RoDiCE: Robust differential protein co-expression analysis for cancer complexome

Yusuke MATSUI[1,*], Yuichi ABE[2] and Kohei UNO[1], Satoru MIYANO[3]

[1] Biomedical and Health Informatics Unit, Department of Integrated Health Science, Nagoya University Graduate School of Medicine, [2] Division of Molecular Diagnostics, Aichi Cancer Center Research Institute. [3] Department of Integrated Data Science, M&D Data Science Center, Tokyo Medical and Dental University

* matsui@met.nagoya-u.ac.jp

Abstract

Motivation: The full picture of abnormalities in protein complexes in cancer remains largely unknown. Comparing the co-expression structure of each protein complex between tumor and normal groups could help us understand the cancer-specific dysfunction of proteins. However, the technical limitations of mass spectrometry-based proteomics and biological variations contaminating the protein expression with noise lead to non-negligible over- (or under-) estimating co-expression.

Results: We propose a robust algorithm for identifying protein complex aberrations in cancer based on differential protein co-expression testing. Our method based on a copula is sufficient for improving the identification accuracy with noisy data over a conventional linear correlation-based approach. As an application, we show that important protein complexes can be identified along with regulatory signaling pathways, and even drug targets can be identified using large-scale proteomics data from renal cancer. The proposed approach goes beyond traditional linear correlations to provide insights into higher order differential co-expression structures.

Availability and Implementation: https://github.com/ymatts/RoDiCE.

Contact: matsui@met.ngaoya-u.ac.jp

Supplementary information: Supplementary data are available online.

1 Introduction

Cancer is a complex system. Many molecular events, such as genomic mutations and epigenetic and transcriptomic dysregulations, were identified as cancer drivers (Hoadley *et al.*, 2018). However, our knowledge of how they characterize the downstream mechanisms with proteomic phenotypes remains scarce (Clark *et al.*, 2019; Liu *et al.*, 2016; Mertins *et al.*, 2016; Zhang *et al.*, 2016). Protein complexes are responsible for most cellar activities. Recent studies (Ori *et al.*, 2016; Romanov *et al.*, 2019; Ryan *et al.*, 2017) have demonstrated that protein subunits tend to show co-expression patterns in proteome profiles; furthermore, the subunits of a complex are simultaneously down-/up-regulated with the genomic mutations (Ryan *et al.*, 2017). However, we know little about the changes in the co-regulatory modes of protein complexes between the tumor and normal tissues.

We propose a novel algorithm for differential co-expression of protein abundances to identify the tumor-specific abnormality of protein complexes. Differential co-expression (DC) analysis is a standard technique of gene expression analysis to find differential modes of co-regulation between conditions, and numerous methods already exist (Bhuva *et al.*, 2019). Correlation is one of the most common measures of co-expression. For example, differential correlation analysis (DiffCorr) (Fukushima, 2013) and gene set co-expression analysis (GSCA) (Choi and Kendziorski, 2009) are two-sample tests of Pearson's correlation coefficients. However, studies report that protein expression levels have greater variability than gene expression levels because of the regulatory mechanism of post-translational modifications (Gunawardana *et al.*, 2015; Liu *et al.*, 2016). This variability can affect the estimation of co-expression as an outlier and can significantly impact DC results.

We developed a robust DC framework, called Robust Differential Co-Expression Analysis (RoDiCE), via two-sample randomization tests with empirical copula. The notable advantage of RoDiCE is noise robustness. Our main contributions are as follows: 1) we develop an efficient algorithm for robust copula-based statistical DC testing; 2) we overcome the computational hurdles of the copula-based permutation test by incorporating extreme value theory; 3) we demonstrate the effective application of copula to cancer complexome analysis; and 4) we develop a computationally efficient multi-thread implementing as R package.

1.1 Motivational example from the CPTAC / TCGA dataset

First, using an actual dataset, we explain why there is a need for robustness in protein co-expression analysis. We analyzed a cancer proteome dataset of clear renal cell carcinoma from CPTAC/TCGA with 110 tumor tissue samples. We measured co-expression using Pearson's correlation coefficient. We compared the correlation coefficients before and after removing the outliers. To identify outlier samples, we applied robust principal component analysis using the R package ROBPCA (Hubert *et al.*, 2005) with default parameters. Among 49,635,666 pairs of 9,964 proteins, the correlation coefficients of 7,541,853 (15.2%) pairs were deviated by more than 0.2 after removing outlier samples (**Figure 1**). This result implied that a non-negligible proportion of protein co-expression would be overestimated or underestimated. To compare the structures of co-expression correctly, it is necessary to compare them while minimizing the over-/under-estimation of co-expression.

2 Methods

Figure 2 describes the outline of RoDiCE. We decompose the expression level of subunits in the protein complex into a structure representing a co-expression and one representing the expression level of each subunit, using a function called an empirical copula (Nelsen, 2010); the empirical copula rank-converts the scale of the original data. Comparing the empirical copula functions with the conditions of statistical hypothesis testing, we derive the p-value as the difference in co-expression structures. We describe our method in detail in the following sections.

2.1 RoDiCE model

Suppose there are n samples, and $g(g = g_1, g_2)$ represents each condition. We compare two conditions and assume that g_1 andg_2 represent the normal group and the tumor group, respectively. Let $\mathbf{X}_g = (X_{1g}, X_{2g}, \dots, X_{Pg})$ be abundances of P subunits in group g. Given a protein complex, we represent the entire behaviors of subunits with a joint distribution $\mathbf{X}_g \sim H_g(x_1, x_2, \dots, x_P)$. The distribution function H_g has two pieces of information as follows: subunit expression levels and the structure of co-expression between subunits. The copula C_g is a function that can decompose those two pieces of information into a form that can be handled separately, as follows:

$$H_g(x_1, x_2, \dots, x_P) = C_g\left(F_{1g}(x_1), F_{2g}(x_2), \dots, F_{Pg}(x_P)\right) \tag{1}$$

The behavior of each subunit $F_{pg}(x_p)$is represented by a distribution function. The copula function itself is a multivariate distribution with uniform marginals. The copula function includes all dependency information among the subunits (Nelsen, 2010; Rémillard and Scaillet, 2009; Seo, 2020).

