BMC Medical Genomics

RESEARCH Open Access

Fuzzy-FishNET: a highly reproducible protein complex-based approach for feature selection in comparative proteomics

Wilson Wen Bin Goh

From 15th International Conference On Bioinformatics (INCOB 2016)
Queenstown, Singapore. 21-23 September 2016

Abstract

Background: The hypergeometric enrichment analysis approach typically fares poorly in feature-selection stability due to its upstream reliance on the *t*-test to generate differential protein lists before testing for enrichment on a protein complex, subnetwork or gene group.

Methods: Swapping the *t*-test in favour of a fuzzy rank-based weight system similar to that used in network-based methods like Quantitative Proteomics Signature Profiling (QPSP), Fuzzy SubNets (FSNET) and paired FSNET (PFSNET) produces dramatic improvements.

Results: This approach, Fuzzy-FishNET, exhibits high precision-recall over three sets of simulated data (with simulated protein complexes) while excelling in feature-selection reproducibility on real data (based on evaluation with real protein complexes). Overlap comparisons with PFSNET shows Fuzzy-FishNET selects the most significant complexes, which are also strongly class-discriminative. Cross-validation further demonstrates Fuzzy-FishNET selects class-relevant protein complexes.

Conclusions: Based on evaluation with simulated and real datasets, Fuzzy-FishNET is a significant upgrade of the traditional hypergeometric enrichment approach and a powerful new entrant amongst comparative proteomics analysis methods.

Keywords: Proteomics, Networks, Bioinformatics, GSEA, QPSP, SNET, FSNET, PFSNET, Renal Cancer

Background

Mass spectrometry (MS)-based proteomics is becoming increasingly important in contemporary biological and clinical research [1]. Yet, despite significant technological advancement marking quantum leaps in protein extraction and spectra-acquisition [2–4], data reliability issues in MS-based proteomics persist: the primary issues being incomplete proteome coverage and inter-sample protein identification inconsistency [5]. These problems are not yet resolved satisfactorily on current proteomics paradigms [6–8]. Moreover, with the advent of brute-force spectra capture strategies e.g. Data-Independent Acquisition (DIA) [9, 10], increased noise becomes an inadvertent consequence, and contribute yet another layer of complexity [2].

Proteomics allows the simultaneous expressional profiling of thousands of proteins (although leaving thousands more which remain undetected). The first order of business is usually to identify proteins which are strongly and consistently differential, with the expectation that these are phenotypically relevant. This process is known as “feature selection”, and helps to concentrate analysis on a smaller feature (protein) set which is easier

Correspondence: goh.informatics@gmail.com
School of Pharmaceutical Science and Technology, Tianjin University, Tianjin, People’s Republic of China

to study, understand and validate experimentally [11]. Unlike animal models or cell lines, clinical samples are highly heterogeneous, reflecting different disease etiologies and genetic backgrounds amongst unique individuals [12]. Heterogeneity, compounded with the fact that different proteins being identified between samples [13], and possible quantification accuracy issues [14] means that in practical deployment, it is difficult to make reliable identification of useful biomarkers or drug targets during analysis of clinical data. Hence, more sophisticated and robust analytical methods are required.

Contextualization at the level of subnets, or more specifically, protein complexes, can resolve proteomic coverage and consistency issues [15–18]. Use of protein complexes as features for feature-selection instead of predicted clusters from reference networks, is a more powerful approach as protein complexes are enriched for biological signal [19]. However, use of protein complexes alone (despite its high biological signal enrichment) is insufficient: the nature of the statistical analysis method is also equally important. The hypergeometric enrichment (HE) test is commonly used in many areas of biological research from testing for functional enrichment [20–25] to testing for over-representation of genes in predicted subnetworks [26]. Yet, despite its wide use, even when used with protein complexes, HE does poorly, particularly in terms of feature-selection stability [16].

HE is actually a two-part test (see Methods). But its reliance on the *t*-test to generate a differential protein list for subsequent enrichment analysis based on the hypergeometric test is a known contributing factor towards its high instability [27, 28], and is demonstrated again in recent work [16]. We may redesign HE using elements of design that have worked well in other techniques.

QPSP, and the rank-based network approaches (RBNAs), SNET (SubNET) [29], FSNET (Fuzzy SNET) and PFSNET (Paired FSNET) [30] have been shown to be highly stable and robust, these techniques are similar in that they use a fuzzy weighting system on proteins ranked by expression [31] (see Methods).

By incorporating the fuzzy weighting system into HE, and doing away with upstream *t*-test differential protein preselection, a new spin on the original HE technique, Fuzzy-FishNET, is introduced here. Its name comes from the incorporation of the fuzzy weighting system in QPSP/FSNET/PFSNET with the one-sided Fisher's exact test (equivalent to the hypergeometric test). Fuzzy-FishNET is evaluated based on precision and recall on three simulated datasets, and also its stability/reproducibility on real data.

Methods

Simulated proteomics datasets — D.1.2, D2.2 and RC1

Two simulated proteomics datasets, D1.2 and D2.2, from the study of Langley and Mayr are used [32]. D.1.2 is obtained from a study of proteomic changes resulting from addition of exogenous matrix metallopeptidase (3 control, 3 test) while D2.2 is obtained from a study of hibernating arctic squirrels (4 control, 4 test). Protein quantification in both studies is based on spectral counts.

For both D1.2 and D2.2, 100 simulated datasets each with 20% randomly generated differential features are generated. The 20% threshold is arbitrary, for D1.2 and D2.2, this corresponds to 177 and 710 differential proteins respectively. For a given feature measured amongst samples derived from two different sample classes A and B, the effect size is the magnitude of the inter-class difference e.g. the differences of the means amongst samples derived from classes A and B. Here, the effect sizes of these 20% differential features are randomly selected from one out of five possibilities or p (20%, 50%, 80%, 100% and 200%), increased in one class and not in the other, and expressed as:

$$SC_{i,j}{}' = SC_{i,j} \ * \ (1 + \mathrm{p})$$

where $\mathrm{SC}_{i,j}$ and $\mathrm{SC}_{i,j}$' are respectively the original and simulated spectral count from the j^{th} sample of protein i.

RC1 comes from the 12 controls from the renal cancer (RC) dataset (see below). As with D1.2 and D2.2, 20% random proteins are randomly selected as differential, an effect size sampled from one of 5 possibilities, and inserted in only half of the controls, thus creating 6 control and 6 artificial test samples. This is also repeated 100 times to generate 100 simulated datasets.

Proteomics dataset — renal cancer (RC)

The renal cancer (RC) study of Guo et al. [2] is derived from six pairs of non-tumorous and tumorous clear-cell renal carcinoma (ccRCC) tissues based on the SWATH spectra-acquisition method. The six sample pairs are examined twice, as two different technical batches.

All SWATH spectra maps are analyzed using OpenSWATH [9] against a spectral library containing 49,959 reference spectra for 41,542 proteotypic peptides from 4,624 reviewed SwissProt proteins [2]. The library is compiled via library search of spectra captured in DDA mode (linking spectra mz and rt coordinates to a library peptide). Proteins are quantified via spectral count.

