Batch effects correction

Anna Łukasik

Abstract

Currently, there is a lot of biological data from many sources. Sequencing is not a problem, but the greater challenge is to analyze and combine data from different studies. Even if the same sequencing methods and machines from the same factory are used to collect the data, problems can arise. This is because of batch effects. They occur when non-biological factors in an experiment cause variability in groups of samples. They can lead to wrong conclusions. Correcting the batch effects is to remove technical variance when combining data from different batches or from different studies. To better understand the problem, I will take an example. Suppose we have chest computed tomography data. Pictures are used to diagnose lung cancer. The photos are from two hospitals: A and B. Hospital A is in a polluted city so the people who live there get sick more often. In addition, the scanner in the hospital A produces brighter images than the machine in the hospital B. This causes the model might learn to associate brighter images with cancer. What if we implement our model for other data? Misclassification may occur.

Reasons for the presence of batch effects include, but are not limited to, laboratory conditions, reagent selection, the person who performed batches, day of the week, time of day or even ozone level in the atmosphere. In order to deal with batch effects at the design stage of the study, be consistent and test all samples at the same time and avoid situations where cancer cells come from laboratory number 1 and healthy cells from laboratory number 2. Unfortunately, we cannot always meet the requirements. Then notes are necessary then. For example, making a note of which sample the person handled, what the date was.

Ways to deal with batch effects on the computer are multilevel models, empirical Bayesian technique (effective for small batches), sva package, or mutual nearest neighbors. The authors of the publication "Unifying cancer and normal RNA sequencing data from different sources"[1] paid attention to combining data from different sources and removing batch effects by sva package. I reconstructed a piece of analysis contained in the Wang et al. paper[1] and added a few new things, e.g. used a different method to remove batch effects. I used three tissues for the analysis: thyroid, breast and bladder. The entire analysis was done by me in the R language (R version 4.0.3).

Introduction

The aim of the authors of the publication on which this paper was based was to combine data from two sources: The Cancer Genome Atlas (TCGA) and The Genotype-Tissue Expression (GTEx). Projects contain genomic, epigenomic, transcriptomic, and proteomic data that helps in the diagnosis and treatment of cancer. These are public resource to study tissue-specific gene expression and regulation. The authors used data from RNA sequencing. Over 800 samples were from TCGA and over 9,000 from GTEx. Cancer and healthy samples were taken from TCGA, and only healthy samples from GTEx. Raw paired-end reads of RNA-seq samples from the TCGA project were taken by authors from the Cancer Genomics Hub (CGHub, https://cghub.ucsc.edu). GTEx samples were obtained from the Genotype and Phenotype Database (dbGaP, http://www.ncbi.nlm.nih.gov/gap). Next they aligned the reads, quantified gene expression, and removed errors specific to each study. Details are in their

publication[1]. I used processed data sets that are on the authors' Github (https://github.com/mskcc/RNAseqDB/tree/master/data). In a folder named "unnormalized", there is data that has not been adjusted for batch effects, while in a folder named "normalized" it is adjusted using the ComBat tool from R package: sva. The names are confusing. It would be better to use "before batch effect correction" and "after batch effect correction", but for consistency, I use the names the authors used.

In order to detect batch effects, you can visualize the data. One method of visually determining if there are undesirable effects in the data is to perform principal component analysis (PCA). Data dimensions are reducing to a few principal components (PC) that explain the greatest variation. When there are batch effects in data, the scatter plot of the top PCs should highlight a separation of the samples due to different batches. There are many methods for dealing with batch effects. In this analysis, I used R functions: 'ComBat' and ‘removeBatchEffect’.

Methods and results

The first step in my analysis, after loading the data, was to check if the data is normalized. The authors used upper quartile normalization to bring expression levels to comparable ranges. Before performing the procedure gene expression data for TCGA mostly ranged from 4 to 10 (log2 of normalized_count), and from 0 to 4 (log2 of RPKM) for GTEx. I created a graph for each tissue (thyroid, breast and bladder). I used the “normalized” and “unnormalized” data. Below I am pasting only thyroid charts. The rest of the plots are included in the supplementary materials. In all the graphs the expression numbers are on comparable levels, no matter they are from TCGA or GTEx.