We use the empirical copula to non-parametrically estimate the copula C_g since it could be widely applicable to various situations. It can be represented using pseudo-copula samples defined via rank-transformed subunit abundance $u_{ip} = \frac{R(x_{ip})}{n} (i = 1,2, \dots, n)$;

$$\hat{C}_g(u_1, u_2, \dots, u_p) = \frac{1}{n}\sum_i I(U_{1g} \leq u_1, U_{2g} \leq u_2, \dots, U_{pg} \leq u_p) \tag{2}$$

where $R(\cdot)$ is a rank-transform function, and we represent transformed pseudo-sample variables as $R(\mathbf{X}_{g_1}) = \mathbf{U}_{g_1}$ and $R(\mathbf{X}_{g_2}) = \mathbf{U}_{g_2}$. The empirical copula is robust to noise because it represents co-expression structures based on rank-transformed subunit expression levels, which is the so called scale invariant property in the context of copula theory (Nelsen, 2010).

To perform DC analysis between group g and g', we consider the following statistical hypothesis:

$$\begin{gathered}\mathcal{H}_0: C_{g_1} = C_{g_2} \\ \mathcal{H}_1: C_{g_1} \neq C_{g_2}\end{gathered} \tag{3}$$

We derive the following Cramér-von Mises type test statistic to perform statistical hypothesis testing (Rémillard and Scaillet, 2009):

$$\boldsymbol{s}(g,g') = \left(\frac{1}{n_{g_1}+n_{g_2}}\right)^{-1} \left\{\frac{1}{n_1^2}\sum_{i=1}^{n_{g_1}}\sum_{j=1}^{n_{g_1}}\prod_{p=1}^{P} \max\left(1-u_{ip}^{(g_1)}, u_{jp}^{(g_1)}\right) - \frac{2}{n_{g_1}n_{g_2}}\sum_{i=1}^{n_{g_1}}\sum_{j=1}^{n_{g_2}}\prod_{p=1}^{P} \max\left(1-u_{ip}^{(g_1)}, u_{jp}^{(g_2)}\right) + \frac{1}{n_{g_2}^2}\sum_{i=1}^{n_{g_2}}\sum_{j=1}^{n_{g_2}}\prod_{p=1}^{P} \max\left(1-u_{ip}^{(g_2)}, u_{jp}^{(g_2)}\right)\right\} \tag{4}$$

where $u_{ip}^{(g)}(i = 1,2, \dots n_g)$ represents pseudo-observation in group g. Note that the computational cost is n^2, where $n^2 \leq n_{g_1}n_{g_2}$; $n = \min(n_{g_1}, n_{g_2})$. For testing the test statistic (4), we also derived the p-value using an algorithm based on Monte Carlo calculations (Rémillard and Scaillet, 2009); however, the computational complexity of the algorithm makes it difficult to apply it to proteome-wide co-expression differential analysis (see the results of the simulation experiments described below).

2.2 Derivation of statistical significance

Using a permutation test, we derive the p-value using the following steps:

(1) Randomizing concatenated variable from the two groups; $\mathbf{W} = (\mathbf{U}_{g_1}, \mathbf{U}_{g_2})$

(2) Constructing a new randomized variable $\mathbf{U}'_{g_1} = (W_{r(1)}, W_{r(2)}, \dots, W_{r(n_1)})$ and $\mathbf{U}'_{g_2} = (W_{r(n_1+1)}, W_{r(n_1+2)}, \dots, W_{r(n_1+n_2)})$ with randomized index $r(i)$.

(3) Replacing copula functions C_{g_1} and C_{g_2} in (3) with re-estimated empirical copula function C'_{g_1} and C'_{g_2} from the randomized samples $\mathbf{X}'_{g_1}$ and $\mathbf{X}'_{g_2}$.

(4) Deriving test statistics $\boldsymbol{s}'(g_1, g_2)$based on (4) with C'_{g_1} and C'_{g_2}.

(5) Steps 2 and 3 are indispensable for deriving the null distribution correctly. Deriving the null distribution by randomizing $\mathbf{W}' = (\mathbf{U}_{g_1}, \mathbf{U}_{g_2})$ alone will distort the distribution, and we will be unable to control for the type I error correctly (Seo, 2020).

2.3 Approximation of p-value

The empirical p-value is derived as follows:

$$p(M) = 1 - \frac{\sum_{i=1}^{M} \mathbf{I}(S_i \leq s_{g,g'})}{M} \tag{5}$$

where M is the number of randomization and S_i is the test statistic from the null distribution of the i-th $(i = 1,2, \dots, M)$ randomization trials. The accuracy of p-value in (5) is bounded by $p(M) \geq 1/M$. As mentioned, calculating the test statistic requires a computational cost of

$O(n^2)$; therefore, an efficient computational algorithm is needed to derive accurate p-values in data with a large number of samples. For instance, proteomic cohort projects such as CPTAC / TCGA have more than n =100 samples. To address this problem, we introduced an approximation algorithm for p-values based on extreme value theory (Knijnenburg *et al.*, 2009) and devised a way to calculate accurate p-values even with a small number of trials.

