ECBM Final Project, , pp. 1–5

doi: ECBM 4060
Date: 16 December 2023

ECBM Final Project: Individual-level analysis of differential expression of genes and pathways for personalized medicine

Sabrina Wang (sw3693), Derek Ning (dn2546) and Christopher Lung (cl4301)

Abstract

Our final project is based upon analysis of a paper published in 2014, titled: "Individual-level analysis of differential expression of genes and pathways for personalized medicine", written by Hongwei Wang and Zheng Guo et al. This paper describes a method known as RankComp whose purpose is to detect differentially expressed genes in individual disease samples. We validated this function using a dataset of 91 samples, consisting of 46 nonsmall cell lung cancer tumor samples and 45 control samples. Using the RankComp function, we identified 3 significantly upregulated genes that corresponded to cancer development and tumorigenesis. However, comparing these results to differential gene expression analysis with edgeR showed no overlap. The methods, analysis, and results of our project are detailed below.

GitHub Repository

The code written and the data generated during the analysis can all be find in the GitHub Repository below:
https://github.com/SabsW/ECBM

Introduction

There are several challenges associated with identifying disease related genes when performing differential gene expression analysis. Intensity and ranking based methods are the two primary methods which entail ranking fold-change values within each pairwise comparison between two types of samples, then calculating a fold-change rank ordering statistic for each gene. Relative to the intensity-based methods like the T-test or Significance Analysis of Microarray (SAM), rank-based tests are robust against outlier values and can offer a higher power test while utilizing smaller sample sizes. However, both of these statistical methods are designed to detect the population-level DE genes and are unable to provide patient-specific differential expression information due to issues with batch effects, which are non-biological artifacts in data related to differences in factors such as laboratory, materials, and personnel. There is a crucial need for differential expression performed at the individual level in order to provide the information needed for personalized medicine.

While data normalization is able to ensure data at the same scale as training samples, it cannot remove batch effects which can result in problems in translating experimental findings to a clinical setting. This paper by Zheng Guo et al. proposes using relative ordering information of gene expression within each sample since it is robust against batch effects and insensitive to data normalization. The method introduced in this paper is able to detect DE genes in individual disease samples by utilizing disrupted ordering in individual disease samples, all while achieving satisfactorily high levels of sensitivity, specificity, and F-score. Taking paired cancer-normal data, this method was able to identify and apply prognostic markers to risk stratification of lung cancer patients according to dysregulation status of signature in each patient rather than having to predefine a risk value. Bypassing the need to set an optimized risk score threshold value to summarize gene expression levels between different sample sets is significant because the existing batch effects can prevent direct comparisons of data sets with technical artifacts, even with the use of normalization.

Background

The primary advantage of the RankComp method is that it takes advantage of the relative ordering of gene expression, which is overall stable in a particular type of normal human tissue but widely disturbed in diseased tissue. The paper shows that this method has excellent performance for individual-level analysis of dysregulated genes and pathways.

Specifically, the relative ordering method would entail ranking each gene expression value low to high and then performing pairwise comparisons for all genes to isolate the stable ordered gene pairs. Each gene pair (G_i, G_j) can have two possible outcomes, the frequency of both are listed as follows:

$$P_{norm}(G_i > G_j) = \frac{1}{n_1}\sum_{t=1}^{1} I[G_{it} > G_{jt}]$$

$$P_{norm}(G_i < G_j) = \frac{1}{n_1}\sum_{t=1}^{1} I[G_{it} < G_{jt}]$$

Gene pairs that are defined as stable are defined as having a P value > 0.99. Reversal gene pairs are next defined by flipping the comparison (greater than/less than). The Fisher's exact test is then used to determine differential expression of

