Journal of Bioinformatics and Computational Biology
Vol. 13, No. 4 (2015) 1550018 (18 pages)
© Imperial College Press
DOI: 10.1142/S0219720015500183

Imperial College Press
www.icpress.co.uk

A quantum leap in the reproducibility, precision, and sensitivity of gene expression profile analysis even when sample size is extremely small

Kevin Lim[*,§], Zhenhua Li[†], Kwok Pui Choi[‡] and Limsoon Wong[*,¶]

[*]*School of Computing*
National University of Singapore
13 Computing Drive, Singapore 117417

[†]*Department of Pediatrics*
National University of Singapore
10 Medical Drive, Singapore 117597

[‡]*Department of Statistics and Applied Probability*
National University of Singapore
6 Science Drive 2, Singapore 117546
[§]*kevinl@comp.nus.edu.sg*
[¶]*wongls@comp.nus.edu.sg*

Received 13 May 2015
Revised 14 May 2015
Accepted 14 May 2015
Published 7 July 2015

Transcript-level quantification is often measured across two groups of patients to aid the discovery of biomarkers and detection of biological mechanisms involving these biomarkers. Statistical tests lack power and false discovery rate is high when sample size is small. Yet, many experiments have very few samples (≤ 5). This creates the impetus for a method to discover biomarkers and mechanisms under very small sample sizes. We present a powerful method, ESSNet, that is able to identify subnetworks consistently across independent datasets of the same disease phenotypes even under very small sample sizes. The key idea of ESSNet is to fragment large pathways into smaller subnetworks and compute a statistic that discriminates the subnetworks in two phenotypes. We do not greedily select genes to be included based on differential expression but rely on gene-expression-level ranking within a phenotype, which is shown to be stable even under extremely small sample sizes. We test our subnetworks on null distributions obtained by array rotation; this preserves the gene–gene correlation structure and is suitable for datasets with small sample size allowing us to consistently predict relevant subnetworks even when sample size is small. For most other methods, this consistency drops to less than 10% when we test them on datasets with only two samples from each phenotype, whereas ESSNet is able to achieve an average consistency of 58% (72% when we consider genes within the subnetworks) and continues to be superior when sample size is large. We further show that the subnetworks identified by ESSNet are highly correlated to many references in the biological literature. ESSNet and supplementary material are available at: http://compbio.ddns.comp.nus.edu.sg:8080/essnet.

Keywords: Microarray data analysis; gene expression profiling; subnetworks; biological pathways; GSEA; SNet; FSNet; PFSNet.

1. Introduction

Over the past decade, many methods have been proposed to find relevant disease-causing mechanisms. A p-value is often associated to these disease-causing mechanisms, depicting the level of significance in which the mechanism is enriched in a phenotype.

Extremely small sample size ($N \leq 5$) limits many of these approaches. For instance, permutation tests cannot reliably compute a p-value in these cases. Scenarios with extremely small sample size are not uncommon. For example, model organisms and cultured cell lines often have two main phenotypes with multiple repeated experiments. Moreover, newer technologies that measure gene expression using RNA-seq can be costly and fewer samples are considered because of the economics of driving a large-scale study.

On the other hand, numerous studies have also shown that large sample sizes are required to maintain high statistical power and low familywise error rate or low false discovery rate.[1–4] For example, one such model requires more than 100 samples to achieve power at 0.9 and false discovery rate at 0.05.[1]

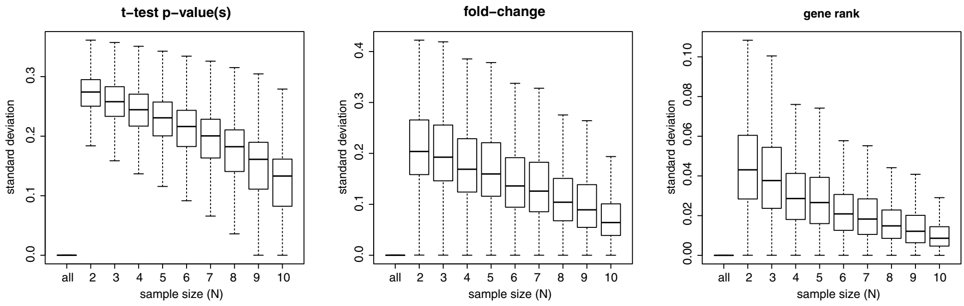
To date, no method is able to identify disease mechanisms in extremely small sample size situations. Even in a moderately large sample size situation, microarray analysis shows low consistency when applied to independent datasets of the same disease phenotypes, i.e. the genes and/or subnetworks identified by these methods are very different in two independent datasets.[5] The rare exceptions are SNet and its refinement PFSNet, which manage to get high agreement across independent datasets.[6,7] However, these works do not study their performance when sample sizes are small.

This raises the question of whether disease mechanisms can be consistently identified under small sample size situations. In this paper, for gene expression analysis methods like GSEA, ORA, and PFSNet, we show that the agreement between two independent datasets drops when a small number of samples is used in the analysis.[7–9] We further present a new method, ESSNet (Extremely Small sample size Subnetworks), which is shown to be consistent across independent datasets, even when sample size is very small.

2. Background

Many methods are available for detecting significant differential gene expression, most of which have reported results on datasets with reasonably large sample sizes.

The earliest techniques test individual genes for differential expression using fold change, t-test, SAM, etc.[10] However, these techniques are extremely sensitive to



(a) (b) (c)

Fig. 1. Effects of sample size. P-values, log fold changes and gene-expression-level rankings are computed using data taken from Ref. 11 and their ranks are recorded, for every sample size (N) considered ranging from 2 to 10. This process is repeated over 100 times and the standard deviations of the respective ranks are measured. (a) t-test p-values are very sensitive to sample size variation, e.g. the standard deviation of the ranks of the p-values of a gene can be as large as 0.35 when $N = 2$. (b) log fold change is also very sensitive to sample size, e.g. the maximum standard deviation of the ranks of the log fold changes of a gene is 0.4 when $N = 2$. (c) gene ranking based on expression level is less sensitive to sample size, the maximum standard deviation is about 0.1 when $N = 2$.

small sample sizes; see Fig. 1. For example, the gene rankings based on log fold-change and t-test p-values have large deviations between different samples when sample size is small. In addition to this, individual gene testing also faces large amounts of false positives due to multiple hypothesis testing.

