様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　６ 月 ２６ 日現在

機関番号：１４３０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ０６６０９

研究課題名（和文）Comprehensive optimization of cell type-specific gene co-expression networks and construction of a cell type-specific co-expression database

研究課題名（英文）Comprehensive optimization of cell type-specific gene co-expression networks and construction of a cell type-specific co-expression database

研究代表者

ＶＡＮＤＥＮＢＯＮ　ＡＬＥＸＩＳ（Vandenbon, Alexis）

京都大学・医生物学研究所・准教授

研究者番号：６０５７０１４０

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：ヒトおよびマウスの様々な組織と細胞種から得られた大規模なRNA-seqデータを用いて、高品質な遺伝子共発現ネットワークの生成を目的としたデータ処理ワークフローの包括的な評価を行った。その結果、できるだけ多くのRNA-seqサンプルの収集、Upper Quartileの正規化、バッチ効果の修正が重要であることが明らかになった。最適な処理ワークフローを使用することで、高品質の遺伝子発現データセットが得られ、他のバイオインフォマティクス解析をサポートする事例を提供した。最後に、このヒトおよびマウスの遺伝子発現と共発現データから新たな知見を見出せるよう共発現ネットワークのデータベースを作成している。

研究成果の学術的意義や社会的意義
Gene co-expression is widely used for the prediction of gene functions and regulatory mechanisms. We here showed how gene expression data can be processed to obtain high-quality co-expression values. This will contribute to improved bioinformatics analyses and new insights into gene regulation.

研究成果の概要（英文）：We used a large collection of RNA-seq data samples covering 68 human and 76 mouse cell types and tissues to conduct a comprehensive evaluation of which data processing workflow results in the highest quality gene co-expression networks. Our results indicate that it is important to collect as many RNA-seq samples as possible. Second, researchers should use using Upper Quartile normalization and correct batch effects. Finally, in general Pearson’s correlation should be used, but in small datasets Spearman’s rank correlation might be preferable. We confirmed that using the optimized processing workflow, we obtained a high-quality gene expression dataset which can be used as a reference. We provided two illustrations of the use of our dataset as a reference to support other bioinformatics analyses. Finally, we are preparing a freely accessible gene co-expression database, which will allow users to inspect gene expression and co-expression in many human and mouse tissues and cell types.

研究分野：bioinformatics

キーワード：bioinformatics　gene expression　gene co-expression　data normalization　batch effect correction　database

１．研究開始当初の背景

Even though almost all cells inside our bodies contain the same genetic material, they develop into a variety of different cell types that conduct different functions inside our body, such as liver cells and brain cells. To a large extent, these different cell types and different functions are defined by the genes that are transcribed within each cell. Improving our understanding of the regulation and functions of genes is one of the key goals of biology.

One approach for predicting gene regulation and function is to compare gene expression patterns under different conditions. Genes that tend to be expressed together are called "co-expressed". Gene co-expression is an important concept in bioinformatics because it serves as a foundation for predicting gene functions and regulatory mechanisms. So far, many groups have studied gene co-expression patterns in many different organisms (reviewed in van Dam *et al.*, Briefings in Bioinformatics, 2018).

However, several open problems remain. First, previous studies have estimated gene co-expression from gene expression data obtained from many different tissues and cell types merged together, ignoring the fact that co-expression patterns are different in different cell types. For researchers interested in – for example – macrophages, gene co-expression estimates based on data obtained from hepatocytes or neuron cells are not relevant. Instead, they are interested in gene co-expression patterns found in data of macrophages subjected to many different conditions. Such data would be suitable for making a macrophage-specific gene co-expression study.

Second, gene expression data nowadays is typically obtained from genome-wide sequencing of RNA molecules (RNA-seq). This data is complex and requires a number of processing and normalization steps. However, it is not clear which data processing and normalization steps result in the most reliable gene expression and co-expression data.

Finally, it is difficult to merge together RNA-seq datasets produced by different labs under different conditions, because of the presence of technical biases between them, which are referred to as "batch effects". Without suitable treatment of batch effects, gene co-expression predictions could be heavily affected by these batch effects, thus reflecting technical sources of correlation of expression, rather than biological co-expression. How to optimally treat these batch effects in order to get high-quality gene expression data and high-quality gene co-expression predictions remains an open question.

