OPEN ACCESS Freely available online

PLOS ONE

DCGL v2.0: An R Package for Unveiling Differential Regulation from Differential Co-expression

Jing Yang[1,2☯], Hui Yu[3,4☯], Bao-Hong Liu[3], Zhongming Zhao[4,5,6], Lei Liu[2], Liang-Xiao Ma[3], Yi-Xue Li[1,3]*, Yuan-Yuan Li[3]*

1 School of Biotechnology, East China University of Science and Technology, Shanghai, P. R. China, **2** Bioinformatics Center, Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, P. R. China, **3** Shanghai Center for Bioinformation Technology, Shanghai Industrial Technology Institute, Shanghai, P. R. China, **4** Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, **5** Departments of Psychiatry, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, **6** Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America

Abstract

Motivation: Differential co-expression analysis (DCEA) has emerged in recent years as a novel, systematic investigation into gene expression data. While most DCEA studies or tools focus on the co-expression relationships among genes, some are developing a potentially more promising research domain, differential regulation analysis (DRA). In our previously proposed R package DCGL v1.0, we provided functions to facilitate basic differential co-expression analyses; however, the output from DCGL v1.0 could not be translated into differential regulation mechanisms in a straightforward manner.

Results: To advance from DCEA to DRA, we upgraded the DCGL package from v1.0 to v2.0. A new module named ''Differential Regulation Analysis'' (DRA) was designed, which consists of three major functions: *DRsort*, *DRplot*, and *DRrank*. *DRsort* selects differentially regulated genes (DRGs) and differentially regulated links (DRLs) according to the transcription factor (TF)-to-target information. *DRrank* prioritizes the TFs in terms of their potential relevance to the phenotype of interest. *DRplot* graphically visualizes differentially co-expressed links (DCLs) and/or TF-to-target links in a network context. In addition to these new modules, we streamlined the codes from v1.0. The evaluation results proved that our differential regulation analysis is able to capture the regulators relevant to the biological subject.

Conclusions: With ample functions to facilitate differential regulation analysis, DCGL v2.0 was upgraded from a DCEA tool to a DRA tool, which may unveil the underlying differential regulation from the observed differential co-expression. DCGL v2.0 can be applied to a wide range of gene expression data in order to systematically identify novel regulators that have not yet been documented as critical.

Availability: DCGL v2.0 package is available at http://cran.r-project.org/web/packages/DCGL/index.html or at our project home page http://lifecenter.sgst.cn/main/en/dcgl.jsp.

Citation: Yang J, Yu H, Liu B-H, Zhao Z, Liu L, et al. (2013) DCGL v2.0: An R Package for Unveiling Differential Regulation from Differential Co-expression. PLoS ONE 8(11): e79729. doi:10.1371/journal.pone.0079729

Editor: Yi Xing, University of California, Los Angeles, United States of America

Received May 19, 2013; **Accepted** October 3, 2013; **Published** November 20, 2013

Funding: This work was supported by grants from National Key Basic Research Program (grant numbers: 2012CB316501, 2013CB910801, and 2012AA022101); National Natural Science Foundation of China (31171268 to LYY, 31000380 to YH, and81272279 to LN); National Scientific-Basic Special Fund (2009FY120100); National High Technology Research and Development Program (863 Program) (No. 2012AA020409) and was partially supported by National Institutes of Health grants (R01LM011177, R03CA167695, R03DE022093). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: yyli@scbit.org (YYL); yxli@scbit.org (YXL)

☯ These authors contributed equally to this work.

Introduction

In the transcriptome analysis domain, differential co-expression analysis (DCEA) is emerging as a unique complement to traditional differential expression analysis. Rather than calculating expression level changes of individual genes, DCEA investigates differences in gene interconnection by calculating the expression correlation changes of gene pairs between two conditions. In the past few years, a large variety of DCEA methods have been developed, such as Log Ratio of Connectivity (LRC) [1], Average Specific Connectivity (ASC) [2], Weighted Gene Co-expression Network (WGCNA) [3,4], Differential Co-expression profile (DCp) [5,6], Differential Co-expression enrichment (DCe) [5,6], ROS-DET [7], Gene Set Co-expression Analysis [8], and others. These methods vary in how they specify expression correlations and quantify differential co-expression; they also differ in the levels they address: genes or gene sets. As a promising alternative to differential expression analysis, DCEA is drawing increasing attention from computational biologists and, thus, is undergoing rapid methodological improvement.

The rationale behind differential co-expression analysis is that changes in gene co-expression patterns between two contrasting