In recent years, gene expression analysis is increasingly performed in the context of pathways or gene sets to circumvent the problem of large false positives in multiple hypothesis testing of individual genes and to improve interpretability of the results. For example, overlap analysis tests whether the proportion of differentially expressed genes in a pathway is significantly different from a random gene set.[9] Direct-group methods like GSEA and FCS compute a p-value representing the significance of the correlation of an entire pathway to phenotypes.[8,12] Network-based methods like NEA, DEAP, SNet, and PFSNet select smaller components (subnetworks) within a pathway and test whether these subnetworks have significant correlation with phenotypes.[6,7,13,14] Model-based methods like SRI and GGEA construct a dynamic model for a pathway using one phenotype and test whether the model is inconsistent in the other phenotype.[15,16]

These approaches do not work well when sample size is small for various reasons: (1) They involve an intermediate step of computing differential expression of genes within a pathway by fold change, t-test, etc., which are very sensitive to sample sizes. (2) Model-based methods cannot learn parameters with very few training samples. (3) Permutation test cannot be reliably computed because of the limited number of class-label permutations.

Even in a moderately large sample size setting, these methods often return results that are irreproducible when applied to independent datasets.[7]

3. Method

The key idea of ESSNet is to fragment large pathways into smaller subnetworks and compute a statistic that discriminates the subnetworks in two phenotypes that is stable even when sample size is small. ESSNet comprises two main steps: subnetwork generation and subnetwork scoring. These steps are described below.

3.1. *Subnetwork generation*

As biological pathway repositories have very little agreement, the choice of the pathway database used affects the results of gene-set-based microarray analysis. We use pathways from PathwayAPI, a database that unifies popular pathway databases — KEGG, Wikipathways, and Ingenuity (www.ingenuity.com) — so that the biological information is as comprehensive as possible.[17–19]

For each patient in phenotype D, we rank the genes by their expression values in decreasing order so that the most highly expressed gene is assigned the rank 1, the second most highly expressed gene the rank 2, and so on. Let $r(g_i, p_j)$ be the rank of gene i in patient j. We tested and found that gene ranks do not fluctuate as much due to sample size variation as fold change or p-values from t-test; see Fig. 1. Each gene is then given a rank based on the average among the patients of phenotype D:

$$\text{rank}_D(g_i) = \sum_{j \in D} \frac{r(g_i, p_j)}{|D|}, \tag{1}$$

where $|D|$ is the number of samples belonging to the phenotype D.

We obtain a gene list extracted from the top $\alpha\%$ of the gene ranks computed in Eq. (1). We chose $\alpha = 10$ in our experiments. Genes not in this list are removed from every pathway, thus fragmenting each pathway into smaller connected components (i.e., the subnetworks). We only consider subnetworks that are of size at least 5. The subnetworks for phenotype $\neg D$ are generated analogously.

3.2. *Subnetwork testing*

Methods that score subnetworks based on individual samples may not be able to do so reliably when sample sizes are small. For example, one might use t-statistics to score a disease subnetwork with a mean value of 15 in phenotype D and 51 in $\neg D$. With two samples in each group: 10, 20 and 50, 52, a simple t-test produces a p-value of 0.077, whereas with more samples: 9, 10, 20, 21 and 49, 50, 52, 53, the p-value drops to 0.0008. This demonstrates that many existing methods may produce dramatically different outcomes when sample sizes are varied.

Our subnetwork score is based on a novel idea. We postulate that, when a subnetwork is irrelevant to the distinction between phenotypes D and $\neg D$, the difference of the expression values of any gene in this subnetwork in any pair of samples of D and $\neg D$ should be very small. Suppose there are k genes in a subnetwork, m samples in phenotype D and n samples in phenotype $\neg D$. Then there are $m * n$ possible pairs of differences for each of the k genes. Let $\delta(g_i, p_j, p'_l) = e(g_i, p_j) - e(g_i, p'_l)$ for each p_j in D, p'_l in $\neg D$ and g_i in subnetwork s, where $e(a, b)$ represents the expression value of gene a in patient b. According to the postulate, if the subnetwork is irrelevant, these $M = k * m * n$ paired differences should be distributed around 0. Returning to our example of using two samples per group, although we have only two samples per phenotype, we can have up to $4 * k$ values in $\delta(g_i, p_j, p'_l)$ if k is the size of the subnetwork.

We propose using the t-statistic formula, $T_s = \frac{\mu}{sd/\sqrt{M}}$, where μ and sd are respectively the mean and standard deviation of $\delta(g_i, p_j, p'_l)$ in the subnetwork s, and $M = k * m * n$, as a test statistic for evaluating whether the set of pair differences $\delta(g_i, p_j, p'_l)$ in the subnetwork s is distributed around 0. Note that, even though we use the t-statistic formula, its significance should not be evaluated based on the t-distribution of M degrees of freedom in the standard way for two reasons. Firstly, the null hypothesis (that the subnetwork s is irrelevant) does not imply the null distribution is a t-distribution. Secondly, the paired differences are not completely mutually independent — the actual degrees of freedom is somewhere between $m + n$ and $k * (m + n)$ depending on how tightly genes in the subnetwork s are co-regulated.

Our conjecture that $\delta(g_i, p_j, p'_l)$ is a distribution around 0 can be tested on a null distribution generated based on a valid null hypothesis, according to the principles of exchangeability. There are two common ways for generating null distributions in gene expression analysis, in which randomized columns or rows of the expression matrix are used to re-compute the statistic over a number of iterations.

The first way assumes the null hypothesis that the subnetwork being tested is irrelevant to distinguishing the two phenotypes. Thus the gene expression profiles of any pair of patients from the two phenotypes are exchangeable for computing points in the null distribution. In other words, class labels are randomly swapped to create new data inputs from which the null distribution is formed. This method is used by GSEA to evaluate the significance of the Kolmogorov–Smirnov test statistic of a pathway when sample size is sufficiently large. Naturally, this method preserves the full gene–gene correlations in each patient. However, when sample size is small there are limited ways for permuting class labels, resulting in a sparse null distribution. This greatly affects the reliability and the granularity of the p-values.