The test statistic that exceeds the range of the accuracy with randomization trials M is regarded as an "extreme value," and its tail of the distribution could be estimated via a generalized Pareto distribution (GPD), as follows:

$$p_{approx} = \frac{N'}{N}\left(1 - G(s(g, g') - t)\right) \tag{6}$$

where N' is the number of the randomized test statistic exceeding the threshold t that has to be estimated via a goodness-of-fit (GoF) test (Knijnenburg *et al.*, 2009) and G is the cumulative distribution function of the generalized Pareto distribution, $G(x) = 1 - \left(1 - \frac{kx}{a}\right)^{\frac{1}{k}}$ for $k \neq 0$ and $G(x) = 1 - e^{-\frac{x}{a}}$ for $k = 0$. To estimate the threshold t in (6), the GoF test determines whether the excess comes from the distribution $G(x)$ via bootstrap based maximum likelihood estimator (Villaseñor-Alva and González-Estrada, 2009). As we do not know a priori the number of samples sufficient to estimate the underlying GPD with threshold t, we must decide the initial number of samples to use. We begin with a large number of samples and increase this number until the GoF test is not rejected, according to (Knijnenburg *et al.*, 2009). As initial samples, we start with those above 80% of quantiles and decrease samples by 1% while the GoF test is rejected.

2.4 Identification of protein complex alteration

As protein complexes show co-expression among multiple subunits (Kerrigan *et al.*, 2011), we hypothesized that the difference in the co-expression structure of the tumor group compared to the normal group is a characteristic quantity of the protein complex abnormality. In previous studies of the cancer transcriptome, differential co-expression analysis has revealed abnormalities associated with protein complexes (Amar *et al.*, 2013; Srihari *et al.*, 2014). Therefore, we define a protein complex as an abnormal protein complex when it is co-expressed in at least one pair of subunits. Thus, we applied RoDiCE to all protein complexes for each subunit pair (p = 2) and identified protein complexes that showed a statistically significant difference in at least one subunit pair as abnormal.

2.5 Protein membership with protein complex

As we do not know which proteins belong to which protein complexes, we must predict the membership via some method. There are two main approaches. One is membership prediction focusing on the modular structure in PPI networks (Adamcsek *et al.*, 2006; Nepusz *et al.*, 2012) and the other is a knowledge-based method using a curated database. We adopt the latter approach, which is based on already validated protein complex membership information, using CORUM (ver. 3.0)(Giurgiu *et al.*, 2019) as a database (see the Supplementary Data for details).

2.6 R implementation with multi-thread parallelization

To further accelerate the computation of test statistic (4) in the randomization steps, we used RcppParallel (Allaire J, 2019). We utilize the portable and high-level parallel function “parallelFor,” which uses Intel TBB of the C++ library as a backend on systems that support it and TinyThread on other platforms.

2.7 Copula-based simulation model for protein co-expression

We provide the outline of a method for simulating co-expressed structures using a copula. We simulated protein expression levels that showed differential co-expression patterns with outliers in the tumor group and the normal group. We represented the co-expression structure by the covariance parameter in the following bivariate Gaussian copula:

$$C_g(u_1, u_2) = \Phi_g\left(\phi^{-1}(x_1), \phi^{-1}(x_2); \Sigma_{p=2}^{(g)}\right) \tag{7}$$

where Φ_g is the p dimensional Gaussian distribution parameterized by a $p \times p$ covariance matrix (or correlation matrix) in the group g, denoted as $\Sigma_p^{(g)} = \left\{r_{ij}^{(g)}\right\}$ and $\phi(x_i)$ is a univariate distribution. Using the model, we generate the dependency structure with two groups; one group has high correlations and the other has low ones, $r_{ij}^{(g_1)} \sim U(0.8, 0.9)$ and $r_{ij}^{(g_2)} \sim U(0.1, 0.2)$, respectively. We then generated co-expression structure using a Gaussian copula with $\phi(x) = N(0, 1)$. We obtained protein expressions via

$$H_g(x_1, x_2) = C_g\left(F_{1g}(x_1), F_{2g}(x_2)\right) \tag{8}$$

where we simply set as $F_{ig} \sim N(\mu, \sigma)$ for $i = 1,2$ and $g = g_1, g_2$ with $\mu \sim N(2,1)$ and $\sigma \sim gamma(2,1)$. Furthermore, we added outliers that could affect the co-expression structure. Using the model in (6) and (7), we set the outlier population in both group as $r'^{(g_1)}_{ij} \sim U(0, 0.05)$ and $F'_{ig} \sim N(2,4)$ for $i = 1,2$ and $g = g_1, g_2$ (Fig3).

3 Results

3.1 Benchmarking RoDiCE with simulation dataset

We now describe the features of RoDiCE using a simulation model. First, to confirm whether RoDiCE could correctly derive the p-value, we performed a test on two groups, with no differences in co-expression structure without outliers, and confirmed the null rejection rate. We performed 100 tests with the proposed method and calculated the null rejection rate at the 1%, 5%, and 10% levels of significance. The same simulation was repeated 10 times to calculate the standard deviations. The results show that the proposed method can control type I errors (Table 1).

Table 1. Type I error controls of the proposed method

Significance Level	Mean	SD
1%	0.03	0.02
5%	0.05	0.02
10%	0.09	0.02

We then simulated a case in which the co-expressed structure between the two groups was different and included outliers, and we examined the sensitivity of the method to identify a broken co-expressed structure in tumor tissue relative to normal tissue. To demonstrate the advantages of the proposed method, we examined the sensitivity of increasing the percentage of outliers in 2% increments from 0% to 20% and compared it further with DiffCorr and GSCA, a two-group co-expression test method based on Pearson's linear correlation (**Figure 4**). For outliers, the proposed method showed robust co-expression test results, with an accuracy of more than 85% up to a percentage of outliers of approximately 15%. Conversely, the sensitivity of the method based on linear correlation starts to decline from the level of 2% of outliers, and for data containing 15% of outliers, the sensitivity drops to around 30%.

To investigate the relationship between sample size and identification accuracy, we simulated the sensitivity of RoDiCE, as we increased the number of samples in increments of 10 from 30 to 100 samples. All other settings were the same as those in **Figure 5**, except that the percentage of outliers was set at 5%.

