BIOINFORMATION
Discovery at the interface of physical and biological sciences

Open access

BIOMEDICAL INFORMATICS

www.bioinformation.net
Volume 12(6)

Hypothesis

# A comparative analytical assay of gene regulatory networks inferred using microarray and RNA-seq datasets

**Fereshteh Izadi*, Hamid Najafi Zarrini, Ghaffar Kiani & Nadali Babaeian Jelodar**

[1]Plant Breeding Department, Sari Agricultural Sciences and Natural Resources, Iran; Fereshteh Izadi - E-mail: izadi1991@yahoo.com; *Corresponding author

Received July 7 2016; Revised August 5, 2016; Accepted August 6, 2016; Published October 12, 2016

**Abstract:**
A Gene Regulatory Network (GRN) is a collection of interactions between molecular regulators and their targets in cells governing gene expression level. Omics data explosion generated from high-throughput genomic assays such as microarray and RNA-Seq technologies and the emergence of a number of pre-processing methods demands suitable guidelines to determine the impact of transcript data platforms and normalization procedures on describing associations in GRNs. In this study exploiting publically available microarray and RNA-Seq datasets and a gold standard of transcriptional interactions in *Arabidopsis*, we performed a comparison between six GRNs derived by RNA-Seq and microarray data and different normalization procedures. As a result we observed that compared algorithms were highly data-specific and Networks reconstructed by RNA-Seq data revealed a considerable accuracy against corresponding networks captured by microarrays. Topological analysis showed that GRNs inferred from two platforms were similar in several of topological features although we observed more connectivity in RNA-Seq derived genes network. Taken together transcriptional regulatory networks obtained by Robust Multiarray Averaging (RMA) and Variance-Stabilizing Transformed (VST) normalized data demonstrated predicting higher rate of true edges over the rest of methods used in this comparison.

**Key words**: gene regulatory network, RNA-Seq, microarray, normalization

**Background:**
Nowadays data mining approaches is a prominent strategy for extracting meaningful information from a growing wealth of biological data such as gene expression profiles **[1]** and elucidating high-fidelity regulatory interactions from transcriptome data is one of the most important applications of computational systems biology. Gene interactions in complex networks lead to cell metabolism thereby understanding functional molecular mechanisms obtained by these interactions is essential for gaining some insights into cellular functions, predicting downstream events and ideally manipulating the process based on desired goals. A Gene Regulatory Network (GRN) is a graph representation of biological units in which nodes represent genes while the edges are the interaction between them **[2].** Causality of regulatory process explicated by identifying and understanding the GRNs has remained as a problem in molecular biology and various methodologies have been proposed to address this issue **[3].** Generally, tools designed for recovering these interactions rely on similarity matrices indirectly measured by correlation matrices or mutual information **[4].** These matrices usually include many indirect links that should be identified and removed for increasing the reliability of GRN inference algorithms hence several sophisticated approaches attempted to remove indirect interactions and detect the causal relationships between gene pairs **[5, 6].** Gaussian Graphical Models (GGM) is one of these approaches that rely on partial correlation and supposes two genes directly related if their expression values remain dependent after removing the effects of all other variables **[5].** Because of linearity assumption between molecular measurements in Gaussian graphical models also drawback in case of thousands of genes in relation to very small number of samples, another methodologies were developed to cope with these problems. For example information-theoretic approaches such as CLR **[7]** and ARACNE **[8]** have been successfully applied for reconstructing GRNs **[9].** In these approaches first a pair-wise mutual information (MI) matrix is being calculated between all possible pairs of genes. Afterward this matrix is being manipulated for identifying regulatory interactions between nodes **[3].** However connectivity between nodes does not mean the causal relationships. Furthermore GRNs based on mutual information are time consuming for several thousand genes and some of MI estimators are biased thereby algorithms introduced by **[4 & 10]** attempted to tackle these fundamental problems and

BIOINFORMATION
Discovery at the interface of physical and biological sciences

Open access

remove overestimated regulation dependencies. While the aforementioned approaches reconstruct GRNs based on bilateral relationships, regression-based methods extract one-to-many interactions between nodes from measurement of gene expression **[11].** In this context GENIE3 infers GRNs by decomposing of network recovery procedure to p steps where p is the number of genes and each step is consisting of identifying genes that regulates a given target gene **[6].**