The second way postulates that any two gene expression values within the same patient are exchangeable to compute the null distribution. This method creates new data inputs by randomly re-labeling genes. This method is used by GSEA to evaluate the significance of the Kolmogorov–Smirnov test statistic of a pathway when the dataset has a small sample size, since a sizeable null distribution can be generated this way. However, this postulate is based on the assumption that the genes' expression are independent of each other, ignoring the correlation between genes. In other words, this method actually tests if the genes in the pathway behave no differently from a random set of genes. But the genes in any pathway are coordinated by nature, whereas a random set of genes is not. Hence this null hypothesis is false. So it has a tendency of being rejected, producing false positives.

We rely instead on a third way to produce a null distribution for our test. It postulates that randomized gene expression profiles that preserve the gene–gene correlation structure in the original dataset are exchangeable with it. This postulate is based on the assumption that genes in any pathway are as coordinated as specified by the pathway, and the pathway is functional when the genes behave — i.e. have correlated expression — as specified by the pathway. Due to exchangeability following this postulate, it is sound to use correlation-preserving randomized gene expression profiles to obtain a null distribution of the test statistic. Array rotation is one of the known techniques for producing a large number of these correlation-preserving randomized gene expression profiles.[20] We use this technique to produce statistically valid p-values for our test statistic. We call this the rotation t-test, to distinguish it from the standard t-test.

The details about the computation of our test statistic as well as other improvements and variants are discussed in the supplementary material sections S1–S3.

4. Results

For each disease type, we use two independent microarray data sets from previously published experiments: Leukemia,[21,22] Childhood Acute Lymphoblastic Leukemia (ALL Subtype)[23,24] and Duchenne Muscular Dystrophy (DMD).[11,25] We use the notation dataset 1 and dataset 2 to refer to the former and latter datasets, respectively.

4.1. *Comparing subnetwork- and gene-level overlap*

We randomly partition the two independent datasets into subsets of smaller sample sizes ranging from 2 to 10 from each phenotype. In order to observe the effect of sample size on various methods, we compare the subnetwork overlap of the corresponding methods with varying sample sizes.

For every sample size (N) considered, we partition the datasets accordingly and use the subnetwork generation procedure mentioned in Sec. 3.1 to generate the subnetworks in one dataset. We then test these subnetworks for statistical significance, under a significance threshold of 5% using the rotation t-test mentioned in Sec. 3.2 on the two datasets independently. The subnetwork overlap is a Jaccard-like agreement, defined as follows: Let the two sets of significant subnetworks identified by dataset 1 and dataset 2 using N samples be SN_1^N and SN_2^N, respectively. Then the subnetwork-level agreement is defined as

$$\frac{|\mathrm{SN}_1^N \cap \mathrm{SN}_2^N|}{|\mathrm{SN}_1^N \cup \mathrm{SN}_2^N|}. \tag{2}$$

There are many ways to partition a dataset of M samples into subsets of N samples. For our experiments, we test the procedure many times and report the average subnetwork-level agreement.

Since ORA and GSEA identifies whole pathways instead of subnetworks, in testing these methods, we measure the pathway-level agreement which is defined analogously.

We also measure the overlap in genes between the predicted subnetworks, which is defined analogously below, where Genes_i^N denotes the set of genes in SN_i^N:

$$\frac{|\mathrm{Genes}_1^N \cap \mathrm{Genes}_2^N|}{|\mathrm{Genes}_1^N \cup \mathrm{Genes}_2^N|}. \tag{3}$$

We compare the subnetwork-level agreement of our method, ESSNet-unweighted, with other gene set methods (ORA-hypergeo, ORA-paired, GSEA, NEA-paired, DEAP, and PFSNet); see Fig. 2.

ORA-hypergeo is the usual overlap analysis method. It tests whether a pathway is significant by intersecting the genes in the pathway with a pre-determined list of differentially expressed genes (here, we use all genes whose t-statistic meets the 5%

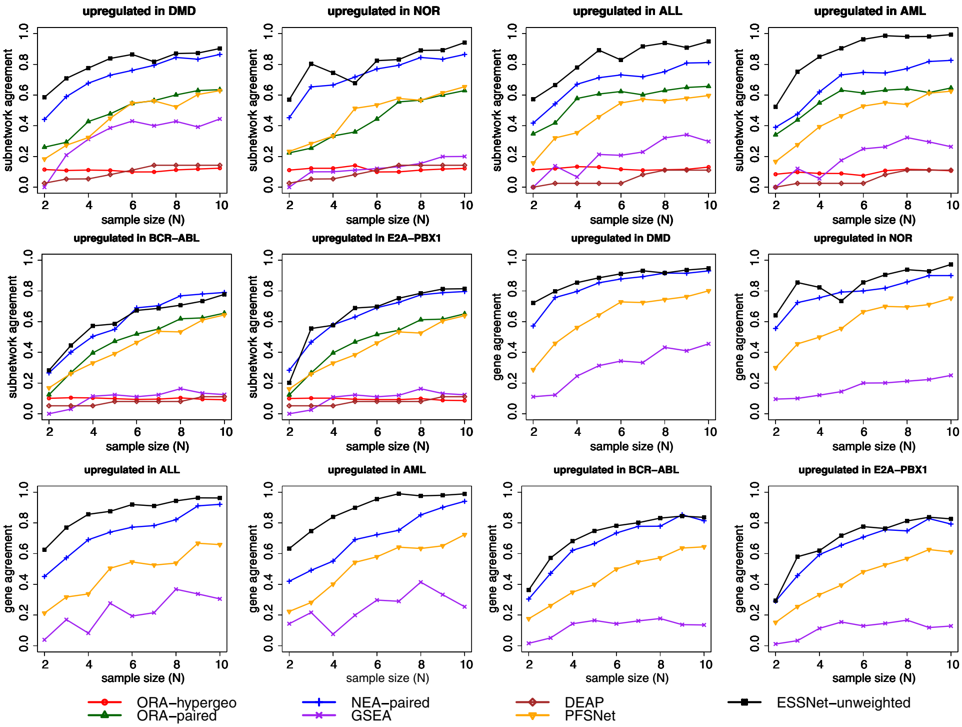


Fig. 2. Consistency of subnetworks and their genes in Leukemia (ALL/AML), ALL Subtype (BCR-ABL/E2A-PBX1) and DMD dataset (DMD/NOR) computed using dataset partitions of smaller sample sizes ranging from 2 to 10 from each phenotype.

