PLOS ONE

RESEARCH ARTICLE

Evaluation of critical data processing steps for reliable prediction of gene co-expression from large collections of RNA-seq data

Alexis Vandenbon[1,2]*

1 Institute for Frontier Life and Medical Sciences, Kyoto University, Kyoto, Japan, **2** Institute for Liberal Arts and Sciences, Kyoto University, Kyoto, Japan

* alexisvdb@infront.kyoto-u.ac.jp

OPEN ACCESS

Citation: Vandenbon A (2022) Evaluation of critical data processing steps for reliable prediction of gene co-expression from large collections of RNA-seq data. PLoS ONE 17(1): e0263344. https://doi.org/10.1371/journal.pone.0263344

Editor: Y-h. Taguchi, Chuo University, JAPAN

Received: September 22, 2021

Accepted: January 16, 2022

Published: January 28, 2022

Data Availability Statement: Both the human and the mouse RNA-seq datasets are available in figshare with DOIs doi.org/10.6084/m9.figshare.14178446.v1 and doi.org/10.6084/m9.figshare.14178425.v1. These datasets include 1) raw read counts of genes in all RNA-seq samples, 2) the same RNA-seq data after UQ normalization and batch effect correction using ComBat, which is in general the best workflow according to our study, 3) annotation data assigning a study ID and cell type or tissue to each sample, and 4) A list of each cell type or tissue included in the dataset along with its sample count. Code used in this study is available in a GitHub repository (https://github.

Abstract

Motivation

Gene co-expression analysis is an attractive tool for leveraging enormous amounts of public RNA-seq datasets for the prediction of gene functions and regulatory mechanisms. However, the optimal data processing steps for the accurate prediction of gene co-expression from such large datasets remain unclear. Especially the importance of batch effect correction is understudied.

Results

We processed RNA-seq data of 68 human and 76 mouse cell types and tissues using 50 different workflows into 7,200 genome-wide gene co-expression networks. We then conducted a systematic analysis of the factors that result in high-quality co-expression predictions, focusing on normalization, batch effect correction, and measure of correlation. We confirmed the key importance of high sample counts for high-quality predictions. However, choosing a suitable normalization approach and applying batch effect correction can further improve the quality of co-expression estimates, equivalent to a >80% and >40% increase in samples. In larger datasets, batch effect removal was equivalent to a more than doubling of the sample size. Finally, Pearson correlation appears more suitable than Spearman correlation, except for smaller datasets.

Conclusion

A key point for accurate prediction of gene co-expression is the collection of many samples. However, paying attention to data normalization, batch effects, and the measure of correlation can significantly improve the quality of co-expression estimates.

com/alexisvdb/rnaseq_coexpression), including an example for normalization, batch effect correction, calculation of correlation, GO term and TFBS analysis, as well as the code used for the linear regression and making the plots for the manuscript.

Funding: A.V. received a KAKENHI Grant-in-Aid for Scientific Research (C) (20K06609) from the Japan Society for the Promotion of Science (JSPS; https://www.jsps.go.jp/english/). 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.

Introduction

Understanding the functions and regulatory mechanisms of genes is one of the central challenges in biology. Gene co-expression is an important concept in bioinformatics because it serves as a foundation for predicting gene functions and regulatory mechanisms, and for more complex network inference methods [1–6]. Several gene co-expression databases have been developed [7–10].

High numbers of samples are needed to accurately infer correlation of expression [11]. Public databases are attractive sources of expression data, but in practice high numbers of samples can only be obtained by aggregating data from different studies conducted by different laboratories. As a result, the input data for gene co-expression analysis often contains considerable technical variability, called batch effects [9,12].

In a recent study, we showed that correcting batch effects improved the quality of gene co-expression estimates significantly [9]. However, our previous study was limited to microarray data, and considered only one data normalization method and one batch correction method, i.e. ComBat [13]. Moreover, other studies have shown that treating batch effects can also result in unwanted artifacts such as exaggerated differences between covariates in gene expression and DNA methylation data [14–17].

Here, we present a systematic analysis of the effects of RNA-seq data normalization, batch effect correction, and correlation measure on the quality of gene co-expression estimates. We applied 50 data processing workflows on data for 68 human and 76 mouse cell types and tissues, resulting in 7,200 sets of genome-wide gene-gene co-expression predictions. Through analysis of the quality of these cell type- and tissue-specific co-expression predictions, we confirmed the importance of large numbers of samples [11]. We also found that some normalization methods (especially UQ normalization) resulted on average in better co-expression predictions than others. In addition, treating batch effects resulted in a significant improvement of the co-expression estimates, especially in larger datasets consisting of samples produced by many different studies. It is imperative that future studies pay attention to batch effects in order to make optimal use of large amounts of public data. Finally, the difference between Pearson's correlation and Spearman's correlation was small, with Spearman working better in small datasets and Pearson better in medium-sized datasets. To the best of our knowledge, this is the first comprehensive study evaluating the importance of batch effect correction for the prediction of gene co-expression from large collections of RNA-seq data.

Results

Overview of this study

The goal of this study is to gain insights into which data processing steps are preferable for obtaining high-quality gene co-expression estimates from large collections of RNA-seq data. To address this issue, we collected a dataset of 8,796 human and 12,114 mouse bulk RNA-seq samples, from 401 and 630 studies, covering 68 human and 76 mouse cell types and tissues (see Methods; S1 and S2 Tables) [18]. On these two datasets, we applied combinations of data normalization approaches and batch effect correction approaches (see Fig 1 for a summary of the workflow). As proxies for batches we used the studies that produced each sample (1 study is 1 batch). We also applied the method ComBat-seq on the raw read count data without any prior normalization [19]. The resulting 25 (6 normalizations x 4 batch effect correction approaches, and ComBat-seq without normalization) human and 25 mouse datasets were used to estimate correlation of expression using Pearson's correlation and Spearman's correlation in the data of each cell type or tissue. This resulted in a total of 7,200 (3,400 human and 3,800

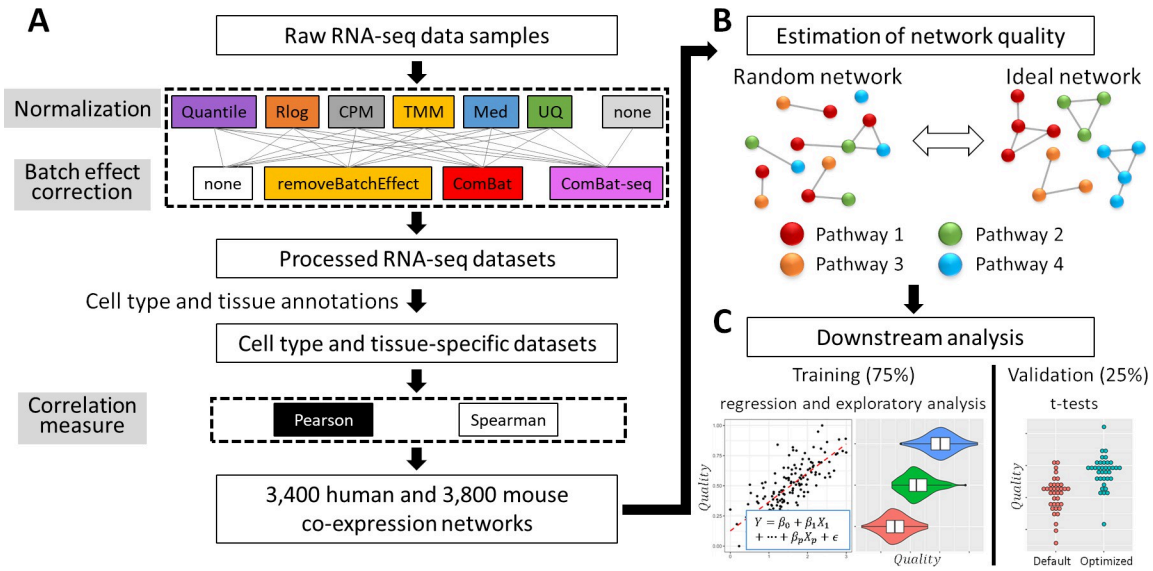

Fig 1. Summary of this study. (A) Raw RNA-seq data was processed with 50 different combinations of normalization, batch effect correction, and correlation measures into 7,200 genome-wide sets of cell type and tissue-specific co-expression predictions, which we refer to as "co-expression networks". **(B)** Quality of co-expression networks was estimated based on the enrichment of functional annotations of correlated genes and regulatory motifs in their promoters. In random co-expression networks no common annotations and motifs are expected to be found among correlated genes. In contrast, in ideal networks such enrichments should be encountered frequently. Here nodes represent genes and edges co-expression. **(C)** Quality measures were processed into 7,200 quality scores, which were used for downstream analysis. We use 75% of cell types and tissues for regression and exploratory analysis, and the remaining 25% for validation.

https://doi.org/10.1371/journal.pone.0263344.g001

mouse) genome-wide sets of cell type or tissue-specific gene co-expression predictions. We will refer to each genome-wide set of cell type or tissue-specific gene co-expression predictions as a "co-expression network". However, the goal of this study is not to analyze network topology. Our focus is to identify the key features that result in accurate co-expression predictions.

