. [41] we addressed the question of mutual enrichment in two ranked lists. A clear benefit in statistically supporting the analysis of ranked lists of genes rather than of fixed sets of genes is due to the fact that results from most functional genomics experiments can be represented as ranked lists. In the work
presented, we further explore the activity of miRNAs by analyzing the behavior of their gene targets in high-throughput experiments. For this task a novel statistical tool was developed that translates information available about targets into ranking. Two ranked lists, one derived from the measurement assay and the other inferred from target information, are jointly analyzed in a data driven manner. This interpretation of the results allowed statistical support not only for miRNA activity but also for statistical support for the level and extent of such activity.

The final main work presented in this thesis - Aure & Steinfeld et al. [46] - further develops the integrated analysis framework to uncover novel driver genes in breast cancer. We analyzed matched datasets of gene copy number and gene expression for the same cohort of breast cancer patients. Several previous studies analyzed integrated data for the purpose of prioritizing genes for their relevance to cancer [47], yet our work is the first that uses prior knowledge about the genes as a factor in the interpretation. We develop a computational framework that consists of a series of conceptual steps where each is aimed to focus on genes likely to drive the cancer process. Our predictions are substantiated in an independent cohort of breast cancer samples as well as by in-vitro experiments on cell-line models.

In summary – the studies presented here demonstrate the use of rigorous statistics to analyze multi-datasets by translating results into ranked lists and by using flexible approaches to analyze and statistically assess properties of the latter.

4.1 Summary discussion sections

In the following section we summarize and discuss each study, emphasizing on merits, main important results and contributions as well as optional extensions for future work.

4.1.1 miRNAs role in breast cancer [42]

In this study we introduce an extensive analysis of miRNA expression in 101 tumor samples taken from breast cancer patients. We show that analysis of miRNA expression alone is sufficient to distinguish between different breast cancer subtypes and other clinical properties (see Figure 1 and Figure 2 of the manuscript). Yet, our major
contribution in this work is the introduction of several new approaches to analyze miRNA and mRNA expression data in an integrated manner using a systems biology approach. Specifically, we take an unsupervised approach that allows us to unveil association of miRNA to the disease. In particular - we assess the enrichment of various gene sets amongst the genes correlated/anti-correlated to the expression levels of a pivot miRNA using rank based statistics. This integrated analysis depicted in Figure 3 of the manuscript enabled us to assemble the miRNA-GO association network (see Figure 5 and section 2.2.1) and thereby to link miRNAs to central cancer related biological processes. This approach led us to additional insight into the role of miRNAs in breast cancer and enabled the identification of key players. In particular, we point to a significant association of many miRNAs to the cell-cycle module. Strikingly, our methodology shows that miRNAs over-expressed in proliferative samples, are correlated to positive regulators of cell-cycle. Whereas, miRNAs with lower expression levels in proliferative samples are correlated with negative regulators of cell-cycle. The relationship between enrichment of these gene sets and proliferation related differential expression is, in general, monotone, as depicted in Figure 5 of the manuscript. This monotonicity further demonstrates the sensitivity of our approach in terms of detecting miRNA association to biological processes. This is in particular relevant when the biological process is challenging to measure (e.g. miR-210 and the glycolysis pathway).

Our integrated analysis also revealed the association of several miRNAs to the immune response module, a major biological process closely associated with cancer progression and development [48]. We found miR-150, as well as miR-155 and miR-142, to have strong positive correlation to the immune response module. Several studies have identified miR-150 to be involved in controlling B-cell differentiation by targeting the transcription factor c-Myb [49]. In our dataset we see an inverse correlation (Pearson’s r = -0.18) between Myb and miR-150 expression levels which might point to a similar regulation of immune response pathways in breast cancer. Immune response was previously linked to clinical outcome in ER-/HER2- samples [50], and in accordance with its strong association to the immune response module we see that miR-150 can predict clinical outcome in ER-/HER2- samples (Supplementary Figure 5 of the manuscript). This makes miR-150 a good marker for the activity of the immune response
in breast cancer samples, as high expression of miR-150 is associated with active immune response and better prognosis. Since we could not find any association of miR-150 to any of the known breast cancer subtypes we propose that its expression can act as a mean for classifying breast cancer samples based on immune response. However, more studies are needed to better elucidate the causal relationship between miR-150 and the immune response as well as between miR-150 and prognosis.