significance based on the standard t-test), and checking the significance of the size of the intersection using the hypergeometric test. ORA-paired is a modification of ORA-hypergeo; it does not use a pre-determined list of differentially expressed genes and the hypergeometric test. Instead, it applies the rotation t-test described in Sec. 3.2 using all the genes in the pathway. GSEA is a direct-group method based on the Kolmogorov–Smirnov test statistic as described in Sec. 3.2. As sample size is small, the gene permutation option is used to evaluate significance. NEA-paired is a network-based method where each gene and its immediate neighborhood form a subnetwork. The subnetworks are subjected to the rotation t-test discussed in Sec. 3.2. DEAP examines all possible maximal linear paths in the pathway and chooses the path with maximum absolute differential expression score. The score given for a path is recursively computed based on the catalytic or inhibitory edges taken as positive and negative summands, respectively.[14] PFSNet is a network-based method as previously mentioned in Sec. 2. ESSNet-unweighted generates subnetworks based on the method discussed in Sec. 3.1 and tests each subnetwork for statistical significance using the rotation t-test from Sec. 3.2.

ORA-hypergeo has very low pathway-level overlap even when sample size is 10; cf. Fig. 2. There are three weaknesses that contribute to its poor performance. Firstly, it amounts to testing whether the entire pathway is significantly differentially expressed. If only a branch of a large pathway is relevant to the phenotypes, the noise from the large irrelevant part of the pathway can mask the signal from that branch. Secondly, it relies on a pre-determined list of differentially expressed genes. This list is sensitive to the choice of threshold that defines which genes are considered as differentially expressed. And, irrespective of the threshold, as shown in Fig. 1, this list lacks consistency when sample size is small. Thirdly, its use of the hypergeometric test corresponds to the null hypothesis that genes in a pathway behave no differently from random sets of genes of the same size as the pathway. As genes in a pathway are coordinated in their behavior to perform the specific function associated with the pathway, this null hypothesis is false.

ORA-paired circumvents the second weakness of ORA-hypergeo since it does not need any list of differentially expressed genes. It also eliminates the third weakness of ORA-hypergeo since it uses a biologically much more plausible null hypothesis that genes in the pathway have similar expression values between the two phenotypes if the pathway is irrelevant to the difference of the two phenotypes. Therefore, ORA-paired performs much better than ORA-hypergeo. The subnetwork-level agreement increases to as high as 65% versus 13% in ORA-hypergeo when $N = 10$ and 34% versus 11% when $N = 2$; cf. Fig. 2. This suggests that the rotation t-test on paired differences is a strategy that works extremely well in a small sample size situation.

A disease could be the result of the dysfunction of a small part of a large pathway. In this situation, most of the genes in this large pathway may not be differentially expressed. Even though ORA-paired has improved on ORA-hypergeo, it is still unlikely to find this large pathway significant. That is, ORA-paired retains the first weakness of ORA-hypergeo. Thus, it makes sense to extract subnetworks from pathways and test these subnetworks individually for significance.

We apply the NEA idea to generate candidate subnetworks from a pathway.[13] The idea is to take every gene and its immediate neighbors in the pathway to be a subnetwork. After this, we apply ORA-paired to determine the significance ones. This NEA-paired approach circumvents all three weaknesses of ORA-hypergeo. Hence it performs even better than ORA-paired. As shown in Fig. 2, the subnetwork-level agreement increases to as high as 85% when $N = 10$ and 43% when $N = 2$. This suggests that the subnetwork generation procedure increases the sensitivity of NEA-paired. We believe this is because paired differences around the neighborhood of selected genes enable the test to correctly reject subnetworks that have no differentially expressed genes within them.

GSEA also suffers less from the second weakness of ORA-hypergeo because it does not need any list of pre-determined differentially expressed genes. Nevertheless, GSEA does not completely escape from this weakness because its Kolmogorov–Smirnov test statistic is based on the rank of the t-statistic values of genes; these ranks are unstable when sample size is small, cf. Fig. 1. Moreover, GSEA still retains the first weakness of ORA-hypergeo. And, when the gene permutation option is used to determine the significance of the Kolmogorov–Smirnov test statistic, as in this paper, it also retains the third weakness of ORA-hypergeo. Therefore, while it outperforms ORA-hypergeo, it is inferior to ORA-paired and NEA-paired. GSEA achieves a maximum pathway level overlap of 45% when N is 10 and 27% when $N = 3$, cf. Fig. 1. We are unable to evaluate GSEA when $N = 2$ because it requires sample size of at least 3.

DEAP partially eliminates the first weakness of ORA-hypergeo because it breaks the pathway into maximal linear paths. However, only the best scoring maximal linear path within a pathway is reported; this considerably reduces its reproducibility because a different maximal linear path may be chosen in another dataset. This problem is further compounded as DEAP scores the paths based on differential gene expression, which we have shown in Fig. 1 to be unstable in small-sample-size situations. Consequently, DEAP has very poor performance at the subnetwork level. Thus, we evalute DEAP at the pathway level, where a pathway is considered to be reported by DEAP if any path within the pathway is reported. Unfortunately, despite this, DEAP still does poorly. As shown in Fig. 2, DEAP achieves a maximum pathway-level overlap of 28% when N is 10 and 6% when N is 2.

PFSNet does not need any list of pre-determined differentially expressed genes, eliminating the second weakness of ORA-hypergeo. It generates subnetworks, using a technique different from NEA, and so eliminates the first weakness of ORA-hypergeo. For each subnetwork and each patient, it computes a pair of scores for that patient based on phenotype D data and phenotype $\neg D$ data, respectively. It postulates very reasonably that, if the subnetwork is irrelevant to the difference between D and $\neg D$, these pairs of scores should be distributed around 0. It then uses class-label permutations to evaluate this null hypothesis. Thus PFSNet also eliminates the third weakness of ORA-hypergeo. However, when sample size is small, the null distribution cannot be properly produced using class-label permutations. Thus PFSNet has good performance when N is reasonably high but inferior performance when N is small. As shown in Fig. 2, PFSNet has an overlap of 65% when $N = 10$ and 21% when $N = 2$.

