BMC Genomics

SOFTWARE

Open Access

PRANA: an R package for differential co-expression network analysis with the presence of additional covariates

Seungjun Ahn[1,2,3*] and Somnath Datta[4]

Abstract

Background Advances in sequencing technology and cost reduction have enabled an emergence of various statistical methods used in RNA-sequencing data, including the differential co-expression network analysis (or differential network analysis). A key benefit of this method is that it takes into consideration the interactions between or among genes and do not require an established knowledge in biological pathways. As of now, none of existing softwares can incorporate covariates that should be adjusted if they are confounding factors while performing the differential network analysis.

Results We develop an `R` package `PRANA` which a user can easily include multiple covariates. The main `R` function in this package leverages a novel pseudo-value regression approach for a differential network analysis in RNA-sequencing data. This software is also enclosed with complementary `R` functions for extracting adjusted *p*-values and coefficient estimates of all or specific variable for each gene, as well as for identifying the names of genes that are differentially connected (DC, hereafter) between subjects under biologically different conditions from the output.

Conclusion Herewith, we demonstrate the application of this package in a real data on chronic obstructive pulmonary disease. `PRANA` is available through the CRAN repositories under the GPL-3 license: https://cran.r-project.org/web/packages/PRANA/index.html.

Keywords Differential network analysis, Pseudo-value regression, RNA-Seq data, Covariate adjustment

*Correspondence:
Seungjun Ahn
seungjun.ahn@mountsinai.org
[1] Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, USA
[2] Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, USA
[3] Institute for Healthcare Delivery Science, Icahn School of Medicine at Mount Sinai, New York, USA
[4] Department of Biostatistics, University of Florida, Gainesville, USA

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

The RNA-sequencing (RNA-Seq) leverages the rapid breakthroughs of the next-generation sequencing platform for profiling high-quality gene expression. Over the span of years, the RNA-Seq has emerged as an alternative to other gold standard techniques in transcriptomes [1, 2]. In contrast to microarrays, RNA-Seq achieves a higher resolution and lower technical variability [2–4] which leads to a higher reproducibility [5]. Another advantage of RNA-Seq relative to previously developed transcriptomic sequencing methods is that it has the ability to track transcriptomic dynamics (or gene expression changes) of tissues during physiological changes [5, 6], which thus allows a comparison of biological samples from patients with or without a specific disease or condition.

In response to these advantages, a vast number of statistical methods have become available to elucidate the genes or biological pathways associated with biological conditions or health outcomes, such as differential expression (DE) analysis [7, 8] and pathway enrichment (PE) analysis [9–11] of read counts (or gene expression) of an RNA-Seq data. However, it can be argued that results of DE analysis may provide limited information with the increased evidence that genes work in conjunction each other [12, 13]. The PE analysis appears to be a useful complement to the analysis of DE. The fundamental hypothesis in a PE analysis is that genes are regulated under common biological processes and clustered as a 'pathway' [13, 14], which borrows *a priori* pathway knowledge from the public repositories, namely, Gene Ontology [15], Kyoto Encyclopedia of Genes and Genomes [16], or Reactome [17]. To put it another way, PE analysis is primarily restricted to its use in a reference collection of well-studied biological processes only. Thereby, the idea of 'network' is introduced to pursue the veiled information that are obscured in those well-defined pathways [18].

The differential network (DN) analysis provides novel insights for identifying changes in gene-gene interactions under different biological conditions [19]. In theory, such changes are assessed through a comparison in characteristics of a network structure (*i.e.*, network topology) between two or more networks that are perturbed by a specific biological condition such as the development of cancer.

Despite the growing popularity, none of existing methods [20–22] fully addresses how to adjust for additional covariates (*e.g.* patient-age, patient-reported family histories, and other comorbidities) that may be associated with network topology.