Defining the quality of co-expression predictions

Next, we evaluated the quality of the co-expression predictions produced by each workflow. Many studies have used enrichment of shared functional annotations among correlated genes or regulatory DNA motifs in their promoter sequences as quality measures for co-expression predictions [10,20–22]. In high-quality co-expression networks, we expect correlated genes to belong to shared pathways or to be controlled by a common regulatory mechanism (Fig 1B). In contrast, in a randomly generated network, correlated genes are expected to lack common functions or regulatory mechanisms. In this study, in each co-expression network, for every gene X, we extracted the set of 100 genes with the highest correlation, which we refer to as set_X. We then defined eight quality measures that are based on how frequently we observed significant enrichment of Gene Ontology (GO) functional annotations and transcription factor binding site (TFBS) motifs among these sets of 100 genes (Table 1, see Methods for more details). In high-quality networks this frequency should be high, and in low-quality networks it should be low. Although the eight quality measures were based on sets of 100 highly correlated genes, using instead the top 50 or top 200 genes resulted in highly consistent quality estimates (S1 Fig).

We collected the eight quality measures of the 7,200 co-expression networks and Principal Component Analysis (PCA) revealed that they are highly consistent and correlated: the first PC explained 81.4% of variability in the quality measures (S2A Fig), and had a high correlation with all eight measures (range 0.77 to 0.96; S2B and S2C Fig). In contrast, the second PC

Table 1. Overview of the eight quality measures used to define the quality of genome-wide gene co-expression networks.

Quality measure	Definition
$Enrichment_{MF}$	fraction of genes for which set_X has one or more enriched GO terms of the domain Molecular Function.
$Enrichment_{BP}$	fraction of genes for which set_X has one or more enriched GO terms of the domain Biological Process.
$Enrichment_{CC}$	fraction of genes for which set_X has one or more enriched GO terms of the domain Cellular Component.
$Enrichment_{TFBS}$	fraction of genes for which the promoters of set_X have one or more significantly enriched TFBS motifs.
$Accuracy_{MF}$	fraction of genes for which known annotations of the domain Molecular Function overlapped with enriched GO terms of set_X.
$Accuracy_{BP}$	fraction of genes for which known annotations of the domain Biological Process overlapped with enriched GO terms of set_X.
$Accuracy_{CC}$	fraction of genes for which known annotations of the domain Cellular Component overlapped with enriched GO terms of set_X.
$Accuracy_{TFBS}$	fraction of genes for which predicted TFBSs in their promoter sequence overlapped with enriched TFBS motifs in the promoter sequences of set_X.

set_X refers to the set of 100 genes with the highest correlation of expression with each gene X in a given co-expression network.

https://doi.org/10.1371/journal.pone.0263344.t001

accounted for only 12.1% of the variance in the data and was did not show consistent correlation with all measures (S2A, S2B and S2D Fig). To facilitate the downstream analysis, we therefore decided to use this first PC as a general quality score (*Quality*, see Methods) after rescaling it to the range 0 (worst networks) to 1 (best networks) (S3A Fig). As an illustration, S3B Fig shows the quality measures for the networks with the lowest (Med + ComBat-seq + Spearman applied on human salivary gland data), 50th percentile (Rlog + ComBat + Pearson applied on human neuron data), and highest (UQ + removeBatchEffect + Spearman applied on mouse liver data) *Quality*. From the lowest-quality network to the highest-quality network, the measures of quality are progressively increasing.

At this point we randomly split the 68 human and 76 mouse cell types and tissues into 2 groups (S1 and S2 Tables). One group (51 human and 57 mouse cell types and tissues; corresponding to 75% of the datasets) will be used for the analysis of features that contribute to high-quality gene co-expression predictions. We refer to this as our training set, and will focus on it in the following sections. The remaining 25% of cell types and tissues (17 human and 19 mouse cell types and tissues) will be used as an independent validation set later (section "The best workflows result in significantly better co-expression estimates").

Fig 2 shows the 50 workflows that we examined, sorted by the average *Quality* of the 108 (51 human and 57 mouse cell types and tissues in the training set) networks that they each resulted in. We observed that the top four workflows used UQ normalization, while Quantile normalization resulted in low average quality. Similarly, the top 15 workflows all include a batch correction step, while many of the worst-performing workflows did not treat batch effects. The top-ranking workflow (UQ + ComBat + Pearson) resulted in an above-average network for 102 (94%) of the 108 training datasets, and for 135 (94%) of all 144 datasets (S4 Fig).

Modeling the quality of co-expression networks

To gain more quantitative insights into what factors contribute to high-quality co-expression estimates, we performed linear regression on the *Quality* scores using as predictors: 1) the

