Adjustment of spurious correlations in co-expression measurements from RNA-Sequencing data

Ping-Han Hsieh[1, 2], Camila Miranda Lopes-Ramos[3], Geir Kjetil Sandve[2], Kimberly Glass[4, 5], and Marieke Lydia Kuijjer[1, 6, *]

[1] *Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway*

[2] *Department of Informatics, University of Oslo, 0316 Oslo, Norway*

[3] *Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA*

[4] *Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA*

[5] *Harvard Medical School, Boston, MA 02115, USA*

[6] *Department of Pathology, Leiden University Medical Center, 2300RC Leiden, the Netherlands*

[*] *corresponding author*

1 Abstract

Gene co-expression measurements are widely used in computational biology to identify coordinated expression patterns across a group of samples, which may indicate that these genes are controlled by the same transcriptional regulatory program, or involved in common biological processes. Gene co-expression is generally estimated from RNA-Seq data, which are generally normalized to remove technical variability. Here, we find and demonstrate that certain normalization methods, in particular quantile-based methods, can introduce false-positive associations between genes, and that this can consequently hamper downstream co-expression network analysis. Quantile-based normalization can, however, be extremely powerful. In particular when preprocessing large-scale heterogeneous data, quantile-based normalization can be applied to remove technical variability while maintaining global differences in expression for samples with different biological attributes. We therefore developed CAIMAN, a method to correct for false-positive associations that may arise from normalization of RNA-Seq data. CAIMAN utilizes a Gaussian mixture model to fit the distribution of gene expression and to adaptively select the threshold to define lowly expressed genes, which are prone to form false-positive associations. Thereafter, CAIMAN corrects the normalized expression for these genes by removing the variability across samples that might lead to false-positive associations. Moreover, CAIMAN avoids arbitrary gene filtering and retains associations to genes that only express in small subgroups of samples, highlighting its potential future impact on network modeling and other association-based approaches in

large-scale heterogeneous data.

2 Introduction

Understanding the cell's regulatory machinery can provide relevant insights into healthy tissues as well as human diseases [1, 2]. While certain experimental techniques, such as chromatin immunoprecipitation sequencing (ChIP-Seq), can map interactions made by regulatory elements, such as transcription factors, it is challenging to directly observe the combined effect of multiple regulators in a systematic way. Previous studies have shown that genes undergoing similar regulatory processes tend to have coordinated expression, also called "co-expression," across samples [3, 4, 5]. Therefore, estimates of gene co-expression are commonly used to help infer associations between genes. Gene co-expression can also be used in combination with other molecular data to improve the detection of regulatory interactions [6, 7, 8, 9, 10].

Most commonly, co-expressed genes are identified using correlation coefficients [11], such as Pearson correlation, Spearman correlation, and mutual information [12, 13]. Another popular approach is to first construct models that predict the expression of one gene, using that of all other genes or potential regulators, and then apply variable selection to identify the dependencies between these genes [14, 15]. Both types of approaches aim at identifying associations between genes based on their expression levels across all samples in a dataset. Therefore, as with standard gene expression analysis, it is essential to preprocess the expression data that is used as input for co-expression analysis [16].

A commonly used approach in transcriptomic data preprocessing is the filtering of lowly expressed genes. For such genes, it is often unclear whether they are truly non-expressed, or not expressed due to, for example, low sequencing depth. Usually, such genes are filtered out using a threshold on their expression level or read count, as well as on the fraction of samples they are expressed in. However, these thresholds are often chosen arbitrarily. Moreover, when conducting analysis on a dataset with multiple experimental conditions, this approach can remove genes that are specifically expressed in a subset of samples, and thus remove potential genes of interest.

To correct for technical variability across samples, various RNA-Sequencing (RNA-Seq) normalization methods have been developed [17, 18]. Most of these methods use global properties to correct for the technical variability. For instance, DESeq computes the median ratio of gene counts relative to the geometric mean across all samples. However, such global properties can vary when considering different subsets of samples in a heterogeneous dataset—*i.e.* a dataset with samples from multiple experimental or biological conditions. Therefore, using such methods could hamper the detection of global shifts in gene expression [19]. This is particularly relevant in co-expression analysis, since the heterogeneity across samples in a dataset is what defines co-expression. To address this issue, a quantile normalization-based method was recently developed that utilizes both group-specific and global properties of the expression distribution to correct for technical variability [20]. This method, called smooth quantile normalization, or *qsmooth*, yields better preservation

of global shifts in expression as well as adequate control over the within-group variability.

Here, we show that quantile-based normalization methods, and in particular smooth quantile normalization, can introduce false-positive associations, or false-positive co-expression values, between genes. This can particularly occur in datasets that have large differences library size across samples. To correct for these false-positive associations, we developed CAIMAN, or **C**ount **A**djustment to **I**mprove the **M**odeling of **A**ssociation-based **N**etworks. CAIMAN uses a Gaussian mixture model to identify genes with low expression signals obscured by the variability introduced by quantile-based normalization methods. This allows CAIMAN to remove false-positive associations between genes, without the need to select an arbitrary threshold or to exclude genes from the analysis.