GRNs are mostly inferred by transcriptomics profiles such as microarrays and RNA-Seq that microarrays has been used intensively **[12].** Rapid advances in Next Generation Sequencing (NGS) techniques has deviated the massive emplyment of microarrays as the main expression data platform to RNA-Seq because of fast improvement of its depth and quality **[13].** In addition to a lower sensitivity because of unavoidable noise coming from the nature of microarrays, RNA-Seq offers several advantages over microarrays **[12, 14]**. RNA-Seq techniques do not depend on genomic pre-knowledge for transcriptome analysis and can be utilized for model and non-model organisms **[15].** While microarrays can cover only the characterized parts of genome, RNA-Seq is able to identify about whole of the transcripts **[16].** Finally detecting the novel transcripts and variant splicing are other capabilities of RNA-Seq **[17, 18].**

Biases in sequencing for example variable sequencing depths and nucleotide decomposition such as GC content and different primers demands normalization strategies to correct several biases in library preparation and any error due to uniformed sampling **[19].** On the other hand, non-specific signals produced by non-perfect match probes or background signals originated from non-similarity between sequences in microarrays are the reasons of arising normalization procedures like RMA and GCRMA to correct the noises and accurately quantification of gene expression. Briefly, expression datasets obtained by microarrays or RNA-Seq contain numerous biases which inhibit accurate quantification of gene expression level, therefore various normalization procedures are used as an essential step in transcriptome analysis **[20, 21, 22].**

The aim of this study is determining whether microarrays or RNA-Seq including different normalization procedures can recover a more accurate GRN in *Arabidopsis thaliana*. In this regard we employed six state-of-the art unsupervised algorithms considering their ability in recognition and removing indirect links between genes as well as the most popular methods of gene expression data normalization.

**Methodology:**

**Used datasets and pre-processing**

In RNA-Seq part of this study we downloaded 60 *Arabidopsis thaliana* Illumina based experiments SRA files from Sequence Read Archive (SRA) database of NCBI **(Table 1).** Collected files were converted to standard FASTQ format using sratoolkit.2.5.2 **[23].** The FASTQ files were filtered for any adapter contamination or low quality reads by cutadapt-1.8.3 **[24].** Here using $Q$-score ≥ 40 options, we only kept high quality reads corresponding to a sequencing error probability of 0.0001%. Obviously shorter reads have a greater chance to map accidentally in multiple positions on reference genome; therefore we filtered out reads shorter than 15 nucleotides to reduce any multi-mapping errors. We then performed the quality check of FASTQ files using FastQC program **[25].** Afterward we downloaded Ensembl source of *A.thaliana* TAIR10 from iGenomes database **[26]** as reference genome on which the reads passed from cutadapt filters were aligned within TopHat v2.1.0 with default parameters **[27].** Finally the BAM files output of TopHat was used as input for featurcounts-1.5.0 **[28]** to provide the raw counts file. Ultimately raw counts were normalized with three normalization methods; R package edgeR **[29]** was used to provide RPKM (Reads Per Kilo-base of gene model per Million mapped reads) values **[30].** We used R package DESeq2 to normalize the raw counts into Variance-Stabilizing Transformation (VST) and regularized logarithm (rlog) values. we as well downloaded 139 raw CEL files from NCBI Gene Expression Omnibus (GEO) database **[31]** and normalized them with four microarray pre-processing procedures; Affymetrix's MicroArray Suite (MAS5) **[32]**, Robust Multiarray Averaging (RMA) **[33],** GeneChip RMA (GCRMA) **[34]** and Variance Stabilizing Normalization (VSN) **[35].** All of the used normalization procedures can be obtained by Bioconductor packages. We checked the quality of downloaded CEL data files using Robin software **[36].** The microarray and RNA-Seq samples covered a wide range of *Arabidopsis* tissues, ages and treatments and we could consider them as analogous samples. Getting together we provided four files from each of microarrays and RNA-Seq datasets as input for GRN inference algorithms and subsequent statistical analysis.

**GRN inference methodologies**

In this study for reconstructing GRNs, we utilized six algorithms considering their ability in recognition and removing of indirect links between genes; matlab implementation of Global Silencing by **[10]** and Network Deconvolution by **[4],** Graphical Gaussian Models (GGM) by (37) using GeneNet R package and R implementation of GEne Network Inference with Ensemble of Trees (GENIE3) by **[6].** We used the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) and Context Likelihood of Relatedness (CLR) using spearman estimator and drawing ROC curves embedded in minet R package **[8].**

**Evaluation statistics**

To assess the impact of two platforms and normalization methods on GRN discovery and proposing the better ones, we drew the ROC curve that plots true positive rate versus the false positive rate. We also computed the below statistics using minet R package; (1) AUPR: The area under the PR curve and higher values higher true positive rate; (2) AUROC: The area under the ROC curve that the values larger than 0.5 showing a higher true positive rate.

BIOINFORMATION
Discovery at the interface of physical and biological sciences

Open access

Furthermore, based on the below formula we computed an overall score to evaluate the prediction of microarray and RNA-Seq platforms and different normalization procedures separately so that the larger values the better performances.