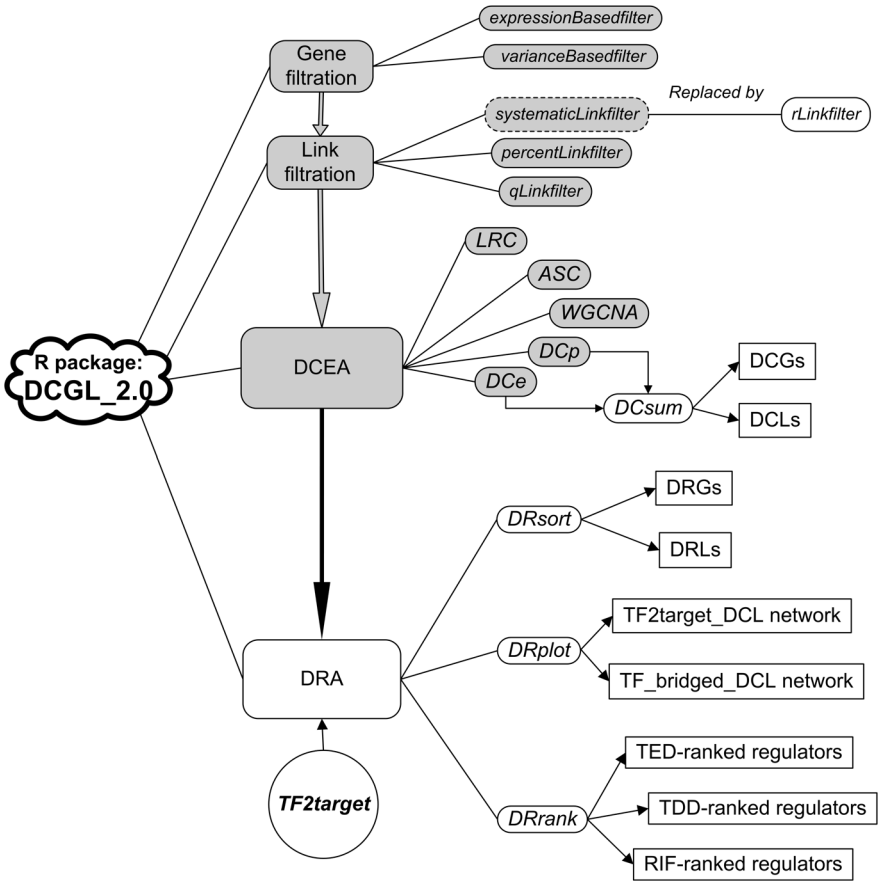


Figure 1. Overall design of DCGL v2.0. Boxes in light grey are modules/functions common to both DCGL v1.0 and v2.0.
doi:10.1371/journal.pone.0079729.g001

phenotypes (e.g., healthy and disease) provide hints regarding the disrupted regulatory relationships or affected regulatory subnetworks specific to the phenotype of interest (in this case, the disease phenotype). Therefore, among the many growing directions of DCEA, there is the so-called "differential regulation analysis" (DRA), which integrates the transcription factor (TF)-to-target information to probe upstream regulatory events that account for the observed co-expression changes. Recently, researchers have integrated differential co-expression and differential expression concepts to propose a novel Regulatory Impact Factor (RIF) that can be used to prioritize disease-causative TFs [9,10]. In addition, researchers have begun to perform differential co-expression analyses of microRNAs [11,12]. These studies are expected to lead to DRA at the post-transcription level. While the algorithm/theory facet of DRA is on the rise, the tool/application facet is lagging. The few existing tools, such as CoXpress [13] and DiffCorr [14], are fine-tuned at the DCEA stage but have not been expanded to the DRA front. Recent DRA methods, such as the RIF metric mentioned above, have not been implemented as practical tools. Currently, a software tool that implements the frontier DRA methods would fill this gap and consequently propagate DRA methods to many more biomedical research fields.

Three years ago, we released the R package DCGL (referred to as DCGL v1.0 hereafter) [6], which was designed to identify differentially co-expressed genes and links (DCGs and DCLs, respectively). The DCGL package facilitated the application of our DCEA algorithms DCp and DCe [5] in a diverse array of disease studies [15–18]. In our current work, we introduce an upgraded version of DCGL (referred to as DCGL v2.0 hereafter), in which we added ample functions to facilitate differential regulation analyses. Specifically, we incorporated the human TF-to-target library into the package, achieved the identification of differentially regulated genes and links (DRGs and DRLs, respectively), enabled a network display of intertwined regulation links and differential co-expression links, and implemented the RIF metric as well as two additional novel regulator prioritizing metrics. We have managed to turn the DCGL package into a comprehensive DRA tool. Our case study in hepatocellular carcinoma gene expression studies demonstrated the usage and applicability of DCGL v2.0 in human diseases.

Methods and Implementations

Modification of the Existing Modules in DCGL v1.0

As illustrated in Figure 1, the previous DCGL package [6], DCGL v1.0, has three functional modules: Gene filtration, Link filtration, and DCEA (short for differential co-expression analysis). The "Gene filtration" module provide two functions, *expressionBasedfilter* [19] and *varianceBasedfilter* [20], to filter out genes whose expression values are extremely low or notably invariable across samples/conditions. The "Link filtration" module includes three functions, *qLinkfilter*, *percentLinkfilter* and *systematicLinkfilter*, which are designed to construct gene co-expression networks. The "DCEA" module has three algorithms previously proposed by others (*LRC* [1], *ASC* [2], and *WGCNA* [3,4]) and two methods (*DCp* and *DCe*

[5,6]) we developed to identify differentially co-expressed genes (DCGs) and differentially co-expressed links (DCLs). These existing modules were improved in DCGL v2.0 as follows. 1) The source codes were re-organized to a more logical and efficient level. 2) A stand-alone function, *rLinkfilter*, was added to the "Link filtration" module, which cuts off links according to their expression correlation value. *SystematicLinkfilter*, used to be in DCGL v1.0, was removed from the update because it is extremely time-consuming [21], and its results require manual interpretation before they are applied to downstream functions. 3) A *DCsum* function was attached to module DCEA in order to summarize a final set of DCGs and DCLs (see the companion vignette for more details, Text S1).

DRA: a New Module in DCGL v2.0

In DCGL v2.0, we designed and implemented a new module, DRA, specifically for differential regulation analysis. The human gene regulatory relationships were developed from the "tfbsConsSites.txt" and "tfbsConsFactors.txt" files extracted from UCSC hg19 (http://genome.ucsc.edu/), and were compiled as the data library *TF2target*. *TF2target* includes 214,607 binary tuples involving 215 human TFs and 16,863 targets (see the companion vignette for more details, Text S1). In order to keep abreast of developments in human regulatory data analysis, we will continue to tidy and promote our *TF2target* library. The DRA module is comprised of three major functions, *DRsort*, *DRplot*, and *DRrank*. Briefly, *DRsort* sorts differentially regulated genes (DRGs) and differentially regulated links (DRLs) from the *DCsum*-outputted DCGs and DCLs. *DRplot* visualizes the networks of intertwined regulation links and DCLs. *DRrank* prioritizes candidate causal regulators using three alternative metrics.