We applied CAIMAN to RNA-Seq data from the Genotype-Tissue Expression (GTEx) Consortium [21] to showcase the problem, and to data from the Mouse Encyclopedia of DNA Elements (ENCODE) [22] to validate the method. We found that CAIMAN effectively removes false-positive associations between genes. In addition to correcting for false-positive associations, CAIMAN calculates the probability of whether genes with low counts are actually expressed in the cell, instead of being artifacts caused by the non-specific alignment of reads or by technical variability introduced during data preprocessing. We anticipate that our method will benefit future co-expression and regulatory network analyses, in particular those that involve the analyses of large-scale heterogeneous RNA-Seq datasets.

3 Materials and Methods

3.1 The CAIMAN algorithm

We specifically designed CAIMAN to be used on smooth quantile normalized data, as this method allows for the normalization of large-scale heterogeneous datasets. Since smooth quantile normalization ensures the expression distribution in the same group to be similar [20], we developed CAIMAN as an adjustment algorithm that uses a Gaussian mixture model to fit the expression distribution of all samples within each group, and that then uses the posterior probability of the mixture model to identify those genes that might form false-positive associations with other genes.

CAIMAN first concatenates the two-dimensional expression array in each group into a group-specific vector of expression. It then adds a pseudo-count of one to the normalized expression data and applies a $\log_2$ transformation to avoid numerical instability and to reduce the variance of the expression distribution (Figure 1, steps 1 and 2). The pseudo count of one is chosen to ensure the values remains positive after applying the $\log_2$ transformation. The CAIMAN algorithm then constructs an augmented group-specific expression profile by concatenating the negative transformed expression values with the original log-transformed expression data (Figure 1, steps 3–4). The distribution of the resulting group-specific expression profile is symmetrical around zero. By combining these two operations, CAIMAN can approximate the distribution of normalized expression using the Gaussian mixture model, without the need to deal with non-negative