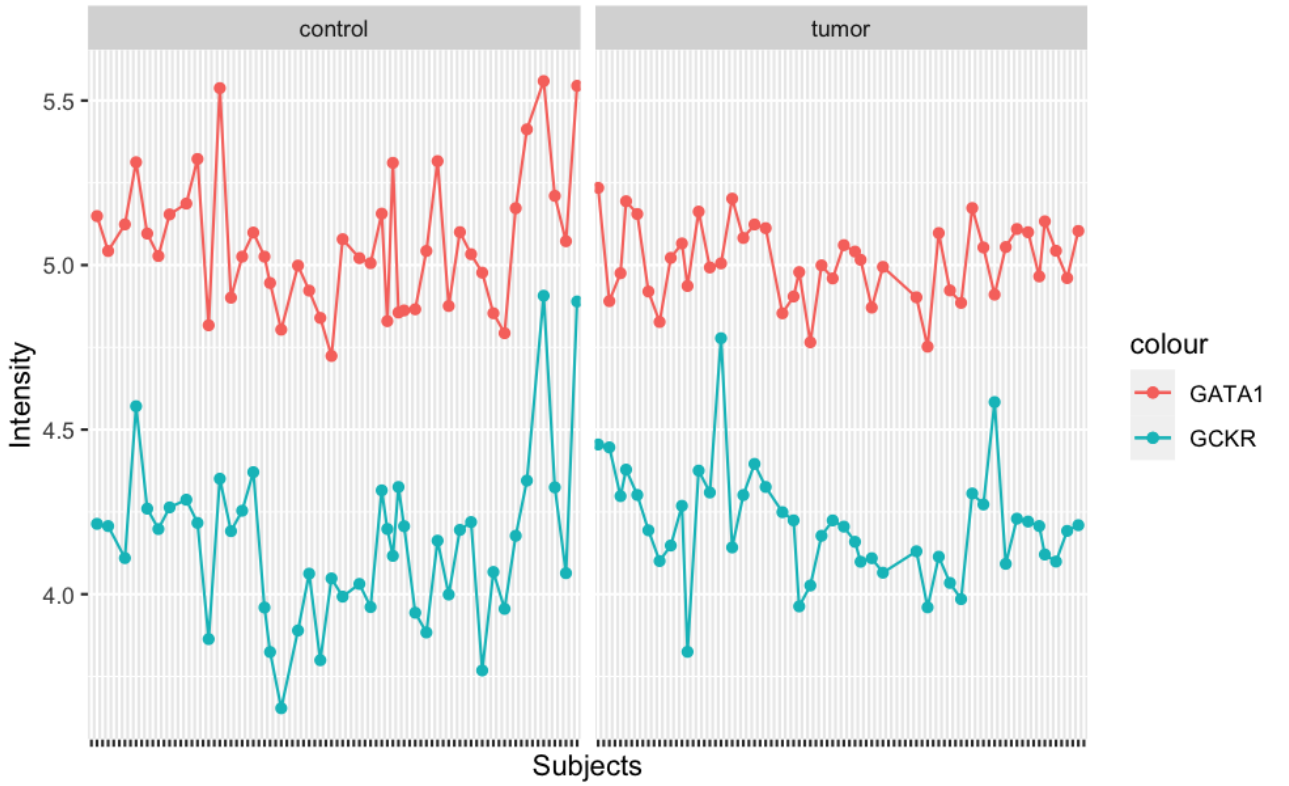

Fig. 1. Expression distributions of 2 selected genes (GATA1, GCKR) in normal control sample vs tumor sample. Note that the relative ordering of GATA1 with higher levels of expression intensity is common to both the control and tumor samples, and this stable expression ordering of gene pairs is a crucial component in the RankComp function used to identify differential expression analysis of genes at the individual level.

a given gene in a specific sample. The null hypothesis tested is that the number of reversal gene pairs that support up-regulation and down-regulation are equal. If the ordering in normal samples and disease samples are consistently opposite of each other, then the gene pair is considered to be supporting up/down regulation of the gene in the sample. The final step of the process is to apply a filter in order to minimize the potential effect of expression changes of other genes. Therefore, only if a gene is still significant after excluding the downregulated/upregulated partner genes involved in the reversal gene pair will it be retained.

Finally, the RankComp Method can be evaluated against the paired disease-normal samples and a precision score ratio for the consistent DE genes to all DE genes can be determined.