DRsort: Sorting Out Potential DRGs & DRLs

As the foremost function of the DRA module, *DRsort* is designed to sift differentially regulated genes (DRGs) and differentially regulated links (DRLs) from the *DCsum*-outputted DCGs and DCLs. In this function, we scrutinize the DCGs and DCLs against the TF-to-target information and highlight a subset of the genes and links that are potentially highly related to the putative differential regulation mechanisms. If a DCG coincides with a TF (A and B in the left table in Figure 2), it is regarded as a differentially regulated gene (DRG, or TF DCG) based on the implication that a differential co-expression of this type of DCGs could be attributed to disrupted regulatory relationships between the TF and its targets. If a DCG is not a TF by itself, but its regulator(s) is/are traceable in *TF2target* (C and D in the left table in Figure 2), this DCG, though not regarded as a DRG, is reserved in a *DRsort* output to ease downstream analyses.

Among all DCLs, we obtained two types of DRLs, namely TF2target_DCLs and TF_bridged_DCLs. "TF2target_DCLs" refers to DCLs that coincide with TF-to-target relations (for example, the edge between A and B in the DRL table in Figure 2), while "TF_bridged_DCLs" refers to DCLs for which both genes share common TF(s) ((B, C) in the right table in Figure 2). Our rationale here is that the disruption of regulatory relations can affect not only the co-expression links between a regulator and its targets (TF2target_DCL), but also the co-expression links among the multiple targets of a TF (TF_bridged_DCL).

DRplot: Visualizing Differential Co-expression and Regulatory Relationships

Given the DRGs' and DRLs' output from *DRsort* as well as the TF-to-target regulatory relationships in *TF2target*, we developed a *DRplot* function to visualize a DRG-highlighted, DRL-centered network. By definition, our DRLs involved differential co-expression links and transcriptional regulation links, consequently leading to a heterogeneous network display. We offer two network plotting options to present the two types of DRLs separately: TF2target_DCLs (Figure 3A) and TF_bridged_DCLs (Figure 3B). In addition, we allow users to delimit a sub-network according to predefined gene(s) of interest, where the predefined genes involving DRLs and regulation links are extracted from the whole network (Figure 3C).

We utilized dataset GSE17967 [22] from GEO (http://www.ncbi.nlm.nih.gov/geo/) to demonstrate the function of *DRplot*. This dataset was also used in the subsequent steps for *DRrank* illustration. The details of data processing and analyses can be found in the section "Results: Assessment of DCGL v2.0."

DRrank: Ranking Regulators

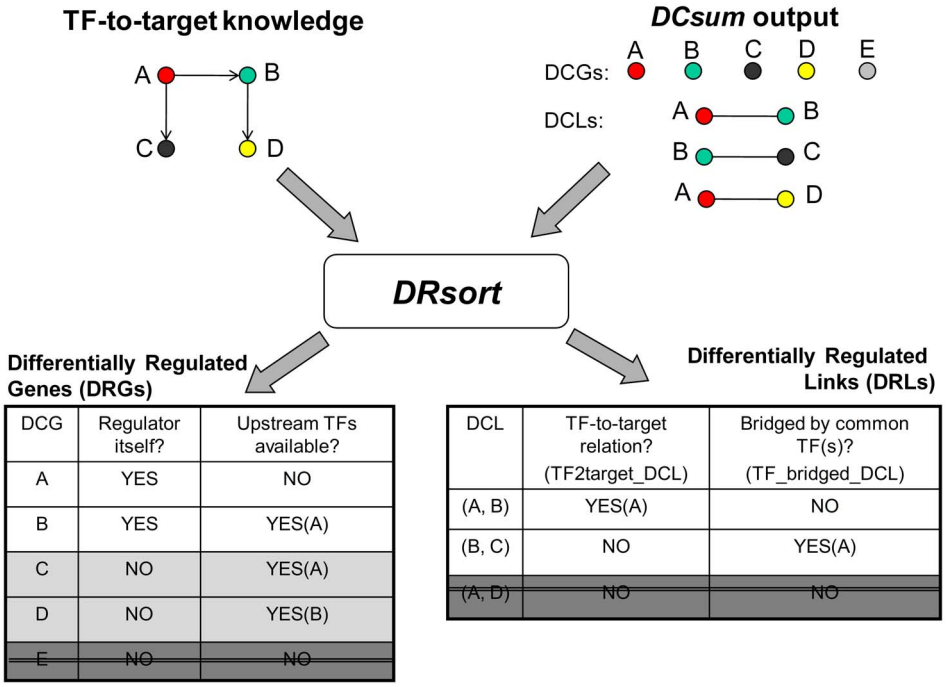
Finally, in *DRrank*, we implemented three alternative metrics for prioritizing regulators that are putatively causative to the phenotype of interest. The TED and TDD scores, short for "Targets' Enrichment Density" and "Targets' DCL Density," respectively, are novel inventions in light of our *DRsort* analysis. In addition, the "Regulatory Impact Factor" (RIF score) established earlier [9,10] was also implemented in our package.