Protein complexes (subnets)

Although subnets or clusters are predictable from large biological networks, real biological complexes are enriched for biological signal, far outperforming predicted complexes/subnets from reference networks [19, 31, 33, 34]. Here, known human protein complexes derived from the CORUM database are used [35].

To avoid high fluctuation in the test statistics used by some of the methods considered here (e.g. QPSP), complexes with at least 3 proteins that were identified and measured in the proteomics screen are retained (1363 complexes)

Hypergeometric-enrichment (HE)

HE is a frequently used form of protein complex/subnetwork evaluation and consists of two steps [5]: First, differential proteins are identified using the two-sample *t*-test [36]. This is followed by a hypergeometric test where given a total of N proteins (with B of these belonging to a complex) and n test-set proteins (i.e., differential), the exact probability P that b or more proteins from the test set are associated by chance with the complex is given by [37]:

$$P(X \geq b) = \sum_{i=b}^{\min(n,B)} \frac{\binom{n}{i}\binom{N-n}{B-i}}{\binom{N}{B}}$$

The sum $P(X \geq b)$ is the p-value of the hypergeometric test.

Gene-set enrichment analysis (GSEA)

The direct-group (DG) analysis approach, Gene-Set Enrichment Analysis, or GSEA is developed as a more powerful alternative to HE, as it obviates the *t*-test-based protein pre-selection step. In GSEA, a complex is tested by comparing the distribution of constituent protein expression between phenotype classes against that of proteins outside the complex using a Kolmogorov-Smirnov (KS) statistic [38].

Denoting proteins in the complex as the set D and proteins outside the complex as the set D', the KS-statistic $KS_{D,D'}$is expressed as:

$$KS_{D,D*} = max_x \left| F_{1,D}(x) - F_{2,D'}(x) \right|$$

where $F_{1,D}(x)$ and $F_{2,D*}(x)$ are respectively the number of proteins in D and D' that whose rank is higher than the rank x. The null hypothesis is rejected at an alpha of 0.05 if

$$KS_{D,D'} \geq c(alpha) * \sqrt{\frac{|D| + |D'|}{|D| * |D'|}}$$

where *c(alpha)* is the critical value at a given alpha level. Here, at an alpha of 0.05, *c(alpha)* = 1.36.

Quantitative proteomics signature profiling (QPSP)

In QPSP, each sample is sorted based on abundance. The most abundant proteins above a certain percentile (the value is denoted as alpha1) are first selected. A second percentile value (defined as alpha2) is then used to extend the protein list [30]. To penalize lower-ranked proteins below alpha1 and above alpha2, proteins are assigned interpolated weights based on their ranks.

Alpha1 and alpha 2 are typically set as the top 10%, and top 20% ranks respectively. Rank-based weighting is achieved via discretization of the ranks from top 10-20% into four bins: 10–12.5% (weight 0.8), 12.5% to 15% (0.6), 15–17.5% (0.4), 17.5 to 20% (0.2). All other proteins beyond alpha2 (viz. remaining proteins) have a weight of 0 (and are thus ignored). Proteins above alpha1 are assigned a full weight of 1.

Given each sample, a vector of hit-rates is generated by considering the overlaps of the proteins (given their weights) against a vector of complexes. Given a sample in class A (S_A) and a vector of complexes of length n, for each complex C_i in the complex vector, the hit-rate is the intersection of proteins in S_A and C_i, modulated by the weight, over the total number of proteins in C_i.

Let the hit-rate for S_A in C_i be $H(S_{A,} C_i)$. Therefore, the vector of hit-rates for sample S_A is $H_{SA} = \langle H(S_{A,} C_1)\ldots, H(S_{A,} C_n)\rangle$. This vector of hit-rates signifies the sample's signature profile based on complexes.

QPSP works with simple two-sample *t*-test. For each complex C in the complex vector, two lists are compared against each other, $HA = \langle H(A_1,C),\ldots, H(A_m,C)\rangle$ and $HB = \langle H(B_1,C),\ldots, H(B_n,C)\rangle$, where A and B are phenotype classes of lengths m and n respectively. The t-statistic, t_score, between HA and HB is computed by:

$$t\_score = \frac{\overline{HA} - \overline{HB}}{S_{HA,HB}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

where

$$S_{HA,HB} = \sqrt{\frac{(m-1)S_{HA}{}^2 - (n-1)S_{HB}{}^2}{m + n - 2}}$$

If t_score for C_i is significant (i.e. its associated p-value falls below 0.05), then C_i is considered differential.

SNET/FSNET/PFSNET

In SNET, given a protein g_i and a tissue p_k, let $fs(g_i,p_k) = 1$, if the protein g_i is among the top alpha percent (default = 10%) most-abundant proteins in the tissue p_k; and = 0 otherwise.

Given a protein g_i and a class of tissues C_j, let

$$\beta(g_i, Cj) = \sum_{p_k \in C_j} \frac{fs(g_i, pk)}{|C_j|}$$

That is, $\beta(g_i, C_j)$ is the proportion of tissues in C_j that have g_i among their top alpha percent most-abundant proteins.

Let *score*(S,p_k,C_j) be the score of a protein complex S and a tissue p_k weighted based on the class C_j. It is defined as:

$$score(S, p_k, Cj) = \sum_{g_i \in S} fs(g_i, pk) * \beta(g_i, Cj)$$

The function $f_{SNET}(S, X, Y, C_j)$ for some complex S is a t-statistic defined as:

$$f_{SNET}(S, X, Y, C_j) = \frac{mean(S, X, C_j) - mean(S, Y, C_j)}{\sqrt{\frac{var(S,X,C_j)}{|X|} + \frac{var(S,Y,C_j)}{|Y|}}}$$

where *mean*(S,#,C_j) and *var* (S,#,C_j) are respectively the mean and variance of the list of scores {*score*(S,*pk*,C_j) | p_k is a tissue in #}.

The complex S is considered differential (weighted based on C_j) in X but not in Y if $f_{SNET}(S,X,Y,C_j)$ is at the largest 5% extreme of the Student t-distribution, with degrees of freedom determined by the Welch-Satterwaite equation.

Given two classes C_1 and C_2, the set of significant protein complexes returned by SNET is the union of {S | $f_{SNET}(S,C_1,C_2,C_1)$ is significant} and {S | $f_{SNET}(S,C_2,C_1,C_2)$ is significant}; the former being complexes that are significantly consistently highly abundant in C_1 but not C_2, the latter being complexes that are significantly consistently highly abundant in C_2 but not C_1.

FSNET is identical to SNET, except in one regard:

For FSNET, the definition of the function $fs(g_i,p_k)$ is replaced such that $fs(g_i,p_k)$ is assigned a value between 1 and 0 as follows: $fs(g_i,p_k)$ is assigned the value 1 if g_i is among the top alpha1% (default = 10%) of the most-abundant proteins in p_k. It is assigned the value 0 if g_i is not among the top alpha2% (default = 20%) most-abundant proteins in p_k. The range between alpha1% and alpha2% is divided into n equal-sized bins (default $n = 4$), and $fs(g_i,p_k)$ is assigned the value 0.8, 0.6, 0.4, or 0.2 depending on which bin g_i falls into in p_k. This tiered weighting system is termed fuzzification.