AUROCscore = i=1nAUROCi1n,

AUROCscore = i=1nAUPRi1n,

Overall Score = AUROCscore+AUROCscore 2

**Results**

**Microarray-based networks versus networks derived from RNA-Seq**

We utilized a relatively large amount of transcriptomics data gathered with microarray and RNA-Seq techniques. Our reasons for selecting and exploiting these datasets are that: I) these data cover multiple tissues, ages and experimental conditions in *Arabidopsis* and II) the predictions can be compared to a gold standard list consist of 4775 experimentally validated gene regulatory interactions in *Arabidopsis* obtained from AGRIS database **[38]** (supplementary file 2 – available with authors). This list was used to construct a reference network required for validate function in R package minet to assess the prediction accuracy of networks resulted by different algorithms and normalization methods. Microarray and RNA-Seq samples were selected as analogous as possible respect to tissue, age and experimental conditions and we used the intersection between expression profile of 32550 and 20922 genes from RNA-Seq and microarray data and 4775 gene interactions in gold set list respectively. The result was 2857 genes that were common between microarray, RNA-Seq data and gold standard and we used the expression values of these genes for recovering GRNs. In order to overviewing of datasets, we drew bar plots of log2 transformed RNA-Seq raw counts (to remove the impact of zero values) using R package DESeq2 and microarrays using R package limma **(Figure 1).** ARACNE algorithm is already sparse then we used whole of its inferred interactions against 3521 edges of reference network. About the other algorithms, we tested the first 10000 highly ranked edges and resulted networks were compared with reference network derived from gold standard. In order to quantifying the prediction accuracy of compared algorithms, we computed the area under the ROC curves (AUROCs) and the area under PR (precision-recall) curves (AUPRs) by R package minet summarized in **Tables 1 and 2.** In the next step we assessed the performance of each of normalization procedures for all of GRNs and each algorithm over each of platforms by averaging geometric mean of corresponding AUROCs and AUPRs in terms of AUROCscores and AUPRscores **[39].** Overall Score is arithmetic mean of AUROCscores and AUPRscores that ranks the compared methods such that methods with larger Overall Score perform better. As the last step of assessment procedure we computed a total AUROC for each platform, algorithm and normalization method by taking intersection between networks (see **Tables 1, 2 and the last row of Table 3**). We intersected the networks inferred by all of the algorithms from each platform and as expected we observed the robustness of RNA-Seq data in generating networks with higher performances than ones by microarrays (**Table 3 entities presented in bold**) moreover as illustrated in Fig.1 in term of ROC (receiver operating characteristic) curves, RNA-Seq derived GRNs are more overlap with gold standard in comparison to microarray derived networks.

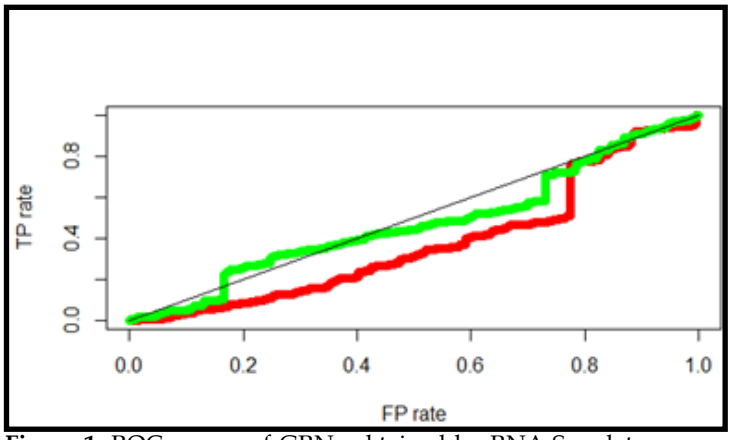


**Figure 1**: ROC curves of GRNs obtained by RNA-Seq data versus corresponding networks derived by Microarrays. Green=RNA-Seq, Red=Microarray, FP rate= false positive rate, TP rate= true positive. Evidently from the figure, GRNs derived by RNA-Seq data contained more true positive compared to corresponding networks from microarray data.