Recently, we have adopted a pseudo-value regression [23] that allows covariate adjustment for the DN analysis while maintaining a high precision and recall values via a Monte Carlo simulation comparing with other methods available in `R` packages such as `DINGO` [20] and `dnapath` [22]. In addition, the computation time of this approach was shown competitive. To date, this is the first attempt of statistical method for the DN analysis with the inclusion of additional covariates.

In this article, we describe the software built as an `R` package, namely `PRANA` (**P**seudo-value **R**egression **A**pproach in **N**etwork **A**nalysis). `PRANA` is tailored to incorporate additional covariates information that may be associated with measures of connectivity of a gene (*i.e.* centrality) and a binary group indicator. This differs from previous statistical framework (or softwares) in DN analysis such as `dnapath` and `DINGO`.

Implementation

Algorithm

The algorithm below summarizes how the pseudo-value regression approach is embedded in a function named with `PRANA`. Briefly, the association measures are marginal quantities, such as degree centralities of each gene. Through the use of jackknife pseudo-values [24], we find the contribution of each individual data point to these quantities. Therefore, we could regress them on additional covariates as shown in studies with multi-state survival data [25, 26]. More details on methodological aspects are fully described elsewhere [23].

Algorithm 1 PRANA: Pseudo-value Regression Approach in Network Analysis

Input:	n_z samples (in rows) $\times$ p expression levels of genes (in columns) RNA-Seq expression data and $n_z \times q$ phenotype data for each group $z = 1, 2$.
Output:	A vector of adjusted p-values (and coefficient estimates and p-values) of the group variable for each gene k with a covariate adjustment.
1:	Estimate $p \times p$ association matrix (a matrix form of a network) via ARACNE [27] from the $n_z \times p$ expression data for each group $z = 1, 2$.
2:	Obtain the group-specific degree centrality by taking the marginal sums of association matrix of each taxa $k \in \{1, \cdots, p\}$.
3:	Repeat the first two steps above but using the association matrix that is re-estimated from the expression data without the $i \in \{1, \cdots, n_z\}$ individual of $n_z \times p$ data.
4:	Calculate a group-specific jackknife pseudo-value for each gene k and individual i based on summary measures of degree centrality from Steps 2 and 3.
5:	For each gene k, a robust regression is fitted with a binary group variable and additional covariates to obtain the p-values of the group variable. In the regression, a binary group variable is the main predictor to declare a gene is DC between two groups under different conditions (or phenotypes).
6:	Lastly, a vector of gene-specific adjusted p-values [28] of the group variable is returned. See the Results section for more demonstration.

Details of functions in PRANA

Main function

The main R function to perform the pseudo-value regression for the DN analysis with additional covariates is PRANA. The PRANA function imports two R scripts for the calculation of (1) total connectivity of an association matrix estimated from an observed expression data (as in thetahats function) and (2) adjusted *p*-values with the empirical Bayes screening procedure (as in EBS function) [28]. A list of three data.frame objects (coefficient estimates, *p*-values, and adjusted *p*-values of each predictor variable included in the regression for each gene) are returned upon the execution of PRANA function.

Supporting functions

For user convenience, we provide six additional R functions for extracting adjusted *p*-values (adjpval, adjpval_specific_var), coefficient estimates (coeff, coeff_specific_var), and genes that are significantly DC (sigDCGnames, sigDCGtab) from the output from PRANA function.

Dependencies

The PRANA package is fully implemented in R statistical programming language. The package depends on the base R packages (parallel, stats) and other R packages from the Comprehensive R Archive Network library (CRAN; dnapath, dplyr, robustbase) and Bioconductor (minet). Of important note, minet package should be directly installed from Bioconductor for a full utilization of PRANA package. This can be done by executing the code below in the R console.

```
> if (!require("BiocManager", quietly = TRUE))
>     install.packages("BiocManager")
>
> BiocManager::install("minet")
```

Results

In this section, we illustrate how PRANA can be applied in practice using the sample dataset available from the package. This case study is the same as the one analyzed in our methodology paper [23].