A test statistic f_{FSNET} is defined analogously to f_{SNET}. Given two classes C_1 and C_2, the set of significant complexes returned by FSNET is the union of {S | $f_{FSNET}(S,C_1,C_2,C_1)$ is significant} and {S | $f_{FSNET}(S,C_2,C_1,C_2)$ is significant}.

For PFSNet, the same $fs(g_i,p_k)$ function as in FSNet is used. But it defines a score *delta*(S,p_k,X,Y) for a complex S and tissue p_k wrt classes X and Y as the difference of the score of S and tissue p_k weighted based on X from the score of S and tissue p_k weighted based on Y. More precisely: *delta*(S,p_k,X,Y) = *score*(S,p_k,X) − *score*(S,p_k,Y).

If a complex S is irrelevant to the difference between classes X and Y, the value of *delta*(S,p_k,X,Y) is expected to be around 0. So PFSNet defines the following one-sample t-statistic:

$$f_{PFSNET}(S,\ X,\ Y,\ Z) = \frac{mean(S, X, Y, Z)}{se(S, X, Y, Z)}$$

where *mean*(S, X, Y, Z) and *se*(S, X, Y, Z) are respectively the mean and standard error of the list {*delta*(S,p_k,X,Y) | p_k is a tissue in Z}. The complex S is considered significantly consistently highly abundant in X but not in Y if $f_{PFSNet}(S,\ X,\ Y,\ X \cup Y)$ is at the largest 5% extreme of the Student t-distribution.

Given two classes C_1 and C_2, the set of significant complexes returned by PFSNet is the union of {S | $f_{PFSNet}(S,C_1,C_2,C_1 \cup C_2)$ is significant} and {S | $f_{PFSNet}(S,C_2,C_1,C_1 \cup C_2)$ is significant}; the former being complexes that are significantly consistently highly abundant in C_1 but not C_2, and vice versa.

Fuzzy-FishNET

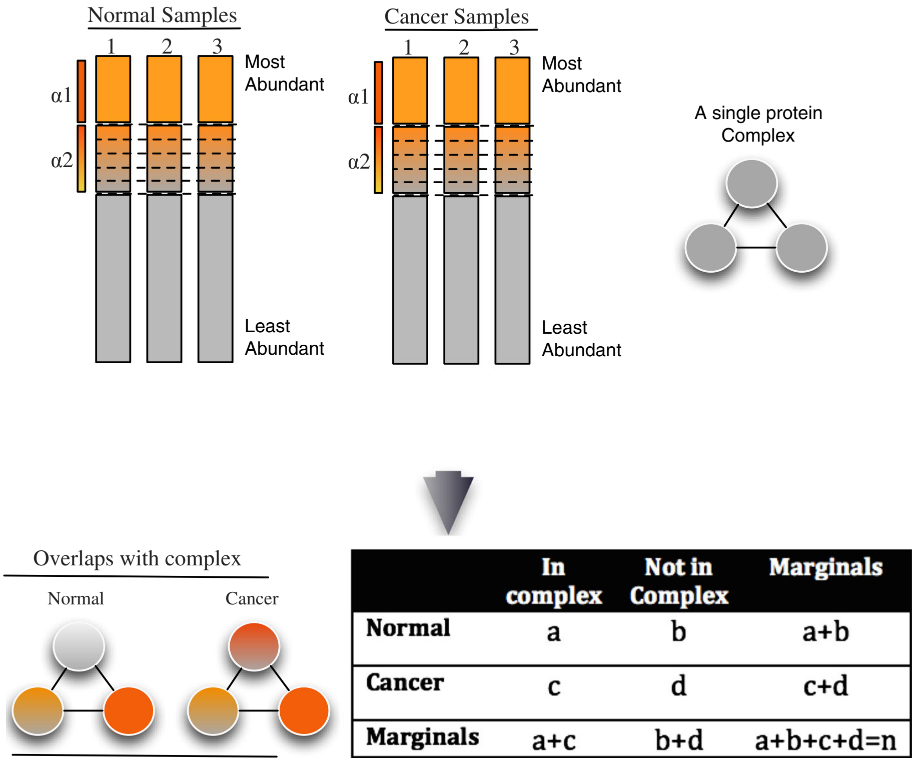
In Fuzzy-FishNET, a gene, gi, in sample pk, is assigned a weight ff(gi,pk) between 5 and 0 as follows: ff(gi,pk) is assigned the weight value 5 if gi is among the top alpha1% (default = 10%) of the most-abundant proteins in pk (Fig. 1). To boost sensitivity, a second alpha level, alpha2 is defined between the range top 10–20%. To account for the higher level of uncertainty for proteins in this region, weights are assigned based on ranks. To do this, proteins within alpha2 are divided into n equal-sized bins (default =4), and ff(gi,pk) and assigned a weight of 4, 3, 2, or 1 depending on which bin gi falls into in pk. Proteins that fall below the top 20% are assigned weights of 0.

For a complex S, and samples in class J, C_j, and samples in class k, C_k, the sum of weights can be expressed in a contingency table (Table 1) as shown below:
where a and c are the sum of weights for samples in class C_j and C_k mappable to proteins within complex S respectively, b and d are the sum of weights across samples in class C_j and C_k that are missed for proteins in complex S respectively. The Fisher exact probability p of obtaining this given set of values is then:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

$$= \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

The Fisher exact probability p is also the hypergeometric probability of observing this particular arrangement of the data, assuming the given marginal totals, on the null hypothesis that both C_j and C_k have similar distributions of top alpha proteins across their class members mappable to constituent proteins belonging to complex S [37].



	In complex	Not in Complex	Marginals
Normal	a	b	a+b
Cancer	c	d	c+d
Marginals	a+c	b+d	a+b+c+d=n

Fig. 1 Schematic of Fuzzy-FishNET A vector of complexes are compared against protein weights (based on expression ranks). For each complex, a contingency table may be constructed based on the sum of weights between normal and cancer samples, and a *p*-value calculated based on the Fisher's exact test

The p-value is calculated in a similar manner as in HE, as the sum of probabilities of obtaining an observation greater than or equal to a.

Performance benchmarks (simulated data)

In simulated data, differential proteins are are defined *a priori* and used to construct pseudo-complexes at various levels of purity (i.e., the proportion of significant proteins in the complex).

Proteins in the same complex are expected to be expressionally correlated. To incorporate this principle in pseudo-complex generation, a Euclidean distance is calculated for all differential protein pairs across all samples. These are then clustered via Ward's linkage. The differential proteins are reordered such that those with similar expression pattern are adjacent to each other. This reordered list is then split at regular intervals to generate 20, 101 and 62 differential pseudo-complexes for D1.2,D2.2 and RC1 respectively. An equal number of non-differential proteins are randomly selected, reordered based on expressional correlation, and then split to generate an equal number of non-differential pseudo-complexes.