**Networks derived from different pre-processing methods**

Our results revealed that in microarray data the assessment values were closer than ones from RNA-Seq. regarding overall scores in microarrays, GCRMA and RMA and in RNA-Seq VST and RPKM were better suited **(see Tables 1 and 2).** Greater overall scores derived by Quantile-based normalization procedures (RMA and GCRMA) is an evident of more possibility of predicting true edges from these normalization methods especially in larger datasets that is inconsistent with results of a study by **[40]** declaring that co-expression networks inferred by MAS5-normalized expression values are more accurate. Although co-expression networks can be considered as the simplest kinds of GRNs that does not attempt to distinguish causal relationships from indirect interactions. Standard deviation obtained by AUROC curves in each of normalization methods was smaller in VST, rlog, VSN and RMA methods. Based on Total AUROC, in microarray data RMA and GCRMA and in RNA-Seq data RPKM and VST out-performed other normalization procedures. However, authors in **[41]** demonstrated that Correlation-based normalization procedures such as RMA and GCRMA specially the former one are not capable of predicting accurate correlation-based GRNs because of overestimating of pairwise correlation and they suggested that MAS5 is more faithful for reconstructing protein-protein

interaction networks by predicting less but more accurate correlations. RMA performed similar to MAS5 even though it is weak. RMA and GCRMA normalization procedures are based on quantile normalization and reached a good accuracy to reduce the variation between arrays. GCRMA has been a popular procedure used to convert raw microarray data into gene expression profiles and it was shown to outperform other normalization procedures in detecting differentially expressed genes **[42].** In agreement, (43) found that a combination of RMA-normalized expression values and Bayesian algorithm can predict faithful genetic networks. They believed that the performance of MAS5 and RMA is depend on the inference algorithm so that MAS5 was more powerful in recovering physical interaction networks such as protein-protein interactions and significant binding motifs while transcriptional networks reconstructed by RMA-normalized data showed consistently better accuracy. Furthermore both RMA and especially GCRMA produced highly significant correlation measurements even in the randomized set although GCRMA introduced an extraordinary number of false positives and performed poorly versus RMA due to its background adjustment step (40). VST-normalized expression values with the smallest standard deviation respect to overall scores performed the best in predicting true edges while based on Total AUROC, outperformed by RPKM normalized values **[40].** Demonstrated the similarity between VST-normalized RNA-Seq data and microarray data respect to inter-sample variation, correlation coefficient distribution and network topological architecture. VST and rlog are within DESeq2 Bioconductor package common for differential gene expression analysis and are based on a negative binomial distribution model with variance and mean linked by local regression **[44].** VST and rlog ignoring the read length, are inter-sample normalization methods that give a scaling factor, which scale sample size for each sample. In this study concerning high divergent samples coming from a wide range of experimental backgrounds, VST was more efficient in inferring GRNs versus rlog.

RPKM as a non-abundance estimation normalization method considering the sample size corrects the impact of gene length on gene counts and despite the failing in removing this bias on count number **[45, 46],** is still used in many practical fields. AUROC calculated from RPKM and raw counts data were closer to each other over the rlog and VST-normalized expression values **[47]** Revealed that the RPKM and raw count data perform similarly in defining low expressed genes as differentially expressed while in this study RPKM showed an obvious difference in detecting transcripts and following inferring regulatory relationships **[48]** showed that RPKM-normalized expression values enable to perform better in estimating Spearman correlation coefficient especially in datasets consist of read lengths of 35 nucleotides and more efficient when alignment accuracy was low and in a low rate of alignment of longer reads, RPKM revealed the least correlation. In this study the average of read length was 48 bp (supplementary file 1 – available with authors), regarding overall scores VST and regarding total AUROC RPKM performed better.

**Interaction between used datasets and algorithms:**

Random forest-based algorithm GENIE3, the best performer in DREAM4 and owning the high score in DREAM5, according to overall score statistic performed better in RNA-Seq data. This approach out-performed the other methods in a comparative study by **[5]** in microarray data. This arrangement in microarray data was ARACNE and CLR respectively. Critically about total AUROC and AUPR we should note that if used algorithms agree on an interaction (an edge between two genes) we can't be ensure that this interaction is true than ones inferred by each of algorithms separately because may one algorithm correctly infers an interaction while others did not infer this edge correctly therefore by taking intersection between networks both edges will be disappeared causing a drop in sensitivity without any increase in prediction accuracy. Hence treating these metrics cautiously, the difference between overall scores and total AUROC and AUPR might be implied on utilizing more precise parameters in GRN assessment procedure than taking a union intersection between networks. However respect to overall score, GENIE3 was more compatible with RNA-Seq data while regarding total AUROC and AUPR this method did not show a significant power over GGM, network deconvolution and global silencing. Overall score introduced ARACNE preferably more powerful in recovering GRNs from microarray data than RNA-Seq meanwhile CLR with the highest total AUROC and AUPR outperformed ARACNE. The higher performance of ARACNE could be due to the large number of datasets used in this study where the minimum recommended number of microarray expression profiles is 100 for estimating reliable mutual information in ARACNE **[8].** Indeed and concerning the less evaluated edges from ARACNE compared to the rest of compared algorithms, information-theoretic approaches especially CLR were more compatible with microarrays likely due to compatibility between correlation networks and correlation-based normalization methods. In order to evaluation of networks topological properties, we calculated two different network centrality parameters **[49]** for ARACNE networks obtained by 8 datasets from two platforms; mean node degree (number of connections) and betweenness centrality (as the percentage of times a node appears on the shortest path between all pairs of nodes in the network). We selected ARACNE because of pre-sparsity and checking all of the links without selecting a number of edges. Topological parameters were estimated using *NetworkAnalyzer* **[50]** Cytoscape plug-in. Results showed that global network properties was very similar between networks even though in RNA-Seq data GRNs tend to be associated with a higher network connectivity than genes in microarray-based networks (Supplementary file 3 – available with authors). These topological characteristics could discover critical genes in different diseases **[51, 52].** Identifying the central nodes by these measures can provide promising essential genes so that genes with higher betweenness centrality and degree