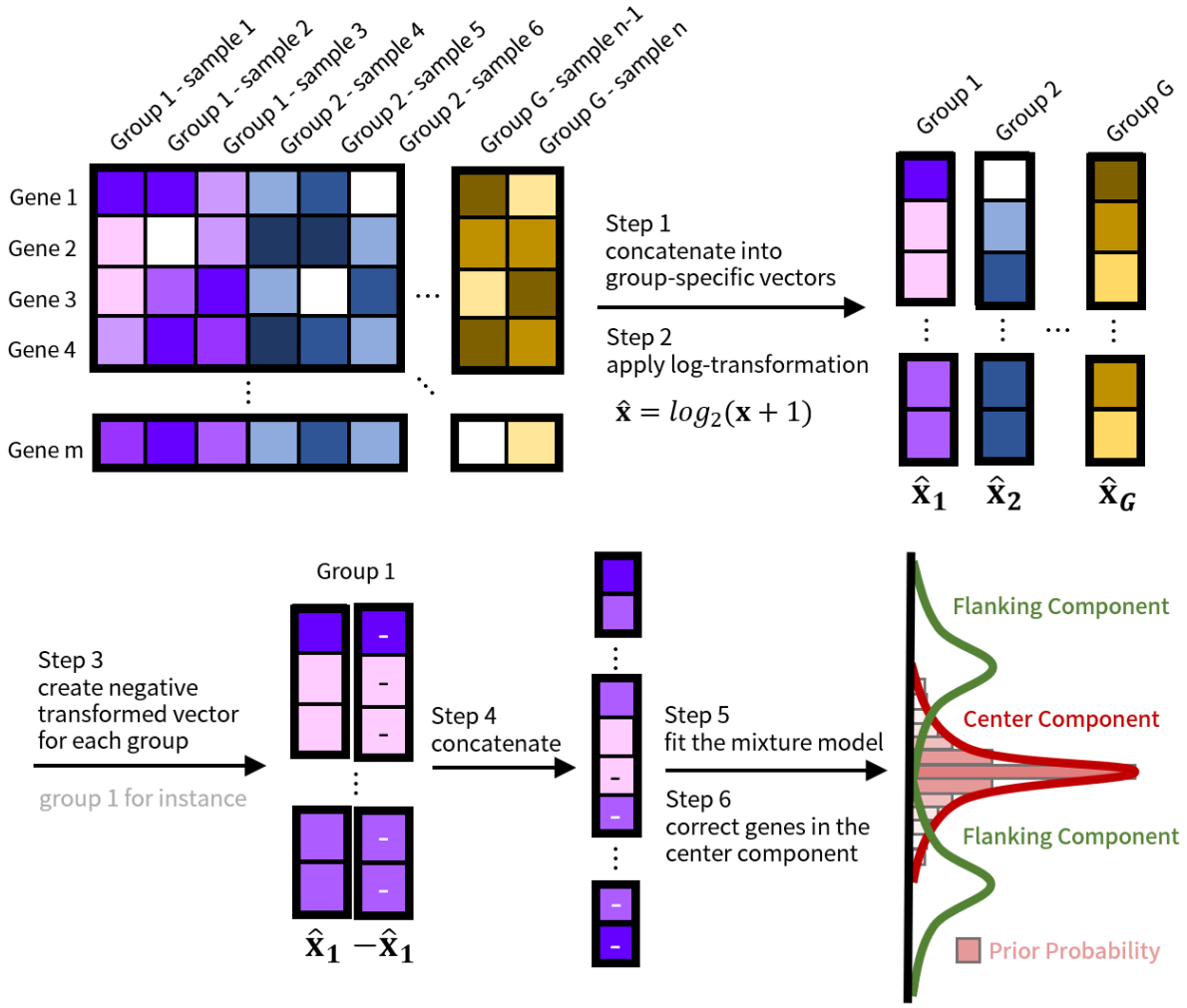


Figure 1: The overall workflow of the CAIMAN algorithm. Step 1: CAIMAN concatenates the two dimensional expression array for samples in the same group into a group-specific vector of expression. Step 2: CAIMAN applies a $\log_2$-transformation with a pseudo count of one. Step 3: CAIMAN applies a negative transformation. Step 4: the method concatenates the two vectors into an augmented group-specific vector of expression profiles. Step 5: the method fits the Gaussian mixture model with the distribution of group-specific expression profile using an EM algorithm. Step 6: CAIMAN corrects the normalized expression levels of genes that form false-positive associations.

constraints when finding the optimal parameters for the mixture component [23], making the approximation process computationally efficient (Supplementary Materials 1, section 6).

Thereafter, the algorithm fits the augmented distribution using a Gaussian mixture model, with a center and flanking components (Figure 1, step 5). The total number of components can be optimized using a likelihood ratio test implemented in the CAIMAN implementation (Supplementary Materials 1, section 3.4). CAIMAN then utilizes an expectation-maximization (EM) algorithm to iteratively estimate the mean and standard deviation for each Gaussian mixture component that maximizes the total likelihood of the model. Using these statistics, CAIMAN predicts the posterior probability of each gene being assigned to each Gaussian mixture component, given the gene's expression level. The center component is then used to model the expression distribution for lowly expressed genes. This component remains zero during the fitting of the expression data. Finally, the expression of genes that have a higher posterior probability of being assigned to the center component, compared to the flanking components, is replaced with zero.

In addition, to ensure computational stability, CAIMAN initializes the means of the flanking components to be symmetrical to zero (Supplementary Materials 1, section 3.2), and makes the absolute values of parameters identical for the positive flanking components and their negative counterpart during the maximization process (Supplementary Materials 1, section 3.3). For a more detailed description of the algorithm, please refer to Supplementary Materials 1, section 3.

3.2 Preparation of GTEx data

We downloaded RNA-Seq count data from the GTEx Consortium V8 release [21]. We selected five tissues previously reported to have the largest number of genes deviating in their expression levels compared to the median expression across all tissues, in particular, *Brain – Cerebellum*, *Lymphoblastoid Cell Lines*, *Liver*, *Testis*, and *Whole Blood* [1]. The resulting dataset consisted of 1,757 samples and 55,878 genes. Combined with the tissue information (*SMTSD*) from the metadata, we then used the Bioconductor package *qsmooth* (version 1.4.0) [20] to perform smooth quantile normalization with default settings. Finally, for the comparison of co-expression levels before and after adjustment with CAIMAN, we applied CAIMAN to the *qsmooth*-normalized data, using the default of two flanking components.

3.3 Preparation of ENCODE data

We used data from ENCODE to validate CAIMAN. We downloaded bulk polyA plus RNA-seq count data consisting of tissues with more than five RNA-Seq experiments (*Embryonic Facial Prominence*, *Forebrain*, *Heart*, *Hindbrain*, *Intestine*, *Kidney*, *Limb*, *Liver*, *Lung*, *Midbrain*, *Neural Tube*, *Stomach*) from the Mouse ENCODE database using Bioconductor package *ENCODExplorer* (version 2.14.0, download date September 11th, 2020) [24]. Among all the available experiments, we extracted those for which External RNA Controls Consortium(ERCC)-only spike-ins (accession: ENCSR884LPM) information was available. The resulting dataset consisted of 126 samples and 43,346 genes.

To establish the ground truth expression, we normalized the read counts by the expression of 96 spike-ins genes (Supplementary Materials 1, section 4). Similar to the preparation of the GTEx dataset (Method 3.2), we used *qsmooth* to perform smooth quantile normalization on the raw read counts with corresponding tissue information of each sample (*biosample_name*) in the metadata. We applied the CAIMAN algorithm using the same parameters described in previous section (Methods 3.2).

3.4 Validation of the CAIMAN algorithm

For both the GTEx and ENCODE datasets, we extracted those genes that were exclusively expressed in one tissue, denoted as *tissue-exclusive genes*. We defined tissue-exclusivity using the following two criteria: (1) the median ground truth expression of the gene is higher than or equal to 10 across all samples from the tissue-of-interest, and (2) the median ground truth expression is lower than or equal to 1 across samples from all other tissues. Note that we used these criteria to facilitate the visualization of the problem. We did not particularly look into the biological insights of these genes in this study.

To showcase the false-positive associations introduced by smooth quantile normalization, we compared the Spearman's rank correlation coefficients for these tissue-exclusive genes with and without adjusting the data with CAIMAN. Since the number of tissue-exclusive genes was significantly higher in $Testis$ compared to the other tissues in the GTEx dataset, we excluded this tissue for better visualization of the problem. Moreover, we excluded tissue-exclusive genes that are expressed in those tissues that had less than five tissue-exclusive genes in total from the analysis. Note that the exclusion of these genes is not necessary when applying CAIMAN. The number of tissue exclusive genes for the two datasets are shown in Supplementary Materials 1, section 5. To evaluate the performance of CAIMAN, we defined two genes to be associated when the absolute value of their Spearman's rank correlation coefficient, based on the ground truth expression in the ENCODE dataset, was higher than or equal to a specific value. We ranged this value from 0.2 to 0.8.