Sample dataset

The PRANA package includes a sample dataset named combinedCOPDdat_RGO with 406 samples. combinedCOPDdat_RGO consists of an RNA-Seq expression data for 28 genes that were spotlighted as associated with the chronic obstructive pulmonary disease (COPD) from a genome-wide association study [29]. It is a subset of the original study stored in the Gene Expression Omnibus (GEO) database with the accession number GSE158699 [30]. In this sample dataset, a phenotype data on six clinical and demographic variables is also available: current smoking status (main grouping variable), smoking pack years, age, gender, race, and FEV1 percent. The user can call the sample data into R by executing the following code:

```
> data(combinedCOPDdat_RGO)
```

Alternatively, the user can also assign the data to an object by running the code below:

```
> combinedCOPDdat_RGO <- data.frame(combinedCOPDdat_RGO)
```

Data processing

The PRANA function requires a user to provide each expression and phenotype data separately.

```
> # A gene expression data part of the sample data in the package.
> rnaseqdat <- combinedCOPDdat_RGO[ , 8:ncol(combinedCOPDdat_RGO)]
> rnaseqdat <- as.data.frame(apply(rnaseqdat, 2, as.numeric))

> # A clinical data with additional covariates sorted by current smoking groups.
> # The first column is ID, so do not need it for the analysis.
> phenodat <- combinedCOPDdat_RGO[order(combinedCOPDdat_RGO$currentsmoking), 2:7]
```

The main predictor variable in this example analysis is the current smoking status. As discussed in the Algorithm subsection, the estimation of association matrices (or networks) and the calculation of jackknife pseudo-values are carried out for each group separately. Hence, we add another step that locates the indices of subjects from each 'current' vs. 'non-current smokers' group. These indices are used to dichotomize expression dataset into 'non-current smokers (Group A)' and 'current (Group B).'

```
> # Locate indices of non-current smoker (namely Group A)
> index_grpA <- which(combinedCOPDdat_RGO$currentsmoking == 0)

> # Locate indices of current smoker (namely Group B)
> index_grpB <- which(combinedCOPDdat_RGO$currentsmoking == 1)
```

Apply PRANA function for DN analysis with additional covariates

Once the data processing is complete, a user can perform a DN analysis with additional covariates. PRANA function takes an expression and phenotype data, separately, in which a user specifies each for RNASeqdat and clindat, respectively. To be more specific, the variables included in phenotype data are included in the regression. In addition, the group-specific indices for the main binary indicator variable are provided as groupA and groupB within the function.

```
> PRANAres <- PRANA(RNASeqdat = rnaseqdat, clindat = phenodat,
                    groupA = index_grpA, groupB = index_grpB)
```