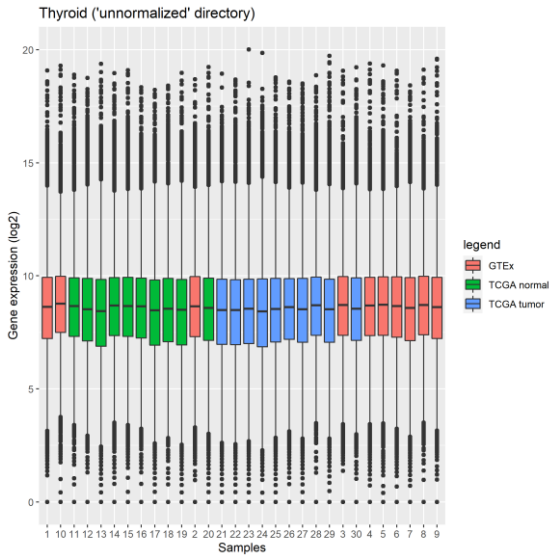

Figure 1. Visualizing data from a folder that contains data that has not been corrected for batch effects.

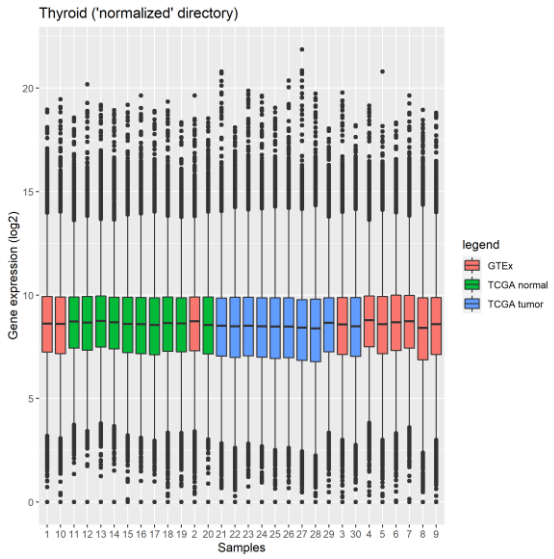

Figure 2. Visualization of data from a folder containing data corrected for batch effects by Wang et al.

In order to speed up the calculations, not all samples were used. I used 10 of them from the GTEx data, 10 from TCGA normal and 10 from TCGA tumor. I also checked which genes are common to all datasets and only these were used for further analysis.

Later it was necessary to create a new structure. A large dataframe. The rows were genes and the columns were samples. The dataframe contained three tissues: thyroid, breast and bladder. I have created an auxiliary table called description. It contained information about the sample number, tissue (for example: 'thyroid GTEx', 'thyroid TCGA', 'breast GTEx', 'breast TCGA',…), whether the sample is normal or cancerous, and batch information. A batch equal to 1 was assigned to the samples from the GTEx data and a batch of 2 was connected with the data from TCGA, as in baseline paper[1]. After creating the dataframes, I was able to use the svd function. I applied it to the first 1,700 genes. Then I plotted two first principal components. I did it in four ways. The first one was that the clusters were marked with ellipses in the picture (Fig.3.). The three consecutive plots contained only dots colored according to tissue type, cancer or 'batch' value. I used the ComBat function to correct the batch effects and drew the same graphs. The difference was especially noticeable for the last types of charts. Before correction, the figures showed visible clusters for batch 1 and batch 2 (Fig.5.). Cancer and normal samples also were separated (see Supplementary Materials), but this was due to all GTEx samples (batch=1) being normal so the batch information and whether the cell was cancerous were related. The tissues were not separated at all (Fig.3.). After correcting the batch effects, batch clusters were not visible (Fig.6.), while the thyroid clearly separated from the rest of the tissues (Fig.4.).