In a previous collaborative study, I constructed the first cell type-specific co-expression database, called Immuno-Navigator [Vandenbon *et al.*, PNAS, 2016; https://genomics.virus.kyoto-u.ac.jp/immuno-navigator/]. However, the data in this database was based on somewhat old technology (microarrays), and covered only 19 human and 24 mouse cell types, limited to cell types of the immune system. Moreover, no comprehensive comparison of different data processing methods was performed. The current proposal aimed to address these issues.

２．研究の目的

The goals of this study were 1) to conduct a comprehensive evaluation of the effect of different data analysis steps (data normalization, batch effect correction, measure of correlation, downstream processing of gene co-expression estimates) on the quality of gene co-expression estimates, 2) to construct an open-access, freely available database of gene co-expression in many cell types in human and mouse, and 3) to provide clear guidelines to other researchers about data processing steps to improve their own co-expression data.

３．研究の方法

<u>(1) Preparation of a large collection of gene expression data</u>

I collected 8,796 human and 12,114 mouse RNA-seq samples from the European Nucleotide Archive (ENA). These samples were produced by 401 and 630 studies, and covered 68 human and 76 mouse cell types and tissues, respectively. This collection of data was merged into 2 large genome-wide datasets, one for human and one for mouse samples. Annotation data was prepared showing for each sample which cell type or tissue it was obtained from, and which study had generated it. The studies were used as proxies for batches, assuming that each study represents one batch.

<u>(2) Data normalization, batch effect correction, and correlation calculation</u>
The human and mouse datasets were normalized using six different normalization approaches: Trimmed Mean of M-values (TMM) [Robinson *et al.*, Genome Biol., 2010], Counts per million (CPM) and Median (Med) [Dillies *et al.*, Brief Bioinform, 2013, Abbas-Aghababazadeh *et al.*, PLoS One, 2018], Upper Quartile (UQ) [Bullard *et al.*, BMC Bioinformatics, 2010], Regularized Logarithm (RLog) [Love *et al.*, Genome Biol, 2014], and Quantile [Ritchie *et al.*, Nucleic Acids Res, 2015]. This resulted in 12 normalized datasets (6 for human and 6 for mouse data). Data was log-transformed.

On the log-transformed data, I applied two batch effect correction methods: ComBat [Johnson *et al.*, Biostatistics, 2007] and the removeBatchEffect function of the limma R package [Ritchie *et al.*, Nucleic Acids Res, 2015]. As biological variable the cell types or tissues were used, and as batch variable the studies. In addition, we also considered using no batch correction (only a normalization step), and also batch effect correction using ComBat-seq which uses non-normalized RNA-seq data as input [Zhang *et al.*, NAR Genom Bioinform, 2020]. This way, we obtained 25 human and 25 mouse datasets in total (6 normalizations x 4 batch corrections, and ComBat-seq without normalization).

Finally, we estimated gene co-expression in each of the 25 human and 25 mouse datasets using two measures of correlation: Pearson's correlation and Spearman's rank correlation. These correlation estimates were calculated in the data of each cell type or tissue separately. Thus, for each cell type or tissue, we obtained 50 different genome-wide sets of gene co-expression values (50 different combinations of normalization, batch effect correction, and correlation measure). Each such set of genome-wide gene co-expression values can be regarded as a gene co-expression network. Since there were 144 cell types and tissues (68 human and 76 mouse) and 50 combinations, this resulted in 7,200 co-expression networks. In the next step, I evaluated and compared the quality of these genome-wide co-expression networks.

<u>(3) Evaluation and comparison of gene co-expression networks</u>
For each of the 7,200 co-expression networks, I calculated 8 measures of quality. For a detailed description, I refer to the corresponding publication [Vandenbon, PLoS One, 2022]. In brief, the measures of quality are based on how well highly co-expressed genes resemble each other, in terms of known functions and DNA motifs in promoter regions. The 8 measures of quality were highly consistent, and were therefore processed into a single quality indicator, rescaled to be in the range of 0 (very low) to 1 (very high). Using linear regression analysis, I analyzed how much the following features contributed to a high/low quality: the number of RNA-seq samples, number of batches, normalization method, batch effect correction method, correlation measure, species (human or mouse). The findings are summarized in Table 1.