Table 1 A typical 2 x 2 contingency table

	In complex S	Not in Complex S	Marginals
Samples in Class J (C_j)	a	b	a + b
Samples in Class K(C_k)	c	d	c + d
Marginals	a + c	b + d	a + b + c + d = n

The purity of the pseudo-complexes is lowered by decreasing the proportion of differential proteins [39]. This makes it harder for a differential pseudo-complex to be detected. So lowering purity tests for robustness and sensitivity. Here, purity is tested at three levels: 100%, 75% and 50%. At 100% purity, simulated complexes are comprised solely of differential proteins. These are randomly swoped out at lower purity levels; e.g. at 75% purity, 25% constituent differential proteins are randomly replaced with non-differential ones.

The differential and non-differential pseudo-complexes are combined into a single complex vector, which can be used for precision and recall-based evaluation of complex-based feature-selection methods:

$$Precision = \frac{TP}{TP + FP}; Recall = \frac{TP}{TP + FN}$$

where TP, FP and FN are the True Positives, False Positives and False Negatives respectively. Since precision and recall are both important performance measures, they can be combined to generate an average. A common way of doing this is the F-score (F_S) which is the harmonic mean between precision and recall:

$$F_s = 2 * \frac{Precision \ * \ Recall}{Precision + Recall}$$

Performance benchmarks (real data)

On real data, differential complexes are not known *a priori,* so direct precision-recall analysis is not possible. Instead, one may test for reproducibility/stability [31, 34, 39].

Reproducibility can be gauged based on the overlaps between technical replicates. To compare the technical replicates in RC, let T_1 and T_2 be the significant complexes selected by independently applying a given network method on the two replicates. Then reproducibility may be measured as overlaps based on the Jaccard coefficient (J):

$$J = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$$

This comparison may not be sufficiently robust and doesn't allow evaluation at small sample sizes. So, one may perform resampling at different levels (resampling sizes of 4, 6 and 8) and generate a binary matrix for each method, where a value of 1 indicates significance at a significance level of 0.05 and 0 otherwise per complex. The binary matrix may be analyzed in 2 ways: Row summation to evaluate the numbers of predicted differential complexes and column summation to evaluate the stability of predicted differential complexes [34].

Cross-validation (real data)

To demonstrate that Fuzzy-FishNET selects class-relevant differential complexes and only works when sample classes are real, cross-validation is performed 1000 times in two scenarios: where real classes exist(A) and where classes are shuffled/randomized (B) in RC. In each instance, half the data is used for feature-selection using Fuzzy-FishNET. A quarter of the remaining data is used for training, and the final quarter, validation. The classifier used is the deterministic Naïve Bayes method [40]. Cross-validation accuracy (CVAccuracy) is defined as:

$$CVAccuracy = \frac{Number \ of \ correct \ class \ assignments}{Total \ size \ of \ validation \ set}$$

A good feature-selection approach will select features that can build highly accurate prediction models when true class labels are present. But if the class labels are shuffled, then this is expected to lead to strong decrease in predictive performance.

Results and discussions

F-score comparisons (Simulated datasets)

The F-score distributions suggest that under noisier conditions (purity below 75%), Fuzzy-FishNET is an improvement over conventional HE and GSEA methods but overall appears to be a weaker method than earlier complex-based feature selection methods such as PFSNET (Fig. 2).

HE typically has the worst F-scores over all methods surveyed but in actuality, does very well in precision (Additional File 1) but falls short largely in recall (Additional File 2). Although GSEA is developed to address HE's reliance on the unstable differential protein pre-selection step (e.g. based on the *t*-test), it is only powerful when purity is at 100%. If purity drops to 75% and below, GSEA's F-score distributions quickly plummets. At purity of 50%, GSEA becomes the second worst method, beating only HE. Since noise and uncertainty in biological data is certainly expected, be it expressional or complexes, GSEA is unlikely a superior alternative to HE.

Fuzzy-FishNET's performance is comparable to existing complex-based methods which also rely on the fuzzification process e.g. QPSP, FSNET, PFSNET. It does however, gain power as sample size increases. In D1.2, where n =3 (per class), Fuzzy-FishNET falls behind SNET even (the earliest incarnation, and least powerful of the RBNAs). But as sample size increases to n = 6 (per class) in RC1, then Fuzzy-FishNET beats most methods, is comparable to FSNET but weaker than PFSNET.

It is fascinating that swapping the differential protein selection step in favour of a fuzzy weighting system based on expressional ranks can greatly improve the precision-recall performance in HE. This is consistently observed over three sets of simulated data. However, there is no gold-standard for generating pseudo-complexes, nor is it known if the lack of biological coherence in the pseudo-complexes unfairly penalizes certain complex-based feature-selection approaches. Therefore it is also essential to consider results based on real data and real complexes for a comprehensive evaluation.

Reproducibility of technical replicates (Real dataset)

Since technical replicates are present in RC. Each complex-based feature-selection method can be applied independently on each replicate. Inter-replicate overlaps is used as an indicator of complex-based feature-selection reproducibility.

Reproducibility is a strength of Fuzzy-FishNET (Table 2). Moreover, it does not make an overly large number of predictions, hence high-overlaps due to feature inflation or test hyper-sensitivity are not likely (misleading) contributors to its good performance.