We note that our comprehensive systems biology approach unveils other aspects of miRNAs participation in cancer related processes. Glycolysis and miR-210 are found to be associated (Figure 5 in section 2.2.1). miR-210 was recently shown to have an important role in regulating the hypoxic response in breast cancer cell-lines [51], and our result confirms this participation in a clinical cohort.

miRNA associations to other processes can be unveiled by using the same methodology that leads to the miR-GO network and applying it to other gene sets, such as sets from MSigDB [39]. One such interesting example is the miRNA-Cytoband association network that associates between miRNAs and genes co-expression residing in specific cytobands. The miRNA-Cytoband network might point to a regulation of miRNAs mediated by genomic aberrations. In summary – the computational and statistical framework we developed in this work supports several novel biological findings. Building upon the statistical setting of our approach, we can assess the results and provide indications of confidence.

4.1.2 Detection of miRNA activity [41]

In Enerly & Steinfeld et al. [42] we worked with ranked lists and fixed set of genes for which we sought to assess enrichment of the set in the top of the ranked list. Our contribution, broadening the mHG approach, was to apply the mHG statistics to hundreds of lists obtained through consideration of all measured miRNAs. In the current study we broaden the mHG approach in a different direction. Specifically - we address the question of mutual enrichment in more than one list. Most results obtained from recent high-throughput measurement technologies are naturally given as ranked lists of genes rather than as fixed sets of genes. The question of mutual enrichment in two
ranked lists of genes is thus highly relevant to the analysis of such datasets. Adequate statistical methods that address the analysis of mutual enrichment in two ranked lists of genes are absent from statistical analysis and from bioinformatics in particular. The \textit{mmHG} approach developed and applied in this study focuses on commonalities in the top ends of the two analyzed ranked lists (see section 2.1.1). In addition to the development of the \textit{mmHG} statistics, we also introduce \textit{miTEA} - a framework for miRNA target enrichment analysis. \textit{miTEA} takes as input a ranked list of genes from the user, and using public information on miRNA targets, it finds miRNAs of which their targets are enriched in the top of the ranked list of genes provided by the user.

As depicted in Figure 1 of the manuscript, the statistical approach used by \textit{miTEA} finds the two thresholds in the top of the two ranked lists of genes that maximize the mutual enrichment. This approach thereby allows the data to define the set of affected genes in the experiment and the set of miRNA targeted genes. As can be observed in the figure, fixing a threshold either to define the downregulated genes in the experiment or to define the genes targeted by the perturbed miRNA in the experiment might result in loss of observed significance. Moreover, a pair of thresholds that optimizes enrichment in one configuration will not be optimal for a different miRNA and expression configuration. Thus, the different thresholds selected in each of the experiments tested much improve the enrichment analysis results, and that improvement further supports the need for a statistical tool that enables flexible threshold selection. This flexibility endows \textit{miTEA} with better accuracy and robustness in detecting activity, as shown in Figure 2 of the manuscript, where \textit{miTEA} is able to find the correct active miRNA in different experiments. Specifically, using Targetscan as the target prediction algorithm, \textit{miTEA} detected the correct activity in 20 out of the 21 tested cases with high significance levels. In comparison, using Spearman’s correlation, that also takes two ranked lists as input, we are able to detect only 17 active miRNAs and with less significance (see Supplementary Table 6 of the manuscript). Furthermore, \textit{miTEA} is able to be more specific and more sensitive in its findings also when compared with other online tools, as shown in Figure 3 of the manuscript.

In addition to establishing \textit{miTEA}’s capabilities in detecting miRNA activity, in this work we also expanded the integromics approach presented in Enerly & Steinfeld \textit{et al.}
al. [42] to develop a framework for dataset specific miRNA activity rather than single experiment activity. This analysis was carried out on a matched mRNA and miRNA dataset of healthy tissues as well as on a similar matched dataset of cancer tissues and cell-lines. Using this approach, we were able to detect statistically significant direct activity of more than 30 unique miRNA families. Interestingly, significantly many miRNAs are found to be directly active in both the healthy and the cancer related datasets as shown in Figure 5 of the manuscript. This result provides the first comprehensive miRNA activity map for nine human tissue types and for 60 cancer cell-lines tested. Moreover, the high prevalence of directly active miRNAs corroborates the involvement of miRNAs in differentiation and development.