To evaluate whether the false-positive associations introduced by quantile normalization-based methods can propagate through downstream network analysis, we constructed sample-specific networks using Bioconductor package lionessR (version 1.2.0-0) [25] with Spearman's rank correlation coefficients as the network reconstruction function. We modelled these networks based on the *qsmooth*-normalized expression and CAIMAN-adjusted expression data. We then used Bioconductor package limma (version 3.44.1) [26] to identify differences in the t-statistics of the distribution of edge weights measured across all sample-specific networks constructed from the ground truth co-expression, against the distribution of edge weights in the CAIMAN-adjusted and the non-adjusted *qsmooth* co-expression networks. For visualization purposes, we excluded 27 tissue-exclusive genes that did not have gene symbol annotations in biomaRt (version 2.36.1) [27].

3.5 Code availability

We developed CAIMAN as a Python package. The implementation of the CAIMAN algorithm and all of the analyses conducted in this study can be reproduced using the Snakemake workflow management system [28]

from the repository `https://github.com/kuijjerlab/CAIMAN`.

4 Results

4.1 Quantile-based normalization methods can introduce false-positive associations in large-scale heterogeneous datasets

In this section, we demonstrate how quantile-based normalization, and in particular smooth quantile normalization, can introduce false-positive associations between genes. To do so, we will present a case study on co-expression analysis between genes that are exclusively expressed in a specific tissue using RNA-Seq data from the Genotype Tissue-Expression (GTEx) [21] (see Methods 3.4). We visualized co-expression levels between these tissue-exclusive genes across different tissues in Figure 2a.

We would expect to only observe co-expression between genes that are expressed in the same tissue and not between genes that are exclusively expressed in different tissues. However, while we do observe high co-expression levels between tissue-exclusive genes within the same tissue when calculating co-expression levels based on the smooth quantile normalized data, relatively high levels of co-expression are also observed between genes that are exclusively expressed in different tissues, in particular between in *Whole Blood*, *Lymphoblastoid Cell Lines (LCL)* and *Liver* (Figure 2a).

To understand how these false-positive associations arise, we dissected the smooth quantile normalization algorithm as it is implemented in the Bioconductor package *qsmooth* [20]. Smooth quantile normalization computes the mean expression level in each quantile, considering only the samples within a given group, as well as the mean expression level in each quantile, considering all samples. Finally, *qsmooth*-normalized expression is defined as the weighted average of the means derived from the given group and derived from the means of the corresponding quantile considering all samples.

One important detail to consider here is the ranking method used to process genes that have the same read count (tied counts) in each sample. The quantiles corresponding to these genes are dependent on the number of genes that have the same read count in that specific sample. Therefore, even if a gene would have the exact same read count in two different samples, the quantiles assigned to these can be drastically different. Especially for zero-inflated RNA-Seq data in heterogeneous datasets that have large differences between the smallest and the largest number of non-expressed genes across samples (Supplementary Materials 1, section 1), lowly expressed genes could, for example, share the same quantile with non-expressed genes in different samples. As the normalized expression levels are dependent on the quantiles of the expression distribution for each sample, this can introduce small technical variability across samples and this can lead to false-positive associations between the normalized expression of these lowly-expressed genes (Supplementary Materials 1, section 1 and 2).