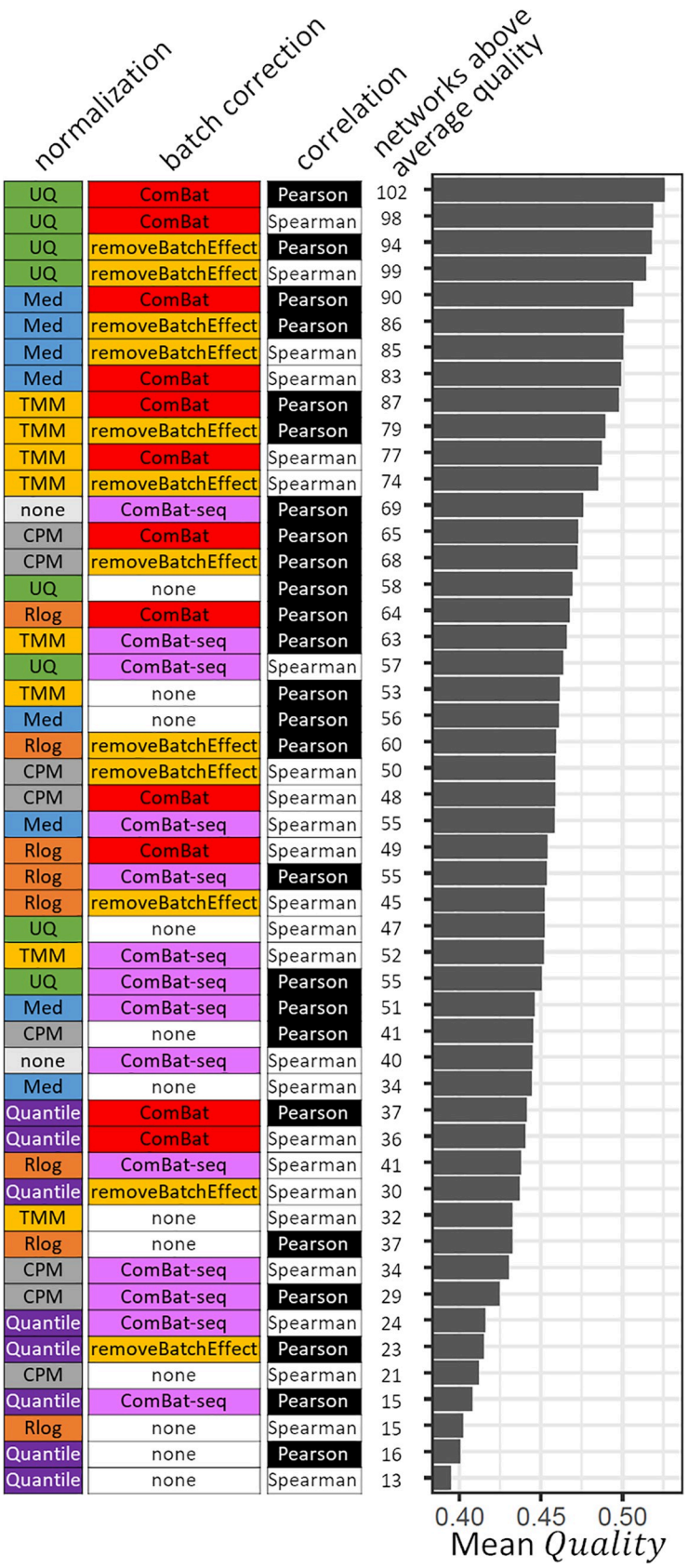
normalization
batch correction
correlation
networks above average quality
UQ ComBat Pearson 102
UQ ComBat Spearman 98
UQ removeBatchEffect Pearson 94
UQ removeBatchEffect Spearman 99
Med ComBat Pearson 90
Med removeBatchEffect Pearson 86
Med removeBatchEffect Spearman 85
Med ComBat Spearman 83
TMM ComBat Pearson 87
TMM removeBatchEffect Pearson 79
TMM ComBat Spearman 77
TMM removeBatchEffect Spearman 74
none ComBat-seq Pearson 69
CPM ComBat Pearson 65
CPM removeBatchEffect Pearson 68
UQ none Pearson 58
Rlog ComBat Pearson 64
TMM ComBat-seq Pearson 63
UQ ComBat-seq Spearman 57
TMM none Pearson 53
Med none Pearson 56
Rlog removeBatchEffect Pearson 60
CPM removeBatchEffect Spearman 50
CPM ComBat Spearman 48
Med ComBat-seq Spearman 55
Rlog ComBat Spearman 49
Rlog ComBat-seq Pearson 55
Rlog removeBatchEffect Spearman 45
UQ none Spearman 47
TMM ComBat-seq Spearman 52
UQ ComBat-seq Pearson 55
Med ComBat-seq Pearson 51
CPM none Pearson 41
none ComBat-seq Spearman 40
Med none Spearman 34
Quantile ComBat Pearson 37
Quantile ComBat Spearman 36
Rlog ComBat-seq Spearman 41
Quantile removeBatchEffect Spearman 30
TMM none Spearman 32
Rlog none Pearson 37
CPM ComBat-seq Spearman 34
CPM ComBat-seq Pearson 29
Quantile ComBat-seq Spearman 24
Quantile removeBatchEffect Pearson 23
CPM none Spearman 21
Quantile ComBat-seq Pearson 15
Rlog none Spearman 15
Quantile none Pearson 16
Quantile none Spearman 13
0.40
0.45
0.50
Mean Quality

Fig 2. Evaluating the quality of co-expression networks. All 50 workflows are shown in order of decreasing average quality of the networks they produced. From left to right are shown: Normalization method, batch effect correction method, and measure of correlation used in each workflow. Next, the number of training datasets (108 in total) in which the workflow resulted in an above-average quality network is shown, and the mean *Quality* of the 108 networks generated using each workflow.

https://doi.org/10.1371/journal.pone.0263344.g002

number of RNA-seq samples which the network was based on ($\log_{10}$ values), 2) the number of batches in the data ($\log_{10}$ values), 3) the species (human or mouse), 4) the data normalization approach, 5) batch correction approach, and 6) the correlation measure. In the next sections, we will focus on workflows that did not include ComBat-seq. ComBat-seq differs from ComBat and RemoveBatchEffect in that it takes integers as input and therefore cannot be used on data that has already been normalized. Networks generated using ComBat-seq will be treated separately in section "ComBat-seq results in lower-quality networks".

The resulting linear model is summarized in Table 2. Despite its simplicity, this model explains 55% of the variability in *Quality* ($R^2 = 0.55$) in the training datasets. The reliability of estimated coefficients was confirmed by 4-fold cross validation (CV), each time leaving out 25% of the cell types and tissues and repeating the same linear regression (S3 Table). In each of the four models, the signs and relative magnitudes of coefficients was consistent. For example, in each case, the coefficient of the number batches was negative, and the ordering of the coefficients of normalizations methods was the same. Below we discuss the roles of sample counts, data normalization, batch effect correction, and correlation measures in more detail.

The importance of sample counts

The most significant predictor for the quality of co-expression estimates was the number of samples they were based on (Table 2). *Quality* follows a roughly linear trend with the logarithm of the sample count (Fig 3A). This is consistent with a previous study [11]. At the same

Table 2. Linear regression analysis of co-expression network quality scores.

	Feature	Estimate	Std. Error	t value	Pr(>|t|)
	(Intercept)	-0.150	0.011	-13.9	5.1E-43
	log10(sample count)	0.2894	0.0072	40.1	8.5E-295
	log10(batch count)	-0.0302	0.0081	-3.7	0.00019
species	**human**	*baseline*			
	mouse	0.0462	0.0035	13.3	3.0E-39
normalization	**Quantile**	*baseline*			
	Rlog	0.0231	0.0060	3.9	0.00012
	CPM	0.0318	0.0060	5.3	1.2E-07
	TMM	0.0540	0.0060	9.0	3.0E-19
	Med	0.0638	0.0060	10.7	3.9E-26
	UQ	0.0782	0.0060	13.1	3.4E-38
batch effect correction	**no correction**	*baseline*			
	removeBatchEffect	0.0412	0.0042	9.7	4.1E-22
	ComBat	0.0468	0.0042	11.1	5.4E-28
correlation measure	**Pearson**	*baseline*			
	Spearman	-0.0107	0.0035	-3.1	0.0019

A linear model was trained on 3,888 networks in our training set, excluding those treated using ComBat-seq. Features, their estimated coefficient, standard error, t value (= estimate/std. error) and p-value of a two-sided t-test with 3,876 degrees of freedom are shown. Qualitative predictors are grouped by species, normalization, batch effect correction and correlation measure.