Finally, we also examined the computational speed, comparing it with the R package TwoCop, which implements the Monte Carlo-based method (ref) used for the two-group comparison of copulas (**Table 2**). The proposed method is 68 times faster than TwoCop and is sufficiently efficient as a copula-based two-group comparison test method. In contrast, the estimation of the copula function required more computational time than the linear correlation coefficient-based method because of the computational complexity of estimating the copula function.

Table 2. Computation time for 10 replicates

Method	#Replications	Execution time (s)	Relative time
DiffCorr	10	0.001	1
GSCA	10	0.152	152
RoDiCE	10	0.373	373
TwoCop	10	25.301	25301

3.2 Application to cancer complexome analysis

We demonstrate RoDiCE with actual data using the clear renal cell carcinoma (ccRCC) published by CPTAC/TCGA (Clark *et al.*, 2019). The data are available from the CPTAC data portal (https://cptac-data-portal.georgetown.edu) in the CPTAC Clear Cell Renal Cell Carcinoma (CCRCC) discovery study. The data labeled "CPTAC_CompRef_CCRCC_Proteome_CDAP_Protein_Report.r1" were used. In the following analysis, only protein expression data that overlap with protein groups in human protein complexes in CORUM and in CPTAC were used. Missing values were completed based on principal component analysis, and the missing values were completed by 10 principal components using the pca function in pca Methods.

For the complete data, RoDiCE was applied to the normal and cancer groups for each protein complex. FDR was calculated by correcting the p-value for each complex using the Benjamini–Hochberg method. We identified anomalous protein complexes in protein expression data from 110 tumor and 84 normal samples; out of 3,364 protein complexes in CORUM, 1,244 complexes contained at least one co-expression difference between subunits with FDR ≤ 5 % (**Supplementary Data**).

The proposed method has identified several protein complexes containing driver genes on regulatory signaling pathways in ccRCC (**Figure 6a**) (Li *et al.*, 2019). The identified pathways included known regulatory pathways important for cancer establishment and progression, starting with chromosome 3p loss, regulation of the cellular oxygen environment (VHL), chromatin remodeling, and disruption of DNA methylation mechanisms (PBRM1, BAP1). They also included abnormalities in regulatory signals involved in cancer progression (AKT1). Moreover, several identified complexes also included key proteins, for example, MET, HGF, and FGFR proteins, which could be inhibited by targeting them with drugs such as Cabozantinib and Lenvatinib directly. Because a previous study reported that sensitivity to knockdowns of several genes was well associated with expression levels of protein complexes (Nusinow *et al.*, 2020), co-expression information on protein complexes containing druggable genes might be useful to optimize drug selection.

A close examination of the above identified protein complexes allows us to partially understand how the dysregulation of protein was a co-expression abnormality between VHL and TBP1. The upregulation of TBP1 is known to induce dysregulation of downstream HIF1A molecules in a VHL-dependent manner (Corn *et al*., 2003). In fact, the protein expression of TBP1 increased in the tumor group. We also examined the PBAF complex containing the driver gene PBRM1, which is thought to occur following VHL abnormalities. Along with a decrease in PBRM1 protein expression, there was a loss of tumor group-specific co-expression structure among many subunits involved with PBRM1 levels.

4 Discussion

In this study, we developed an algorithm of robust identification for protein complex aberrations based on differential co-expression structure using protein abundance. Protein expression data measured through LC/MS/MS contains a non-negligible percentage of outliers due to technical limitations and variation due to biological reasons such as post-translational modifications. This causes the problem of over- (or under-) estimation of co-expression. The copula-based DC approach is a powerful statistical framework as a solution to this problem.

In addition to noise robustness, this study does not include several other key properties of the copula that are important in capturing the co-expression structure. The first is self-equitability (Chang *et al.*, 2016; Ding *et al.*, 2017). Copulas can capture nonlinear structures between variables, and self-equitability allows us to evaluate the degree of dependency equally between variables in linear and nonlinear relations. Therefore, copula allows us to compare a much broader range of co-expressed structures than conventional linear and nonlinear correlations.

Second, we can also model simultaneous co-expression structures between three or more proteins. Although this study only identified pairwise co-expression differences,–equation (4) allows the identification of simultaneous co-expression differences across three or more proteins. However, high-dimensional estimation of the copula remains limited, and at present, in our simulations, the comparison of simultaneous co-expressed structures of 15 proteins is a performance limitation for about 100 samples.

As described, the copula-based co-expression analysis approach is a powerful modeling method for data sets where noise is expected, although there remain challenges in high-dimensional estimation. In particular, it could be useful for modeling proteome-wide protein expression patterns. The proposed approach is useful for understanding the abnormalities in the protein complexes of cancer. Studies focusing on protein complexes in large-scale cancer proteomics are in their infancy. We believe that this approach will provide valuable insights into the molecular mechanisms of cancer and the search for new drug targets.

Acknowledgements

The super-computing resource was provided by the Human Genome Center (the Univ. of Tokyo).

Funding

This work was supported by MEXT Program for Promoting Researches on the supuercomputer Fugaku, JSPS KAKENHI Grant Numbers JP18H04899, JP18K18151, and JP20H04282.

Conflict of Interest: none declared.

Data Availability Statement

The data underlying this article are available in CPTAC data portal, and CORUM3.0. The datasets were derived from sources in the public domain: https://cptac-data-portal.georgetown.edu and http://mips.helmholtz-muenchen.de/corum/.

References

Adamcsek, B. et al. (2006) CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics, 22(8), 1021-1023.

Allaire J., Francois, R., Ushey, K., Vandenbrouck, G., Geelnard, M. Intel (2019) RcppParallel: Parallel Programming Tools for 'Rcpp'. R package version 4.4.4.

Amar, D. et al. (2013) Dissection of regulatory networks that are altered in disease via differential co-expression. PLoS Comput Biol, 9(3), e1002955.