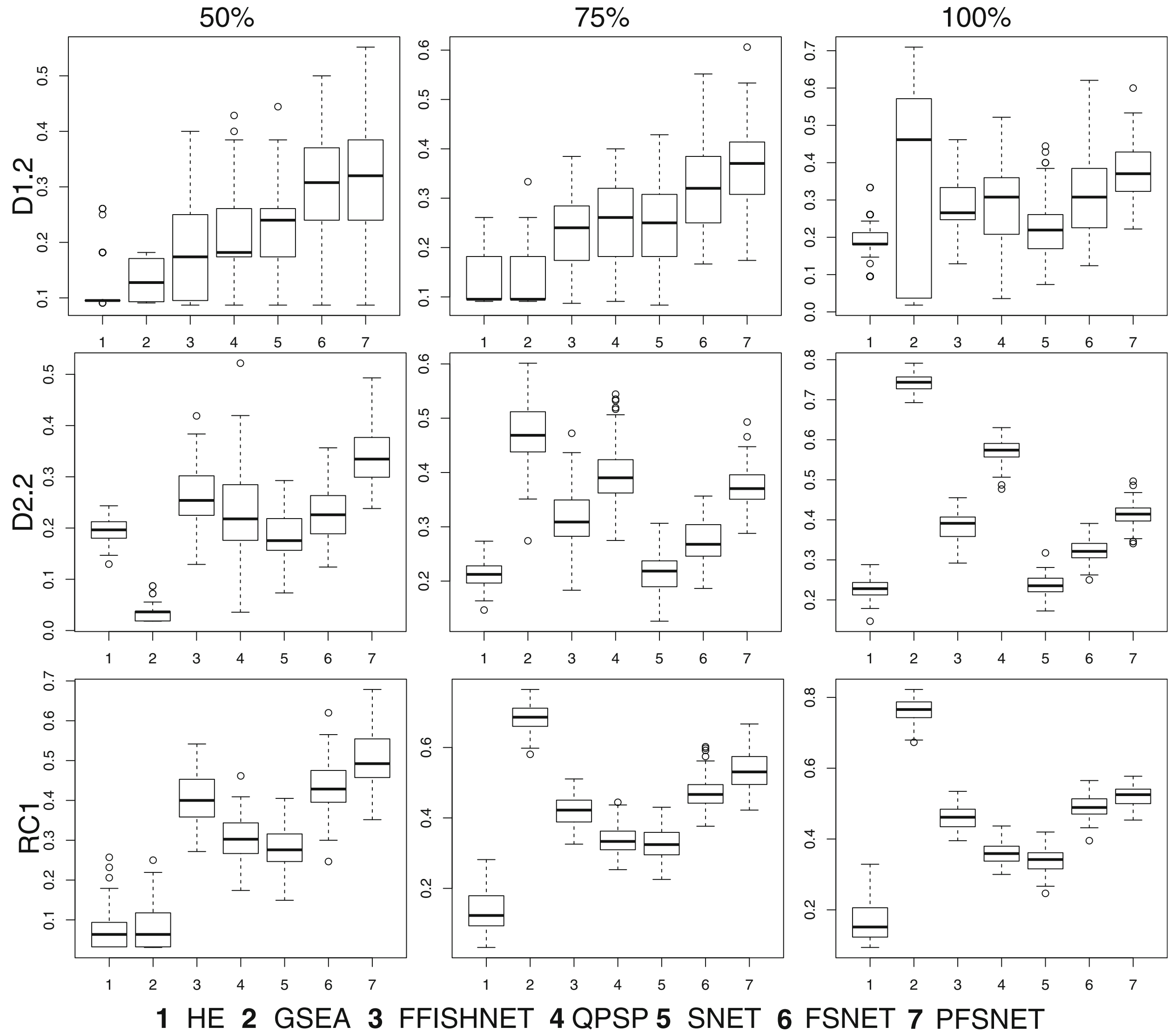


Fig. 2 F-score distributions. The F-score distributions for several network-based methods are shown for three simulated datasets (D1.2, D2.2 and RC1) over three levels of purity (50, 75 and 100%)

Feature-selection stability (Real dataset)

Inter-replicate overlap is a good and simple way of evaluating reproducibility given that technical noise should be the only source of variability (and not due to biological/clinical heterogeneity). But resampling at various levels to evaluate feature-selection stability is also possible. This is useful, as it also allows explicit evaluation of feature-selection stability in the small sample-size scenario (most feature-selection methods do not work well when sample sizes are very small [41]).

Given the full RC dataset, random resamplings of sizes 4, 6 and 8 (representing small to moderate size sample size scenarios) followed by feature-selection are performed 1000 times. Two aspects are considered: the number of selected features at each resampling level, and the stability over all selected features.

It is observed that HE, GSEA, Fuzzy-FishNET are particularly stable even as resampling size increases (Fig. 3 Top). While this is a good sign, it does not necessarily mean that the same features are selected each resampling round. Figure 3 (Bottom) shows that for HE and GSEA, feature-selection stability is particularly low. This

Table 2 Selected features for Replicate 1 and 2 are shown alongside their intersections

	HE	GSEA	FFISHNET	QPSP	SNET	FSNET	PFSNET
Replicate 1	4	1	27	86	34	38	45
Replicate 2	6	2	28	75	32	39	46
Overlap	0.25	0.50	0.96	0.66	0.83	0.88	0.93

is especially so for GSEA, possibly due to the presence of noise and uncertainty in real data. This supports the simulation observations.

On the other hand, Fuzzy-FishNET's feature-selection stability is second only to that of PFSNET's. However, PFSNET is more affected by sampling size increments, and it also selects considerably more features than Fuzzy-FishNET. There is a possibility that PFSNET may suffer from higher hyper-sensitivity and therefore feature-inflation issues than Fuzzy-FishNET.

Comparing Fuzzy-FishNET with PFSNET

Both PFSNET and Fuzzy-FishNET do very well on reproducibility. In the previous section, as PFSNET selects more features and appears to be more affected by sampling size increments, it is possible it is relatively hyper-sensitive, thus leading to feature-selection inflation.

To determine if this is likely, significant features selected by PFSNET and Fuzzy-FishNET are compared (Fig. 4a), revealing deep overlaps (and therefore high agreements) between both methods. Since there are many more PFSNET complexes than Fuzzy-FishNET's, the former's *p*-values distributions for intersecting and non-intersecting complexes are compared, revealing that Fuzzy-FishNET selects higher quality complexes (Fig. 4b).

Unlike PFSNET, Fuzzy-FishNET doesn't assign network scores for each complex per patient sample. However, class-discrimination analysis can still be performed using identified protein expressions found within significant complexes. Via hierarchical clustering (Ward's linkage; Euclidean Distance) coupled to bootstrap resampling [42], the constituent proteins (found within Fuzzy-FishNET selected complexes) are highly discriminative between sample classes (Fig. 4c; red boxes signify highly stable branches within the tree structure).

Many Fuzzy-FishNET differential complexes are associated with ribosomal complexes [43], although some are also associated with the cytoskeleton [44], proteasome [45] and TNF-alpha complexes [46] (Additional File 3). Although these are consistent with previous observations [16], in the absence of actual experimental validation, it is better to withhold judgement based solely on expected functionalities with renal cancer.

Fuzzy-FishNET selected complexes that are also class-relevant, i.e., it doesn't select features that are weakly associated with sample classes. The distribution of cross-validation (CV) accuracies when class labels are real, and when class labels are shuffled, reveals strong differences where the prediction model is far more accurate in the former than in the latter. If the feature-selection method selects a large number of irrelevant or weakly associated complexes (hyper-sensitive), then it is expected there will be little to no differences in CV accuracy between real and shuffled class labels.

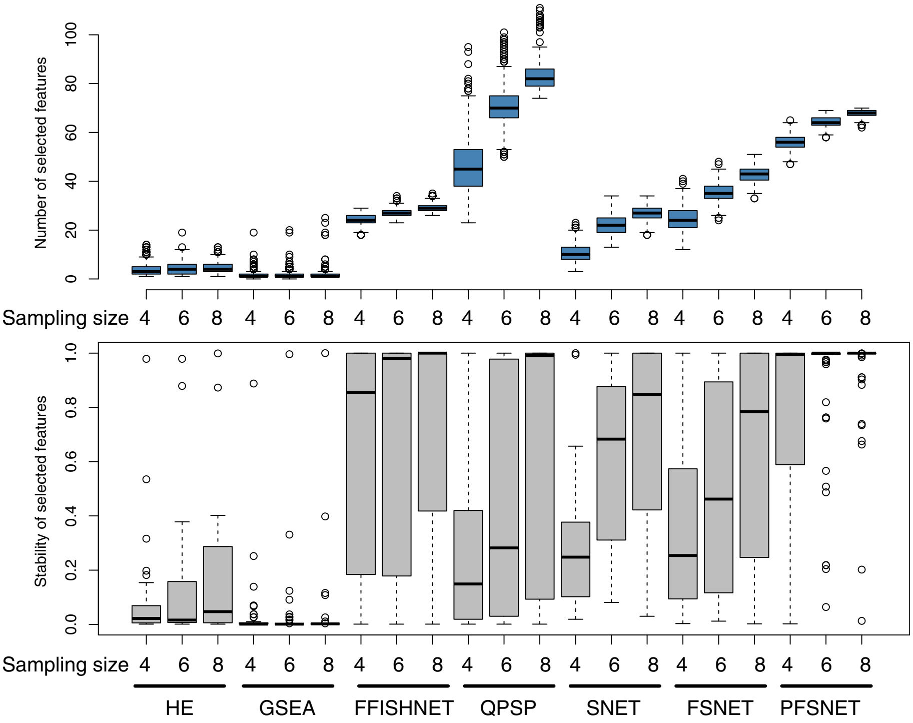


Fig. 3 Differential complex-selection size and complex-selection stability (*Top*) Number of selected differential complexes across resampling sizes of 4, 6 and 8. (*Bottom*) Corresponding feature-selection stability of differential complexes at each resampling size