Table 1. Top 10 Differentially Expressed Genes Identified by RankComp

Downregulated genes	Upregulated genes
TCF21	TOP2A
TCF21	NEK2
INMT	TTK
PHACTR1	GINS1
CLIC5	BUB1
CLEC3B /// EXOSC7	ASPM
ADAMTS8	E2F8
FGR	KIF14
PEBP4	CENPF
GLIPR2	TOP2A

Results and Analysis

Following the successful execution of the RankComp algorithm on the data set, two sets of matrix data were generated: a matrix representing the type of gene regulation for each sample and each gene, and another matrix that displayed the p-values resulting from the Fisher's exact test applied to the data. The null hypothesis used in the statistical test to calculate the P value is that the number of genes up-regulated or down-regulated relative to the tested gene has no association with the presence or absence of tumor/disease. In other words, the numbers of upregulation and downregulation reversal gene pairs are equal. The tested genes were classified into either up-regulated, down-regulated, and non-dysregulated, which are represented by +1, -1, and NA respectively in the first matrix. In order to identify the top up-regulated genes from our results, we looked at the rows which had a +1 up-regulation and sorted the genes by the lowest p-value that indicated statistically significant differences.

Table 1 shows the top 10 most down and up-regulated genes identified by the RankComp method, which are the up or down-regulated genes that have the lowest p-values.

In this discussion, we focused on identifying the up-regulated gene pathways. The top 3 up-regulated genes identified by RankComp after the filtering process are TOP2A, NEK2, and TTK. Figure 2 shows the expression levels of those 3 genes in the tumor samples in comparison to normal samples, demonstrating that the RankComp method was successful