https://doi.org/10.1371/journal.pone.0263344.t002

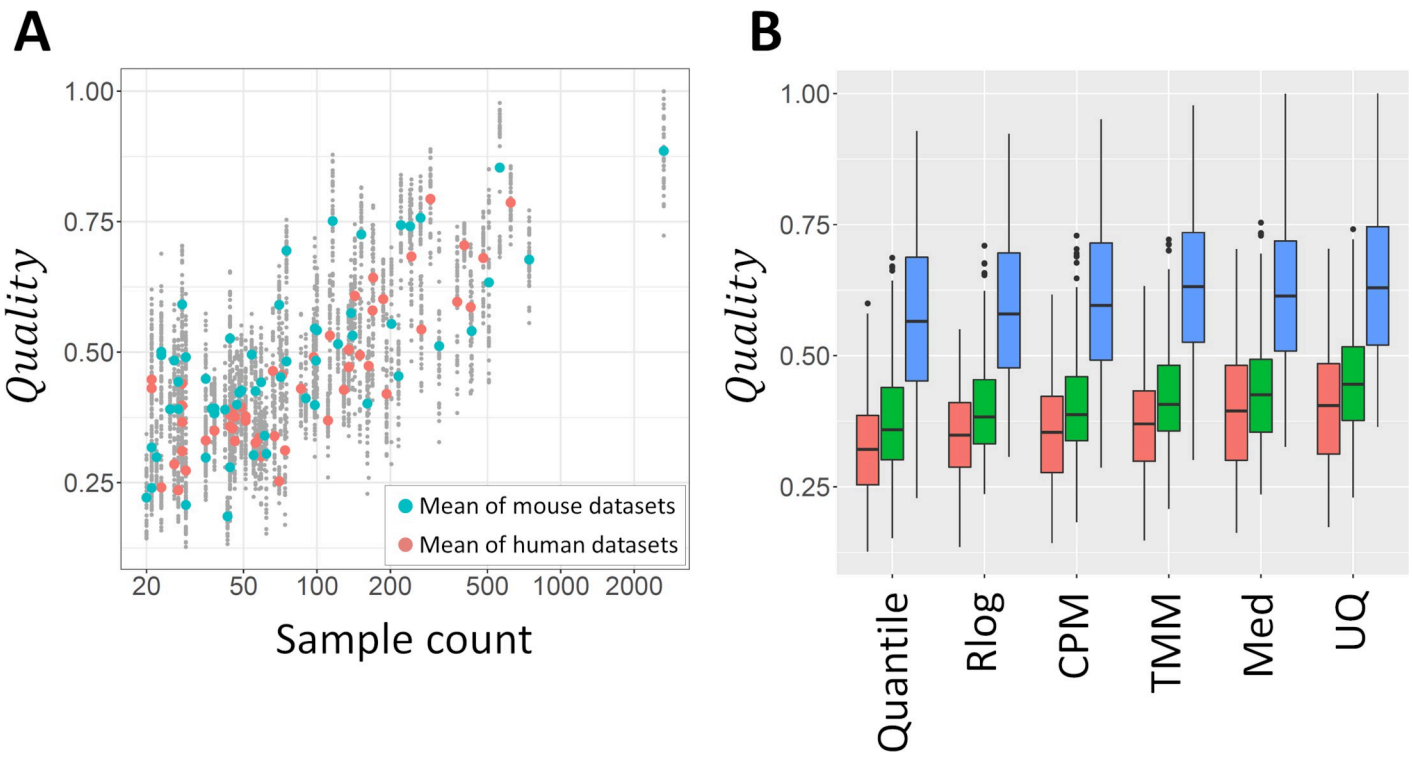

Fig 3. Importance of sample numbers and normalization approaches. (A) Sample count vs quality of co-expression networks. The quality of individual networks of each cell type and tissue generated by using different workflows are indicated by small points (forming a vertical pattern). Larger points are averages for each dataset. Blue: Mouse, red: Human datasets. **(B)** Boxplots of the quality of networks made using each of the six normalization methods, in function of dataset size. Datasets were divided into three sets of 36 datasets according to size. Red: Small datasets (20 to 44 samples); Green: Medium-sized datasets (45 to 111 samples); Blue: Large datasets (113 to 2,644 samples).

https://doi.org/10.1371/journal.pone.0263344.g003

time, an increase in the number of batches results in a small decrease in *Quality* (Table 2). A number of samples obtained from a small number of batches is expected to result in better co-expression predictions than an equal number of samples generated by many smaller batches. It should be pointed out that there is a strong correlation between the number of samples and the number of batches for each cell types and tissue (S5 Fig; Pearson correlation 0.84) which can result in instability of estimated coefficients. However, in our 4-fold CV analysis, the coefficients of sample count ($\log_{10}$) and batch count ($\log_{10}$) were consistently positive and negative, respectively (S3 Table).

Quality being roughly linearly related to the logarithm of the sample count implies that an ever-increasing number of additional samples is needed to achieve the same improvement in quality. Collecting hundreds or thousands of additional samples is practically impossible under most circumstances. Therefore, it makes sense to look for data processing steps that can maximize the quality of co-expression predictions even in the absence of an increase in samples.

The importance of data normalization approach

Regression analysis revealed clear differences in the average quality of networks generated using the six different normalization methods (Table 2), confirming the tendencies observed in Fig 2. Med and UQ resulted in an average increase of 0.064 and 0.078 in *Quality* compared to the baseline (here Quantile normalization, the worse performing method), respectively. These improvements are equivalent to a 66% and 86% increase in sample count. Additional

exploratory analysis revealed interactions exist between sample counts and normalization methods: the performance of normalization methods depends on the size of datasets. We divided our datasets according to sample counts into three sets of 36 cell types and tissues each. Fig 3B shows the *Quality* of networks based on small (20 to 44 samples), medium-sized (45 to 111 samples), and large (113 to 2,644 samples) datasets. While all normalization methods showed progressively higher performance with larger dataset sizes, UQ performed relatively well not only on the large, but also on the small and medium-sized datasets.

Correcting batch effects in general improves co-expression quality

Correction of batch effects by removeBatchEffect or ComBat resulted in better networks, increasing the *Quality* on average by respectively 0.041 and 0.047 compared to no correction (Table 2). These improvements are equivalent to a 39% and 45% increase in sample count, respectively. However, here too, the improvement depends on sample counts, and on the number of batches in the dataset. The improvement in quality appears to increase roughly with the sample count, for both removeBatchEffect (Fig 4A) and ComBat (Fig 4B). Especially ComBat consistently resulted in higher-quality networks in larger datasets (Fig 4A and 4B).

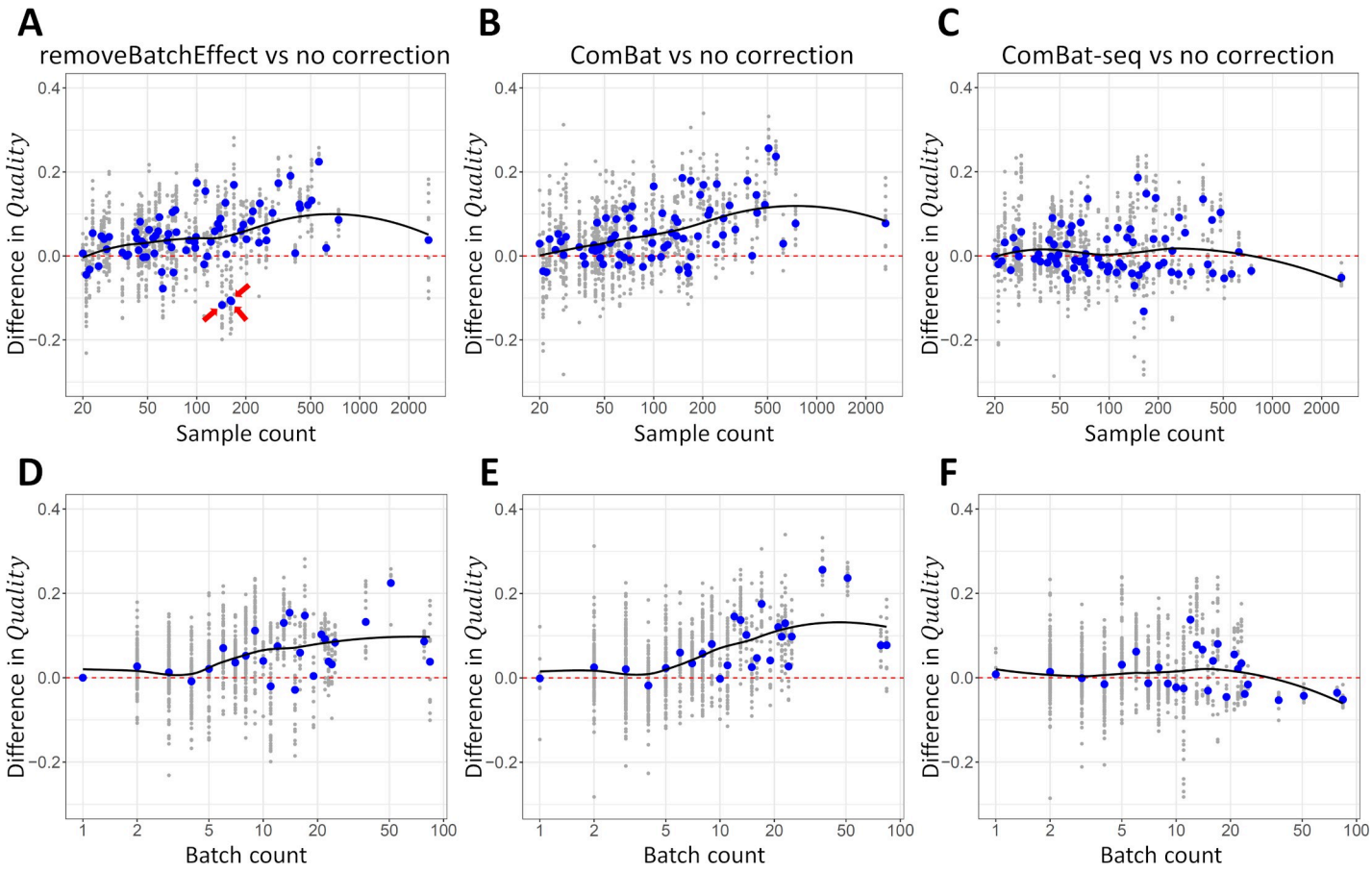

Fig 4. The number of samples and batches affect the advantage of batch effect correction. (A-C) Difference in quality of co-expression networks based on data with and without treating batch effects in function of sample count, for data treated with removeBatchEffect **(A)**, ComBat **(B)**, and ComBat-seq **(C)**. Positive values indicate an increase in quality in the batch-treated networks. **(D-F)** Difference in quality of co-expression networks based on data with and without treating batch effects in function of the number of batches in each dataset, for removeBatchEffect **(D)**, ComBat **(E)**, and ComBat-seq **(F)**. The average difference in quality is shown in blue for each sample count and for each batch count. A smoothed pattern (loess) is included in each plot.

https://doi.org/10.1371/journal.pone.0263344.g004

For datasets with >200 samples, ComBat's average improvement in *Quality* exceeded 0.10, equivalent to a >120% increase in sample count. removeBatchEffect failed to correct batch effects in a few datasets, resulting in somewhat worse overall quality (Fig 4A, indicated datasets). Batch effect correction offered no clear advantage when a dataset contained few batches (e.g. less than 5, Fig 4D and 4E). However, for datasets containing 5 or more batches, using ComBat or removeBatchEffect resulted in general in better networks.