Finally, we apply the same sets of tests to ESSNet-unweighted, which selects subnetworks as described in Sec. 3.1 and tests these subnetworks for significance using the rotation t-test in Sec. 3.2. Clearly, ESSNet-unweighted also eliminates all three weaknesses of ORA-hypergeo in a manner analogous to NEA-paired. It has excellent performance, superior to all other methods studied here. We get generally higher subnetwork overlap of up to 99% when $N = 10$ and 58% when $N = 2$; cf. Fig. 2. In addition, ESSNet continues to be superior even when a large dataset is used; see supplementary material section S4. We believe ESSNet-unweighted performs better than other methods because of the following additional reasons.

Even though NEA-paired performs quite well, its subnetwork is based on a seed gene and its immediate neighboring genes in that pathway, regardless of whether those neighboring genes are themselves differentially or highly expressed. This can potentially cause a loss in signal, especially when the seed gene has a large number of immediate neighbors. Moreover, such a subnetwork cannot capture a long causal chain of genes. These two issues are rectified in ESSNet-unweighted which forms a subnetwork in a pathway based on a connected component comprising entirely of highly expressed genes and, as shown earlier in Fig. 1, relying on gene ranking based on expression level (rather than differential expression level) is more robust to sample-size variation.

4.2. *Precision and recall*

As ESSNet-unweighted attains very high subnetwork overlap when the sample size is large, it is possible to define a set of gold-standard subnetworks as follows, to estimate the false-positive and false-negative subnetworks induced by small samples:

$$G = \mathrm{SN}_1^{\mathrm{all}} \cap \mathrm{SN}_2^{\mathrm{all}}, \tag{4}$$

where $\mathrm{SN}_i^{\mathrm{all}}$ is the set of significant subnetworks produced by ESSNet-unweighted based on the entire dataset i.

The precision and recall are defined respectively as:

$$\mathrm{precision} = \frac{|\mathrm{SN}^N \cap G|}{|\mathrm{SN}^N|}, \quad \mathrm{recall} = \frac{|\mathrm{SN}^N \cap G|}{|G|}, \tag{5}$$

where SN^N is the set of significant subnetworks produced by ESSNet-unweighted using an N-sample subset of one entire dataset.

It is surprising that precision does not drop much even when smaller sample sizes are considered. For example, we get a precision of about 90%, 85%, and 88% even when $N = 2$. On the other hand, the maximum recall when $N = 2$ is about 50%; cf.

Table 1. Precision and recall of ESSNet-unweighted.

		Precision						Recall					
		DMD		ALL		BCR		DMD		ALL		BCR	
		D	$\neg D$	D	$\neg D$	D	$\neg D$	D	$\neg D$	D	$\neg D$	D	$\neg D$
Sample size (N)	2	0.96	0.88	0.87	0.95	0.93	0.91	0.45	0.31	0.34	0.25	0.19	0.17
	3	0.93	0.86	0.99	0.89	0.90	0.87	0.56	0.45	0.56	0.41	0.21	0.16
	4	0.88	0.88	0.97	0.92	0.91	0.87	0.67	0.50	0.51	0.53	0.35	0.48
	5	0.89	0.88	0.94	0.90	0.89	0.90	0.73	0.52	0.74	0.55	0.36	0.38
	6	0.82	0.88	0.93	0.92	0.89	0.91	0.78	0.62	0.74	0.62	0.44	0.44
	7	0.85	0.86	0.95	0.93	0.90	0.87	0.75	0.59	0.66	0.64	0.55	0.53
	8	0.84	0.89	0.97	0.94	0.90	0.92	0.81	0.69	0.74	0.66	0.61	0.66
	9	0.88	0.90	0.94	0.92	0.89	0.89	0.90	0.67	0.76	0.74	0.65	0.67
	10	0.88	0.93	0.97	0.92	0.90	0.90	0.86	0.84	0.89	0.74	0.66	0.73

Table 1. Thus, more bona fide subnetworks are missed from the predictions when N is very small, while few false positives are produced. This is reasonable as a small sample may not have captured all the causes underlying a phenotype.

4.3. *Comparing the number of predicted subnetworks using negative control data*

It is also possible to test whether ESSNet is robust to false positives. We conduct in-silico testing by randomly generating matrices of gene expression data; for each gene we sample from a random normal distribution, using the same mean and standard deviation in both phenotypes. The purpose of the test is to see if ESSNet detects any subnetworks as significant when it should not. On these random input matrices, ESSNet reports very small number of false subnetworks (typically about 3), well within that expected from the p-value threshold and much fewer than other methods; cf. Fig. 3.

4.4. *Informative subnetworks*

While biological pathways provide a wealth of information to explain disease phenotype, large pathways offer little biological insight. On the other hand, subnetworks may narrow down the biological cause of a disease but very small subnetworks are trivial and non-informative. In order to assess how informative our significant subnetworks are, we compare the size of the significant subnetworks identified by ESSNet with those subnetworks induced from individual genes declared significant by t-test.

When subnetworks are induced using significant individual genes from t-test, the genes are scattered over the pathways and have very few edges with other significant genes in the pathway. This results in very-small-sized subnetworks that contains little useful biological information. In contrast, the subnetworks detected by ESSNet are bigger and thus more informative; cf. Fig. 4.