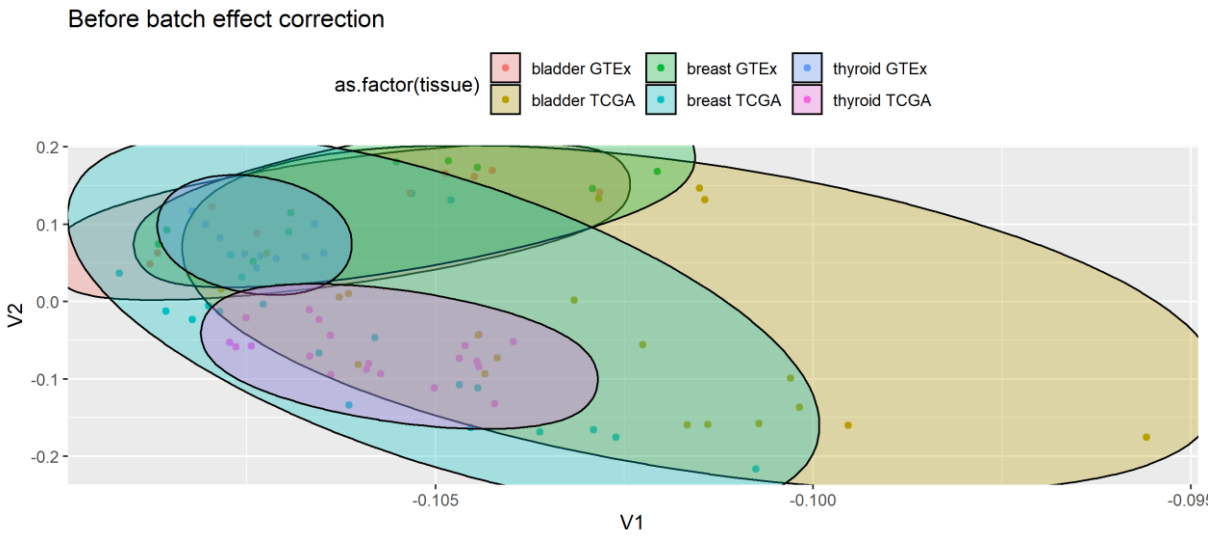

Figure 3. Presentation of the first two components before batch effects correction. Tissues close together.

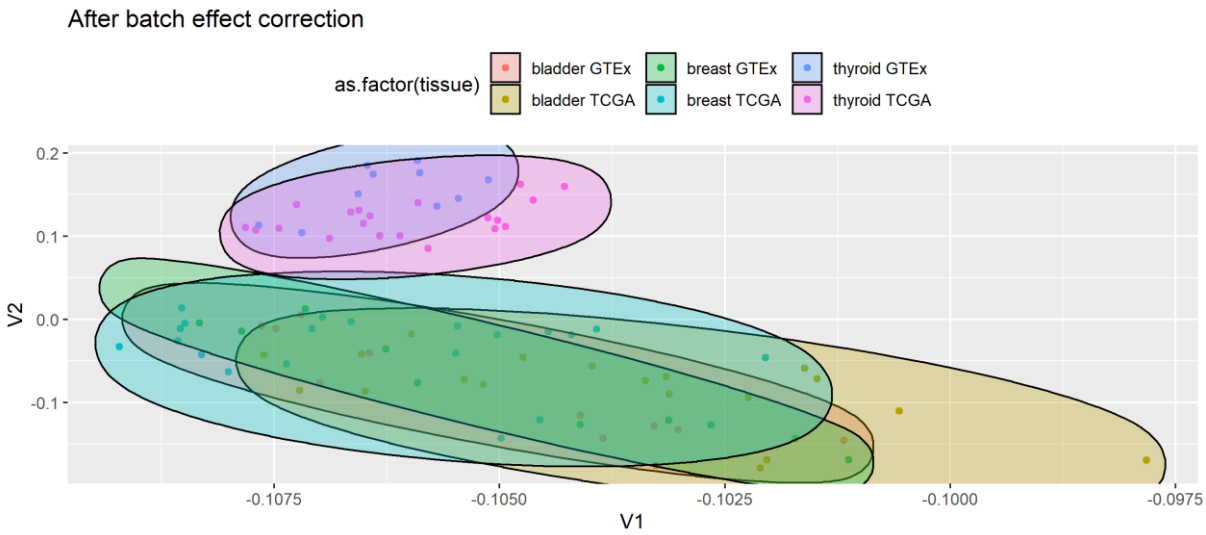

Figure 4. Presentation of the first two components after batch effects correction. Thyroid clearly separated.