Spearman correlation is preferred for small datasets

Using Spearman's correlation instead of Pearson correlation resulted on average in a 0.011 decrease in *Quality* (equivalent to a 8.2% decrease in sample count) (Table 2). However, for small datasets (sample count < 30) Spearman's correlation had on average an advantage (Fig 5). For medium-sized datasets (roughly 30 to 100 samples) Pearson's correlation lead in general to better co-expression networks, but the difference became smaller with higher sample counts.

These results make intuitive sense. Spearman's correlation, which is based on ranks and not on raw values, is less sensitive to extreme values. In small datasets, extreme values have a strong influence on correlation, adversely affecting Pearson's correlation. However, in medium-sized

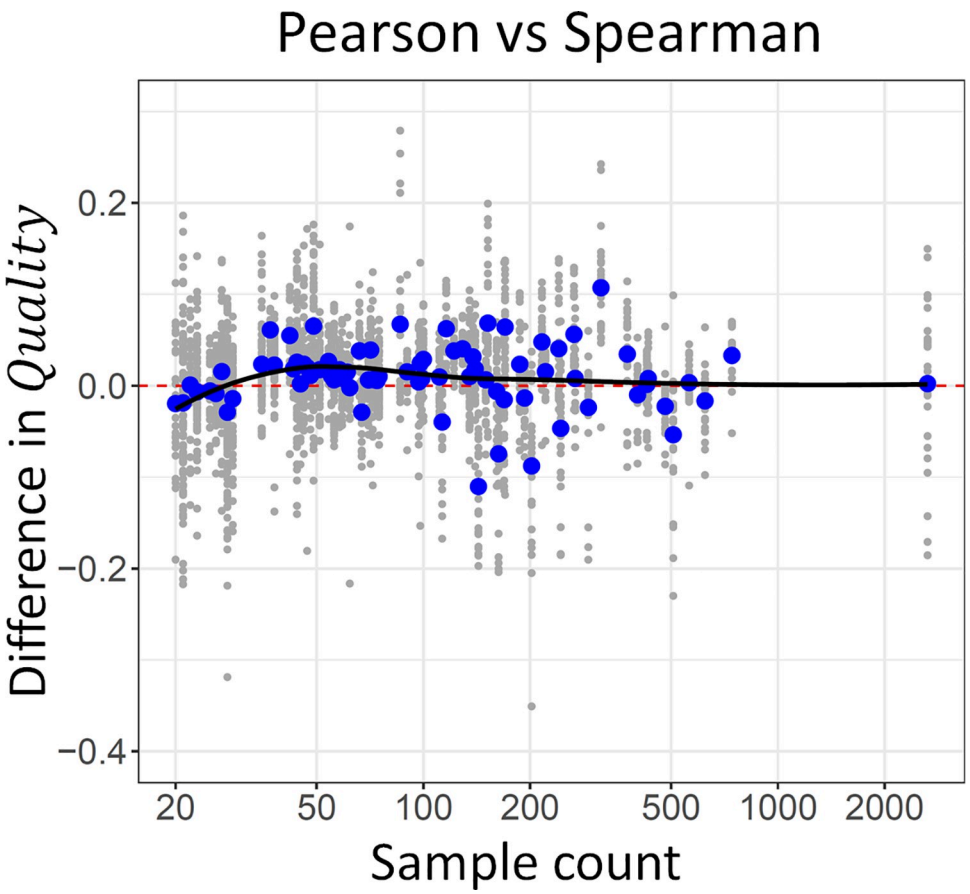

Fig 5. The preferred correlation measure depends on the number of samples. Difference in the quality of co-expression networks based on Pearson's correlation and networks based on Spearman's correlation is shown in function of the number of samples in the datasets. Positive values indicate an advantage for Pearson's correlation. The average difference in quality is shown in blue for each dataset. A smoothed pattern (loess) is included in the plot.

https://doi.org/10.1371/journal.pone.0263344.g005

datasets, the influence of extreme values decreases and Pearson's correlation appears to be better able to capture biological signals.

ComBat-seq results in lower-quality co-expression estimates

On average, ComBat-seq did not result in high-quality networks compared to ComBat and removeBatchEffect (Fig 2). Adding a normalization step following the correction by ComBat-seq did not improve qualities but rather reduced them (Fig 2). Zhang and colleagues themselves noted that on some datasets ComBat-seq did not outperform ComBat [19]. In our data, ComBat-seq resulted in lower-quality networks even in the datasets with more samples (Fig 4C) or more batches (Fig 4F). Although the reason for this failure is not clear, we did note that ComBat-seq returned unrealistically high read counts in a substantial subset of samples. For example, in 821 human samples (out of a total of 8,796) the total read count after correction exceeded 10 billion reads. A subset of genes had extremely high reads counts in some samples. For example, *Prh1* had corrected read counts > 1e100 in a subset of human salivary gland samples. These observations suggest that ComBat-seq adjusted a subset of the data to negative binomial distributions that are extremely skewed, negatively affecting the quality of co-expression networks.

The best workflows result in significantly better co-expression estimates

Finally, we returned our attention to the eight raw quality measures, and the validation datasets. We compared the networks produced by a default workflow (Rlog + no batch correction + Pearson) with those produced by optimized workflows (UQ + ComBat + Spearman for datasets with < 30 samples, and UQ + ComBat + Pearson otherwise) (Fig 6). For all eight measures, using the optimized approaches resulted in a significant improvement compared to the default workflow (one-sided paired t-tests; all eight p-values < 0.01; improvements seen in 25 to 33 of the 36 validation datasets). The optimized workflows lead to co-expressed genes

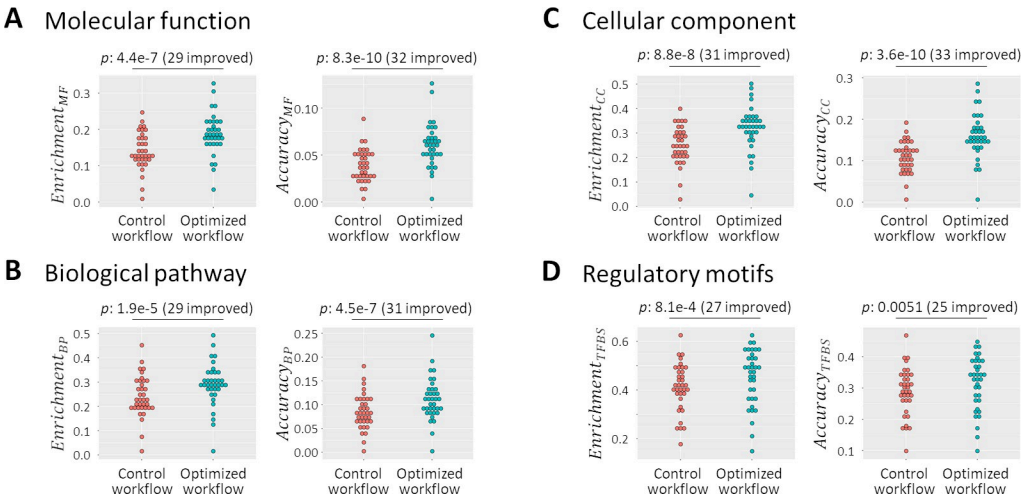

Fig 6. Optimized workflows lead to a significant improvement in all eight quality measures. (A-D) (left) Dotplots showing the fraction of genes with enrichment of GO terms and regulatory motifs in networks produced by a default workflow (red) and the optimized workflow (blue). (right) Dotplots showing the fraction of genes with an annotation that fit with enrichment GO terms, and with promoters that contain an instance of an enriched regulatory motif. Each dot represents one of the 36 (17 human and 19 mouse cell types and tissues) validation datasets. P-values are based on one-sided paired t-tests. The number of validation datasets in which an improvement was observed is indicated between brackets.

https://doi.org/10.1371/journal.pone.0263344.g006

sharing common functional annotations more frequently (Fig 6A–6C left side), as well as shared annotations fitting with the known annotations of genes more frequently (Fig 6A–6C right side). The same was true for regulatory motifs in promoter sequences (Fig 6D).