The output of the `PRANA` function is a list containing three `data.frame` objects for coefficient estimates, *p*-values, and adjusted *p*-values of all covariates included in the fitted model for each gene. Results are shown as following:

```
> PRANAres
$beta_hat
       currentsmoking       packyrs           age        gender          race      FEV1perc
10370    -0.015923185  1.272127e-04  1.884305e-04 -0.0003448328  5.311220e-04 -5.826295e-05
10420    -0.017520087  3.758589e-06  5.567182e-05 -0.0002452846  1.384601e-03 -1.562595e-05
1306     -0.001199981 -3.696866e-05  3.087665e-06 -0.0006514587 -9.729100e-04 -2.816952e-05
155185    0.032747279  9.815991e-06  2.823639e-06 -0.0006648456  9.509333e-05  4.436619e-06
158158   -0.003688330 -1.293570e-04 -2.800379e-04 -0.0008169064 -7.055233e-03 -1.798243e-04
1653      0.052229992  1.673516e-04 -8.586810e-04 -0.0031473933  1.551101e-02  1.296290e-04
1762     -0.039504702 -3.296516e-05  1.151257e-04  0.0042993327 -3.245991e-03 -4.655511e-05
23389    -0.029471437 -1.911291e-04 -1.091513e-03 -0.0068881178 -5.445772e-03  2.912158e-04
253461   -0.015654090 -2.183082e-05 -2.101935e-04  0.0002418143 -1.537160e-03  6.989872e-05
26112     0.006796306  4.224185e-04  2.284288e-04  0.0035982066  1.237162e-02  1.293891e-04
27436     0.263922890  4.920743e-05 -2.591953e-04 -0.0068087712  3.108772e-03  5.590200e-05
3308     -0.046753366 -2.112484e-04  7.026322e-05 -0.0021921008  2.667842e-03 -1.979934e-04
3696     -0.022265826 -2.464874e-05 -1.223302e-04  0.0003193570 -2.248239e-04  3.409178e-06
374739    0.047504274  1.021637e-05 -5.005521e-05  0.0005436326  2.017402e-03 -1.213636e-06
3842      0.018684831  4.506823e-05 -2.569539e-04 -0.0044633798  2.337416e-03  2.923333e-05
406      -0.073622065  1.403537e-05  2.899945e-04  0.0004705523 -2.805929e-04 -2.657994e-05
56986     0.074511137  7.438047e-06 -2.327374e-04 -0.0044879749  9.779275e-04 -3.414017e-05
57188    -0.189838433  3.672060e-05 -2.143454e-04  0.0002165781 -1.060657e-02 -1.025648e-04
6239     -0.028301163 -1.070732e-04 -1.232089e-04 -0.0037386434  5.506294e-03  3.845861e-04
7067     -0.022673542 -6.418046e-05 -1.618867e-04 -0.0018703457 -7.020795e-03  3.248006e-05
7871     -0.048841485 -8.521528e-05  4.224567e-04 -0.0008603334  3.020690e-04 -7.829072e-05
79961    -0.028555433  8.651445e-05 -1.527294e-04 -0.0018326148 -5.355818e-03 -1.095733e-04
79991     0.044442935  1.242848e-04 -2.601397e-04  0.0071609694 -1.732144e-03 -1.508376e-04
8224      0.053191007 -1.948290e-05 -7.986941e-05 -0.0017460934  1.204092e-05 -1.309136e-05
8853     -0.077543049  8.401266e-05 -4.137692e-04 -0.0045733001 -1.126188e-03 -9.023869e-05
8870      0.013084836  1.676543e-05 -8.365854e-05  0.0022132956  2.158166e-04 -3.518894e-05
9258     -0.043918475  3.071681e-05  1.030171e-04  0.0008799247  3.058057e-03 -1.129862e-05
9686     -0.002303434 -2.649319e-04  2.784477e-04 -0.0027024727 -9.294526e-03  3.663285e-05

$p_values
       currentsmoking   packyrs       age    gender       race  FEV1perc
10370    7.732482e-03 0.2837789 0.5861065 0.9469778 0.93177198 0.5771873
10420    6.483446e-15 0.9281961 0.6535006 0.8963903 0.52868933 0.6847284
1306     3.845696e-01 0.1618687 0.9697622 0.5952988 0.50060997 0.2657330
155185   1.341347e-40 0.8012608 0.9811419 0.7131627 0.96367596 0.9047306
158158   7.380534e-01 0.5749667 0.6558980 0.9345270 0.54825412 0.3706859