Bhuva, D.D. et al. (2019) Differential co-expression-based detection of conditional relationships in transcriptional data: comparative analysis and application to breast cancer. Genome Biol, 20(1), 236.

Chang, Y. et al. A robust-equitable copula dependence measure for feature selection. In: Arthur, G. and Christian, C.R., editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Proceedings of Machine Learning Research: PMLR; 2016. p. 84-92.

Choi, Y. and Kendziorski, C. (2009) Statistical methods for gene set co-expression analysis.

Bioinformatics, 25(21), 2780-2786.

Clark, D.J. et al. (2019) Integrated proteogenomic characterization of clear cell renal cell carcinoma. Cell, 179(4), 964-983 e931.

Corn, P.G. et al. (2003) Tat-binding protein-1, a component of the 26S proteasome, contributes to the E3 ubiquitin ligase function of the von Hippel–Lindau protein. Nat Genet, 35(3), 229-237.

Ding, A.A. et al. (2017) A robust-equitable measure for feature ranking and selection. J Mach Learn Res, 18(1), 2394-2439.

Fukushima, A. (2013) DiffCorr: an R package to analyze and visualize differential correlations in biological networks. Gene, 518(1), 209-214.

Giurgiu, M. et al. (2019) CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res, 47(D1), D559-D563.

Gunawardana, Y. et al. (2015) Outlier detection at the transcriptome-proteome interface. Bioinformatics, 31(15), 2530-2536.

Hoadley, K.A. et al. (2018) Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell, 173(2), 291-304.e296.

Hubert, M. et al. (2005) ROBPCA: A new approach to robust principal component analysis. Technometrics, 47(1), 64-79.

Kerrigan, J.J. et al. (2011) Production of protein complexes via co-expression. Protein Expr Purif, 75(1), 1-14.

Knijnenburg, T.A. et al. (2009) Fewer permutations, more accurate P-values. Bioinformatics (Oxford, England), 25(12), i161-i168.

Li, Q.K. et al. (2019) Challenges and opportunities in the proteomic characterization of clear cell renal cell carcinoma (ccRCC): A critical step towards the personalized care of renal cancers. Semin Cancer Biol, 55, 8-15.

Liu, Y. et al. (2016) On the dependency of cellular protein levels on mrna abundance. Cell, 165(3), 535-550.

Mertins, P. et al. (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. Nature, 534(7605), 55-62.

Nelsen, R.B. An introduction to copulas. Springer Publishing Company, Incorporated; 2010.

Nepusz, T. et al. (2012) Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods, 9(5), 471-472.

Nusinow, D.P. et al. (2020) Quantitative proteomics of the cancer cell line encyclopedia. Cell, 180(2), 387-402.e316.

Ori, A. et al. (2016) Spatiotemporal variation of mammalian protein complex stoichiometries. Genome Biol, 17(1), 47.

Rémillard, B. and Scaillet, O. (2009) Testing for equality between two copulas. J Multivar Anal, 100(3), 377-386.

Romanov, N. et al. (2019) Disentangling genetic and environmental effects on the proteotypes of individuals. Cell, 177(5), 1308-1318.e1310.

Ryan, C.J. et al. (2017) A compendium of co-regulated protein complexes in breast cancer reveals collateral loss events. Cell Syst, 5(4), 399-409 e395.

Seo, J. (2020) Randomization tests for equality in dependence structure. J Bus Econ Stat, 1-35.

Srihari, S. et al. (2014) Complex-based analysis of dysregulated cellular processes in cancer. BMC Syst Biol, 8(4), S1.

Villaseñor-Alva, J.A. and González-Estrada, E. (2009) A bootstrap goodness of fit test for the generalized Pareto distribution. Comput Stat Data Anal, 53(11), 3835-3841.

Zhang, H. et al. (2016) Integrated proteogenomic characterization of human high-grade serous ovarian cancer. Cell, 166(3), 755-765.

Fig 1. Actual example of effects of outliers on co-expression. Difference in Pearson's correlation before and after removing outlier samples; the left panel shows a histogram of the difference in correlation differences. The right panel shows a scatter plot of the original correlation against one without outlier samples.

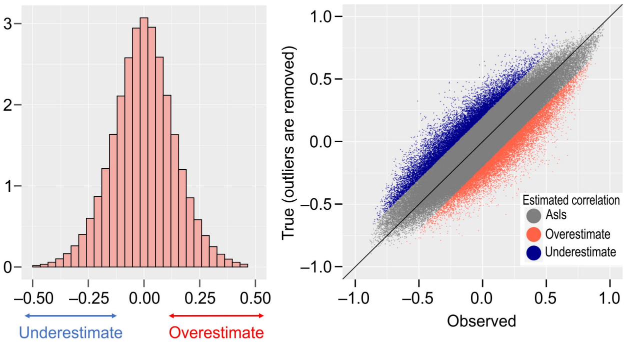
Fig 2. Overview of RoDiCE. **a) Objective of the analysis via RoDiCE.** The proposed method aims to identify abnormal protein complexes by comparing two abnormal groups. An abnormal complex is one where the co-expressed structure is different in at least two subunits. **b) Protein co-expression and outliers.** The protein expression levels measured through LC/MS/MS contain some outliers because of the addition of noise from several sources. These can cause over- (or under-) estimation in the co-expression structure. **c) Copula decomposition.** The RoDiCE model decomposes the observed joint distributions of protein expression into a marginal distribution representing the behavior of each protein and an empirical copula function representing the latent co-expression structures between proteins. This allows us to extract potential co-expressed structures and compare them robustly against outliers. The figure shows an example where the co-expressed structure estimated by copula is actually the same for two apparently different joint distributions of protein expression. d) Copula robustness. A copula is a function that expresses a dependency on a rank-transformed space of data scales. One advantage of transforming the original scale into a space of rank scale is that it is robust to outliers. The example in the figure compares Pearson's linear correlations with Pearson's linear correlations in the space converted to a rank scale by a copula function (Spearman's linear correlations). Pearson's linear correlation underestimates from 0.74 to 0.44 due to outliers, whereas the linear correlation on the rank scale has a relatively small effect (0.72 to 0.62). **e) RoDiCE is a copula-based two-sample test.** RoDiCE is an efficient method for testing differences in copula functions between two groups. Rather than a summary measure such as correlation coefficients, we compare copula functions expressing overall dependence between groups. This allows us to robustly identify differences in complex co-expression structures between two groups of protein complexes to outliers.