Discussion

We presented a systematic analysis of 50 workflows for processing large collections of RNA-seq data into gene co-expression predictions. We applied the workflows on data for 68 human and 76 mouse tissues and cell types, and estimated the quality of the resulting 7,200 sets of genome-wide gene co-expression datasets ("co-expression networks"). We used linear regression analysis to gain understanding of the important factors for obtaining high-quality co-expression networks. Our aim was to re-analyze existing large RNA-seq expression datasets, that have already been trimmed, aligned to a reference genome, and counted using a standardized pipeline. We focused on the steps of RNA-seq data normalization, batch effect correction, and measure of correlation of expression. Other studies have compared between read trimming and alignment approaches in related contexts [23].

We found that co-expression network quality is to a large degree determined by the number of samples which it is based on, as has been reported before [11]. It is therefore important to gather as many samples as possible. However, in practice the number of available samples is always limited. Moreover, co-expression network quality appeared to be roughly a linear function of the logarithm of the sample count. This means that for cell types and tissues with abundant data, adding hundreds of additional samples might result in only modest improvements in quality. It therefore makes sense to optimize the data processing workflow to obtain high-quality networks even in the absence of large sample counts.

Treating batch effects in general lead to better networks. On average, ComBat performed better than limma's removeBatchEffect function. ComBat-seq however performed considerably worse than ComBat and removeBatchEffect. In our analysis, correcting batch effects using ComBat resulted in improvements to network quality equivalent to a 45% increase in sample count on average. In larger datasets, the advantage was even more pronounced, equivalent to roughly a doubling in sample count. Unfortunately, gene co-expression studies still often ignore batch effects. Clearly, more attention needs to be paid to the issue of batch effects in order to extract the maximum potential out of ever-increasing public gene expression datasets.

We found that some data normalization approaches lead to better co-expression estimates than others. Especially UQ performed well. UQ has also been found to perform relatively well compared with total count normalization (equivalent to CPM in our study) and quantile normalization in the context of predicting differentially expressed genes [24]. Other comparisons of RNA-seq normalization methods (outside of the context of co-expression) have come to different and conflicting conclusions [25–27].

The measure of correlation appeared to be less crucial, but Pearson's correlation seems to have a slight advantage, except when there are only small numbers of samples (<30). In the latter case, Spearman's correlation seems better.

Although no workflow dominated all others, UQ + ComBat + Pearson (or Spearman for small datasets) resulted in the best quality overall, and in above-average co-expression networks in $>90\%$ of the tissues and cell types we examined (Fig 2).

In addition to the findings described above, the dataset collection and the workflows used in this study are valuable resources. The collection of raw human and mouse samples and their annotation data have been made public, together with the data processed using UQ normalization and ComBat batch effect correction (see section "Data and code availability"). Scripts and

an example workflow have been made public in a GitHub repository. We hope that together this data and code can serve as a basis for future studies.

Methods

Gene expression data collection and normalization

We used the RNASeq-er REST API of the European Bioinformatics Institute [18] to obtain read count data for 8,796 human and 12,114 mouse RNA-seq samples, produced by 401 and 630 studies, covering 68 human and 76 mouse cell types and tissues, respectively (see S1 Methods and S1 and S2 Tables). On these two large datasets of human and mouse samples, we applied the following six normalization approaches:

Trimmed Mean of M-values (TMM). For all genes, log ratios are calculated versus a reference sample [28]. The most highly expressed genes, and genes with high log ratios are filtered out. The mean of the remaining log ratios is used as a scaling factor. This normalization method is the default normalization method of the edgeR function calcNormFactors [28,29]. In this study, we first removed genes that have less than 1 read per million reads in all samples prior to normalizing the remaining genes.

Counts per million (CPM). The number of reads per gene is divided by the total number of mapped reads of the sample and multiplied by 1 million [25,30].

There are several variations on CPM, describe below, including Upper Quartile and Median.

Upper Quartile (UQ). Counts are divided not by total count but by the upper quartile of non-zero values of the sample [24].

Median (Med). Counts are divided not by total count, but by the median of non-zero values of the sample [25].

Regularized Logarithm (RLog or RLE). A regularized-logarithm transformation is applied which is similar to a default logarithmic transformation, but in which lower read counts are shrunken towards the genes' averages across all samples. We applied this normalization using the R package DESeq2 [31].

Quantile. All samples are normalized to have the same quantiles. We applied this normalization using the function normalizeQuantiles of the limma R package [32].

Note that methods that correct for differences in gene length (RPKM and FPKM) are not relevant here, since they don't affect correlation values. In this study, these methods would be equivalent to CPM normalization.

We thus obtained 12 normalized datasets (6 each for the human and the mouse data). Each dataset was transformed to log values after addition of a small pseudo count (defined as the 1% percentile of non-zero values in the normalized dataset).

Batch effect correction using ComBat and limma's removeBatchEffect function

On the 12 log-transformed datasets we applied two batch effect correction methods: ComBat (function ComBat in the sva R package) and the removeBatchEffect function of the limma R package [13,32]. Both ComBat and removeBatchEffect allow users to specify batch covariates to remove from the data ("batch" parameter in both functions). Here, studies were used as substitutes for batches. Users can also give biological covariates to retain ("mod" parameter in ComBat and "design" parameter in removeBatchEffect). In this study, biological covariates are the cell type or tissue from which the samples were obtained. Both ComBat and removeBatchEffect were used using default parameter settings.

To be able to treat batch effects in a dataset there can be no confounding between technical and biological covariates. In practice, studies often focus on a single cell type, which makes confounding of batches and cell types highly probable. Both the human and mouse datasets could be divided into several subsets with no shared cell types or tissue annotations. Therefore, batch effects were corrected for each of these subsets of samples separately, and finally the treated datasets were merged again into one dataset.

In addition to correcting the data using ComBat and removeBatchEffect we also considered 2 other options. One is to ignore batch effects and use the normalized data directly for estimating gene co-expression. Another is to use ComBat-seq.

Batch effect correction and normalization using ComBat-seq

ComBat-seq differs from ComBat and limma's removeBatchEffect function in that both its input and output are integer counts, making it more suitable for RNA-seq read count data [19]. To correct batch effects using ComBat-seq (function Combat_seq in the sva R package), we therefore gave as input the raw count data (without any normalization step applied to it), as well as the same batch covariates and biological covariates as we used for ComBat and removeBatchEffect. ComBat-seq was used with default parameter settings. The output read counts were transformed to log values after adding a pseudocount of 1. These log-transformed data were used for estimating correlation of expression (see next section) without an additional normalization step.

However, quality estimates of the co-expression networks generated using ComBat-seq were relatively low compared to ComBat and removeBatchEffect, which use normalized data as input (Figs 2, 4C and 4F). Therefore, to avoid an unfair comparison, we also applied the 6 normalization approaches (TMM, CPM, UQ, Med, Rlog and quantile) on the ComBat-seq output data. Rlog normalization failed because of the extremely high reads counts in a subset of the data (see also section "ComBat-seq results in lower-quality co-expression estimates"). We therefore trimmed all read counts >1e9 to 1e9 before conducting the Rlog normalization.

Together, this resulted in 7 datasets which had been processed using ComBat-seq (i.e. 6 with a normalization step, and 1 without).