Similar to the inferences made in previous works regarding the relationship between TFs and differentially expressed genes (DEGs) [23–25], we speculate that a TF must be more subject-relevant or even causative if more of its targets are DCGs. Based on this speculation, TED evaluates the enrichment of a particular TF's targets in DCGs using the binomial probability model. While an overall set of K DCGs are determined from a total of N genes with available expression data, out of which N_0 genes (K_0 DCGs) are covered by *TF2target* library, a TF (TF_i) with Ti targets out of N_0 genes should by chance have $Ti*K_0/N_0$ targets fall within the K_0 DCGs. If the actual number of its DCG targets, Ti', is significantly larger than the expected number, $Ti*K_0/N_0$, as judged by the cumulative density function of the binomial probability model (Eq. 1), we tend to rank TF_i higher in the regulator prioritization list. That is, if more DCGs are enriched in the targets, then the regulator is more prioritized. Of note, TED is applicable to any TF as long as the expression level of its targets are measured, which means its own expression information is not required.

$$\begin{aligned} TED(TF_i) &= -\log_2(CDF.pbinom(N,K,T_i,T_i')) \\ &= -\log_2\left(\sum_{x=T_i'}^{T_i}\binom{T_i}{x}\left(\frac{K}{N}\right)^x\left(-\frac{K}{N}\right)^{T_i-x}\right) \end{aligned} \quad (1)$$

Still using GSE17967 as an example, this dataset tested a total of 12,632 genes, out of which 1,052 genes were identified as DCGs. Taking the simplified scenario of 13 genes and 23 links in Figure 3C as an example, the TF *EGR1* (*Egr-1*) has four regulatory targets covered by GSE17967. By chance, *EGR1* should have 4*1,052/12,632 DCG targets; however, *EGR1*'s real number is three. According to Eq. 1, we have $TED(EGR1) = -\log_2\left(\sum_3^4\binom{4}{3}\left(\frac{1052}{12632}\right)^3\left(1-\frac{1052}{12632}\right)^{4-3}\right) = 14.34351$.

TDD is designed to prioritize TFs whose targets form DCLs (i.e., "common TFs" of the TF_bridged_DCLs). Bearing the same heuristic approach as in TED, we speculate that a TF of higher importance should have more of its targets forming DCLs. Based on this speculation, we borrowed the "clustering coefficient"



DCG	Regulator itself?	Upstream TFs available?
A	YES	NO
B	YES	YES(A)
C	NO	YES(A)
D	NO	YES(B)
~~E~~	~~NO~~	~~NO~~

DCL	TF-to-target relation? (TF2target_DCL)	Bridged by common TF(s)? (TF_bridged_DCL)
(A, B)	YES(A)	NO
(B, C)	NO	YES(A)
~~(A, D)~~	~~NO~~	~~NO~~

Figure 2. *DRsort* functionality: sifting the DRGs and DRLs with regulation knowledge. Given the TF-to-target knowledge (top left) as a reference, *DRsort* highlights a subset of differentially co-expressed genes and links (top right) as either differentially regulated genes (bottom left) or differentially regulated links (bottom right). As they are sorted, some DCGs/DCLs are discarded (dark grey rows with double strikethrough), while some DCGs, though they are not termed DRGs themselves, are reserved to ease downstream analysis (light grey rows in the bottom left table).
doi:10.1371/journal.pone.0079729.g002

formula [26] to measure the relevance of a TF(Eq. 2). For TF_i, if we identify N targets that have available expression data, among which k DCLs are formed, then TDD is essentially a normalized number of TF_i-bridged DCLs (Eq. 2). As with TED, TDD can rank those TFs that are not tested in an expression dataset as long as their targets' expression information is available.

$$TDD(TF_i) = CC(TF_i) = \frac{2k}{N(N-1)} \quad (2)$$

Again, based on GSE17967, *EGR1* has four targets with expression data, among which three DCLs are formed (Figure 3C). According to Eq. 2, we have $TDD(EGR1) = 2*3/4(4-1) = 0.5$.

The regulatory impact factor (RIF) was recently proposed and demonstrated as efficient in a proof-of-concept study of bovine Piedmontese myostatin mutants [9,10]. The RIF measurement simultaneously integrates three sources of information: (i) the extent of differential expression; (ii) the abundance of differentially expressed genes; and, (iii) the differential co-expression between a TF and its differentially expressed target genes (Eq. 3). In other words, the RIF algorithm assigns a high score to those TFs that are "cumulatively most differentially wired to the abundant most differentially expressed genes" [9]. In combination, these factors are assumed to contribute to the relevance of a TF in relation to the phenotype under research.

$$RIF_i = \frac{1}{n_{de}} \sum_{j=1}^{j=n_{de}} [(e1_j \times r1_{ij})^2 - (e2_j \times r2_{ij})^2] \quad (3)$$

In Eq. 3, n_{de} is the number of the differentially expressed gene (DEG); $e1_j$ or $e2_j$ denotes the expression value of DEG_j in an experimental condition (1 or 2); $r1_{ij}$ or $r2_{ij}$ designates the correlation between TF_i and DEG_j [27].

To evaluate the statistical significance of TED and TDD scores, we implement a permutation test to provide p-values as well as false discovery rate (FDR) values in conjunction with the TED/TDD scores. We randomly designate an unchanged number of pseudo targets to each TF and calculate a pseudo TED or TDD score. The number of repeated permutations can be chosen by the user (0 by default). A large number (e.g., 1,000) of pseudo TED or TDD statistics form an empirical null distribution from which the p-value can be estimated and FDR value can be obtained accordingly.