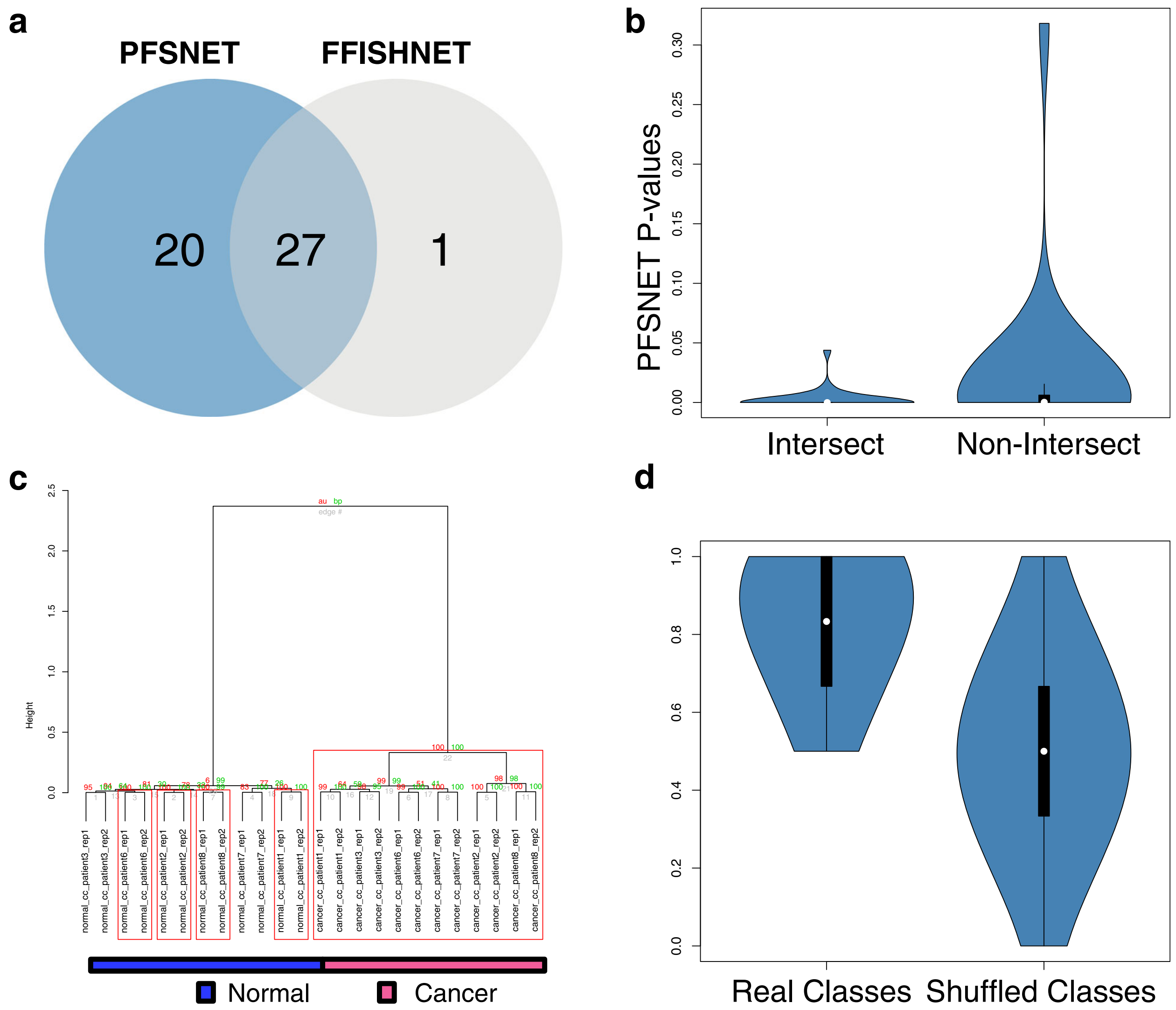


Fig. 4 Differential complex agreement between Fuzzy-FishNET and PFSNET **a** Selected complex features in PFSNET and Fuzzy-FishNET (FFISHNET) overlap strongly. **b** Overlapping components are more significant given the *p*-values (**c**) The constituent proteins (found within Fuzzy-FishNET selected complexes) are highly discriminative between sample classes; red boxes signify highly stable branches based on bootstrap resampling (**d**) FFISHNET selected features are properly class-discriminative, i.e., provides a high accuracy when actual class differences exist

Determining the contribution of Rank Weights (Fuzzification) towards signal stability

The positive impact of incorporating fuzzification in Fuzzy-FishNET is not known. It is possible the method may work just as well with a uniform weight of 1 across the top alpha%. So, an unweighted version of Fuzzy-FishNET, FishNET is tested. Note that Fuzzy-FishNET is analogous to SNET's uniform weight of 1 for the top alpha proteins, and 0 for all others.

Figure 5 shows the impact of fuzzification on Fuzzy-FishNET over three benchmarks: A/the frequency distribution of feature-selection stability, B/The pairwise feature-selection similarity based on the Jaccard distance and C/the frequency distribution of false positive rates based on random class assignment of RC's normal samples into two pseudo-classes followed by feature-selection. Benchmarks A and B are shown over resamplings of sizes 4, 6 and 8. HE is included as a point of reference since it is a primordial version of Fuzzy-FishNET (and FishNET). FishNET is not a strong improvement over HE (given its weaker feature-selection stability), and thus it is clear that fuzzification has a very strong positive impact on feature-selection stability, as well as robustness against false positives. The most informative rank shifts lies within the most highly ranked proteins, and assigning higher weights to these, improves signal-to-noise ratios.