Estimating correlation of expression

For each of the 25 data processing combinations (6 normalization methods x 4 batch correction methods, and ComBat-seq without normalization), we calculated correlation of expression between each pair of genes in the log-transformed data for each cell type or tissue. We did this using Pearson's correlation and Spearman's rank correlation. Before calculating correlation coefficients in the expression data of a tissue or cell type, genes with general low levels of expression (less than 10 mapped reads in >90% of samples, or less than 10 reads in >80% of samples and fewer than 50 reads in all samples) or with no variation in expression (standard deviation = 0) were removed. The number of samples and genes used for calculating co-expression in each tissue and cell type are listed in S1 and S2 Tables.

Evaluation of gene co-expression network quality

The processing steps described above resulted in 7,200 (3,400 human and 3,800 mouse) sets of genome-wide cell type or tissue-specific gene co-expression predictions, which we refer to as "co-expression networks". The main goal of this study is to gain understanding into what are the critical features that distinguish good co-expression networks from bad ones. Because there is no gold standard co-expression network available, we first defined eight measures of quality (see also Table 1) based on the enrichment of biologically meaningful features

(functional annotations of genes and regulatory DNA motifs in promoter sequences) among co-expressed genes. For each gene X in a co-expression network, we define set_X to be the 100 genes with the highest correlation of expression with X (excluding X itself). Our quality measures are based on the enrichment of GO annotations and TFBS motifs in the set_X of every gene X in each genome-wide co-expression network. These measures should not be interpreted as strict measures of accuracy of for example functional annotation predictions, but rather as rough indicators of quality of the inferred gene co-expression values.

GO enrichment frequency. Genes involved in the same biological process are expected to be co-expressed more frequently than unrelated sets of genes. In a high-quality co-expression network, we would expect the genes in set_X to share a functional annotation (Fig 1B). In contrast, in a low-quality network (e.g. a randomly generated network) we expect genes in set_X to have a random set of annotations. We define $Enrichment_{MF}$, $Enrichment_{BP}$, and $Enrichment_{CC}$ as the fraction of genes in a network for which set_X contained one or more significantly enriched GO terms (after correction for multiple testing) for Molecular Function (MF), Biological Process (BP) and Cellular Component (CC) GO terms. The number of tested GO terms in each dataset is shown in S1 and S2 Tables.

GO enrichment accuracy. Where we found set_X to have enriched GO terms, we checked if the enriched terms overlapped with the GO terms of gene X. $Accuracy_{MF}$, $Accuracy_{BP}$, and $Accuracy_{CC}$ were defined as the fraction of genes in the network for which this was the case for MF, BP, and CC GO terms.

TFBS enrichment frequency. Genes with similar expression profiles are likely to be under the control of a shared regulatory mechanism, including regulation by a similar set of transcription factors (TFs). In a high-quality co-expression network, we would therefore expect the genes in set_X to contain a shared set of transcription factor binding sides (TFBSs). We define $Enrichment_{TFBS}$ as the fraction of genes in a network for which the promoter sequences of set_X contained one or more significantly enriched TFBSs.

TFBS enrichment accuracy. Where we found set_X to have enriched TFBSs, we checked if the promoter of gene X contained one or more of those TFBSs. $Accuracy_{TFBS}$ was defined as the fraction of genes in the network for which this was the case.

Correlation between the eight quality measures was high (range 0.60 to 0.98). To facilitate comparison between co-expression networks, the eight measures were combined into a single quality score, *Quality*, which is based on PCA of the eight measures. PCA was conducted using the function prcomp in R, after standardizing the eight quality measures to mean 0 and standard deviation 1. Analysis of the Principal Components (PCs) revealed that 81.4% of the total variation in the quality measures could be explained by the first PC (S2A Fig). We decided to use this first PC as the general quality score, *Quality*, after rescaling it to values between 0 (worst network) to 1 (best network). The correlation between *Quality* and each of the quality measures was high (range 0.77 to 0.96; S2C Fig). The p-values of Pearson correlation coefficients as shown in S1 and S2 Figs are based on a Student's t-distribution with n-2 degrees of freedom [33].

To evaluate the sensitivity of our quality measures with regard to the number of genes they are based on, we also calculated the eight quality measures using the top 50 and top 200 genes (instead of the top 100). We did this for 200 randomly selected co-expression networks (out of the total of 7,200). We used scatterplots of the results (S1 Fig) and Pearson correlation to evaluate the consistency of the eight quality measures based on the top 50, 100, and 200 genes.

Linear regression analysis

We randomly split the 68 human and 76 mouse cell types and tissues into four parts, each representing 25% of the human and mouse cell types and tissues. We used three parts to form the

training set (representing 75% of cell types and tissues; 51 human and 57 mouse cell types and tissues), and the remaining part was used as a validation set (representing the remaining 25%; 17 human and 19 mouse cell types and tissues).

We conducted least squares regression using the lm function in R using the 3,888 networks of the training set, excluding networks where batch effects were treated using ComBat-seq. As response variable we used *Quality*, and as predictors we used 1) the number of RNA-seq samples on which the co-expression network was based ($\log_{10}$ values), 2) the number of batches in the data ($\log_{10}$ values), 3) the species (human or mouse), 4) the data normalization approach, 5) batch correction approach, and 6) the correlation measure. For the categorical predictors (i.e. species, data normalization approach, batch correction approach and correlation measure) the lm function uses dummy variables with values 0 and 1. The baseline levels of these categorical variable were set in a way that facilitates interpretation of the results. The resulting model as shown in Table 2 is the output of the R function lm, including the estimated coefficients, their standard errors, t values (= estimated coefficient/standard error), and p-value of a two-sided t-test. In this case, the degree of freedom in the t-test is 3,876 (= the number of observations in the dataset minus the number of coefficients to estimate = 3,888–12).

To evaluate the stability of estimated coefficients of the above model, we step-by-step left out each of the three parts used to form the original training set, and used the other parts (the two remaining parts of the training set and the original validation set) to fit the same linear model, effectively implementing a 4-fold cross validation (CV) analysis. Coefficients of the resulting four models are shown in S3 Table.

Supporting information

S1 Fig. Consistency of the eight quality measures with regard to the number of top correlated genes they are based on. Scatterplots are shown for the eight quality measures (see main manuscript and Methods section) based on the top 100 highly correlated genes (X axes) and the top 50 highly correlated genes **(A)** or the top 200 highly correlated genes **(B)** (Y axes). Scatterplots show data for 200 randomly selected co-expression networks (out of a total of 7,200 networks). Pearson correlations coefficients (and p-values) are indicated. For each quality measure a high correlation (PCCs between 0.93 and 0.98) was observed, suggesting that the quality measures are robust with regard to the number of highly correlated genes they are based on.
(DOCX)

S2 Fig. Principal Component Analysis of the eight quality measures. (A) Proportion of the variance in the eight quality measures explained by the principal components (PCs). The first and second PCs explain 81.4% and 12.1% of the total variance, respectively. **(B)** Bar plot of the loadings of the first and second PC. **(C-D)** Scatterplots of PC1 **(C)** and PC2 **(D)** (in the X-axes) versus each of the eight individual quality measures (Y-axes). Each plot shows 7,200 dots, each representing a genome-wide gene-gene co-expression network for a cell type or tissue. The Pearson correlation coefficient (PCC) and its p-value are indicated in each plot.
(DOCX)

S3 Fig. Distribution of the *Quality* score. (A) The distribution of the 7,200 general quality scores, *Quality*. **(B-C)** The 8 quality measures of the worst (red), the median (green), and the best (blue) network for GO Molecular Function, Biological Process, Cellular Component and Regulatory motifs in promoter sequences. **(B)** shows the frequency of enrichment and **(C)** the accuracy.
(DOCX)

S4 Fig. Overview of the performance of each workflow on each of the 144 datasets. For each of the 50 workflows the relative performance on each of the 144 datasets is visualized. Workflows are ordered in order of average overall performance as in Fig 2 in the main text. Datasets are ordered according to sample counts. Colors indicate the relative performances of the 50 workflows on each dataset.
(DOCX)

S5 Fig. Scatterplot of the sample count and batch count for the 144 cell types and tissues. The Pearson correlation between the sample count and batch count (both in $\log_{10}$ values) is 0.84.
(DOCX)