Results: Assessment of DCGL v2.0

Validation of Differential Regulation Analysis Methods

We utilized dataset GSE17967 [22] to demonstrate the utility of the new functions in DCGL v2.0. GSE17967 was designed to detect gene expression in cirrhotic tissues with (sample number = 16) and without (sample number = 47) hepatocellular carcinoma (HCC). First, a total of 1,052 DCGs and 787,150 DCLs were summarized by *DCsum* based on *DCp* and *DCe* results. The 787,150 DCLs involved 7,533 genes. These DCGs and DCLs were then used as inputs for the differential regulation analysis (DRA) pipeline, and we obtained the following major results.

DRsort identified 10 DRGs, 751 TF2target_DCLs (*i.e.* Type I DRLs), and 215,897 TF_bridged_DCLs (*i.e.* Type II DRLs). The total 216, 648 DRLs involved 6,068 genes. We found that *DRsort* here achieved a significant enrichment of the human cancer-related genes (obtained from "Cancer Gene Census," http://

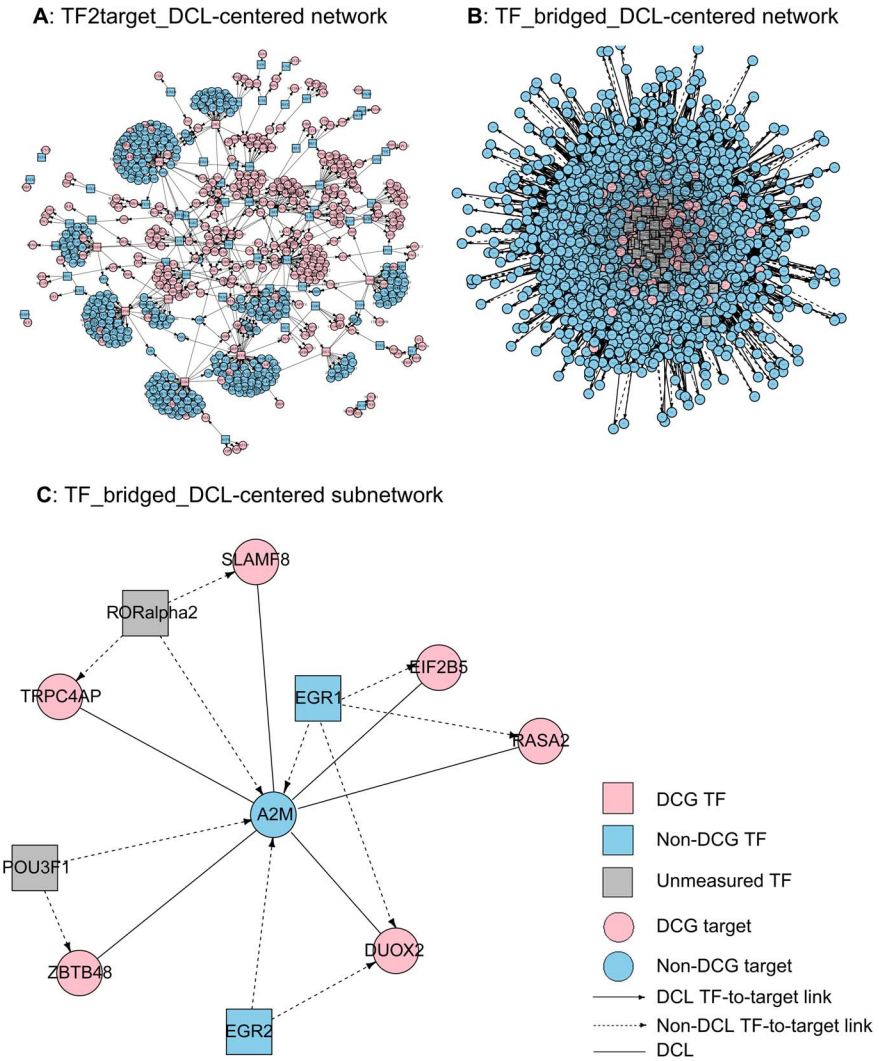


Figure 3. Example DRL-centered heterogeneous networks produced using the *DRplot* function. GSE17967 was used as the sample dataset. Nodes denote genes, and edges denote DCLs or TF-to-target links (see symbol illustration). A, TF2target_DCL-centered network contains 663 genes and 751 links. B, TF_bridged_DCL-centered network contains 6,207 genes and 294,117 links. C, A subnetwork out of the TF_bridged_DCL-centered network surrounding the predefined gene *A2M*.
doi:10.1371/journal.pone.0079729.g003

www.sanger.ac.uk/genetics/CGP/Census/) when it sifted DRGs/DRLs from DCGs/DCLs (Table 1).

DRplot plotted two types of networks, TF2target_DCL-centered (Figure 3A) and TF_bridged_DCL-centered (Figure 3B). A subnetwork of the TF_bridged_DCL-centered network is displayed in Figure 3C, which was determined using an HCC relevant gene, *A2M* [28].

Of the total 215 TFs included in the *TF2target* library, 131 had expression data available in our expression dataset. As a consequence, TED and TDD produced rankings of all 215 TFs,

Table 1. Numbers of *DCsum*/*DRsort* result items and the enrichment of cancer genes from *DCsum* result to *DRsort* result.

Comparison	Result items	Total Number	Cancer Gene Total	Cancer Gene Enrichment*
From DCGs to DRGs	DCGs	1,052	27	1.2×10^{-4}
	DRGs	10	3	
From DCLs to DRLs	Genes in DCLs	7,533	216	7.1×10^{-4}
	Genes in DRLs	6,068	191	

*Cancer Gene enrichment is shown as a p-value resulting from a binomial probability model using the four total numbers in the left columns.
doi:10.1371/journal.pone.0079729.t001

while RIF gave its ranking of the 131 TFs with available expression data.