1653     4.277080e-03 0.6355931 0.4242131 0.8496803 0.42626394 0.7027843
1762     4.079048e-13 0.7462144 0.6991162 0.3425724 0.54749793 0.6139273
23389    5.767745e-02 0.5533291 0.2327674 0.6229866 0.74056014 0.3136595
253461   2.099401e-05 0.7432837 0.3222342 0.9395823 0.68458124 0.2844497
26112    5.242799e-01 0.0435240 0.7261802 0.7077250 0.28045988 0.5159566
27436    5.990881e-71 0.7996742 0.6614024 0.4506606 0.76994599 0.7677969
3308     8.295055e-04 0.4478664 0.9302960 0.8601277 0.85704212 0.4353824
3696     1.884030e-16 0.6177023 0.3987125 0.8843605 0.93166199 0.9409488
374739  1.239949e-119 0.5905722 0.3569484 0.5174993 0.03970556 0.9417403
3842     1.442885e-03 0.6984548 0.4362995 0.3837444 0.69600119 0.7765790
406      1.646660e-35 0.8866099 0.3279592 0.9170204 0.95803496 0.7642985
56986    1.097939e-27 0.9506012 0.5074842 0.4078940 0.87814365 0.7580619
57188    3.217633e-24 0.9177676 0.8313784 0.9888367 0.54732207 0.7458750
6239     2.680984e-01 0.8361306 0.9327978 0.8692856 0.83351508 0.4067122
7067     1.283914e-14 0.2225225 0.3261859 0.4488176 0.01584216 0.5117362
7871     2.430468e-14 0.4783881 0.2258711 0.8732380 0.96236055 0.4846421
79961    6.168738e-03 0.6809502 0.8004045 0.8441911 0.62800877 0.5791109
79991    2.708993e-09 0.3734978 0.5409022 0.2671561 0.81936379 0.2596152
8224     2.082235e-85 0.5434599 0.4169679 0.2406212 0.99481843 0.6721167
8853     1.016950e-18 0.6037731 0.3900027 0.5309755 0.89507641 0.5350331
8870     1.963830e-08 0.7000305 0.5310892 0.2681548 0.92727982 0.3951455
9258     4.718550e-51 0.4800908 0.4317508 0.6668133 0.20129646 0.7846253
9686     7.807454e-01 0.1167067 0.5739855 0.7200415 0.28758605 0.8109360

$adjp_values
       currentsmoking packyrs   age gender  race FEV1perc
10370           0.004   0.998 1.000      1 1.000        1
10420           0.000   1.000 1.000      1 1.000        1
1306            0.714   0.990 1.000      1 1.000        1
155185          0.000   1.000 1.000      1 1.000        1
158158          0.864   1.000 1.000      1 1.000        1
1653            0.002   1.000 1.000      1 1.000        1
1762            0.000   1.000 1.000      1 1.000        1
23389           0.020   1.000 0.998      1 1.000        1
253461          0.000   1.000 0.998      1 1.000        1
26112           0.804   0.874 1.000      1 1.000        1
27436           0.000   1.000 1.000      1 1.000        1
3308            0.000   1.000 1.000      1 1.000        1
3696            0.000   1.000 1.000      1 1.000        1
374739          0.000   1.000 1.000      1 0.632        1
3842            0.000   1.000 1.000      1 1.000        1
406             0.000   1.000 0.998      1 1.000        1
56986           0.000   1.000 1.000      1 1.000        1
57188           0.000   1.000 1.000      1 1.000        1
6239            0.436   1.000 1.000      1 1.000        1
7067            0.000   0.998 0.998      1 0.338        1
7871            0.000   1.000 0.998      1 1.000        1
79961           0.004   1.000 1.000      1 1.000        1
79991           0.000   1.000 1.000      1 1.000        1
8224            0.000   1.000 1.000      1 1.000        1
8853            0.000   1.000 1.000      1 1.000        1
8870            0.000   1.000 1.000      1 1.000        1
9258            0.000   1.000 1.000      1 0.998        1
9686            0.864   0.980 1.000      1 1.000        1
```