(a) Smooth quantile normalized

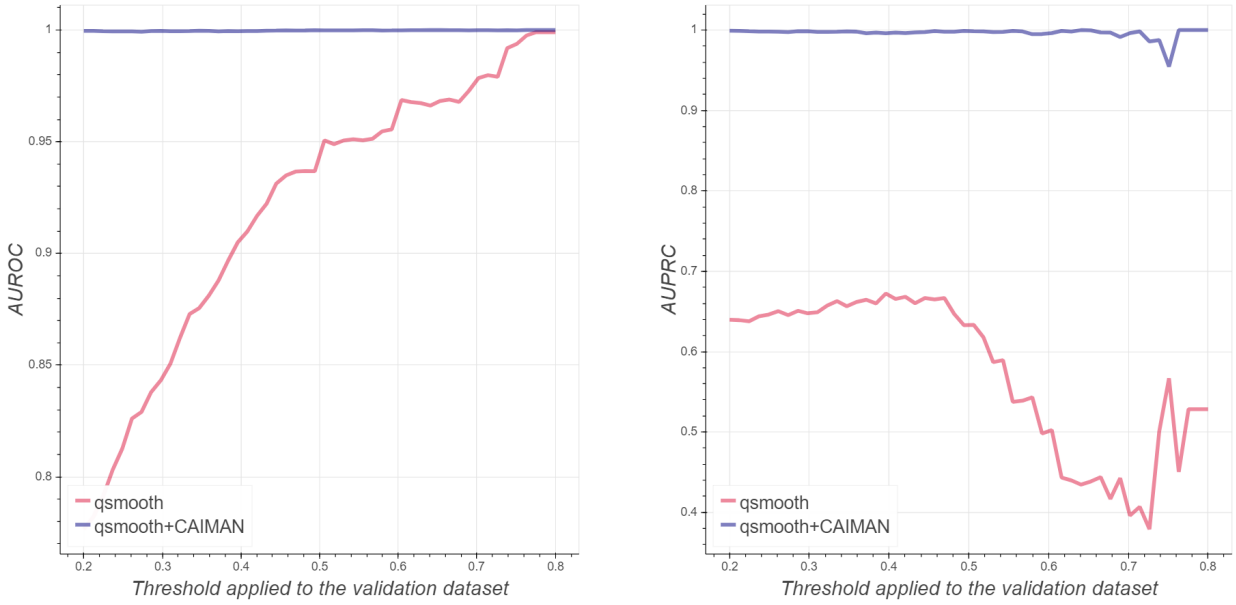
(b) After CAIMAN adjustment

Figure 2: Spearman's rank correlation coefficients between tissue-exclusive gene pairs, based on smooth quantile normalized data from GTEx. (a) False-positive associations are detected between tissue-exclusive genes from different tissues, as can be seen in off-diagonal blocks of co-expression. (b) After applying CAIMAN to the normalized data, most false-positive associations are removed.

4.2 Reducing false-positive co-expression levels with CAIMAN

We want to be able to take advantage of quantile-based normalization methods, and in specific smooth quantile normalization, so that we can retain global expression differences in the data. However, we also need to ensure the identified co-expression signals are reliable. We therefore developed a new algorithm that can remove these false-positive associations by correcting the expression levels of lowly expressed genes that might be obscured by the technical variability introduced by quantile-based normalization methods (see Methods 3.1 and Supplementary Materials 1, sections 1 and 2).

We applied CAIMAN to adjust gene expression levels in the GTEx data and then repeated the co-expression analysis described in the previous section (see Methods 3.4). For each tissue, the fitted distributions of the expression values and their posterior probabilities to the center component obtained with CAIMAN are shown in Supplementary Materials 2. From these probabilities, we found that different thresholds need to be applied for different tissues or groups, to identify those genes that might form false-positive associations. CAIMAN automatically infers the thresholds that best fit the expression profile in each group. Comparing the Spearman's rank correlation coefficients obtained in the corrected with those obtained in the uncorrected data, we find that CAIMAN is capable of removing false-positive associations introduced by smooth quantile normalization (Figure 2b).



(a) Area under receiver-operator curve.

(b) Area under precision-recall curve.

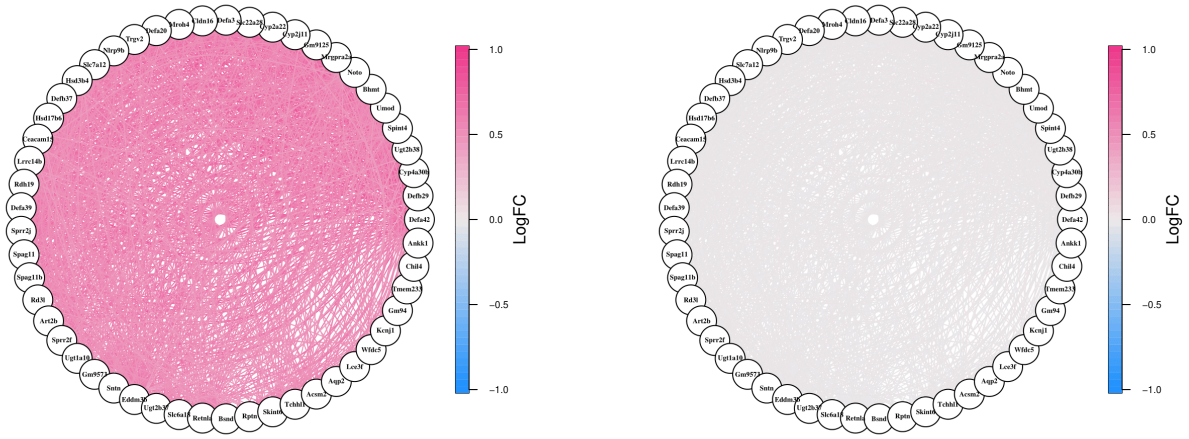
Figure 3: CAIMAN can effectively reduce the number of false-positive associations in co-expression analysis. The x-axis denotes the thresholds of absolute Spearman's rank correlation coefficient based on ground truth expression that define true associations between genes, while the y-axis corresponds to the area under the Receiver Operator Curve (AUROC) and Precision Recall Curve (AUPRC).

To provide more evidence supporting the capability of CAIMAN to reduce false-positive associations introduced by normalization methods, we applied our correction method to RNA-Seq data from the Mouse ENCODE database, which includes spike-ins (Methods 3.3). We used the expression of spike-ins genes to normalize the read count (Supplementary Materials 1, section 4), creating the ground truth expression. Comparing with Spearman's rank correlation coefficients derived from the ground truth expression for every gene pair (see Method 3.4), we observed that the root mean square error (RMSE) between the correlation coefficients decreases from 0.389 to 0.015 after applying CAIMAN.