According to the heuristic approach underlying DCEA, it is assumed that the following TFs are more likely to be implicated in the putative differential regulation mechanisms: Type I TFs, i.e. TFs that are DCGs by themselves (DRG, or TF DCGs); Type II TFs, i.e. TFs that change co-expression links with their targets (regulators involved in TF2target_DCLs); and Type III TFs, i.e. TFs whose targets form DCLs ("common TF" involved in TF_bridged_DCLs). In this case, we found 10 Type I TFs, 72 Type II TFs, and 215 Type III TFs. The 10 Type I TFs were *MYC, EP300, LMO2, FOXO1, EGR1, ZIC1, NR3C1, FOSB, GCGR,* and *STAT6*. Of them, *MYC, EP300,* and *LMO2* are annotated as cancer genes in "Cancer Gene Census" (http://www.sanger.ac.uk/genetics/CGP/Census/). Literature mining informed us that *FOXO1* [29] and *EGR1* [30] have been implicated in HCC, and *ZIC1* is down-regulated in gastric cancer [31] and has been proved to be a tumor suppressor gene in colorectal cancer [32,33]. *NR3C1* is identified as an epigenetically deregulated gene in colorectal tumorigenesis [34]. As for the 72 Type II TFs, they were shown as significantly enriched for "KEGG_PATHWAY:hsa05200:pathways in cancer" by DAVID [35] (FDR = 4.76×10^{-7}) (see Text S3 for KEGG enrichment analysis result of Type II TFs). The high bias of Type I and Type II TFs towards cancer genes in this case study supported our heuristic assumption regarding these particular TFs. Since Type III TFs spanned all 215 TFs included in *TF2target*, they were ignored in the functional enrichment analysis.

Next, we investigated the ranks of the above plausibly relevant TFs in the prioritization lists by TED, TDD, and RIF, respectively (see Text S2 for *DRrank* results for Type I, II and III TFs). It was found that Type I TFs and Type II TFs were significantly highly ranked in the 215-gene list when utilizing TED and TDD, yet this was not the case using RIF (column "Type I TF" and "Type II TF" in Table 2). Since Type III TFs cover all 215 TFs, it is impossible to carry out a comparative evaluation of the three regulator-ranking metrics based on them.

We then extracted the 27 cancer genes from the 215 TFs and discovered that these 27 genes were also significantly highly ranked in the 215-gene list when utilizing TED and TDD, yet this was not the case using RIF (column "Cancer Genes" in Table 2). These observations establish the validity of the TED and TDD designs.

Evaluation of Computational Efficiency Promotion

In DCGL v2.0, the source codes of pre-existing functions were refined/re-organized into a more logical and efficient form. We tested to see if the coding optimization enhanced computational efficiency. For the convenience of backward comparison, dataset GSE3068 [36], adopted as the benchmark dataset in DCGL v1.0, was utilized to demonstrate the promoted computational efficiency. We performed a series of numerical experiments over varied subsets of GSE3068 using the shared functions from DCGL v1.0 and v2.0, respectively. The computation time used by DCGL v2.0 functions was significantly reduced (Table 3).

Table 3. Computing time of shared functions implemented in DCGL v1.0 and DCGL v2.0, tested on different subsets of gene expression datasetGSE3068.

	Number of genes						
Function	**1,000**	**2,000**	**3,000**	**4,000**	**5,000**	**6,000**	**7,000**
DCp.percent.v1.0	0.27	1.35	2.50	4.05	5.23	8.62	12.02
DCp.percent.v2.0	0.27	0.79	1.79	3.83	4.97	7.29	10.29
DCp.qth.v1.0	0.40	2.06	3.78	6.55	8.74	13.78	19.89
DCp.qth.v2.0	0.38	1.32	3.66	5.89	7.72	13.33	19.28
DCe.percent.v1.0	0.54	2.45	4.81	9.65	13.40	18.27	25.10
DCe.percent.v2.0	0.29	1.42	4.34	6.73	9.62	15.90	18.13
DCe.qth.v1.0	0.46	1.46	3.74	6.73	11.95	15.27	25.93
DCe.qth.v2.0	0.12	1.03	3.11	5.67	9.62	11.22	16.67

Different subsets, with a gradually increasing number of genes, were taken from GSE3068 by selecting the upper rows of the full dataset. The computing platform was a Linux system with five nodes, each of which had a dual quad-core Intel Xeon 2.33GHZ CPU and a RAM of 16 GB. Execution time was averaged over three repetitive runs each.

doi:10.1371/journal.pone.0079729.t003

Discussion

Identifying the regulators that are relevant or even causative to a phenotypic change is a challenging and worthwhile goal for both experimental and computational biologists. Unfortunately, this problem cannot be solved using differential expression analysis alone. The first reason for this limitation is that causal signals are always submerged within a large amount of differentially expressed genes. More importantly, however, a causal regulator is not necessarily differentially expressed. For example, if a mutation occurs to the activation domain of a TF, the TF, while at its original expression level, can no longer activate its target genes. Another similar case is the regulation of a TF at the post-translational level, which can hamper the TF's functionality but not its expression level. In either a TF's missense mutation or its post-translational modification, the expression correlation between the TF and its targets can be affected; this phenomenon might be captured using the Differential Co-Expression Analysis (DCEA) and Differential Regulation Analysis (DRA) methodologies [1–7].

Table 2. Wilcox test p-values of particular genes' top-ranking in alternative regulator prioritization lists.

Prioritization Metrics	Type I TFs (10)	Type II TFs (72)	Type III TFs (215)[a]	Cancer Genes (27)
TED	0.026*	0.004*	–	0.007*
TDD	0.041*	0.040*	–	0.006*
RIF	0.076	0.375	–	0.489

Investigated was the positioning of four types of important genes in three Prioritization lists (by metrics TED, TDD, and RIF). Numbers of considered genes are shown in brackets.