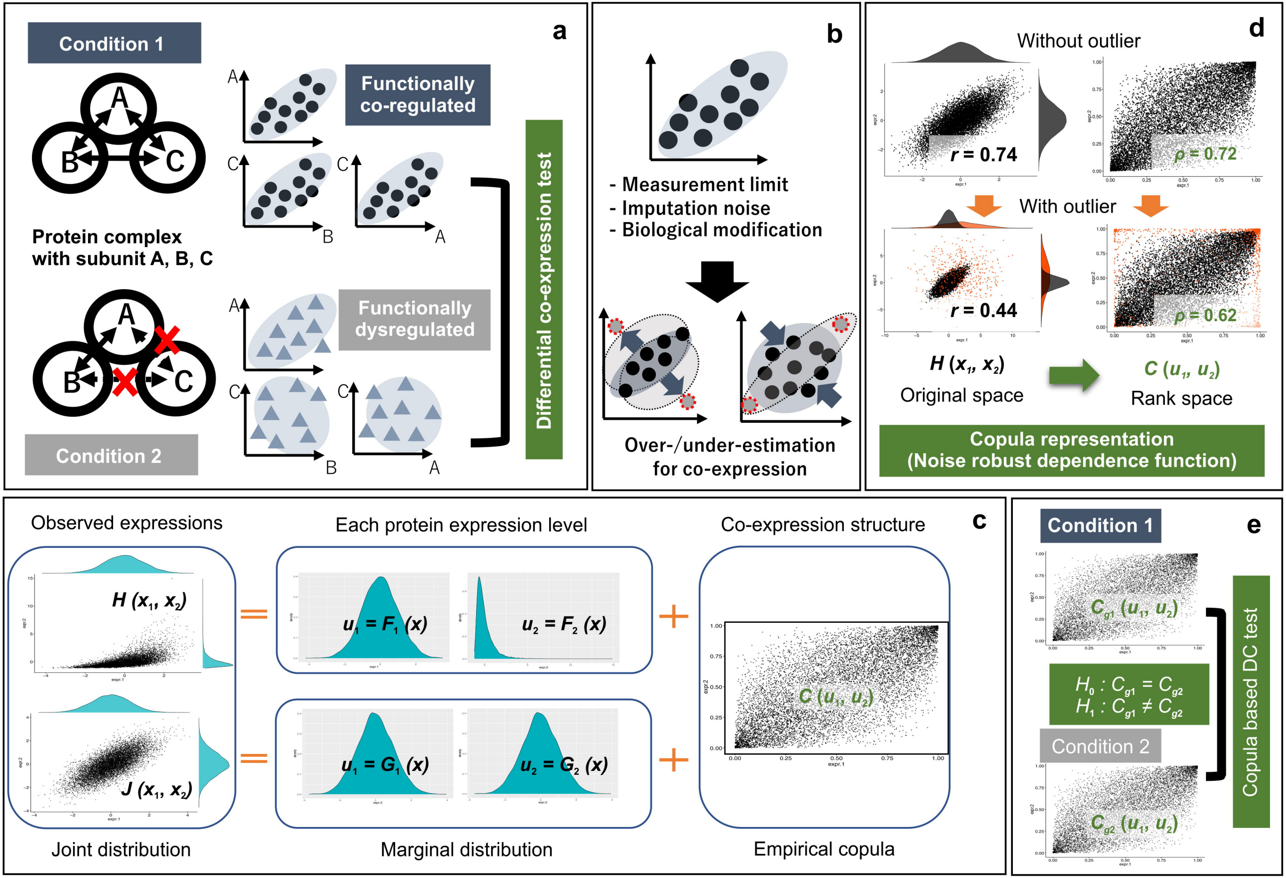
Fig 3. Simulated dataset. Generated samples in the numerical experiments for the bivariate case. To mimic the noises in proteome abundance dataset, the outlier population was assumed other than the that of tumor and normal population.

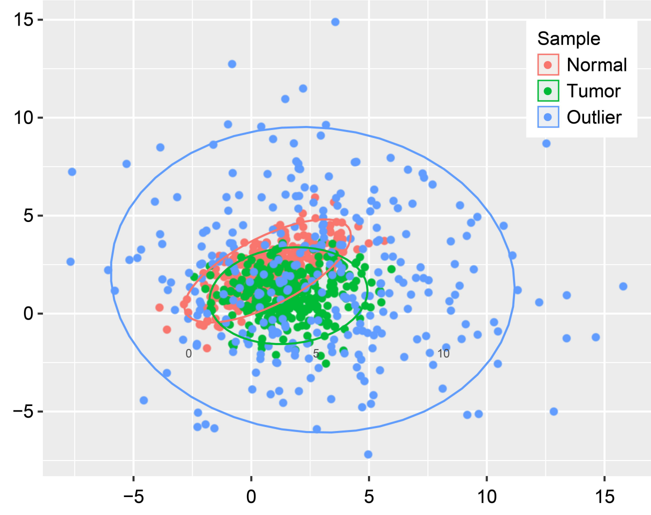
Fig 4. Sensitivities and ratio of outliers. The percentage of outliers is taken on the horizontal axis, and the sensitivity of the co-expression differences by each method (5% level of significance) is shown on the vertical axis.