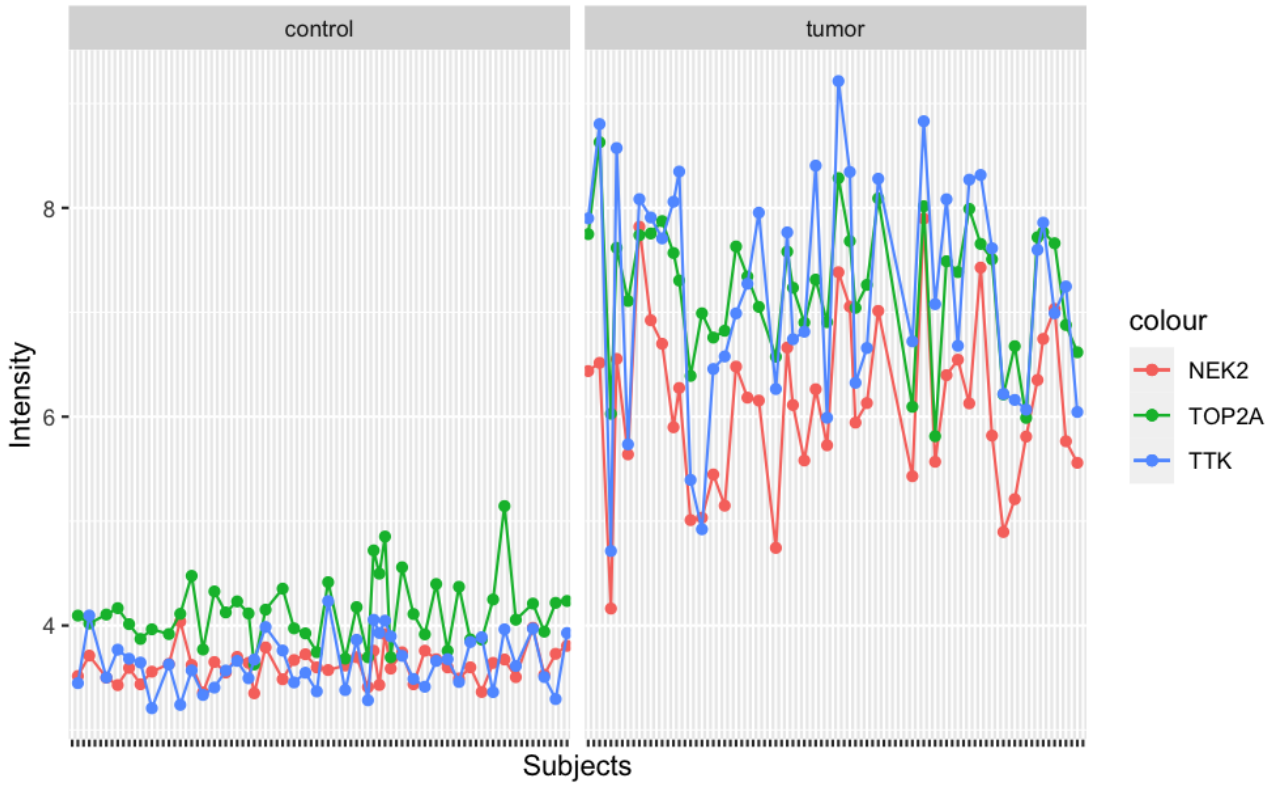

Fig. 2. Expression level of the top 3 upregulated genes in the tumour samples compared to healthy samples as identified by the RankComp method. The expression level of all 3 genes is significantly elevated in the tumour samples compared to the healthy samples.

in identifying those 3 genes as differentially expressed and significantly upregulated.

These results were consistent with our expectations for the types of genes that will be up-regulated with regards to cancer. In particular, all these genes are related to enzymes and protein kinases that control cell developmental regulation and checkpoint mechanisms in cell proliferation/mitosis. TOP2A is a gene for a DNA enzyme that is responsible for monitoring and altering DNA states during transcription. Mutations of this specific gene usually is associated with development of drug anticancer resistance. NEK2, a gene that encodes for the protein kinase necessary for mitotic regulation, is known to exhibit up-regulation in various cancer cell lines. Lastly, the TTK gene is a protein kinase that is crucial for cell proliferation and chromosome alignment during cell division, and a degradation in this protein may lead to tumorigenesis and anomalies in mitotic spindles. From the results of these top up-regulated genes and the normalized log intensity for for those 3 genes in Figure 2, it was evident that the expression was clearly distinct and stronger in intensity relative to the control healthy subjects. The subjects were aligned along the x-axis of that figure, with each dot representing a single subject. Clearly, each and every subject had higher levels of expressions of NEK2, TOP2A, and TTK for the tumor samples as compared to control. This observation was significant in our study to apply the RankComp method to our data set because it demonstrated the ability to conduct differential expression analysis at the individual-specific level and achieve a consistent output of the differentially expressed genes using the disrupted ordering method.

When comparing our methods and results to the original research paper, there are some similarities and differences. Here in Figure 1, we demonstrated the clear advantage of stable ordering of gene pairs for the GATA1 and GCKR genes between tumor and control for our lung tissue samples. In the paper, they compared the intensity of these same two pairs of genes that exhibited stable expression ordering from lung tissue samples across multiple different datasets. We decided to focus on only one specific dataset to analyze in this study because our data already contained large amounts of samples, and to pre-process, apply RankComp, and analyze the results required adequate time to test and discuss. Regardless, the main idea that the relative ordering of gene expression is unaffected by distribution variations was maintained between both of these representations, which is essential to the operating principle of why the RankComp method works.

Additionally, we were able to qualitatively identify top up-regulated and down-regulated genes like the original paper suggested, by ranking it by logFC and p-value. Reasonably, we obtained differentially expressed genes that are closely related to tumorigenesis and cell proliferation regulation, so our main objective of applying RankComp to identify DE genes on an individual-basis was achieved. Overall, we were successful in identifying prognostic lung cancer markers in this study based on the dysregulation signature status for each patient, which aligned with the same objective described in the Bioinformatics article.

Lack of Overlap with edgeR Analysis

While our RankComp analysis returned promising results in terms of identifying genes clearly upregulated in tumor samples, comparison to edgeR analysis showed no overlap between the top genes identified by highest logFC value. This could be attributed to differences in the statistical test performed between the RankComp and the edgeR statistical analysis. We hypothesized that another issue could involve the differences in dataframe manipulation required to run the data on edgeR versus RankComp.