*Statistical significance ($p<0.05$).

[a]Since Type III TFs coincide with all TFs, they are ignored in this analysis.

doi:10.1371/journal.pone.0079729.t002

Indeed, a differential wiring analysis of expression data succeeded in identifying the gene containing the causal mutation in bovine Piedmontese myostatin mutants [9]. Although we cannot routinely identify those causal regulators at the current stage, differential co-expression analysis has gained wide acknowledgement as a promising method to solve this problem [37]. From a practical viewpoint, developing efficient differential regulation analysis methods and implementing the currently available algorithms is crucial.

The present package DCGL v2.0, upgraded from the earlier version DCGL v1.0 [6], has realized a differential co-expression analysis and, furthermore, a differential regulation analysis pipeline. This upgrade enabled the identification of DCGs and DCLs, the scrutinization of DRGs and DRLs, and, more importantly, the prioritization of potential causal regulators in terms of their relevance or causativeness to a specific phenotype. We have implemented not only the recently proposed RIF method [9], but also two other self-proposed novel ones, TED and TDD. Last, but not least, we created a user-friendly graphic view of the differential co-expression/differential regulation networks. To the best of our knowledge, DCGL v2.0 is the first R package that provides convenient and practical DRA functionalities.

The prioritization of candidate regulators, or the identification of critical regulators, is the toughest and most intriguing part of differential regulation analysis. If one can properly integrate the expression information and regulatory knowledge in a biologically relevant manner, there will be a greater chance to identify true causal factors. Taking the RIF measure as an example [10], by combining the extent of differential expression, the abundance of differentially expressed genes, and differential co-expression between TFs and differentially expressed targets, this approach could capture those regulators that are cumulatively most differentially wired to the abundant most differentially expressed genes. As an efficiency-validated metric, RIF is implemented in DCGL for users' convenient utilization.

In our design of DRG/DRL selection and regulator prioritization approaches, we also aimed to make full use of the expression information and regulatory knowledge available between TFs and targets. On one hand, the plausibly relevant TFs, TF DCGs, TFs in DCLs, and TFs shared by DCL gene pairs were catalogued in our *DRsort* output for potential intensive examination, as demonstrated in our hepatocellular carcinoma case study. On the other hand, when prioritizing the candidate regulators, TED and TDD examine different aspects of differential regulation. TED assigns a high score to those regulators that regulate more DCGs, while TDD attributes a high score to those whose targets form more DCLs. In our case study, our novel metrics TED and TDD seemed to outperform RIF since they were better at prioritizing phenotype (cancer) related genes. Additionally, our two novel metrics are unique in that they can work on any TFs as long as their target genes' expression information is available. In contrast, RIF requires the expression information of the regulator itself. A systematic comparative evaluation of the regulator-prioritization metrics remains for future continuous study. However, we decided to implement all of these three measures in DCGL v2.0 since different approaches identify different sets of genes that may contribute to different parts of the process of interest.

In conclusion, DCGL v2.0 implements valuable differential co-expression analysis and differential regulation analysis methodologies. It has universal applicability and is suitable for both microarray data and RNA-seq data. With the present update, DCGL could be used to systematically identify novel TFs contributing to phenotypic change that have not yet been documented as critical, thereby significantly increasing the biological knowledge that could be derived from expression data.

Availability and Requirements

Project name: DCGL_2.0.
Project stable release: http://cran.r-project.org/web/packages/DCGL/index.html.
Project home page: http://lifecenter.sgst.cn/main/en/dcgl.jsp.
Operating system(s): Platform Independent.
Programming language(s): R.
Other requirement(s): R, R packages (igraph, limma).
License: GPL (>2).

Supporting Information

Text S1 DCGL v2.0 Vignette.
(PDF)

Text S2 *DRrank* result for the GSE17967 dataset.
(TXT)

Text S3 KEGG enrichment analysis result of Type II TFs for the GSE17967 dataset.
(TXT)

Acknowledgments

We thank Rebecca H. Posey for proofreading and editing an earlier draft of the manuscript.

Author Contributions

Conceived and designed the experiments: YYL HY JY YXL. Performed the experiments: JY HY BHL. Analyzed the data: JY HY. Contributed reagents/materials/analysis tools: YYL JY HY LXM LL. Wrote the paper: YYL HY JY ZZ.

References

1. Reverter A, Ingham A, Lehnert SA, Tan SH, Wang Y, et al. (2006) Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer. Bioinformatics 22: 2396–2404.
2. Choi JK, Yu U, Yoo OJ, Kim S (2005) Differential coexpression analysis using microarray data and its application to human cancer. Bioinformatics 21: 4348–4355.
3. van Nas A, Guhathakurta D, Wang SS, Yehya N, Horvath S, et al. (2009) Elucidating the role of gonadal hormones in sexually dimorphic gene coexpression networks. Endocrinology 150: 1235–1249.
4. Mason MJ, Fan G, Plath K, Zhou Q, Horvath S (2009) Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. BMC Genomics 10: 327.
5. Yu H, Liu BH, Ye ZQ, Li C, Li YX, et al. (2011) Link-based quantitative methods to identify differentially coexpressed genes and gene pairs. BMC Bioinformatics 12: 315.
6. Liu BH, Yu H, Tu K, Li C, Li YX, et al. (2010) DCGL: an R package for identifying differentially coexpressed genes and links from gene expression microarray data. Bioinformatics 26: 2637–2638.
7. Kayano M, Takigawa I, Shiga M, Tsuda K, Mamitsuka H (2011) ROS-DET: robust detector of switching mechanisms in gene expression. Nucleic Acids Res 39: e74.
8. Choi Y, Kendziorski C (2009) Statistical methods for gene set co-expression analysis. Bioinformatics 25: 2780–2786.
9. Hudson NJ, Reverter A, Dalrymple BP (2009) A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. PLoS Comput Biol 5: e1000382.
10. Reverter A, Hudson NJ, Nagaraj SH, Perez-Enciso M, Dalrymple BP (2010) Regulatory impact factors: unraveling the transcriptional regulation of complex traits from expression data. Bioinformatics 26: 896–904.