We note that the threshold optimization process obviously introduces statistical multiple testing. We address multiple testing correction, through a combination of an exact approach to $mHG$ [40] and use of a union bound, and provide a rigorous bound on the p-value of the resulting enrichment. Further characterization of the distribution of the $mmHG$ statistics over the null model (a uniform measure over $S_N$ - the group of all permutations) will significantly contribute to improved analysis of mutual enrichment in two ranked lists. This is a line of work that the Yakhini group intends to further pursue.

4.1.3 Identifying driver genes in breast cancer [46]

In this study we extend the integrative methodology developed in Enerly & Steinfeld et al.[42] and in Steinfeld et al. [41] to address a question which is central in cancer research – the identification of altered genomic regions that affect important disease mechanisms. These regions harbor gene that may be highly relevant to cancer. In this study we explore cancer gene relevance in a cohort of breast cancer by analyzing gene expression data and DNA copy number data in an integrative manner. An overview of the analysis workflow is depicted in Figure 1 of the manuscript, and as shown, the methods proposed here are aimed at identifying genes subject to selection in breast cancer by detecting commonly aberrant genes (genes aberrant in many tumors) for which expression seems to be affected by the observed aberration. That is, gene for which a significant in-cis correlation is observed between their copy number and their expression
levels. Furthermore, the method requires the thus far identified in-cis genes to be correlated with a cancer related biological process, represented by a GO term. These genes, which seem to be driven by the cancerous process and in turn have an observable effect on a biological process, we term in-trans process associated and cis-correlated (iPAC) genes.

In the field of identifying cancer gene relevance our approach is not novel in exploiting the integration of matched copy number and expression data. Woo et al. [52] worked with an integrated copy number and expression data set and used the prognostic significance of genes to guide the selection process. Akavia et al. [53] also utilized this sort of integration in their CONEXIC algorithm to try to associate probable driver genes with a set of correlated genes.

Our iPAC method has a similar rationale as CONEXIC. However, our method differs from the approaches described above in several aspects. First, our approach considers the effect observed on the measured gene expression that is originated by the gene’s local aberration status. As can be seen in Figure 3 of the manuscript this effect can be substantial. Thus, we use residual expression of the gene from its local aberration status when assessing the in-trans correlation between two genes. As can be seen in Figure 4 of the manuscript we are thus able to bypass the potential confounder effect of co-occurring copy numbers. Second, we propose a biological role by which the candidate driver gene manifests its activity and give statistical support for this hypothesis. The step in the analysis where we look for this support is depicted in Figure 6 of the manuscript where we analyze the genes clustered around the iPAC gene and using robust enrichment analysis we are able to significantly associate it with a shift in cancer related transcriptional programs. Together with our use of the residual expression, the modulator properties of the iPAC genes are more robustly captured.

Many factors can drive the expression of an iPAC gene. We consider the copy number in our analysis as we believe it is the main one but we also acknowledge that there might be other factors such as post-transcriptional regulation. Indeed, we see in Figure 6 of the manuscript that not in all samples where the iPAC gene is upregulated it is a result of a high local copy number. It will be of interest to consider in future analyses other factors as well.
As shown in Table 1 of the manuscript the list of obtained 56 breast cancer iPAC genes is enriched in known breast cancer relevant genes (e.g. ATAD2) as well as general cancer relevant genes (e.g. ECT2). More interestingly are those breast cancer iPAC genes that were not previously associated to cancer (e.g. PSMD4). In this respect our work not only proposes involvement in breast cancer to cancer related genes but also to genes not previously linked with cancer. We also put their activity in a biological context. In order to substantiate our results we analyzed a similar dataset yet on an independent cohort with different genetic background, collected by a different research group and measured by a different measurement platform. Interestingly, as shown in Figure 8 and Figure 9 of the manuscript the breast cancer iPAC genes found in our cohort retain their properties of being in-cis correlated genes as well as in-trans process associated with the same process in the other cohort of samples. Thus, in addition to having a significant statistical support we show that our results can be reproduced on other independent datasets.

As proof-of-concept, we selected three iPAC genes for siRNA knockdown experiments in MCF7 breast cancer cell-lines and measured viability following knockdown. Out of these, silencing of ECT2 led to significant decrease in cell viability and silencing of PSMD4 led to a less significant yet substantial decrease in cell viability. While ECT2 was previously been associated with cancer, its corroboration in breast cancer is nicely validated. In the case of PSMD4, which is a component of the proteosome, this is in effect a novel association with cancer and with breast cancer in particular; interestingly, our results associated it with the DNA damage module, which is clearly linked with cancer progression, is also validated in this experiment.