Challenges

One major challenge we faced was learning how to run the RankComp package developed by the original authors of the paper. Since the paper was published in 2015, the RankComp package was based off of version 2 of R which was not compatible with RStudio. Newer versions of the RankComp package were also based upon Linux only which was a key limitation. We chose to troubleshoot and learned how to successfully install R 2.15.3 as well as navigate the console and software without RStudio.

Another challenge was processing and running the dataset. The documentation for the RankComp package is relatively slim and we needed to understand how to properly format and splice the data in order to run RankComp properly (Refer to Source Data and Data Processing Section). We also did not anticipate constraints regarding the memory and time requirements. The resulting program performed RankComp comparisons at a rate of around 2 per second adding up to >54,000 calculations. We had to work around limited memory constraints and running the dataset took in total almost 24 hours to complete.

The final major challenge involved exporting the data into a format that we could analyze. The resulting RankComp output is in a custom RCtest class and methods such as catch.output and sink() did not work in properly exporting the test results. After trial and error we managed to export our data successfully by indexing the RCtest class and building our own dataframe. The resulting analysis was completed in RStudio.

Methods

Source Data

The dataset used in the results and analysis is sourced from this paper published in 2010 titled "Gene expression profiling revelas novel biomarkers in nonsmall cell lung cancer" by Sanchez-Palencia and Farez-Vidal et. al. This dataset describes microannalysis on a set of 46 tumor samples and 45 paired nontumor samples of nonsmall cell lung cancer (NSCLC) totalling 91 samples. This dataset was chosen because it was also used by the author of the original paper as normal sample data in order to identify stable gene pairs. The data was obtained from Gene Expression Omnibus (GEO), series GSE18842.

Data Processing

The data was obtained using Bioconductor's GeoQuery package in Rstudio. The data extracted for the purpose of this analysis included sample metadata, gene expression data, and corresponding Entrez gene IDs and gene symbol names.

There are three data structures required to run the RankComp function: 1 data frame of gene expression, 1 vector of corresponding Entrez gene IDs, and 1 vector of sample labels (tumour vs. healthy). The source data obtained from GeoQuery is then cleaned up and structured in the required way.

RankComp Package and R 2.15.3

As mentioned in the Challenges Section, we had issues running and exporting our analysis. Since the paper was published in 2015, the subsequent RankComp package the authors utilized was also based off of R version 2. A more updated package was available on GitHub, but was only functional on Julia. Subsequently we downloaded R 2.15.3 and installed the RankComp package located in the supplementary information.

The RankComp package consists of a single function, titled RankComp with arguments expdata, label, gene, and freq.

$$\text{Function Input: } RankComp(expdata, label, gene, freq)$$

$$\rightarrow \text{Function Output: } [[Pvalue], [Direction]]$$

Expdata is a numeric matrix of data values while gene corresponds to a numeric vector of Entrez gene IDs corresponding to the number of rows in expdata. Finally, label is a numeric vector of values where '0' represents a control sample label and '1' represents disease sample. The output of this RankComp function produced a list with class "RCtest" containing two elements: Pvalue represented the p-value of the Fisher's exact test, and Direction represented the dysregulation direction for each gene in each disease sample. We processed the data obtained from GEO accordingly and ran the function, which took 24 hours to process >54,000 entries. This data was exported as a CSV file into Rstudio and analyzed.

Differentially Expressed Gene Analysis using edgeR

We used Bioconductor's edgeR package to analyse differentially expressed genes in the dataset to compare with the results obtained using the RankComp methods. The dgelist object was created using the expression data and the metadata obtained from the GSE18842 dataset using GeoQuery.