All above processing and statistical analysis was done using the R programming language.

<u>(4) Applications of our processed dataset as a high-quality reference</u>
After identifying the optimal data processing combination (see above), we obtained a high-quality RNA-seq dataset, which can be used as a gene expression reference dataset. To further illustrate the usefulness of this dataset, we applied it to two additional studies.

First, in one study we analyzed gene expression in mouse liver tissues using single-cell (scRNA-seq) and spatial transcriptomics (10X Genomics Visium platform) [Vandenbon *et al.*, Commun Biol, 2023], in control mice and breast cancer-bearing mice. We used our reference dataset to support cell type annotation and the analysis of gene co-expression in different parts of the liver tissues.

Second, we used our reference dataset to study differentially expressed genes. In one example, we used 1,958 mouse samples from different brain-related tissues to predict genes with different levels of expression [Vandenbon and Diez, bioRxiv, 2022].

(5) Database construction

We are constructing a freely accessible gene co-expression database. We are using the optimally processed gene expression data for both human and mouse samples as input, as well as promoter sequences and functional annotations of genes. We are implementing the database using Flask, a web framework written in Python, and SQLite.

４．研究成果

(1) Evaluation and comparison of gene co-expression networks

We used linear regression to evaluate how each aspect of data and processing steps contributes to the quality of the resulting gene co-expression networks (Table 1). For a detailed explanation I refer to the publication [Vandenbon, PLoS One, 2022]. Each aspect will be briefly discussed below.

Table 1. Linear regression model of the co-expression network quality scores. The table summarizes a linear model of using co-expression quality scores as response variable. Predictors, their estimated coefficient, standard error, t value (= estimate divided by std. error) and p-value are shown. Qualitative predictors are grouped by species, normalization, batch effect correction and correlation measure.

	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

First, the result suggests that the most important point is the number of RNA-seq samples on which the gene co-expression estimates are based. This was true for both mouse and human samples, and, moreover, the same result was found for any combination of data normalization and batch effect correction. However, the quality of co-expression networks is roughly linearly related to the logarithm of the sample counts. This logarithmic trend means that an ever-increasing number of samples is needed to obtain the same improvements in quality. In practice, it is impossible to always collect thousands of RNA-seq samples. This means that it makes sense to optimize also other aspects, such as the normalization and batch effect correction steps.

Next, the results suggest a clear difference in the quality caused by different normalization approaches. In particular, Upper Quartile (UQ) performed well. The use of UQ instead of Quantile normalization (used as baseline here) is roughly equivalent to an 86% increase in sample counts. UQ performed well not only on dataset with many samples, but also on datasets with few samples (not shown here).

The analysis also suggested that the quality of the co-expression networks is negatively related to the number of batches. In other words, keeping all other variables constant, the quality of co-expression estimates is expected to decrease if the underlying gene expression data was obtained from many different studies. This reflects the existence of batch effects. Indeed, treating batch effects using the removeBatchEffect or ComBat approaches lead in general to an improvement in quality, especially in larger datasets containing data obtained from many studies (i.e., many batches). However, this was less the case for the ComBat-seq approach, which resulted in lower quality (not shown here).

Finally, we observed that the use of Pearson's correlation resulted in a slightly higher average quality score compared to Spearman's rank correlation. However, Spearman's rank correlation performed better than Pearson's correlation on datasets with few samples (i.e., datasets with < 30 samples; not shown here).

(2) General guidelines for obtaining high-quality gene co-expression data

Taken together, the linear regression analysis of a large number of workflows suggests the following guidelines for obtaining high-quality gene co-expression estimates: 1) researchers should attempt to collect as many RNA-seq samples as possible. 2) In general, Upper Quartile (UQ) normalization resulted in high-quality networks. 3) Batch effects should be corrected using – by preference – ComBat. 4) Finally, Pearson's correlation is in general to be preferred, but on smaller datasets (< 30 samples) Spearman's rank correlation is in general better.