4.2 Future work

This thesis describes various approaches for integrated analysis of matched datasets using ranked based statistics. It is clear that while these methodologies can shed light on the underlying biology that is involved there is still much room for extending and further exploring some of the methods used.
4.2.1 Cooperative and complimentary miRNAs target enrichment analysis (CoCo-miTEA)

The miTEA (miRNA target enrichment analysis) tool described in section 3.2 takes as input a ranked list of genes that represents the results of biological measurement assay. Using mmHG statistics (section 2.1.1), miTEA searches for significant mutual enrichment between the input ranked list of genes and the ranked lists of genes obtained through considering targeting by different miRNAs. A significant mutual enrichment is a statistical indication for the activity of the miRNA under consideration in the assayed biological condition. A prominent extension of this approach is to further examine cooperative and complimentary activity of miRNAs by testing the activity of a set of miRNAs in a biological experiment. The set of miRNAs is represented by a single ranked list of genes that should represent the cooperativity or complimentarity of the miRNAs under consideration in regulating their target genes. The miRNA set ranked list of genes is then used in a similar manner as in miTEA to test for mutual enrichment with the user ranked list of genes. Sets of miRNAs to be considered under this approach should be rather small to maintain relevance and interpretability.

There are several ways to take multiple miRNA ranked lists and generate a single representative ranked list based on target prediction pertaining to all miRNA members under consideration. In the case of miRNA complimentarity of action we can assign each gene with a new score ($MIN$) by taking its top rank amongst the ranked lists of genes. Ranking of the genes according to this new $MIN$ score gives preference in the top to genes that are likely targeted by at least one miRNA from the set of miRNAs under consideration. On the other hand, miRNAs cooperativity can be represented by the rank consistency score ($RCoS$, see Appendix 5.2 [34]) that gives preference in the top to genes that are likely targeted by many of miRNAs from the set of miRNAs under consideration. In this case, preliminary results suggest a strong cooperativity between miRNAs (Figure 6).
Figure 6. The expression profiling of HeLa cells following transfection of miR-16 [54]. Enrichment plots are depicted for different sets of miRNAs. Y-axis represent –log transformation of enrichment p-value. miR-15b (of the miR-16 family) has optimal enrichment for mono miRNA set. As can be seen this enrichment is improved more than 20 orders of magnitude in significance when considering cooperation with more miRNAs.

4.2.2 Complexity of the mmHG and mRS scores computation

In both of the mutual enrichment approaches described in previous sections, mmHG (section 2.1.1) and mRS (section 2.1.3), we examine a multiple threshold selection space of $N$ thresholds and optimize our result in a data driven manner. Notice that the multiple threshold testing needs to be corrected to derive the actual level of significance of our result. For a computation of the mmHG statistic, derivation of the $mHG$ p-value for each threshold is needed. As shown above, the complexity of computing each $mHG$ p-value is $O(N^2)$, which yields a total complexity of $O(N^3)$. There are efficient bounds on the $mHG$ p-value, but still the $mHG$ statistic has to be computed. Each $mHG$ run takes $O(N)$ which gives us a total complexity of $O(N^2)$. In the case of the mRS statistic we have shown that $RS$ p-value computation takes $O(N^4)$ which yields a total computation of $O(N^4)$. However, as noted above, for the mRS statistic there is a good approximation of the statistic to the normal distribution. In this case, by sorting the
ranks of one list according to the second list we can get the RS statistic for each threshold in one pass. In this case our total complexity is \( O(N \cdot \log N) \).

The \( mmHG \) relies on the \( mHG \) statistic and as such, it is more robust to extreme outliers in the data. However, relying on efficient approximations the complexity of the \( mRS \) statistic has better complexity and thus execution of the \( mRS \) will be mandatory for large datasets. Further work is needed to better characterize the \( mmHG \) distribution and to improve the both p-value estimation efficiency of computation.

4.2.3 Identifying alternative mechanism for cancer promotion

In the work of Aure & Steinfeld et al. [46] we analyzed a matched breast cancer dataset of gene copy number and gene expression. The goal of this work was to identify genes that are amplified or deleted and as a result are over or under expressed in a manner which leads to the activation of specific cancer related pathways. This approach integrated matched copy number data and expression data with gene annotation databases to hypothesize a means of cancer promotion. An alternative approach would be to harness the existing rich high-throughput information available from multiple levels of profiling technologies. Such rich datasets like the TCGA [55] have extensive information on various cohorts of different cancer types that include copy number and expression data but also methylation, epigenetic marks, proteomics, miRNA, lincRNA, mutation and protein binding information. It will be of interest to expend the approach presented in this thesis to be able to analyze such richer multi-datasets.