Conclusion

Scientific Findings

Unlike previous analysis methods such as intensity-based and ranks-based tests used to identify differentially expressed genes in the population level, the proposed method of detecting diseased samples on an individual level using RankComp and relative ordering of gene expressions was explored and applied in this study. The relative ordering idea is effective because it is observed to be stable in normal human tissues but drastically perturbed in diseased tissue. By combining this method with the cancer-normal paired data that we inputted into the algorithm, we were able to isolate dysregulated genes and pathways for each individual. Additionally, this approach has the advantage of mitigating batch effects from differences in laboratory experimental setup and data collection between different datasets, because the relative ordering is insensitive to data normalization.

As seen in the results and methods above, we started out with a large dataset of normal and diseased lung tissue samples found on the Gene Expression Omnibus database, and processed the data before passing all the expression and label data into the input of RankComp. From the function, two important sets of outputs were generated: the regulation list

that indicated the type of gene regulation and a p-value list that displayed the statistical measurement of the Fisher's exact test. Filtering and organizing this data, we summarized our findings of the top 10 up-regulated and down-regulated genes into Table 1. From the qualitative analysis and discussion of the top 3 up-regulated genes, it was evident that these genes are directly correlated to the tumor sample, as those genes were closely associated with normal cell development and proliferation. In the end, we were able to achieve the desired project goal of using relative ordering to identify DE genes on an individual scale, while validating that there is a significant difference between the tumor and control samples when focusing on the intensity of particular genes.

In addition to the individual-level differentially expressed gene identification using RankComp, another application that can be tested in future studies is to identify DE genes at the subpopulation level, a feature that other outlier detection methods are capable of. The RankComp method can be adapted to detect DE genes in a more broader subpopulation level rather than individual level, and the accuracy and performance of this algorithm to detect potential genetic outliers in the data can be compared to the existing methods for further study.

Most importantly, the results of this individual analysis can translate to personalized medicine that better targets the patient based on the presence of the dysregulation status of each patient's signature genome. In this way, the heterogeneity of an individual's genome can be explored in depth, rather than overlooked as seen in the regular DE analysis at the population level.

Project Adaptions

In our original project proposal, we discussed the goals for our project as well as some forseeable challenges. Throughout the course of this project we learned a lot by adapting from our original goal to reach a meaningful conclusion.

One issue we anticipated was that the dataset itself might contain genes with similar differential expressions that would be too specific for RankComp. We overcame this by deciding to selected a dataset used by the authors of the original paper and were able to find 3 genes with significant upregulation.

Initially, we intended to apply the RankComp algorithm to a class dataset and compare the results from RankComp with results from the differential gene analysis we learned in class. However, due to issues in data processing and limitations in memory/time, we decided to instead use a dataset from the paper and compare the RankComp algorithm to the differential gene analysis performed in class. It was valuable going through the trial and error process testing different datasets from class and from the paper in order to better understand the RankComp package.

Competing interests

No competing interest is declared.

Author contributions statement

All authors contributed equally to the analysis and the manuscript.

Acknowledgments

The authors thank Professor Wei Yi Cheng and Keren.

References

Li, X Cai, H Wang, X Ao, L Guo, Y He, J Gu, Y Qi, L Guan, Q Lin, X & Guo, Z (2019) A rank-based algorithm of differential expression analysis for small cell line data with statistical control. *Briefings Bioinf.*, *20*(2) 482–491. https://doi.org/10.1093/bib/bbx135

pathint. (2022a) RankCompV3.jl [Online; accessed 15. Dec. 2022■] https://github.com/pathint/RankCompV3.jl

pathint. (2022b) reoa [Online; accessed 15. Dec. 2022■] https://github.com/pathint/reoa/tree/master/bin

Sanchez-Palencia, A Gomez-Morales, M Gomez-Capilla, J A Pedraza, V Boyero, L Rosell, R & Fárez-Vidal, M E (2011) Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *Int. J. Cancer*, *129*(2) 355–364. https://doi.org/10.1002/ijc.25704

Wang, H Sun, Q Zhao, W Qi, L Gu, Y Li, P Zhang, M Li, Y Liu, S.-L & Guo, Z (2015) Individual-level analysis of differential expression of genes and pathways for personalized medicine. *Bioinformatics*, *31*(1) 62–68. https://doi.org/10.1093/bioinformatics/btu522