We next defined two genes to be associated if the ground truth Spearman's rank correlation coefficients exceeded a certain threshold, which we ranged from 0.2 to 0.8. We conducted receiver-operator curve (ROC) and precision-recall curve (PRC) analyses and reported the area under the two curves. We found that CAIMAN can reduce the false-discovery rate in co-expression analysis, regardless of the strength of the correlation signal (Figure 3). Note that when the true association is more strictly defined (correlation coefficient above 0.70), the small number of positive associations (28) causes a fluctuation in the AUPRC.

4.3 CAIMAN reduces false-positives in downstream network analysis

We next wanted to evaluate whether the false-positive associations introduced by quantile-normalized methods also affect downstream network analysis. To explore this, we used the LIONESS algorithm [29] to infer sample-specific networks from co-expression measurements. We compared the distribution of edge weights across all sample-specific networks constructed on ground truth co-expression to the distribution of edge weights from the co-expression networks constructed on (a) smooth quantile normalized and on (b) CAIMAN adjusted expression, using a t-test. Figure 4a shows that the false-positive associations propagate through downstream network analysis, creating 631 false-positive edges from a total of 1,484 potential gene pairs between genes exclusively expressed in different tissues. After adjusting the *qsmooth*-normalized input data with CAIMAN, no edges significantly differ from the network built on the ground truth expression (FDR adjusted P-value ≤ 0.05) (Figure 4b). This analysis shows that CAIMAN can effectively remove false-positive associations between genes, and can thus improve the reliability of downstream network analysis.



(a) Smooth quantile normalized

(b) After CAIMAN adjustment

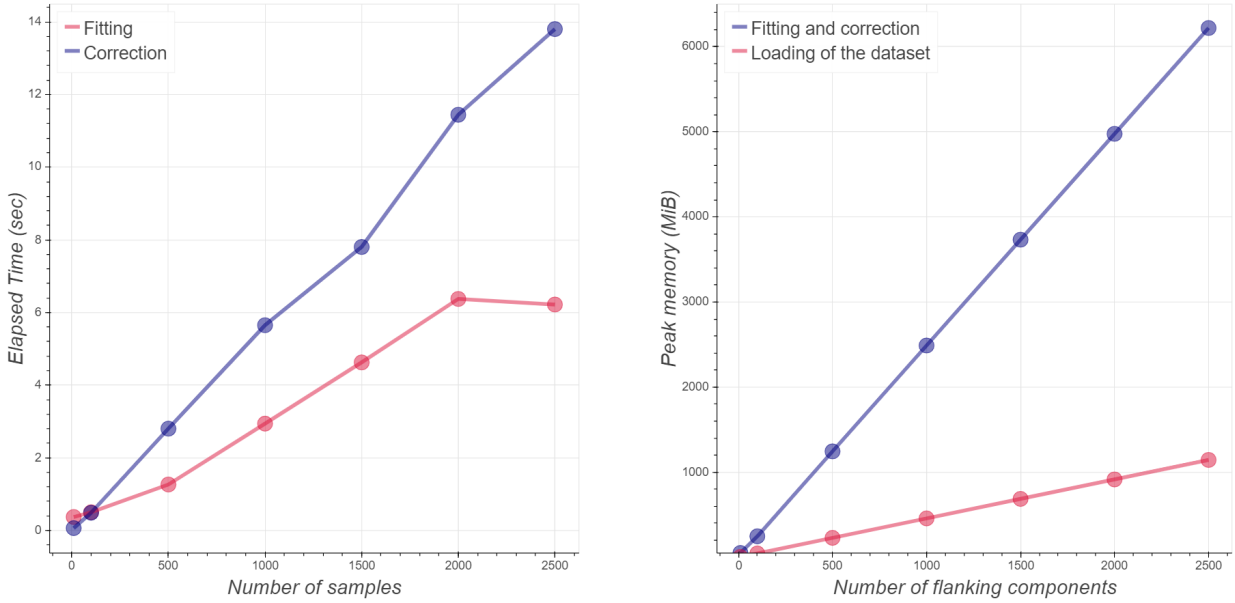
Figure 4: Log-transformed fold changes on the mean edge weights comparing the network constructed on ground truth expression with the networks constructed on (a) the smooth quantile normalized expression and on (b) the CAIMAN-adjusted expression levels. All edges with FDR adjusted P-value ≤ 0.05 are shown. The log-transformed fold change between the edge weight is color-coded with magenta indicating over-estimation and green indicating under-estimation of the edge weight.

4.4 Time complexity and memory usage of CAIMAN

There are two parameters in the CAIMAN model that affect the time complexity—the number of iterations to reach convergence I and the number of flanking components K in the Gaussian mixture model. Consider the expression dataset consisting of M genes, G groups, and a total number of N samples. The time complexity

to fit the model is then $O(IKGM)$. Generally, the number of iterations I and the number of flanking components K is significantly smaller than the number of genes (default values: $I = 10, K = 2$), and these can thus be considered constants. Since the adjustment of normalized expression takes $O(MN)$ to complete, the overall time complexity is $O(MN)$. To evaluate the computational resources that are required to run CAIMAN, we created pseudo datasets by sampling from the ENCODE dataset with different numbers of samples. The computational time for datasets with different sample sizes is shown in Figure 5a.