Some supporting functions

The package offers some auxiliary features. A user can get a table of adjusted *p*-values and coefficient estimates for all variables with `adjptab` and `coeff` functions as following:

```
# PRANAres is an object for the results after PRANA function.
adjptab <- adjpval(PRANAres)
coefftab <- coeff(PRANAres)
```

Suppose, for instance, we are interested in looking at the adjusted *p*-values for the current smoking status variable instead of a table with all variables. `adjpval_specific_var` function is available for that purpose:

```
# Call the table with adjusted p-value for all variables.
> adjptab <- adjpval(PRANAres)
# Next, use adjpval_specific_var() function to
> adjpval_specific_var(adjptab = adjptab, varname = "currentsmoking")
       currentsmoking
10370           0.004
10420           0.000
1306            0.714
155185          0.000
158158          0.864
1653            0.002
1762            0.000
23389           0.020
253461          0.000
26112           0.804
27436           0.000
3308            0.000
3696            0.000
374739          0.000
3842            0.000
406             0.000
56986           0.000
57188           0.000
6239            0.436
7067            0.000
7871            0.000
79961           0.004
79991           0.000
8224            0.000
8853            0.000
8870            0.000
9258            0.000
9686            0.864
```

Similarly, `coeff_specific_var` function can be executed to return a coefficient estimate for a specific variable (current smoking status in the example below). A cautionary note is that the user must provide the name of a variable as in `varname` within each `adjpval_specific_var` or `coeff_specific_var` functions.

```
# Call the table with coefficient estimates for all variables.
> coefftab <- coeff(PRANAres)
> coeff_specific_var(coefftab, varname="currentsmoking")
       currentsmoking
10370    -0.015923185
10420    -0.017520087
1306     -0.001199981
155185    0.032747279
158158   -0.003688330
1653      0.052229992
1762     -0.039504702
23389    -0.029471437
253461   -0.015654090
26112     0.006796306
27436     0.263922890
3308     -0.046753366
3696     -0.022265826
374739    0.047504274
3842      0.018684831
406      -0.073622065
56986     0.074511137
57188    -0.189838433
6239     -0.028301163
7067     -0.022673542
7871     -0.048841485
79961    -0.028555433
79991     0.044442935
8224      0.053191007
8853     -0.077543049
8870      0.013084836
9258     -0.043918475
9686     -0.002303434
```

Additionally, `sigDCGtab` and `sigDCGnames` functions take a `data.frame` object as an input, defined by `adjpval` function earlier, to output the names of DC genes (*i.e.* NCBI Entrez gene IDs in the first column) for the main binary grouping variable utilized for the DN analysis, as well as corresponding adjusted *p*-values. `sigDCGnames` returns the names of DC genes only. A user may adjust the level of significance (`alpha`), which is set to 0.05 by default. Please see the following commands below:

```
# Adjusted p-values and names of significantly DC genes for current smoking status.
> sigDCGtab(adjptab = adjptab, groupvar = "currentsmoking", alpha = 0.05)
       currentsmoking
10370           0.004
10420           0.000
155185          0.000
1653            0.002
1762            0.000
23389           0.020
253461          0.000
27436           0.000
3308            0.000
3696            0.000
374739          0.000
3842            0.000
406             0.000
56986           0.000
57188           0.000
7067            0.000
7871            0.000
79961           0.004
79991           0.000
8224            0.000
8853            0.000
8870            0.000
9258            0.000

# Only the names of significantly DC genes for current smoking status.
> sigDCGnames <- sigDCGnames(adjptab = adjptab, groupvar = "currentsmoking", alpha = 0.05)
> sigDCGnames
 [1] "10370"  "10420"  "155185" "1653"   "1762"   "23389"  "253461" "27436"
 [9] "3308"   "3696"   "374739" "3842"   "406"    "56986"  "57188"  "7067"
[17] "7871"   "79961"  "79991"  "8224"   "8853"   "8870"   "9258"
```

As a result, `PRANA` identified 23 genes that are significantly DC between current and non-current smokers while accounting for additional covariates such as smoking pack years, age, gender, race, and FEV1 percent.