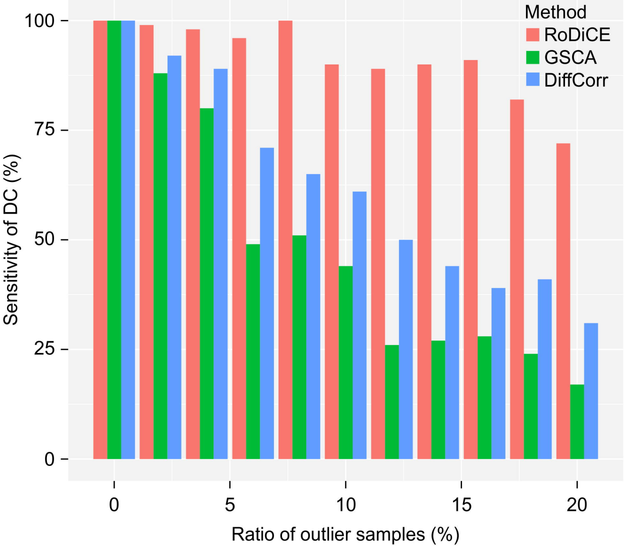
Fig 5. Sensitivities and sample size. The horizontal axis shows the sample size, while the vertical axis shows the sensitivity of the co-expression differences by each method (5% level of significance).

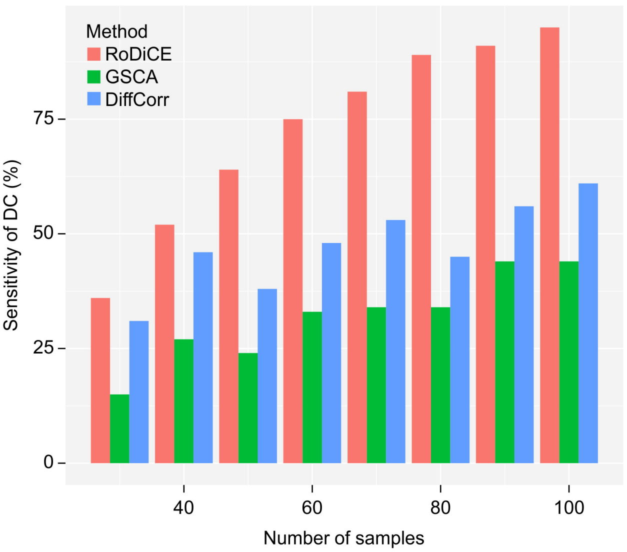
Fig 6.Identified protein complexome related to driver genes.a) Dysregulated protein complex with known driver and druggable genes. The red shows the pairs with differential co-expression between the subunits of the protein complexes (5% level of significance). The thickness of the line is proportional to -log10(p-value). The blue lines are the non-significant pairs. The yellow nodes represent proteins whose expression was actually measured by LC/MS/MS in this study, and the gray ones represent proteins that were not measured. **b) Examples of VHL- TBP1-HIF1A complex and PBAF complex with the co-expression structure**. Blue and red represent the tumor and normal groups, respectively, and the density distribution of protein expression is shown on the diagonal. In the lower diagonal, the co-expression pattern before copula-transformations is illustrated. The co-expression pattern after copula-transformations is illustrated in the upper diagonal.