In addition to loading the dataset into memory, CAIMAN requires memory to store statistics such as the posterior probabilities and log-likelihoods for the expression level of each gene. For the memory usage of CAIMAN, loading of the ENCODE dataset took 599.68Mib on a laptop with Intel Core i7-8750H processor (in single thread) with 32Gib of memory. Figure 5b shows that the memory usage increases linearly when applying on datasets with more samples. On the other hand, CAIMAN applies a sampling procedure to reduce computational time required to fit the Gaussian mixture model. This makes CAIMAN scalable to datasets with large number of samples (see Supplementary Material 1, section 3.1).



(a) Time complexity

(b) Memory usage

Figure 5: The computational time and memory required for fitting and correction for CAIMAN algorithm. Analyses were run on a laptop with Intel Core i7-8750H processor (in single thread) with 32Gib of memory.

5 Discussion

Here, we showed that the use of quantile-based normalization approaches—and in specific smooth quantile normalization—to RNA-Seq data can introduce false-positive associations between genes, and that this can propagate to and affect downstream network analyses. We found that false-positive associations particularly

arise when there is a large difference between the smallest and the largest number of non-expressed genes across the samples in the dataset. This can for example occur when dealing with RNA-Seq datasets collected from large-scale projects that include heterogeneous data, such as data from The Cancer Genome Atlas (TCGA) [30], ENCODE, and GTEx.

A frequently applied strategy that attempts to remove potential false-positive associations is filtering out genes with low read counts across a certain number of samples. However, thresholds used for filtering are often chosen arbitrarily, and can remove genes that are specifically expressed in a subset of samples, such as the tissue-exclusive genes that we described in our example of network analysis in data from GTEx. Therefore, arbitrary filtering is not ideal if one aims to compare gene associations or networks derived from different subgroups of samples. Moreover, it would be ideal to include all genes in large-scale network analysis, as certain network reconstruction algorithms make use of the entire distribution of gene expression and thus filtering out genes may remove some signal from the input dataset [6].

To address this issue, we developed CAIMAN, a method that uses an Expectation-Maximization based approach to fit normalized expression data to a Gaussian mixture model. CAIMAN then adaptively identifies lowly expressed genes that might cause false-positive associations. This means that no arbitrary threshold needs to be defined. In addition, without having to first identify the specific genes to correct for, CAIMAN does not exclude any gene from the analysis, preserving the potential associations between genes that only expressed in small subgroups of samples and that would otherwise not be detected. Apart from reducing the false-discovery rate in gene associations, CAIMAN provides an additional feature to model the posterior probability of non-expressed genes, given their expression distribution across groups of samples. Thus, the method can be used to determine whether genes with low counts are actually expressed in the cell.

CAIMAN utilizes the posterior probability of each genes being assigned to the center component to identify potential targets that might be affected by quantile-based normalization approaches. One limitation of this is that the method cannot correct for such false-positive associations if the expression distribution is best fitted without any flanking components. However, this is unlikely to happen in co-expression analysis, as expression measured using RNA-Seq usually shows high levels of dispersion.

While we specifically focused our examples on modeling co-expression across different tissues in this work, the issue of false-positive associations can also arise when comparing other biological conditions that show large differences in expression profiles under certain experimental settings, such as when comparing networks for males and females [31]. We also envision that other methods that are based on correlations, such as eQTL studies, could potentially include quantile-based normalization-introduced false-positives. We would thus recommend using CAIMAN as a default step in adjusting data from normalized large-scale heterogeneous datasets.

Finally, heterogeneous datasets with increasing numbers of samples and conditions will likely be published in the near future, and new methods for combining data from different studies [32] will result in the emergence of even larger and more heterogeneous datasets. As these datasets will become available for anal-

ysis, CAIMAN will be an important tool that will allow for more precise analyses of large-scale data with network-based approaches.

6 Funding

PHH and MLK are supported by the Norwegian Research Council, Helse Sør-Øst, and University of Oslo through the Centre for Molecular Medicine Norway (NCMM). KG is supported by a grant (K25HL133599) from the US National Heart, Lung, and Blood Institute.

7 Acknowledgments

The authors would like to thank the Kuijjer and Mathelier groups for helpful discussions and Elisa Bjørgø and Ingrid Kjelsvik for administrative support.

8 Author contributions

Conceptualization, MLK, CML, KG; Methodology, PHH, MLK; Software, PHH; Formal Analysis, PHH; Writing — Original Draft, PHH, CML, MLK; Writing — Review Editing, CML, GKS, KG, MKK; Supervision, MKK; Funding Acquisition, KG, MKK.

References

[1] Sonawane, A. R., Platig, J., Fagny, M., Chen, C.-Y., Paulson, J. N., Lopes-Ramos, C. M., DeMeo, D. L., Quackenbush, J., Glass, K., and Kuijjer, M. L. (2017) Understanding tissue-specific gene regulation. *Cell reports,* **21**(4), 1077–1088.

[2] Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017) An expanded view of complex traits: from polygenic to omnigenic. *Cell,* **169**(7), 1177–1186.

[3] Marco, A., Konikoff, C., Karr, T. L., and Kumar, S. (2009) Relationship between gene co-expression and sharing of transcription factor binding sites in Drosophila melanogaster. *Bioinformatics,* **25**(19), 2473–2477.