(3) Applications of our processed dataset as a high-quality reference

Through the optimization of the workflow, we have obtained a high-quality gene expression dataset, covering a wide variety of cell types and tissues in human and mouse. Such a dataset is valuable as a reference. We illustrated this through two applications. In one study, we used our dataset to support the analysis of single-cell and spatial transcriptomics analysis of liver tissues [Vandenbon *et al.*, Commun Biol, 2023]. Our dataset was used to annotated cell types in single-cell data, and for dissecting spatial expression patterns in the spatial transcriptomics data. In a second study, we used our reference dataset to successfully predict differentially expressed genes in various brain-related samples [Vandenbon and Diez, bioRxiv, 2022]. Traditionally, such analysis has been difficult because of small samples numbers and the existence of batch effects. However, here we could leverage the large amounts of samples in our reference data, as well as its high quality.

(4) A high-quality cell type-specific gene co-expression database

Finally, we are implementing a freely accessible gene co-expression database, which allows users to search for genes of interest, and visualize their expression and co-expression in the data obtained from the 68 human and 76 mouse cell types and tissues (Fig. 1). This database will be made freely accessible as soon as possible.

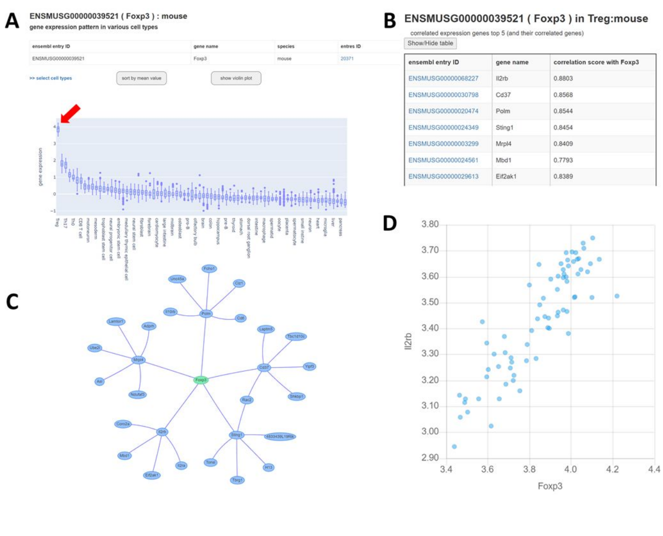

Figure 1: Example usage of the gene co-expression database. (A) Visualization of the expression levels of Foxp3 in mouse cell types. Indicated is the high expression in Tregs. (B) Genes with high correlation of expression with Foxp3 in Tregs. (C) A small co-expression network of Foxp3. (D) Scatterplot showing high correlation of expression between Foxp3 and Il2rb in Treg-derived samples. Several other functions are being implemented (not shown here).

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　2件／うち国際共著　2件／うちオープンアクセス　3件）

１．著者名 Vandenbon Alexis	４．巻 17
２．論文標題 Evaluation of critical data processing steps for reliable prediction of gene co-expression from large collections of RNA-seq data	５．発行年 2022年
３．雑誌名 PLOS ONE	６．最初と最後の頁 e0263344
掲載論文のDOI（デジタルオブジェクト識別子） 10.1371/journal.pone.0263344	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Vandenbon Alexis、Mizuno Rin、Konishi Riyo、Onishi Masaya、Masuda Kyoko、Kobayashi Yuka、Kawamoto Hiroshi、Suzuki Ayako、He Chenfeng、Nakamura Yuki、Kawaguchi Kosuke、Toi Masakazu、Shimizu Masahito、Tanaka Yasuhito、Suzuki Yutaka、Kawaoka Shinpei	４．巻 6
２．論文標題 Murine breast cancers disorganize the liver transcriptome in a zonated manner	５．発行年 2023年
３．雑誌名 Communications Biology	６．最初と最後の頁 97
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s42003-023-04479-w	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Vandenbon Alexis、Diez Diego	４．巻 -
２．論文標題 A universal differential expression prediction tool for single-cell and spatial genomics data	５．発行年 2022年
３．雑誌名 bioRxiv	６．最初と最後の頁 1-32
掲載論文のDOI（デジタルオブジェクト識別子） 10.1101/2022.11.13.516355	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 Vandenbon Alexis
２．発表標題 Evaluation of critical data processing steps for reliable prediction of gene co-expression from large collections of RNA-seq data
３．学会等名 第１１回生命医薬情報学連合大会
４．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

We are preparing a freely accessible gene co-expression database now. We will make it public as soon as possible.

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関