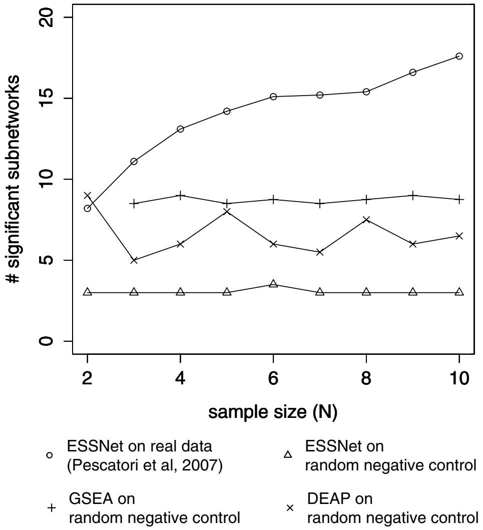


Fig. 3. Comparing various methods on random negative control data (gene expression sampled from the same random normal distributions from both phenotypes). ESSNet predicts fewer false subnetworks than other methods.

Another way to determine how informative our predicted subnetworks are, is to see if they overlap with results produced by other methods. We select significant subnetworks predicted by ESS and test them using GSEA. While GSEA often does not declare a pathway to be significant when the entire pathway is supplied as input, it often declares the subnetworks identified by ESSNet in that pathway to be significant. Specifically, GSEA is able to recover 100%, 51%, and 54% of the subnetworks in the DMD, Leukemia and ALL Subtype dataset, respectively. When PFSNet is included in the analysis, the percentages increased to 100%, 90%, and 91%, respectively. This demonstrates subnetworks predicted by ESSNet can be recovered by other methods (provided these methods are supplied the subnetworks as input, and

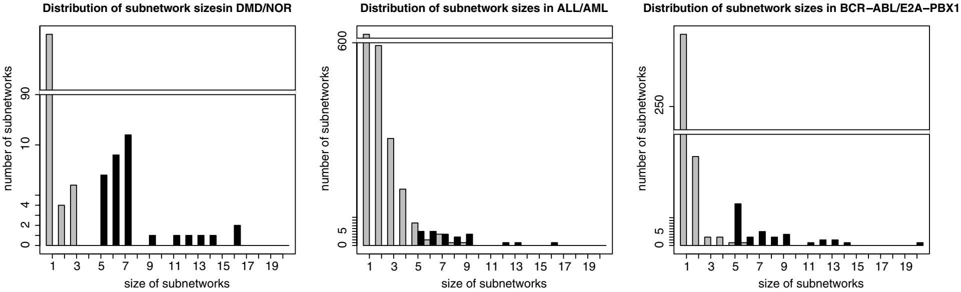


Fig. 4. The distribution of subnetwork sizes. The gray bars correspond to subnetworks induced by the significant genes using t-test, the black bars correspond to subnetworks from ESSNet-unweighted.

Table 2. Biologically relevant subnetworks predicted by ESSNet.

DMD ($N = 2$)	Leukemia ($N = 2$)	ALL Subtype ($N = 4$)
PI3K/Akt signaling	ERK/MAPK signaling	Antigen processing
PTEN signaling	Toll-like receptor signaling	IFNG signaling
ECM receptor	Apotosis signaling	Wnt signaling
Actin cytoskeleton signaling	JAK/STAT signaling	IL-4 signaling
Striated muscle contraction	Antigen processing	JAK/STAT signaling
Integrin signaling	Metab. of xenobiotics by P450	T-Cell receptor

not the entire pathways they come from), and also suggests the plausibility that they are useful and pertinent.

4.5. *Biologically significant subnetworks*

The subnetworks predicted by ESSNet have very strong biological relevance even when a small sample size is used. We consider sample sizes of 2, 2, and 4 for the DMD, Leukemia, and ALL Subtype datasets respectively as these sample sizes give roughly the same subnetwork agreement; cf. Fig. 2. As there are many different predictions since there are many ways to partition the data into subsets of smaller sample sizes from the entire dataset, we report the subnetworks that are detected most frequently in Table 2. Examples of these subnetworks are found in the supplementary material section S5, Figs. F3–F6.

For DMD, striated muscle contraction and actin cytoskeleton signaling are the main cause of the disease.[26,27] ESSNet is not only able to detect these two subnetworks but also other biologically significant signaling pathways that might be the trigger for these main pathways. For example, PTEN signaling contributes to PI3K/Akt signaling which in turn affects the DMD gene found in striated muscle contraction.[28,29] ECM receptor interaction has also been implicated in DMD.[30]

For Leukemia, numerous works have reported the involvement of ERK/MAPK signaling, Toll-like receptor signaling and JAK/STAT signaling in interfering with apoptosis.[31–33] Other subnetworks like antigen processing and metabolism of xenobiotics by cytochrome P450 have also been linked to Leukemia.[34,35]

Similarly, for ALL Subtype, the various subnetworks identified also have biological support, including antigen processing, IFNG signaling, Wnt signaling, IL-4 signaling, JAK/STAT signaling, and T-Cell receptor signaling.[36–40]

5. Conclusion

In this paper, we discuss how extremely small sample size ($N \leq 5$) can affect gene expression analysis. We have demonstrated that many existing methods perform poorly when sample size is small. An ideal method should be able to pick out all relevant factors underlying the phenotypes that are present in a given sample set and should not report any irrelevant factors. It follows from this ideal that we can expect a good method to satisfy these three hallmarks: (i) The selected subnetworks are reproduced when applied to new batches of data that are sufficiently representative of the phenotypes. (ii) The selected subnetworks from a large dataset should be a superset of those chosen from a subset of the dataset. (iii) The relevant subnetworks can be identified using as small a dataset as possible.

We are able to reproduce similar subnetworks in independent batches of data, this is evident in the high subnetwork-level agreement; cf. Fig. 2. ESSNet also demonstrates very good precision, when compared against a set of gold-standard subnetworks derived from the full datasets; cf. Table 1 and supplementary material, Table T1. This suggests that most of the subnetworks predicted using a small sample size are also detected in the large dataset and further implies that it does not produce a lot of false positives even when sample size is small. On the other hand, ESSNet misses out on some gold-standard subnetworks because the small number of samples are unable to capture all the underlying phenotypic differences. However, ESSNet is also superior to other methods for large sample sizes; see supplementary material section S4.