[4] Gu, Q., Nagaraj, S. H., Hudson, N. J., Dalrymple, B. P., and Reverter, A. (2011) Genome-wide patterns of promoter sharing and co-expression in bovine skeletal muscle. *BMC genomics,* **12**(1), 23.

[5] Guo, Y., Alexander, K., Clark, A. G., Grimson, A., and Yu, H. (2016) Integrated network analysis reveals distinct regulatory roles of transcription factors and microRNAs. *RNA,* **22**(11), 1663–1672.

[6] Glass, K., Huttenhower, C., Quackenbush, J., and Yuan, G.-C. (2013) Passing messages between biological networks to refine predicted interactions. *PloS one,* **8**(5), e64832.

[7] Petralia, F., Wang, P., Yang, J., and Tu, Z. (2015) Integrative random forest for gene regulatory network inference. *Bioinformatics,* **31**(12), i197–i205.

[8] Reiss, D. J., Plaisier, C. L., Wu, W.-J., and Baliga, N. S. (2015) cMonkey2: Automated, systematic, integrated detection of co-regulated gene modules for any organism. *Nucleic acids research,* **43**(13), e87–e87.

[9] Nicolle, R., Radvanyi, F., and Elati, M. (2015) CoRegNet: reconstruction and integrated analysis of co-regulatory networks. *Bioinformatics,* **31**(18), 3066–3068.

[10] Kuijjer, M. L., Fagny, M., Marin, A., Quackenbush, J., and Glass, K. (2020) PUMA: PANDA Using MicroRNA Associations. *Bioinformatics,* **36**(18), 4765–4773.

[11] Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics,* **9**(1), 559.

[12] Meyer, P. E., Kontos, K., Lafitte, F., and Bontempi, G. (2007) Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics and systems biology,* **2007**, 1–9.

[13] Lachmann, A., Giorgi, F. M., Lopez, G., and Califano, A. (2016) ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics,* **32**(14), 2233–2235.

[14] Irrthum, A., Wehenkel, L., Geurts, P., et al. (2010) Inferring regulatory networks from expression data using tree-based methods. *PloS one,* **5**(9), e12776.

[15] Haury, A.-C., Mordelet, F., Vera-Licona, P., and Vert, J.-P. (2012) TIGRESS: trustful inference of gene regulation using stability selection. *BMC systems biology,* **6**(1), 145.

[16] Silverman, E. K., Schmidt, H. H., Anastasiadou, E., Altucci, L., Angelini, M., Badimon, L., Balligand, J.-L., Benincasa, G., Capasso, G., Conte, F., et al. (2020) Molecular networks in Network Medicine: Development and applications. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine,* p. e1489.

[17] Robinson, M. D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology,* **11**(3), 1–9.

[18] Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Nature Precedings,* pp. 1–1.

[19] Evans, C., Hardin, J., and Stoebel, D. M. (2018) Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in bioinformatics,* **19**(5), 776–792.

[20] Hicks, S. C., Okrah, K., Paulson, J. N., Quackenbush, J., Irizarry, R. A., and Bravo, H. C. (2018) Smooth quantile normalization. *Biostatistics,* **19**(2), 185–198.

[21] Consortium, G. et al. (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science,* **348**(6235), 648–660.

[22] Stamatoyannopoulos, J. A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D. M., Groudine, M., Bender, M., Kaul, R., Canfield, T., et al. (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome biology,* **13**(8), 1–5.

[23] Yin, S., Wang, X., Jia, G., and Xie, Y. (2020) MIXnorm: normalizing RNA-seq data from formalin-fixed paraffin-embedded samples. *Bioinformatics,* **36**(11), 3401–3408.

[24] Beauparlant, C., Lemacon, A., and Droit, A. (2015) ENCODExplorer: a compilation of ENCODE metadata. *R package version,* **1**(3).

[25] Kuijjer, M. L., Hsieh, P.-H., Quackenbush, J., and Glass, K. (2019) lionessR: single sample network inference in R. *BMC cancer,* **19**(1), 1003.

[26] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research,* **43**(7), e47–e47.

[27] Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols,* **4**(8), 1184.

[28] Köster, J. and Rahmann, S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics,* **28**(19), 2520–2522.

[29] Kuijjer, M. L., Tung, M. G., Yuan, G., Quackenbush, J., and Glass, K. (2019) Estimating sample-specific regulatory networks. *Iscience,* **14**, 226–240.

[30] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. (2013) The cancer genome atlas pan-cancer analysis project. *Nature genetics,* **45**(10), 1113.

[31] Lopes-Ramos, C. M., Chen, C.-Y., Kuijjer, M. L., Paulson, J. N., Sonawane, A. R., Fagny, M., Platig, J., Glass, K., Quackenbush, J., and DeMeo, D. L. (2020) Sex Differences in Gene Expression and Regulatory Networks across 29 Human Tissues. *Cell reports,* **31**(12), 107795.

[32] Collado-Torres, L., Nellore, A., Kammers, K., Ellis, S. E., Taub, M. A., Hansen, K. D., Jaffe, A. E., Langmead, B., and Leek, J. T. (2017) Reproducible RNA-seq analysis using recount2. *Nature biotechnology,* **35**(4), 319–321.