likely to be a key regulator or modulator controlling a wide range of essential cellular functions in a specific process from which *Arabidopsis* embryonic-essential genes showed a higher degree than the rest of transcriptome **[53]**. However these parameters in our study were observed to be highly dataset specific.

**Discussion:**

The availability of Massive and complex genome-wide gene expression data produced by High-throughput technologies and existing various normalization methods, researchers are interested in choosing an efficient combination of expression profiling platforms and normalization procedures as a key step toward the reconstruction of genetic regulatory networks by which they are able to understand how genes are connected and operate within intricate biological networks. RNA-Seq as a relatively novel platform for gene expression profiling allows us to study transcriptome with more precise compared to microarrays. Detecting novel transcripts, isoforms and sequence variations such as single nucleotides variations (SNV) and mapping the boundaries of exons and introns and differential splicing are some of the capabilities of RNA-Seq technology while this transcriptome platform does not suffer from microarray specific bottlenecks such as background noises arise from cross or no-specific hybridization and incorrect annotation of probes **[19].** The stated benefits in addition to providing for more sensitive detection of transcripts by RNA-Seq than microarrays are likely to be the reasons of the ability of this technology for detecting low expressed genes while microarrays fail to differentiate between very low expressed and Non-expressed genes **[54].** The same was demonstrated by **[55]** that RNA-Seq datasets (FPKM-normalized) were more sensitive to detect the low abundance transcripts versus microarray (MAS5-normalized) even though two technologies should not be considered as competitors and can overlap each other in transcriptome analysis. Their observations introduced the transcript abundance as substantial and highly statistically significant components of variations between two platforms that RNA-Seq appeared to be more sensitive for detecting this source of variation. Furthermore transcript GC content was the distinct characteristic of inter-sample variation in RNA-Seq data. Following the previously mentioned reasons, **[56]** declared two times more capacity of RNA-Seq data to predict GRNs over microarrays (RMA and Quantile-normalized data) although they noted that the GRNs in interaction levels of view are highly dataset-specific while functionally they are very similar. In their experiment CLR networks derived from Oligo gene expression dataset revealed the more fractions of cancer associated significant biological processes versus CLR captured by RNA-Seq data that is close to our results where CLR networks by microarray was more reliable. Accordingly for microarray derived GRNs, **[39]** using experimentally verified protein-protein interactions as gold standard revealed that microarray data (MAS5-normalized) was more suitable for exploring regulatory structures although RNA-Seq data was able to analyze more dynamic range networks from entire of transcriptome **[46]** has demonstrated that although RNA-Seq data is more efficient in detecting low intensity expressed gene which microarray is unable to detect their expression easily, RNA-Seq is significantly more transcript abundance-dependent than that from microarrays across multiple normalization methods. Based on their findings, microarray may show systematic biases in low abundance transcripts, possibly due to the cross hybridization. These findings may be statistically relevant with the better performance of VST-normalized data to correct the high variance introduced by low abundance transcripts. Comparing to RNA-Seq data microarray was able to identify the higher percentage of differentially expressed genes where **[52]** did not consider this characteristic due to microarray power to limit the false discovery rate. The same with previous researches as it is evident from Table 2, networks derived by RNA-Seq data allowed to predict the higher percentage of true positive rates subsequently the more accurate regulatory relationships where one reason might be the capability of RNA-Seq technology in accurate measurement of the dynamic range of low and highly expressed genes **[57]** and giving a better resolution of relationship between genes. Challenging data storage, complex data analysis procedures and being more expensive than microarray are some barriers that by overcoming, RNA-Seq platform is expected to be a prominent alternative of microarray in transcriptome analysis **[19].** Despite the expected advantages of RNA-Seq data for inferring GRN, we observed that our inferred GRNs were highly data specific even for well-known algorithms with high performance therefore using RNA-Seq platform and robust normalization methods although can't be advised as a promising and general way to infer a more accurate GRN but obviously will increase the fidelity of GRN reconstruction.