11. Staehler CF, Keller A, Leidinger P, Backes C, Chandran A, et al. (2012) Whole miRNome-wide differential co-expression of microRNAs. Genomics Proteomics Bioinformatics 10: 285–294.
12. Bhattacharyya M, Bandyopadhyay S (2013) Studying the Differential Co-expression of MicroRNAs Reveals Significant Role of White Matter in Early Alzheimer's Progression. Molecular BioSystems: Accepted.
13. Watson M (2006) CoXpress: differential co-expression in gene expression data. BMC Bioinformatics 7: 509.
14. Fukushima A (2013) DiffCorr: an R package to analyze and visualize differential correlations in biological networks. Gene 518: 209–214.
15. Qu Z, Miao W, Zhang Q, Wang Z, Fu C, et al. (2013) Analysis of crucial molecules involved in herniated discs and degenerative disc disease. Clinics (Sao Paulo) 68: 225–230.
16. Diao H, Li X, Hu S, Liu Y (2012) Gene expression profiling combined with bioinformatics analysis identify biomarkers for Parkinson disease. PLoS One 7: e52319.
17. Liu M, Hou X, Zhang P, Hao Y, Yang Y, et al. (2013) Microarray gene expression profiling analysis combined with bioinformatics in multiple sclerosis. Mol Biol Rep 40: 3731–3737.
18. Li G, Han N, Li Z, Lu Q (2013) Identification of transcription regulatory relationships in rheumatoid arthritis and osteoarthritis. Clin Rheumatol.
19. Prieto C, Risueno A, Fontanillo C, De las Rivas J (2008) Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. PLoS One 3: e3911.
20. Simon RaL A (2006) BRB Array Tools Users Guide. Technical Reports. Biometric Research Branch, National Cancer Institute. Available: http://linus.nci.nih.gov/~brb/download_full_new.html.
21. Elo LL, Jarvenpaa H, Oresic M, Lahesmaa R, Aittokallio T (2007) Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process. Bioinformatics 23: 2096–2103.
22. Archer KJ, Mas VR, David K, Maluf DG, Bornstein K, et al. (2009) Identifying genes for establishing a multigenic test for hepatocellular carcinoma surveillance in hepatitis C virus-positive cirrhotic patients. Cancer Epidemiol Biomarkers Prev 18: 2929–2932.
23. Essaghir A, Toffalini F, Knoops L, Kallin A, van Helden J, et al. (2010) Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. Nucleic Acids Res 38: e120.
24. Liu Q, Tan Y, Huang T, Ding G, Tu Z, et al. (2010) TF-centered downstream gene set enrichment analysis: Inference of causal regulators by integrating TF-DNA interactions and protein post-translational modifications information. BMC Bioinformatics 11 Suppl 11: S5.
25. Sohler F, Zimmer R (2005) Identifying active transcription factors and kinases from expression data using pathway queries. Bioinformatics 21 Suppl 2: ii115–122.
26. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393: 440–442.
27. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3: Article3.
28. Gangadharan B, Antrobus R, Dwek RA, Zitzmann N (2007) Novel serum biomarker candidates for liver fibrosis in hepatitis C patients. Clin Chem 53: 1792–1799.
29. Calvisi DF, Ladu S, Pinna F, Frau M, Tomasi ML, et al. (2009) SKP2 and CKS1 promote degradation of cell cycle regulators and are associated with hepatocellular carcinoma prognosis. Gastroenterology 137: 1816–1826 e1811–1810.
30. Hao MW, Liang YR, Liu YF, Liu L, Wu MY, et al. (2002) Transcription factor EGR-1 inhibits growth of hepatocellular carcinoma and esophageal carcinoma cell lines. World J Gastroenterol 8: 203–207.
31. Wang LJ, Jin HC, Wang X, Lam EK, Zhang JB, et al. (2009) ZIC1 is downregulated through promoter hypermethylation in gastric cancer. Biochem Biophys Res Commun 379: 959–963.
32. Gan L, Chen S, Zhong J, Wang X, Lam EK, et al. (2011) ZIC1 is downregulated through promoter hypermethylation, and functions as a tumor suppressor gene in colorectal cancer. PLoS One 6: e16916.
33. Gan LH, Pan J, Chen SJ, Zhong J, Wang LJ (2011) DNA methylation of ZIC1 and KLOTHO gene promoters in colorectal carcinomas and its clinicopathological significance. Zhejiang Da Xue Xue Bao Yi Xue Ban 40: 309–314.
34. Lind GE, Kleivi K, Meling GI, Teixeira MR, Thiis-Evensen E, et al. (2006) ADAMTS1, CRABP1, and NR3C1 identified as epigenetically deregulated genes in colorectal tumorigenesis. Cell Oncol 28: 259–272.
35. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44–57.
36. Hu Z, Willsky GR (2006) Utilization of two sample t-test statistics from redundant probe sets to evaluate different probe set algorithms in GeneChip studies. BMC Bioinformatics 7: 12.
37. de la Fuente A (2010) From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. Trends Genet 26: 326–333.