Our method, ESSNet, is unlike other previously discussed methods because we do not greedily select genes to be included based on differential expression but rely on gene-expression-level ranking within a phenotype, which is shown to be stable even under extremely small sample sizes. In addition, our conjecture that $\delta(g_i, p_j, p'_l)$ is a distribution around 0, is tested on a null distribution obtained by array rotation; this preserves the gene–gene correlation structure and is suitable for datasets with small sample size. This allows us to consistently predict relevant subnetworks even when sample size is small. We have also provided various other options in our ESSNet software, which can be downloaded at http://compbio.ddns.comp.nus.edu.sg:8080/essnet/to allow ESSNet to change the type of test statistic and to be more robust to the threshold used; these are described in the supplementary material sections S2–S3.

The subnetworks that we discover using ESSNet are supported by relevant biological literature and have the potential to allow biologist further insights to the mechanism behind the diseases. Examples of these subnetworks are illustrated in the supplementary material section S5.

One possible shortcoming of ESSNet, and also of PFSNet[7] and SNet,[6] is that these methods are designed for relatively homogeneous phenotypes. If the phenotype D is actually composed of multiple subtypes D_1, $D_2, \ldots$, one should apply these methods to analyze each D_h versus $\neg D$ separately, for $h = 1, 2, \ldots$. There are two reasons for this shortcoming. The first is that, if a subnetwork only behaves differently between a D_h and $\neg D$, and not between other $D_{k \neq h}$ and $\neg D$, the genes in this subnetwork may only be highly expressed in D_h but not in other $D_{k \neq h}$. These genes may have lower average rank computed over the entire D, when the subtypes are not analyzed separately, and thus the subnetwork may not be generated. The second is that, even when the subnetwork is generated, as this subnetwork is not differentially expressed between $D_{k \neq h}$ and $\neg D$, many of the $\delta(g_i, p_j, p'_l)$ where $p_j \in D_{k \neq h}$ and $p'_l \in \neg D$, may be close to zero, thus diluting the test statistic. We plan to address this shortcoming in our future work.

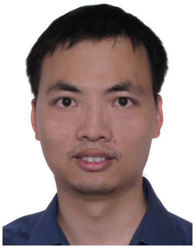
References

1. Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher JP, Calculating sample size estimates for RNA sequencing data, *J Comput Biol* **20**(12):970–978, 2013.
2. Jung SH, Bang H, Young S, Sample size calculation for multiple testing in microarray data analysis, *Biostatistics* **6**(1):157–169, 2005.
3. Lin WJ, Hsueh HM, Chen J, Power and sample size estimation in microarray studies, *BMC Bioinformatics* **11**(1):48, 2010.
4. Tibshirani R, A simple method for assessing sample sizes in microarray experiments, *BMC Bioinformatics* **7**(1):106, 2006.
5. Zhang M, Zhang L, Zou J, Yao C, Xiao H, Liu Q, Wang J, Wang D, Wang C, Guo Z, Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes, *Bioinformatics* **25**(13):1662–1668, 2009.
6. Soh D, Dong D, Guo Y, Wong L, Finding consistent disease subnetworks across microarray datasets, *BMC Bioinformatics* **12**(Suppl 13):S15, 2011.
7. Lim K, Wong L, Finding consistent disease subnetworks using PFSNet, *Bioinformatics* **30**(2):189–196, 2014.
8. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, *Proc Nat Acad Sci USA* **102**(43):15545–15550, 2005.
9. Khatri P, Drăghici S, Ontological analysis of gene expression data: Current tools, limitations, and open problems, *Bioinformatics* **21**(18):3587–3595, 2005.
10. Tusher VG, Tibshirani R, Chu G, Significance analysis of microarrays applied to the ionizing radiation response, *Proc Nati Acad Sci USA* **98**(9):5116–5121, 2001.
11. Pescatori M, Broccolini A, Minetti C, Bertini E, Bruno C, Damico A, Bernardini C, Mirabella M, Silvestri G, Giglio V, Modoni A, Pedemonte M, Tasca G, Galluzzi G, Mercuri E, Tonali PA, Ricci E, Gene expression profiling in the early phases of DMD: A constant molecular signature characterizes DMD muscle from early postnatal life throughout disease progression, *FASEB J* **21**(4):1210–1226, 2007.
12. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC, A global test for groups of genes: Testing association with a clinical outcome, *Bioinformatics* **20**(1):93–99, 2004.
13. Sivachenko AY, Yuryev A, Daraselia N, Mazo I, Molecular networks in microarray analysis, *J Bioinformatics Comput Biol* **5**(2b):429–456, 2007.
14. Haynes WA, Higdon R, Stanberry L, Collins D, Kolker E, Differential expression analysis for pathways, *PLoS Comput Biol* **9**(3):e1002967, 2013.
15. Zampieri M, Legname G, Segr D, Altafini C, A system-level approach for deciphering the transcriptional response to prion infection, *Bioinformatics* **27**(24):3407–3414, 2011.
16. Geistlinger L, Csaba G, Küffner R, Mulder N, Zimmer R, From sets to graphs: Towards a realistic enrichment analysis of transcriptomic systems, *Bioinformatics* **27**(13):i366–i373, 2011.
17. Soh D, Dong D, Guo Y, Wong L, Consistency, comprehensiveness, and compatibility of pathway databases, *BMC Bioinformatics* **11**(1):449, 2010.
18. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M, KEGG for integration and interpretation of large-scale molecular data sets, *Nucl Acids Res* **40**(D1):D109–D114, 2012.
19. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, Pico AR, Wikipathways: Building research communities on biological pathways, *Nucl Acids Res* **40** (D1):D1301–D1307, 2012.
20. Dørum G, Snipen L, Solheim M, Saebø S, Rotation testing in gene set enrichment analysis for small direct comparison experiments, *Stat Appl Genet Mol Biol* **8**:Article34, 2009.

21. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science* **286** (5439):531–537, 1999.
22. Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ, MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia, *Nat Genet* **30**(1):41–47, 2002.
23. Ross ME, Mahfouz R, Onciu M, Liu HC, Zhou X, Song G, Shurtleff SA, Pounds S, Cheng C, Ma J, Ribeiro RC, Rubnitz JE, Girtman K, Williams WK, Raimondi SC, Liang DC, Shih LY, Pui CH, Downing JR, Gene expression profiling of pediatric acute myelogenous leukemia, *Blood* **104**(12):3679–3687, 2004.
24. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimodi SC, Relling MV, Patel A, Cheng C, Campana D, Wilkins D, Zhou X, Li J, Liu H, Pui CH, Evans WE, Naeve C, Wong L, Downing JR, Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, *Cancel Cell* **1**:133–143, 2002.
25. Haslett JN, Sanoudou D, Kho AT, Bennett RR, Greenberg SA, Kohane IS, Beggs AH, Kunkel LM, Gene expression comparison of biopsies from Duchenne muscular dystrophy (DMD) and normal skeletal muscle, *Proc Nat Acad Sci USA* **99**(23):15000–15005, 2002.
26. Goldstein JA, McNally EM, Mechanisms of muscle weakness in muscular dystrophy, *J Gen Physiol* **136**(1):29–34, 2010.
27. Krans JL, The sliding filament theory of muscle contraction, *Nat Edu* **3**(9):66, 2010.
28. Dogra C, Changotra H, Wergedal JE, Kumar A, Regulation of phosphatidylinositol 3-kinase (PI3K)/Akt and nuclear factor-kappa B signaling pathways in dystrophin-deficient skeletal muscle in response to mechanical stretch, *J Cell Physiol* **208**(3):575–585, 2006.
29. Feron M, Guevel L, Rouger K, Dubreil L, Arnaud MC, Ledevin M, Megeney LA, Cherel Y, Sakanyan V, PTEN contributes to profound PI3K/Akt signaling pathway deregulation in dystrophin-deficient dog muscle, *Am J Pathol* **174**(4):1459–1470, 2009.
30. Vidal B, Ardite E, Suelves M, Ruiz-Bonilla V, Janu A, Flick MJ, Degen JL, Serrano AL, Muoz-Cnoves P, Amelioration of duchenne muscular dystrophy in mdx mice by elimination of matrix-associated fibrin-driven inflammation coupled to the M2 leukocyte integrin receptor, *Hum Mole Genet* **21**(9):1989–2004, 2012.
31. Das Gupta S, Halder B, Gomes A, Gomes A, Bengalin initiates autophagic cell death through ERK-MAPK pathway following suppression of apoptosis in human leukemic U937 cells, *Life Sci* **93**(7):271–276, 2013.
32. Dimicoli S, Wei Y, Bueso-Ramos C, Yang H, Dinardo C, Jia Y, Zheng H, Fang Z, Nguyen M, Pierce S, Chen R, Wang H, Wu C, Garcia-Manero G, Overexpression of the toll-like receptor (TLR) signaling adaptor MYD88, but lack of genetic mutation, in myelodysplastic syndromes, *PLoS ONE* **8**(8):e71120, 2013.
33. Furqan M, Mukhi N, Lee B, Liu D, Dysregulation of jak-stat pathway in hematological malignancies and jak inhibitors for clinical application, *Biomarker Res* **1**(1):5, 2013.
34. Hruak O, Porwit-MacDonald A, Antigen expression patterns reflecting genotype of acute leukemias, *Leukemia* **16**(7):1233–1258, 2002.
35. Kanagal-Shamanna R, Zhao W, Vadhan-Raj S, Nguyen MH, Fernandez MH, Medeiros LJ, Bueso-Ramos CE, Over-expression of CYP2E1 mRNA and protein: Implications of xenobiotic induced damage in patients with de novo acute myeloid leukemia with inv(16) (p13.1q22), *Int J Environ Res Public Health* **9**(8):2788–2800, 2012.

36. Giunta M, Pucillo C, BCR-ABL rearrangement and HLA antigens: A possible link to leukemia pathogenesis and immunotherapy, *Rev Bras Hematol Hemoter* **34**(5):323–324, 2012.
37. Kim DH, Kong JH, Byeun JY, Jung CW, Xu W, Liu X, Kamel-Reid S, Kim YK, Kim HJ, Lipton JH, The IFNG (IFN-gamma) genotype predicts cytogenetic and molecular response to imatinib therapy in chronic myeloid leukemia, *Clin Cancer Res* **16**(21):5339–5350, 2010.
38. Ress A, Moelling K, Bcr is a negative regulator of the Wnt signalling pathway, *EMBO Rep* **6**(11):1095–1100, 2005.
39. Cardoso BA, Martins LR, Santos CI, Nadler LM, Boussiotis VA, Cardoso AA, Barta JT, Interleukin-4 stimulates proliferation and growth of T-cell acute lymphoblastic leukemia cells by activating mTOR signaling, *Leukemia* **23**(1):206–208, 2008.
40. Mumprecht S, Claus C, Schurch C, Pavelic V, Matter MS, Ochsenbein AF, Defective homing and impaired induction of cytotoxic T cells by BCR/ABL-expressing dendritic cells, *Blood* **113**(19):4681–4689, 2009.

Kevin Lim is a Research Fellow at Duke-NUS. He received his Ph.D. and B.Comp. degree from the National University of Singapore in 2015 and 2009, respectively. His Ph.D. thesis involved incorporating gene expresison analysis with pathway databases to derive meaningful disease subnetworks. His current interests are in the analysis of whole genome sequencing in cancer genomics.



Zhenhua Li is a Research Fellow at the Department of Paediatrics, National University of Singapore. He received his Ph.D. in Bioinformatics and Computational Biology from Nanyang Technological University in 2013. Before that, he studied computer science in Wuhan University where he was awarded B.Eng. and M.Eng. degrees in 2007 and 2009, respectively. His current research interests include medical data analysis, bioinformatics, and data mining.

Kwok Pui Choi received the B.Sc. (first class) degree from the University of Hong Kong and the M.Sc. and Ph.D. degrees from the University of Illinois at Urbana-Champaign. He is now an Associate Professor in the Department of Statistics and Applied Probability at the National University of Singapore (NUS). He has a joint appointment with the Department of Mathematics at NUS. His research interests include probability and computational biology.

Limsoon Wong is a Professor in the School of Computing and the Yong Loo Lin School of Medicine at the National University of Singapore. He currently works mostly on knowledge discovery technologies and their application to biomedicine. He is a Fellow of the ACM, named in 2013 for his contributions to database theory and computational biology. He serves in the editorial boards of several journals, including the *Journal of Bioinformatics and Computational Biology*.