**Conclusion:**

Our study was an attempt to describe the impact of gene expression platforms and normalization methods on inferring GRNs using a wide range of GRN inference algorithms and normalization procedures. Hence using publically available microarray and RNA-Seq datasets in *Arabidopsis* and a gold standard of transcriptional interactions between gene products as a reference for evaluation of predicted edges, we observed a higher prediction accuracy of RNA-Seq derived GRNs over microarray derived networks presumably due to the ability of RNA-Seq technique in detecting low expressed genes. Moreover RMA, GCRMA, VST and RPKM normalization methods performed better and regarding stability and dispersion, the AUROC values calculated from VSN and VST-normalized data especially in RNA-Seq data with smaller standard deviation were in a closer range. Briefly, in our study used algorithms were highly data-specific and our reconstructed GRNs being transcriptional networks and considering evidences of introducing high rate of false positives by GCRMA-normalized data, we propose RMA as a more suitable microarray pre-processing procedure before inferring GRNs. However because of overestimating of pairwise correlation, we also should be cautious about data obtained by correlation-based

normalization procedures for inferring widely used and powerful correlation networks such as ARACNE and CLR. Respect to RNA-Seq data, considering a high rate of variation between samples and mean read length less than 50 bp, VST and RPKM could be good options depend of the origin of samples, reads length and alignment rate.

**Competing interests**
The authors declared that they have no competing interests.

**Acknowledgement:**
We thank Dr. Nooshin Omranian, scientific staff in Systems Biology and Mathematical Modelling Group, Max Planck Institute for Molecular Plant Physiology, Potsdam, Germany for her precious assistances.