S1 Table. Human datasets. The cell type or tissue, the number of RNA-seq samples, the number of genes included in the final co-expression network, and the number of GO terms tested for the estimation of the network quality is shown. The last column indicates which datasets were included in the validation set.
(DOCX)

S2 Table. Mouse datasets. The cell type or tissue, the number of RNA-seq samples, the number of genes included in the final co-expression network, and the number of GO terms tested for the estimation of the network quality is shown. The last column indicates which datasets were included in the validation set.
(DOCX)

S3 Table. Coefficients and p-values of linear models trained by 4-fold cross-validation. Human and mouse cell types and tissues were randomly divided into 4 folds. Each fold was left out and a linear regression model was trained on the remaining 3 folds. Model 1 is equivalent to the model shown in Table 2 in the main manuscript. For each model, the estimated coefficient (and corresponding p-value) for each parameter is shown. The coefficients estimated in each model are in general consistent with each other.
(DOCX)

S1 Methods.
(DOCX)

Acknowledgments

We thank Dr. Diego Diez (Osaka University), Prof. Yoshio Koyanagi and the members of the Lab. of Systems Virology (Kyoto University), Prof. Kenta Nakai and the members of the Lab. of Functional Analysis in silico (Tokyo University), and Prof. Wataru Fujibuchi (Kyoto University) for helpful discussions and advice.

Author Contributions

Conceptualization: Alexis Vandenbon.

Data curation: Alexis Vandenbon.

Formal analysis: Alexis Vandenbon.

Funding acquisition: Alexis Vandenbon.

Investigation: Alexis Vandenbon.

Project administration: Alexis Vandenbon.

Resources: Alexis Vandenbon.

Validation: Alexis Vandenbon.

Visualization: Alexis Vandenbon.

Writing – original draft: Alexis Vandenbon.

Writing – review & editing: Alexis Vandenbon.

References

1. Eisen MB, Spellman, Paul T., Brown, Patrick O., Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998; 95: 14863–14868. https://doi.org/10.1073/pnas.95.25.14863 PMID: 9843981
2. Wolfe CJ, Kohane IS, Butte AJ. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinformatics. 2005; 6: 1–10. https://doi.org/10.1186/1471-2105-6-1 PMID: 15631638
3. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4. https://doi.org/10.2202/1544-6115.1128 PMID: 16646834
4. Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, et al. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ. 2009; 32: 1633–51. https://doi.org/10.1111/j.1365-3040.2009.02040.x PMID: 19712066
5. Serin EAR, Nijveen H, Hilhorst HWM, Ligterink W. Learning from co-expression networks: Possibilities and challenges. Front Plant Sci. 2016; 7: 444. https://doi.org/10.3389/fpls.2016.00444 PMID: 27092161
6. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene-disease predictions. Brief Bioinform. 2018; 19: 575–592. https://doi.org/10.1093/bib/bbw139 PMID: 28077403
7. Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W. GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. Plant Physiol. 2004; 136: 2621–2632. https://doi.org/10.1104/pp.104.046367 PMID: 15375207
8. van Dam S, Craig T, de Magalhães JP. GeneFriends: a human RNA-seq-based gene and transcript co-expression database. Nucleic Acids Res. 2015; 43: D1124–D1132. https://doi.org/10.1093/nar/gku1042 PMID: 25361971
9. Vandenbon A, Dinh VH, Mikami N, Kitagawa Y, Teraguchi S, Ohkura N, et al. Immuno-Navigator, a batch-corrected coexpression database, reveals cell type-specific gene networks in the immune system. Proc Natl Acad Sci U S A. 2016; 113: E2393–E2402. https://doi.org/10.1073/pnas.1604351113 PMID: 27078110
10. Obayashi T, Kagaya Y, Aoki Y, Tadaka S, Kinoshita K. COXPRESdb v7: A gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. Nucleic Acids Res. 2019; 47: D55–D62. https://doi.org/10.1093/nar/gky1155 PMID: 30462320
11. Ballouz S, Verleyen W, Gillis J. Guidance for RNA-seq co-expression network construction and analysis: Safety in numbers. Bioinformatics. 2015; 31: 2123–2130. https://doi.org/10.1093/bioinformatics/btv118 PMID: 25717192
12. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010; 11: 733–9. https://doi.org/10.1038/nrg2825 PMID: 20838408
13. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007; 8: 118–27. https://doi.org/10.1093/biostatistics/kxj037 PMID: 16632515
14. Harper KN, Peters BA, Gamble M V. Batch effects and pathway analysis: Two potential perils in cancer studies involving DNA methylation array analysis. Cancer Epidemiol Biomarkers Prev. 2013; 22: 1052–1060. https://doi.org/10.1158/1055-9965.EPI-13-0114 PMID: 23629520
15. Nygaard V, Rødland EA. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics. 2016; 17: 29–39. https://doi.org/10.1093/biostatistics/kxv027 PMID: 26272994
16. Price EM, Robinson WP. Adjusting for batch effects in DNA methylation microarray data, a lesson learned. Front Genet. 2018; 9: 1–7. https://doi.org/10.3389/fgene.2018.00001 PMID: 29387083

17. Zindler T, Frieling H, Neyazi A, Bleich S, Friedel E. Simulating ComBat: How batch correction can lead to the systematic introduction of false positive results in DNA methylation microarray studies. BMC Bioinformatics. 2020; 21: 1–15. https://doi.org/10.1186/s12859-019-3325-0 PMID: 31898485

18. Petryszak R, Fonseca NA, Füllgrabe A, Huerta L, Keays M, Tang YA, et al. The RNASeq-er API-A gateway to systematically updated analysis of public RNA-seq data. Bioinformatics. 2017; 33: 2218–2220. https://doi.org/10.1093/bioinformatics/btx143 PMID: 28369191

19. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics Bioinforma. 2020; 2: 1–10. https://doi.org/10.1093/nargab/lqaa078 PMID: 33015620

20. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpresion analysis of human genes across many microarray data sets. Genome Res. 2004; 14: 1085–1094. https://doi.org/10.1101/gr.1910904 PMID: 15173114

21. Verleyen W, Ballouz S, Gillis J. Measuring the wisdom of the crowds in network-based gene function inference. Bioinformatics. 2015; 31: 745–752. https://doi.org/10.1093/bioinformatics/btu715 PMID: 25359890

22. Ballouz S, Weber M, Pavlidis P, Gillis J. EGAD: Ultra-fast functional analysis of gene networks. Bioinformatics. 2017; 33: 612–614. https://doi.org/10.1093/bioinformatics/btw695 PMID: 27993773

23. Corchete LA, Rojas EA, Alonso-López D, De Las Rivas J, Gutiérrez NC, Burguillo FJ. Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. Sci Rep. 2020; 10: 1–15. https://doi.org/10.1038/s41598-019-56847-4 PMID: 31913322

24. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010; 11: 94. https://doi.org/10.1186/1471-2105-11-94 PMID: 20167110

25. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013; 14: 671–683. https://doi.org/10.1093/bib/bbs046 PMID: 22988256

26. Li P, Piao Y, Shon HS, Ryu KH. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. BMC Bioinformatics. 2015; 16: 347. https://doi.org/10.1186/s12859-015-0778-7 PMID: 26511205

27. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016; 17. https://doi.org/10.1186/s13059-016-0881-8 PMID: 26813401

28. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11. https://doi.org/10.1186/gb-2010-11-3-r25 PMID: 20196867

29. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012; 40: 4288–97. https://doi.org/10.1093/nar/gks042 PMID: 22287627

30. Abbas-Aghababazadeh F, Li Q, Fridley BL. Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. PLoS One. 2018; 13: e0206312. https://doi.org/10.1371/journal.pone.0206312 PMID: 30379879

31. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15: 1–21. https://doi.org/10.1186/s13059-014-0550-8 PMID: 25516281

32. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43: e47. https://doi.org/10.1093/nar/gkv007 PMID: 25605792

33. Yamamoto H, Fujimori T, Sato H, Ishikawa G, Kami K, Ohashi Y. Statistical hypothesis testing of factfior loading in principal component analysis and its application to metabolite set enrichment analysis. BMC Bioinformatics. 2014; 15. https://doi.org/10.1186/1471-2105-15-51 PMID: 24555693