4.2.3.1 Internal Rank Consistency Score

One such approach that is under current development is a generalization of the \( RCoS \) methodology (Appendix 5.2 [34]) that deals with a special case of the multi-datasets scenario. In this case our k-fold multi-datasets, \( M_1, \ldots, M_k \), share not only the same set of samples but also the same set of genes. That is \( M_i \) has \( N \) genes and \( m \) samples for every \( 1 \leq i \leq k \). Our goal is to develop a statistical framework that finds genes that are consistently ranked high in each sample in at least one of the datasets. The motivation behind this approach is that if a gene is important for a certain biological process in question then it will result (will rank high) in at least one of the platform
tested. For example, if a gene is a tumor suppressor the cancer will benefit by its removal and thus will choose to either delete it, modify histone epigenetic marks or DNA methylation to inhibit it, repress its expression or mutate it. In either case the activity of this gene will be inhibited and only by observing all parallel levels could we detect it.

To carry out this analysis we propose the Internal Rank Consistency Score (IRCoS). Consider the rank matrices $R_1, \ldots, R_k$, where $R_i$ will hold the gene ranks for each experiment in dataset $i$. That is, $R_i[g,e]$ will hold the rank of gene $g$ in the sample $e$ for dataset $i$. Notice that in this case the highest rank is 1 and the lowest rank is $N$. Next, we generate the high-rank-matrix $R$ to hold for each gene and sample its minimal gene ranks in all datasets. That is, $R(g,e) = \min_{1 \leq i \leq k} R_i(g,e)$. We can now define the IRCoS score $I(g)$ to be the RCoS score for the gene $g$ in the matrix of ranks, $R$.

### 4.2.4 Enrichment network visualization

We are currently working, in a collaboration with Agilent Laboratories, to encapsulate the enrichment network approach, described in Section 2.2.1, as a Cytoscape plug-in [56]. This plug-in is called ENViZ. Release of this freely accessible and user friendly plug-in will make the methods genuinely useful for the research community. See Figure 7 for an example of the enrichment network visualization in the ENViZ environment.
Figure 7. Enrichment network visualization. Visualization of enrichment network (See section 2.2.1) by ENViZ, a freely accessible Cytoscape plug-in to be released in the near future. The enrichment network built from mRNA and miRNA data from Enerly & Steinfeld et al. [42], using WikiPathway annotation [57] is depicted. Results are represented as a bi-partite graph: yellow to red nodes represent pathways and gray nodes represent miRNAs. Edges represent significant associations between pathways and a miRNAs where red edges represent positive correlation and blue edges represent negative correlation.
5 Appendix – Partial list of publications (during my PhD)

We conducted and reported several other studies during my PhD work. The papers that involve aspects central in my thesis, that is integrated analysis and ranked based statistics, are pointed to as supplementary material and a short summary is provided.

5.1 GOrilla: a Tool for Discovery and Visualization of Enriched GO Terms in Ranked Gene Lists [58]

In this study we made use of the \(mHG\) statistics to develop a tool, we named GOrilla, that discovers and visualizes enrichment results of gene ontology (GO) terms in ranked lists. GOrilla is a web-based application that takes as input a ranked list of genes and outputs those GO terms that are found to be enriched in this ranked list in an effective visual manner. GOrilla is currently highly accessible by the scientific community with more than 100 unique users per day on average. GOrilla’s high usage is a result of a) rigorous statistical background that leads to statistically relevant results, b) efficient implementation that leads to reduced running time, and c) result and data driven visualization that leads to better user interface experience. This work currently has more than 200 citations.

5.2 Novel Rank-Based Statistical Methods Reveal MicroRNAs with Differential Expression in Multiple Cancer Types [34]

In this study we introduce a new ranked based statistical framework to handle the identification of consistently over-expressed genes in multiple experiments. In general, this statistics, which we named \(RCoS\) (rank consistency score), takes as input multiple ranked lists and for each element calculates a score that reflects how likely it is to be that highly ranked given a random set of ranked lists of elements. Using dynamic programming, \(RCoS\) also evaluates the significance of this result. In this study the \(RCoS\) statistical tool was employed to analyze a multiple cancer type miRNA expression dataset. The result of this study was the identification of those miRNAs that are up/down regulated in many cancer types. This work currently has more than 50 citations.
5.3 Developmental Programming of CpG Island Methylation Profiles in the Human Genome [59]