**References:**
**[1]** Khosravi P *et al. Algorithms Mol Biol.* 2015 **11(10):** 25 [PMID: 26265933].
**[2]** Dong X *et al. Bioinform Biol Insights.* 2015 **29(9):** 61 [PMID: 25983554].
**[3]** Zhang X *et al. Bioinformatics.* 2012 **28:** 98 [PMID: 22088843].
**[4]** Feizi S *et al. Nat Biotechnology.* 2014 **31:** 726.
**[5]** Omranian N *et al. Scientific Reports.* 2016 **6:** 20533.
**[6]** Huynh-Thu VA *et al. PLoS ONE.* 2010 **5(9):** e12776 [PMID: 0927193].
**[7]** Faith JJ *et al. PLoS Biology.* 2007 **e8** [PMID: 17214507].
**[8]** Margolin AA *et al. Nat Protoc.* 2006 **1:** 662.
**[9]** Altay G & Emmert-Streib F, *Bioinformatics* 2010 **26:** 1738.
**[10]** Barzel B & Barabási AL, *Nat Biotechnology* 2013 **31:** 720.
**[11]** Linde J *et al. EXCLI Journal.* 2015 **14:** 346. [PMID: 27047314].
**[12]** Mantione KJ *et al. Med Sci Monit Basic Res.* 2014 **20:** 138. [PMCID: 4152252].
**[13]** Morin R *et al. Biotechniques.* 2008 **45(1):** 81 [PMID: 18611170].
**[14]** Wang Z *et al. Nat. Rev Genet.* 2009 **10:** 57 [PMID: 19015660].
**[15]** Balakrishnan CN *et al. Genomics.* 2012 **100:** 363.
**[16]** Giorgi FM *et al. BMC Bioinformatics.* 2010 **11:** 553 [PMID: 21070630].
**[17]** Roberts A *et al. Bioinformatics.* 2011 **27:** 2325 [PMID: 21697122].
**[18]** Richard H *et al. Nucleic Acids Res.* 2010 **38:** e112 [PMID: 20150413].
**[19]** Zhao S *et al. PLoS ONE.* 2014 **9(1):** e78644 [PMID: 24454679].
**[20]** Lee S *et al. Nucleic Acids Res.* 2011 **39(2):** e9 [PMID: 21059678].
**[21]** Piao Y *et al. Bioinformatics.* 2012 **28(24):** 3306.
**[22]** Li F *et al. Osong Public Health Res Perspect.* 2014 **5(5):** 279. [PMID: 25389514].
**[23]** http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software.
**[24]** Martin M *EMBnet.journal*. 2011 **17:** 10.
**[25]** http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
**[26]** http://support.illumina.com/sequencing/sequencing_software/igenome.html.
**[27]** Trapnell C *et al. Bioinformatics.* 2009 **25:** 1105 [PMID: 19289445].
**[28]** http://subread.sourceforge.net/.
**[29]** Robinson MD *et al.* Bioinformatics. 2010 26, pp. -1.
**[30]** Mortazavi A *et al Nat. Methods.* 2008 **5:** 621 [PMID: 18516045].
**[31]** Edgar R *et al. Nucleic Acids Res.* 2002 **30:** 207 [PMID: 11752295].
**[32]** Hubbell E *et al. Bioinformatics.* 2002 **18:** 1585 [PMID: 12490422].
**[33]** Irizarry RA *et al. Biostatistics.* 2003 **4:** 249 [PMID: 12925520].
**[34]** Wu Z *et al. J. Am. Stat. Assoc.* 2004 **99:** 909.
**[35]** Huber W *et al.* Bioinformatics. 2002 **18: (Suppl. 1)** S96.
**[36]** Lohse M *et al. Plant Physiol.* 2010 **153:** 642 [PMCID: 2879776].
**[37]** Schäfer J & Strimmer K, *Bioinformatics* 2005 **21:** 754.
**[38]** http://arabidopsis.med.ohio-state.edu/.
**[39]** Marbach D *et al. Nat Methods.* 2012 **9:** 796 [PMID: 22796662].
**[40]** Giorgi *et al. Bioinformatics.* 2013 **29(6):** 717 [PMID: 23376351].
**[41]** Califano A *et al. ISMB/ECCB.* 2007 page i282 [PMID: 17646307].
**[42]** Wu Z *et al. Johns Hopkins University, Dept. of Biostatistics Working Papers.* Working Paper 1 2004.
**[43]** Ooi BNS & Phan TT, *Theoretical Biology and Medical Modelling* 2011 **8**: 13 [PMID: 21535890].
**[44]** Anders S & Huber W, *Genome Biol* 2010 **11(10):** R106 [PMCID: 3218662].
**[45]** Bullard JH *et al. BMC Bioinformatics.* 2010 **11:** 94 [PMID: 20167110].
**[46]** Robinson DG *et al. Nucleic Acids Research.* 2015 **43(20):** 131 [PMID: 24192834].
**[47]** Dillies *et al. Briefings in Bioinformatics* 2013 **14(6):** 671 [PMID: 22988256].
**[48]** Li P *et al. BMC Bioinformatics.* 2015 **16:** 347 [PMID: 23281963].
**[49]** Koschützki D & Schreiber F, *Gene. Regul. Syst.Bio* 2008 **2:** 193 [PMID: 19787083].
**[50]** Assenov Y *et al. Bioinformatics.* 2008 **24(2):** 282 [PMID: 18006545].
**[51]** McDermott JE *et al. J Comput Biol.* 2009 **16(2):** 169 [PMID: 19178137].
**[52]** Diamond DL *et al. PLoS Pathog.* 2010 **6(1):** e1000719 [PMID: 20062526].
**[53]** Mutwil M *et al. Plant Physiol.* 2010 **152:** 29 [PMID: 19889879].
**[54]** Sultan M *et al. Science.* 2008 **321:** 956 [PMID: 18599741].
**[55]** Mooney M *et al. PLoS ONE.* 2013 **8(4):** e61088 [PMID: 23593398].
**[56]** Simoes DM *et al. BMC Systems Biology.* 2015 **9:** 21 [PMID: 25971253].
**[57]** Malone J *et al. BMC Biol.* 2011 **31(9):** 34 [PMID: 21627854].

**Edited by P Kangueane**
**Citation**: **Izadi *et al.*** Bioinformation 12(6): 340-346 (2016)

BIOINFORMATION
Discovery at the interface of physical and biological sciences

Open access

**Table 1**: Statistics for comparison of the difference between microarray and RNA-Seq platforms for different normalization methods on inferring GRNs using microarrays datasets. We used three metrics based on ROC computed by R package minet.