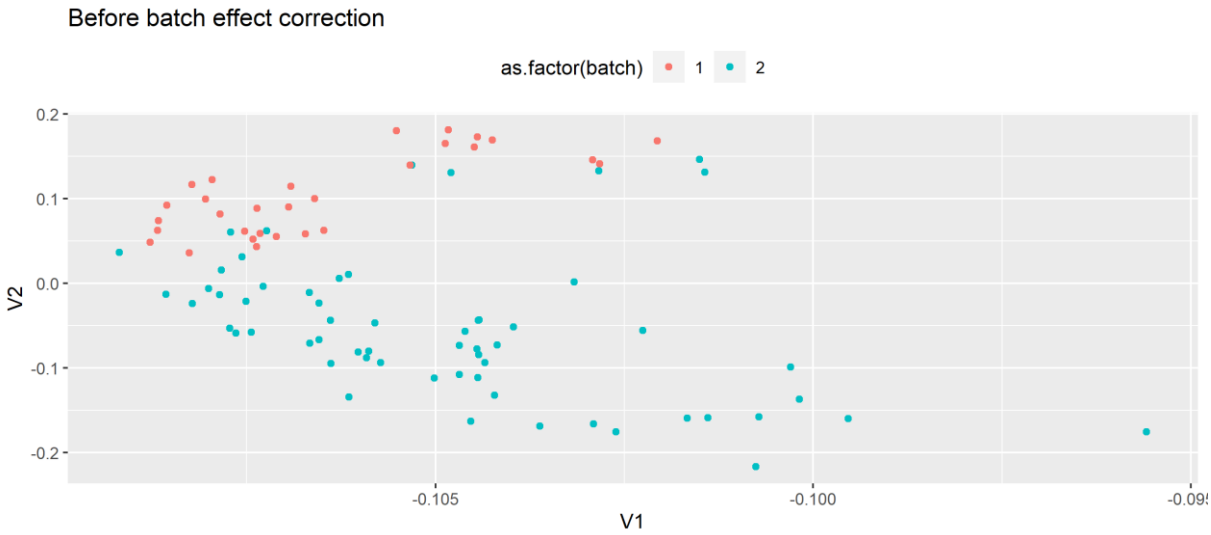

Figure 5. Presentation of the first two components before batch effects correction colored by batch.

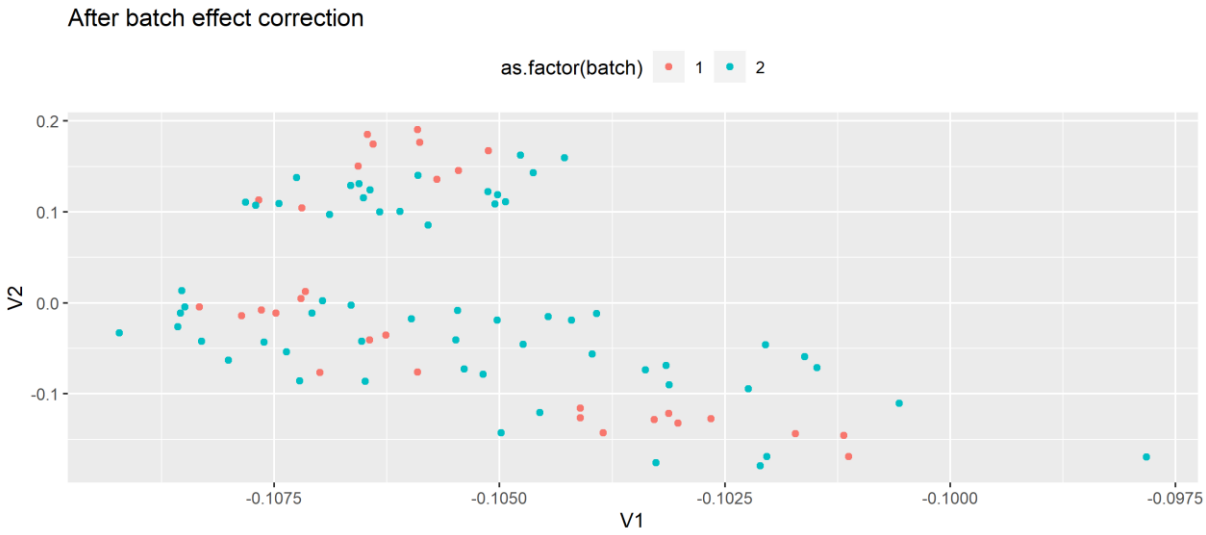

Figure 6. Presentation of the first two components after batch effects correction colored by batch.

I made the same graphs (see Supplementary Materials) for the data from the "normalized" folder, in which the authors of the publication 'Unifying cancer and normal RNA sequencing data from different sources'[1] placed the data that they previously corrected for batch effects. They look like the graphs I created after removing the batch effects with the combat function, so the analysis was successful. The 'removeBatchEffect' function from the 'limma' package was also used to correct batch effects. The results were the same as for the ComBat function.

Discussion

It is thought-provoking why the correction of the batch effects caused separation the thyroid from other tissues, but the bladder and breast were still close together. In the original work, the authors analyzed the prostate instead of the breast. All three tissues were then separated from each other, but the thyroid was much more distant from the other two organs. This was explained by the fact that the prostate and the bladder are similar in terms of their functions. Perhaps the breast and bladder also share some biological properties such that the clusters do not clearly separate from each other.

In conclusion, when analyzing RNA sequencing data and other data from different sources, it is important to remove batch effects. Otherwise it can lead to misinterpretation and incorrect classification. This is especially important in medical applications. There must be no mistakes in determining whether someone is healthy or sick.

Bibliography

[1] Wang, Qingguo et al. "Unifying cancer and normal RNA sequencing data from different sources." Scientific data vol. 5 180061. 17 Apr. 2018, doi:10.1038/sdata.2018.61