In this study we introduce a proprietary dataset of comprehensive methylation profiling of CpG islands in the human genome. As CpG islands are areas in the genome that are protected from DNA methylation their comprehensive characterization in many different tissue types holds potential to improve our understanding of the role of DNA methylation in development and differentiation. In this work I was involved in the assay design but mainly in the development of computational tools to analyze this dataset. As part of this analysis we used ranked based statistics to identify DNA motif elements that distinguish between constitutively methylated islands and constitutively un-methylated islands. Using these statistically supported features and using machine-learning methodologies, we developed a predictor for CpG methylation. Our results in effect provided a redefinition for the concept of CpG islands, which we named under-methylated-regions. Adapting an assay design to cover these under-methylated-regions we were able to uncover thousands new areas in the genome that were found to be protected against methylation and that might have important roles in regulating developmental programming. This work currently has more than 150 citations.

5.4 EGF Decreases the Abundance of MicroRNAs That Restrain Oncogenic Transcription Factors [60]

In this work we explored how epidermal growth factor (EGF) affects the miRNA regulatory programs in cell-lines. EGF is known to regulate programs of gene expression that are frequently deregulated in cancer and as miRNAs are also broadly implicated in cancer their regulation by EGF is of importance. Using comprehensive miRNA profiling on Agilent microarrays we discovered programs of miRNAs that are inhibited in an early stage after EGF stimulus. Using ranked based statistics and integrated analysis of the miRNA data to matched mRNA data we were able to show that this observed effect on miRNA is consistently observed in primary tumors. Specifically, we showed that miRNAs that we observe to be downregulated as a response to EGF stimulus are also repressed, with high statistical significance, according to our rank base approach, in terms of expression in EGFR and HER2 amplified samples within brain and breast tumor
cohorts. Thus, we showed that these miRNAs act as attenuators of growth factor signaling and oncogenesis. This work currently has more than 30 citations.
6 References


ניהלה נחונ אדם מבקריה המשלבים סוגים
שונם של מזרדות כל גוניית

ישראל שטיינפלד
ניתוח-transparentים מעודכנים ומשלים שוניה של מריית כל גנומית

זר blir על מתקרר

לשמ מיילים שלкар של תרומות לكلف הקבוצה התוראי

documor הפילוסופיה

ישראל שטיינמפלד

הוגשת לקמיה כלכלי - מרכז טכנולוגיה בישראל

אב 5573, דיוג, אוגוסט 2013
המאתקר נעשתה בהנחייתו של ד"ר ז"ר ירכי ממחולקה למגфи מחשב

אני מודה לاذכירות על התוכנות המ mụיפה וнимаנה לולמחולקה שמחולקה

Technion - Computer Science Department - Ph.D. Thesis PHD-2013-15 - 2013
1.1 microRNAs

1.3 HG

1.3.1 (HG)

1.3.2 (mHG)

1.3.3 (RS)

1.4

1.4.1

1.4.2

2

2.1

2.1.1 mmHG or m²HG

2.1.2 (QmHG)

2.1.3 (mRS)

2.2

2.2.1

2.2.2

2.2.3

3

3.1 miRNA-mRNA

3.2

3.3

4

4.1

4.1.1

4.1.2
תקציר


(Clusters) של נגירה, היא צורת העברת עצים כתובות מחברות של נגירה, היא צורת העברת עצים כתובות מחברות של נגירה, היא צורת העברת עצים כתובות מחברות של נגירה, היא צורת העברת עצים כתובות מחברות של נגירה, היא צורת העברת עצים כתובות מחברות של נגירה, היא צורת העברת עצים כתובות מחברות של נגירה, היא צורת העברת עצים כתובות מחברות של נגירה, היא צורת העברת עצים כתובות מחברות של נגירה, היא צורת העברת עצים כתובות מחברות של נגירה, היא צורת העברת עצים כתובות מחברות של נגירה, היא צורת העברת עצים כתובות מחברות של נגירה, היא צורת העברת עצים writman

 uniq

minimal-

hypergeometric

hypergeometric

hypergeometric

hypergeometric

minimal-
micro-RNA

Radium
miR29
הכרטוניטיות עלתה. גישה זו איפשרה להבחין לא רק במוצר ה丏伸び אלא גם בתהליך Миיניוו
ש湲שה ויתר לעקרוב החורשים בתוך ערים הкажמים בשפרות.