| Methods | RMA | | | GCRMA | | | MAS5 | | | VSN | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Number Of edges | AU-PR | AU-ROC | Number Of edges | AU-PR | AU-ROC | Number Of edges | AU-PR | AU-ROC | Number Of edges | AU-PR | AU-ROC | Total AU-ROC |
| GGM | 4079796 | 0.0012 | 0.392 | 4074085 | 0.0012 | 0.429 | 4079796 | 0.0008 | 0.397 | 4079796 | 0.0013 | 0.415 | 0.384 |
| ARACNE | 6198 | 0.001 | 0.576 | 6143 | 0.0009 | 0.56 | 5947 | 0.0009 | 0.544 | 6195 | 0.0009 | 0.569 | 0.375 |
| CLR | 1635287 | 0.0009 | 0.548 | 1668894 | 0.0006 | 0.311 | 1632241 | 0.0011 | 0.574 | 1621582 | 0.0008 | 0.45 | 0.471 |
| GENIE3 | 4079796 | 0.0008 | 0.456 | 4032070 | 0.0008 | 0.471 | 4079796 | 0.0008 | 0.531 | 4079796 | 0.0012 | 0.6 | 0.384 |
| Global Silencing | 4082653 | 0.0007 | 0.385 | 4076940 | 0.0008 | 0.365 | 4082653 | 0.001 | 0.466 | 4082653 | 0.0008 | 0.441 | 0.384 |
| Network Deconvolution | 4079796 | 0.0009 | 0.36 | 4074085 | 0.0009 | 0.365 | 4079796 | 0.0008 | 0.351 | 4079796 | 0.001 | 0.479 | 0.384 |
| Overall-scores | | 0.425 | | | 0.432 | | | 0.22 | | | 0.246 | | |
| sd (AUROC) | | 0.087 | | | 0.089 | | | 0.088 | | | 0.074 | | |
| Total AUROC | | 0.387 | | | 0.382 | | | 0.376 | | | 0.381 | | |

**Table 2:** Statistics for comparison of the difference between microarray and RNA-Seq platforms for different normalization methods on inferring GRNs using RNA-Seq datasets. We used three metrics based on ROC computed by R package minet

| Methods | rlog | | | VST | | | RPKM | | | Raw counts | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Number Of edges | AU-PR | AU-ROC | Number Of edges | AU-PR | AU-ROC | Number Of edges | AU-PR | AU-ROC | Number Of edges | AU-PR | AU-ROC | Total AU-ROC |
| GGM | 4079796 | 0.0009 | 0.511 | 4079796 | 0.0016 | 0.606 | 4079796 | 0.0006 | 0.554 | 4079796 | 0.0004 | 0.549 | 0.643 |
| ARACNE | 6539 | 0.0018 | 0.629 | 6492 | 0.0016 | 0.623 | 6320 | 0.0016 | 0.617 | 5272 | 0.0009 | 0.548 | 0.474 |
| CLR | 1678481 | 0.0054 | 0.636 | 1651629 | 0.0033 | 0.629 | 1630702 | 0.0004 | 0.469 | 1630702 | 0.0008 | 0.45 | 0.601 |
| GENIE3 | 4079368 | 0.0013 | 0.605 | 4062138 | 0.0012 | 0.578 | 4064310 | 0.0016 | 0.627 | 4061552 | 0.0018 | 0.61 | 0.662 |
| Global Silencing | 4082653 | 0.0021 | 0.625 | 4082652 | 0.0023 | 0.631 | 4082653 | 0.0009 | 0.529 | 4082653 | 0.0009 | | 0.643 |
| Network De-convolution | 4079796 | 0.0022 | 0.641 | 4079796 | 0.0004 | 0.531 | 4079796 | 0.0004 | 0.476 | 4079781 | 0.0004 | 0.451 | 0.643 |
| Overall-Score | | 0.141 | | | 0.299 | | | 0.271 | | | 0.13 | | |
| sd (AUROC) | | 0.049 | | | 0.039 | | | 0.067 | | | 0.062 | | |
| Total AUROC | | 0.621 | | | 0.651 | | | 0.654 | | | 0.579 | | |

**Table 3:** Overall Score, standard deviation, AUROC and AUPR of each algorithms over each platform and Total AUROC & AUPR for each platform presented in bold

| Methods | Microarray | | | | RNA-Seq | | | |
|---|---|---|---|---|---|---|---|---|
| | Overall Score | Sd (AUROC) | AU-ROC | AU-PR | Overall Score | sd (AUROC) | AU-ROC | AU-PR |
| GGM | 0.205 | 0.017 | 0.383 | 0.0006 | 0.277 | 0.039 | 0.643 | 0.0020 |
| ARACNE | 0.281 | 0.013 | 0.374 | 0.0001 | 0.302 | 0.038 | 0.474 | 0.0004 |
| CLR | 0.229 | 0.119 | 0.470 | 0.0003 | 0.272 | 0.096 | 0.601 | 0.0012 |
| GENIE3 | 0.256 | 0.039 | 0.383 | 0.0006 | 0.303 | 0.02 | 0.641 | 0.0020 |
| Global Silencing | 0.207 | 0.047 | 0.383 | 0.0006 | 0.29 | 0.054 | 0.643 | 0.0020 |
| Network Deconvolution | 0.193 | 0.06 | 0.383 | 0.0006 | 0.26 | 0.071 | 0.643 | 0.0020 |
| Total AUROC & AUPR | | | **0.374** | **0.0001** | | | **0.474** | **0.0004** |