As an additional step, a user can utilize `rename_genes` function from the dependency package (`dnapath`) to rename results with Entrez gene IDs into gene symbols. See below for the demonstration in R console. Results are summarized in Table 1.

```
> dnapath::rename_genes(sigDCGnames, to = "symbol", species = "human")
- saving gene info to /var/folders/6q/jp1685v542nbdz7zfc78hdsh0000gn/T//RtmpOZzY1U/entrez_to_hsapiens.rds
  [1] "CITED2"    "TESK2"    "AMZ1"     "DDX1"     "DMWD"     "MED13L"   "ZBTB38"
  [8] "EML4"      "HSPA4"    "ITGB8"    "TEPP"     "TNPO1"    "BMAL1"    "DTWD1"
 [15] "ADAMTSL3" "THRA"     "SLMAP"    "DENND2D"  "STN1"     "SYN3"     "ASAP2"
 [22] "IER3"     "MFHAS1"
```

Table 1 Results of DC genes obtained from PRANA. The sample dataset contains the NCBI Entrez gene IDs, so does the resulted DC genes (first column). `dnapath::rename_genes` is utilized to rename Entrez gene IDs to gene symbol (second column)

Entrez ID	Gene symbol
10370	CITED2
10420	TESK2
155185	AMZ1
1653	DDX1
1762	DMWD
23389	MED13L
253461	ZBTB38
27436	EML4
3308	HSPA4
3696	ITGB8
374739	TEPP
3842	TNPO1
406	BMAL1
56986	DTWD1
57188	ADAMTSL3
7067	THRA
7871	SLMAP
79961	DENND2D
79991	STN1
8224	SYN3
8853	ASAP2
8870	IER3
9258	MFHAS1

Discussion

The `R` package `PRANA` has been published in the CRAN (https://cran.r-project.org/web/packages/PRANA/index.html). This package has no operating system dependencies. A vignette is available on this package at https://cran.r-project.org/web/packages/PRANA/vignettes/UserManualPRANA.html or can be accessed by typing in an `R` console (`browseVignettes(package="PRANA")`). In this package, the sample dataset is provided with COPD-related genes, as well as clinical and demographic variables. The source code of the package can be found in GitHub: https://github.com/sjahnn/PRANA.

`PRANA` has some plans for future development. Firstly, although a user may attempt a classical regression-based variable selection such as stepwise selection, we have not yet validated this through a statistical simulation experiment. Secondly, the names of genes provided in the sample dataset are Entrez gene IDs. Further extension will include a function that convert from these gene IDS to gene symbols (*i.e.* 10370 to CITED2) and vice versa for user convenience.

In conclusion, `PRANA` is a user-friendly and novel regression-based method that accounts for additional covariates along with the main binary grouping variable for the DN analysis.

Conclusions

The differential network analysis identifies changes in measures of associations between genes under different biological conditions. Although there has been increasing volume of work in this subject, overall covariate

adjustment remains underexplored. In this paper, we present `PRANA`, the first `R` package that adjusts for additional covariates for the differential network analysis. As a brief note on the usage, `PRANA` takes RNA-sequencing and phenotype data (metadata) as inputs and in return tables containing DC gene names and their corresponding adjusted *p*-values are produced for a main binary grouping variable to be adjusted with the presence of additional covariates. This software is easy to install and user-friendly.

Availability and requirements

Project name: PRANA

Project home page: https://cran.r-project.org/web/packages/PRANA/index.html

Operating system(s): Platform independent

Programming language: `R`

Other requirements: Install `dnapath`, `dplyr`, `robustbase`, and `minetR` packages

License: GNU GPL-3

Any restrictions to use by non-academics: No restrictions

Abbreviations

COPD	Chronic obstructive pulmonary disease
DC	Differentially connected
DN	Differential network
GEO	Gene Expression Omnibus

Acknowledgements

The authors would like to thank the maintainers at the Comprehensive R Archive Network (CRAN). It is our special thanks to the investigators (Wang Z and Castaldi P) who have graciously shared their data publicly available in the GEO database.

Authors' contributions

S.D. conceived the original idea of integrating jackknife pseudo-values to differential co-expression network analysis. S.A. developed the methodology, completed statistical programming in R and performed data analyses of the study. S.A. drafted the manuscript. S.D. provided suggestions when writing the manuscript. All authors have reviewed and edited the manuscript.