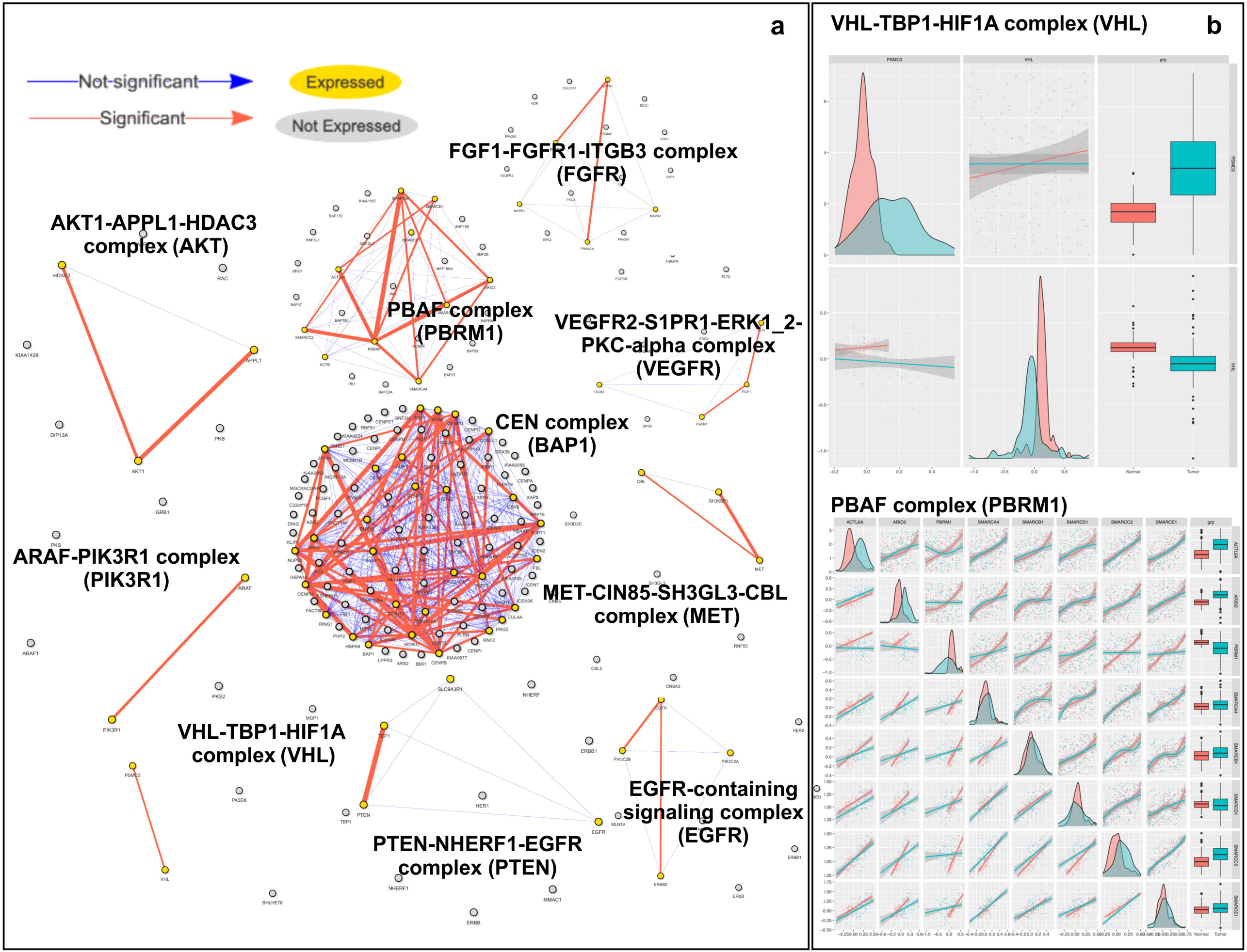

3
2
1
0
−0.50
−0.25
0.00
0.25
0.50
Underestimate
Overestimate
1.0
0.5
0.0
−0.5
−1.0
True (outliers are removed)
−1.0
−0.5
0.0
0.5
1.0
Observed
Estimated correlation
AsIs
Overestimate
Underestimate


a
Condition 1
A
B
C
Functionally co-regulated
Protein complex with subunit A, B, C
Functionally dysregulated
Condition 2
Differential co-expression test
b
- Measurement limit
- Imputation noise
- Biological modification
Over-/under-estimation for co-expression
d
Without outlier
r = 0.74
ρ = 0.72
With outlier
r = 0.44
ρ = 0.62
H (x1, x2)
Original space
C (u1, u2)
Rank space
Copula representation (Noise robust dependence function)
c
Observed expressions
Each protein expression level
Co-expression structure
H (x1, x2)
J (x1, x2)
=
u1 = F1 (x)
u2 = F2 (x)
u1 = G1 (x)
u2 = G2 (x)
+
C (u1, u2)
Joint distribution
Marginal distribution
Empirical copula
e
Condition 1
Cg1 (u1, u2)
H0 : Cg1 = Cg2
H1 : Cg1 ≠ Cg2
Condition 2
Cg2 (u1, u2)
Copula based DC test
expr.1
expr.2
density


15
10
5
0
−5
−5
0
5
10
15
0
5
10
Sample
● Normal
● Tumor
● Outlier


Method
RoDiCE
GSCA
DiffCorr
100
75
50
25
0
Sensitivity of DC (%)
0
5
10
15
20
Ratio of outlier samples (%)


Method
RoDiCE
GSCA
DiffCorr
Sensitivity of DC (%)
0
25
50
75
40
60
80
100
Number of samples


a
Not-significant
Significant
Expressed
Not Expressed
FGF1-FGFR1-ITGB3 complex (FGFR)
AKT1-APPL1-HDAC3 complex (AKT)
PBAF complex (PBRM1)
VEGFR2-S1PR1-ERK1_2-PKC-alpha complex (VEGFR)
CEN complex (BAP1)
ARAF-PIK3R1 complex (PIK3R1)
MET-CIN85-SH3GL3-CBL complex (MET)
VHL-TBP1-HIF1A complex (VHL)
PTEN-NHERF1-EGFR complex (PTEN)
EGFR-containing signaling complex (EGFR)
HDAC3
APPL1
AKT1
RAC
KIAA1428
DIP13A
PKB
GRB1
PKS
ARAF
PIK3R1
ARAF1
PKS2
PSMC3
VHL
PASD8
MOP1
BHLHE78
TBP1
PTEN
TEP1
SLC9A3R1
HER1
EGFR
NHERF1
NHERF
MMAC1
ERBB
ERBB1
ERBB2
PIK3C2B
PIK3C2A
MLN19
CNSA3
HER2
CBL2
RNF55
CBL
SH3KBP1
MET
SH3GL3
SH3D2C
GP3A
ITGB3
FGF1
FGFR1
MAPK1
MAPK3
PRKACA
PKCA
VEGFR2
S1P1
EDG1
ERK1
ERK2
PRKM1
PRKM2
PRKM3
KDR
CHEDG1
FGFBR
HBGFR
FLT2
SMARCB1
SMARCD1
SMARCC2
SMARCA4
SMARCE1
ACTL6A
ARID2
PBRM1
ACTB
BAF170
BAF155
BAF180
BAF190A
BAF47
BAF53
BAF53A
BAF57
BAF60A
BRG1
KIAA1557
SNF5L1
SNF2B
SNF2L4
INI1
INO80K
PB1
BAP1
CENPA
CENPB
CENPC1
CENPI
CENPH
CENPM
CENPO
CENPU
KIAA0977
BMI1
ARS2
LPPR3
HSPA8
PHP2
RING1
FACT80
HSPA1B
NUP98
KLIP1
DING
C22orf18
MGCRACGAP
WCRP135
MCM21R
PANE1
ZFPM
KIAA0224
KIAA0853
RNF51
RNF2
RNF1
PCGF4
SNF2H
WTAP
CUL4A
PRG2
FBL
ICEN36
ICEN37
ICEN7
ICEN2
KIAA0095
DDX38
FACTP140
SUPT1
XAP8
PRP16
CBX8
KIAA0505
WDR15
WDR1
FLRN
b
VHL-TBP1-HIF1A complex (VHL)
PSMC3
VHL
grp
Normal
Tumor
PBAF complex (PBRM1)
ACTL6A
ARID2
PBRM1
SMARCA4
SMARCB1
SMARCD1
SMARCC2
SMARCE1
grp
Normal
Tumor