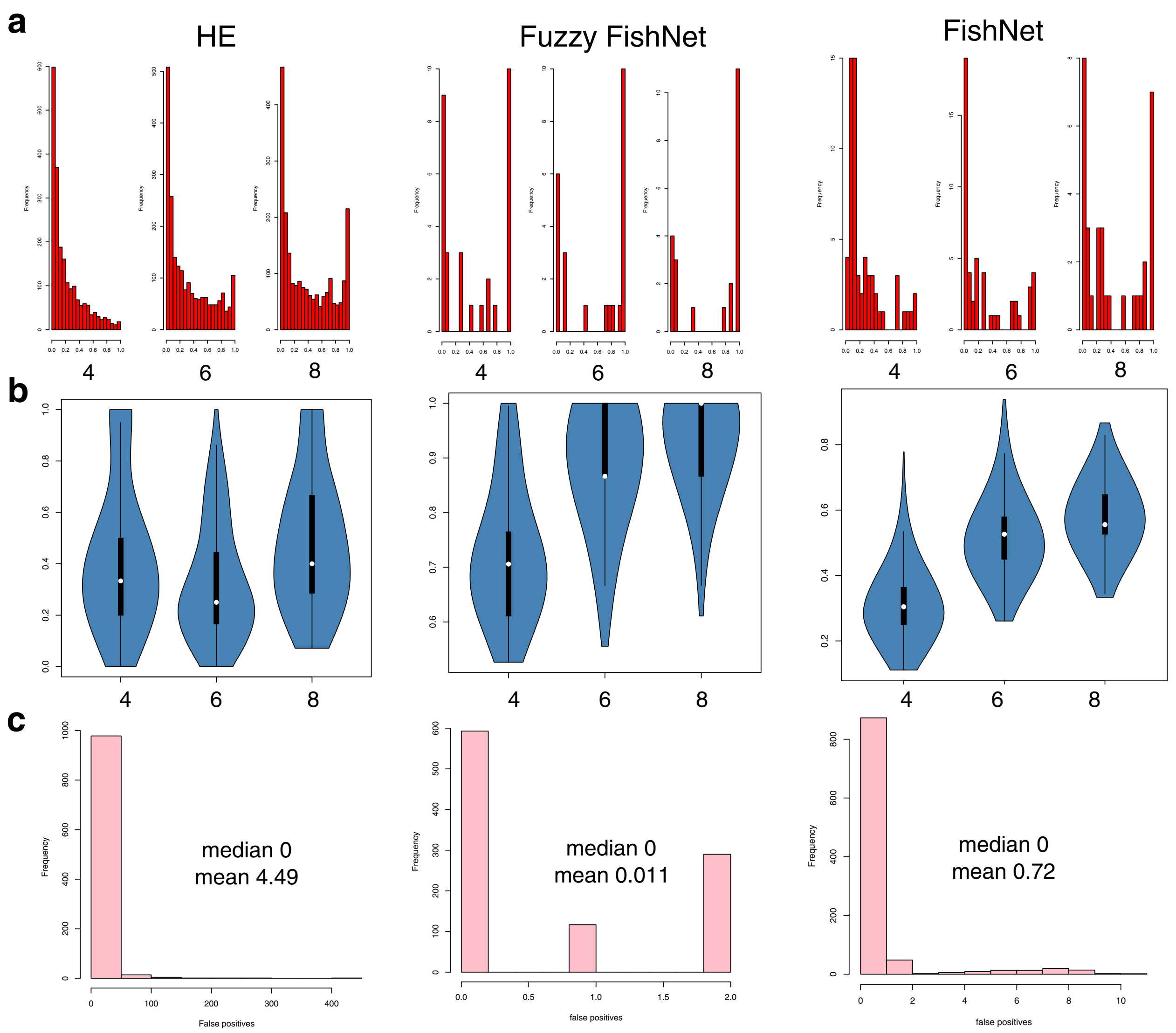


Fig. 5 Comparing Fuzzy-FishNET with its non-fuzzy variant, FishNET and HE (**a**) Feature-selection stability for Fuzzy-FishNET and its non-weighted variant. HE is added on the left side for comparisons. **b** Pairwise feature-selection stability (**c**) False positive evaluations

Robustness towards alpha adjustment

As with the RBNAs, a valid concern is that alpha adjustments may lead to highly different differential complexes being selected. Parameterization of alphas however is not fixed: Increasing alpha from top 10% onwards can increase sensitivity, but comes at the cost of introducing more false positives as signal from lower ranked proteins are introduced into the complex scores (which is why lower weights for alpha2 are assigned). However, we would expect the top scoring complexes to be stable.

To examine this concern, the top ranked complexes generated from top alphas of 10, 20 and 30% is compared against the default Fuzzy-FishNET setting of alpha1 (top 10%) and alpha2 (top 10–20%) based on overlaps (A∩B/min(A,B).

The results are generally stable, with overlaps of 62% at alpha = 10, 64% at alpha = 20, and 69% at alpha = 30 respectively. However, alpha should not be set too low at the onset, as this will likely introduce many poorer quality complexes into the significant complex-based feature set too early into preliminary analysis.

Conclusions

Fuzzy-FishNET is a powerful improvement over its predecessor, the hypergeometric enrichment (HE) approach. It differs only in the differential protein pre-selection step yet exhibits high precision-recall in simulated data while being the most reproducible over evaluations on real data. Based on cross-validations, it selects relevant features. Given these properties, Fuzzy-FishNET is a potentially powerful new entrant amongst complex-based feature-selection methods.

Additional files

Additional file 1: Precision distributions for several network-based methods for three simulated datasets (D1.2, D2.2 and RC1) at three levels of purity (50, 75 and 100%). (PDF 98 kb)

Additional file 2: Recall distributions for several network-based methods for three simulated datasets (D1.2, D2.2 and RC1) at three levels of purity (50, 75 and 100%). (PDF 97 kb)

Additional file 3: Top 25 Fuzzy-FishNET complexes. (PDF 141 kb)

Acknowledgements
WWBG would like to express his gratitude towards Limsoon Wong and Kevin Lim for inspiring the work undertaken in this manuscript.

Declarations
This article has been published as part of BMC Medical Genomics Volume 9 Supplement 3, 2016. 15th International Conference On Bioinformatics (INCOB 2016): medical genomics. The full contents of the supplement are available online https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-9-supplement-3.

Funding
Publication charges for this article have been funded by an education grant (290–0819000002) to WWBG from the School of Pharmaceutical Science and Technology, Tianjin University, China.

Availability of data and material
All data generated or analysed during this study are included in this published article under Additional files. An early version of this manuscript was published in BioXriv (doi: http://dx.doi.org/10.1101/024430).

Authors' contributions
WWBG designed, implemented the bioinformatics method and pipeline, performed analysis, and wrote the manuscript.