Funding

Research reported in this publication was supported in part by the National Cancer Institute Cancer Center Support Grant [NIH P30CA196521-01] awarded to the Tisch Cancer Institute of the Icahn School of Medicine at Mount Sinai and used the Biostatistics Shared Resource Facility. The content is solely the responsibility of S.A. and does not necessarily represent the official views of the National Institutes of Health.

Availability of data and materials

PRANA is freely available at (https://cran.r-project.org/web/packages/PRANA/index.html). The COPDGene data is available in the GEO database with accession number GSE158699 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158699). Please reach out to the maintainer (Seungjun Ahn, seungjun.ahn@mountsinai.org) if you have any further inquiries on the data or code.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 2 June 2023 Accepted: 6 November 2023
Published online: 16 November 2023

References

1. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harb Protoc. 2015;11:951–69.
2. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS ONE. 2014;9:78644.
3. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18(9):1509–17.
4. Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced Applications of RNA Sequencing and Challenges. Bioinform Biol Insights. 2015;9(Suppl 1):29–46.
5. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.
6. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5:621–8.
7. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
8. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
9. Ulgen E, Ozisik O, Sezerman OU. pathfindR: an R package for comprehensive identification of enriched pathways in omics data through active subnetworks. Front Genet. 2019;10:858.
10. 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. 2005;102(43):15545–50.
11. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284–7.
12. de la Fuente A. From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. Trends Genet. 2010;26(7):326–33.
13. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):1002375.
14. Stanford BCM, Clake DJ, Morris MRJ, Rogers SM. The power and limitations of gene expression pathway analyses toward predicting population response to environmental stressors. Evol Appl. 2020;13(6):1166–82.
15. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium Nat Genet. 2000;25(1):25–9.
16. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
17. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005;33(Database issue):428–32.
18. Creixell P, Reimand J, Haider S, Wu G, Shibata T, Vazquez M, et al. Pathway and network analysis of cancer genomes. Nat Methods. 2015;12(7):615–21.
19. Shojaie A. Differential network analysis: a statistical perspective. WIREs Comput Stat. 2021;13:1508.
20. Ha MJ, Baladandayuthapani V, Do KA. DINGO: differential network analysis in genomics. Bioinformatics. 2015;31(21):3413–20.
21. McKenzie AT, Katsyv I, Song WM, Wang M, Zhang B. DGCA: a comprehensive R package for differential gene correlation analysis. BMC Syst Biol. 2016;10(1):106.

22. Grimes T, Potter SS, Datta S. Integrating gene regulatory pathways into differential network analysis of gene expression data. Sci Rep. 2019;9(1):1–12.
23. Ahn S, Grimes T, Datta S. A Pseudo-Value Regression Approach for Differential Network Analysis of Co-Expression Data. BMC Bioinformatics. 202;24(1):8.
24. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Philadelphia: Chapman & Hall/CRC; 1993.
25. Logan BR, Zhang MJ, Klein JP. Marginal models for clustered time-to-event data with competing risks using pseudovalues. Biometrics. 2011;67:1–7.
26. Klein JP, Gerster M, Andersen PK, Tarima S, Perme MP. SAS and R functions to compute pseudo-values for censored data regression. Comput Methods Programs Biomed. 2008;89:289–300.
27. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006;7(Suppl 1):S7.
28. Datta S, Datta S. Empirical Bayes screening of many *p*-values with applications to microarray studies. Bioinformatics. 2005;21(9):1987–94.
29. Sakornsakolpat P, Prokopenko D, Lamontagne M, Reeve N, Guyatt A, Jackson V, et al. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. Nat Genet. 2019;51(3):494–505.
30. Wang Z, Masoomi A, Xu Z, Boueiz A, Lee S, Zhao T. Improved prediction of smoking status via isoform-aware RNA-seq deep learning models. PLoS Comput Biol. 2021;17(10):1009433.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.