Competing interests
The author declares that he has no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Published: 5 December 2016

References

1. Ebhardt HA, Root A, Sander C, Aebersold R. Applications of targeted proteomics in systems biology and translational medicine. Proteomics. 2015;15(18):3193–208. doi:10.1002/pmic.201500004.
2. Guo T, Kouvonen P, Koh CC, Gillet LC, Wolski WE, Rost HL, et al. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. Nat Med. 2015;21(4):407–13. doi:10.1038/nm.3807.
3. Bruderer R, Bernhardt OM, Gandhi T, Miladinovic SM, Cheng LY, Messner S, et al. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. Mol Cell Proteomics. 2015;14(5):1400–10. doi:10.1074/mcp.M114.044305.
4. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, et al. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014;513(7518):382–7. doi:10.1038/nature13438.
5. Goh WW, Lee YH, Chung M, Wong L. How advancement in biological network analysis methods empowers proteomics. Proteomics. 2012;12(4–5):550–63. doi:10.1002/pmic.201100321.
6. Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcaino JA. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. Proteomics. 2015;15(5–6):930–49. doi:10.1002/pmic.201400302.
7. Keich U, Kertesz-Farkas A, Noble WS. Improved False Discovery Rate Estimation Procedure for Shotgun Proteomics. J Proteome Res. 2015;14(8):3148–61. doi:10.1021/acs.jproteome.5b00081.
8. Chick JM, Kolippakkam D, Nusinow DP, Zhai B, Rad R, Huttlin EL, et al. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat Biotechnol. 2015;33(7):743–9. doi:10.1038/nbt.3267.
9. Rost HL, Rosenberger G, Navarro P, Gillet L, Miladinovic SM, Schubert OT, et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat Biotechnol. 2014;32(3):219–23. doi:10.1038/nbt.2841.
10. Gillet LC, Navarro P, Tate S, Rost H, Selevsek N, Reiter L, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics. 2012;11(6):O111 016717. doi:10.1074/mcp.O111.016717.
11. Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. J Mach Learn Res. 2003;3:1157–82. doi:citeulike-article-id:167555.
12. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? Nat Rev. 2012;12(5):323–34. doi:10.1038/nrc3261.
13. Webb-Robertson B-JM, Wiberg HK, Matzke MM, Brown JN, Wang J, McDermott JE, et al. Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. J Proteome Res. 2015;14(5):1993–2001. doi:10.1021/pr501138h.
14. Sandberg A, Branca RM, Lehtio J, Forshed J. Quantitative accuracy in mass spectrometry based proteomics of complex samples: the impact of labeling and precursor interference. J Proteomics. 2014;96:133–44. doi:10.1016/j.jprot.2013.10.035.
15. Goh WW, Fan M, Low HS, Sergot M, Wong L. Enhancing the utility of Proteomics Signature Profiling (PSP) with Pathway Derived Subnets (PDSs), performance analysis and specialised ontologies. BMC Genomics. 2013;14:35. doi:10.1186/1471-2164-14-35.
16. Goh WW, Guo T, Aebersold R, Wong L. Quantitative proteomics signature profiling based on network contextualization. Biol Direct. 2015;10(1):71. doi:10.1186/s13062-015-0098-x.
17. Goh WW, Lee YH, Ramdzan ZM, Sergot MJ, Chung M, Wong L. Proteomics signature profiling (PSP): a novel contextualization approach for cancer proteomics. J Proteome Res. 2012;11(3):1571–81. doi:10.1021/pr200698c.
18. Goh WW, Lee YH, Zubaidah RM, Jin J, Dong D, Lin Q, et al. Network-Based Pipeline for Analyzing MS Data: An Application toward Liver Cancer. J Proteome Res. 2011. doi:10.1021/pr1010845.
19. Goh WW, Sergot MJ, Sng JC, Wong L. Comparative network-based recovery analysis and proteomic profiling of neurological changes in valproic Acid-treated mice. J Proteome Res. 2013;12(5):2116–27. doi:10.1021/pr301127f.
20. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25(8):1091–3. doi:10.1093/bioinformatics/btp101.
21. Zheng Q, Wang XJ. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. Nucleic Acids Res. 2008;36(Web Server issue):W358–63. doi:10.1093/nar/gkn276.
22. Huang WL, Tung CW, Ho SW, Hwang SF, Ho SY. ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. BMC Bioinformatics. 2008;9:80. doi:10.1186/1471-2105-9-80.
23. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics. 2005;21(16):3448–9. doi:10.1093/bioinformatics/bti551.
24. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, et al. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics. 2004;20(18):3710–5. doi:10.1093/bioinformatics/bth456.
25. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol. 2003;4(4):R28.
26. Sivachenko AY, Yuryev A, Daraselia N, Mazo I. Molecular networks in microarray analysis. J Bioinform Comput Biol. 2007;5(2B):429–56.

27. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods. 2015;12(3):179–85. doi:10.1038/nmeth.3288.
28. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput Biol. 2011;7(10):e1002240. doi:10.1371/journal.pcbi.1002240.
29. Soh D, Dong D, Guo Y, Wong L. Finding consistent disease subnetworks across microarray datasets. BMC Bioinformatics. 2011;12 Suppl 13:S15. doi:10.1186/1471-2105-12-S13-S15.
30. Lim K, Wong L. Finding consistent disease subnetworks using PFSNet. Bioinformatics. 2014;30(2):189–96. doi:10.1093/bioinformatics/btt625.
31. Goh WW, Wong L. Evaluating feature-selection stability in next-generation proteomics. J Bioinform Comput Biol. 2016;14(5):16500293. doi:10.1142/S0219720016500293.
32. Langley SR, Mayr M. Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics. J Proteomics. 2015;129:83–92. doi:10.1016/j.jprot.2015.07.012.
33. Goh WW, Wong L. Integrating Networks and Proteomics: Moving Forward. Trends Biotechnol. 2016. doi:10.1016/j.tibtech.2016.05.015.
34. Goh WW, Wong L. Design principles for clinical network-based proteomics. Drug Discov Today. 2016;21(7):1130–8. doi:10.1016/j.drudis.2016.05.013.
35. Ruepp A, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Stransky M, et al. CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res. 2008;36(Database issue):D646–50. doi:10.1093/nar/gkm936..
36. Raju TN. William Sealy Gosset and William A. Silverman: two "students" of science. Pediatrics. 2005;116(3):732–5. doi:10.1542/peds.2005-1134.
37. Fisher RA. The Logic of Inductive Inference. J R Stat Soc. 1935;98(1):39–82. doi:10.2307/2342435.
38. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50. doi:10.1073/pnas.0506580102.
39. Goh WWB, Wong L. Advancing clinical proteomics via analysis based on biological complexes: A tale of five paradigms. J Proteome Res. 2016. doi:10.1021/acs.jproteome.6b00402.
40. Rish I, editor. An empirical study of the naive Bayes classifier. IJCAI-01 workshop on "Empirical Methods in AI". 2011.
41. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14(5):365–76. doi:10.1038/nrn3475.
42. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006;22(12):1540–2. doi:10.1093/bioinformatics/btl117.
43. Hager M, Haufe H, Alinger B, Kolbitsch C. pS6 Expression in normal renal parenchyma, primary renal cell carcinomas and their metastases. Pathol Oncol Res. 2012;18(2):277–83. doi:10.1007/s12253-011-9439-y.
44. Beise N, Trimble W. Septins at a glance. J Cell Sci. 2011;124(Pt 24):4141–6. doi:10.1242/jcs.087007.
45. de Martino M, Hoetzenecker K, Ankersmit HJ, Roth GA, Haitel A, Waldert M, et al. Serum 20S proteasome is elevated in patients with renal cell carcinoma and associated with poor prognosis. Br J Cancer. 2012;106(5):904–8. doi:10.1038/bjc.2012.20.
46. Harrison ML, Obermueller E, Maisey NR, Hoare S, Edmonds K, Li NF, et al. Tumor necrosis factor alpha as a new target for renal cell carcinoma: two sequential phase II trials of infliximab at standard and high dose. J Clin Oncol. 2007;25(29):4542–9. doi:10.1200/JCO.2007.11.2136.