APPLICATION OF QUANTILE DISCRETIZATION AND BAYESIAN NETWORK ANALYSIS TO PUBLICLY AVAILABLE CYSTIC FIBROSIS DATA SETS[a]

Kiyoshi Ferreira Fukutani and Thomas H. Hampton
Geisel School of Medicine, Dartmouth College, 1 Rope Ferry Road
Hanover, 03755, NH, USA
Email: kiyoshi.ferreira.fukutani@dartmouth.edu and Thomas.H.Hampton@dartmouth.edu

Carly A. Bobak
Research Computing and Data Services, Dartmouth College, 1 Rope Ferry Road
Hanover, 03755, NH, USA
Email: carlybobak@dartmouth.edu

Todd A. MacKenzie
The Dartmouth Institute for Health Policy and Clinical Practice, Dartmouth College, 1 Rope Ferry Road
Hanover, 03755, NH, USA
Email: Todd.A.MacKenzie@dartmouth.edu

Bruce A. Stanton
Geisel School of Medicine, Dartmouth College, 1 Rope Ferry Road
Hanover, 03755, NH, USA
Email: Bruce.A.Stanton@dartmouth.edu

The availability of multiple publicly-available datasets studying the same phenomenon has the promise of accelerating scientific discovery. Meta-analysis can address issues of reproducibility and often increase power. The promise of meta-analysis is especially germane to rarer diseases like cystic fibrosis (CF), which affects roughly 100,000 people worldwide. A recent search of the National Institute of Health's Gene Expression Omnibus revealed 1.3 million data sets related to cancer compared to about 2,000 related to CF. These studies are highly diverse, involving different tissues, animal models, treatments, and clinical covariates. In our search for gene expression studies of primary human airway epithelial cells, we identified three studies with compatible methodologies and sufficient metadata: GSE139078, Sala Study, and PRJEB9292. Even so, experimental designs were not identical, and we identified significant batch effects that would have complicated functional analysis. Here we present quantile discretization and Bayesian network construction using the Hill climb method as a powerful tool to overcome experimental differences and reveal biologically relevant responses to the CF genotype itself, exposure to virus, bacteria, and drugs used to treat CF. Functional patterns revealed by cluster Profiler included interferon signaling, interferon gamma signaling, interleukins 4 and 13 signaling, interleukin 6 signaling, interleukin 21 signaling, and inactivation of CSF3/G-CSF signaling pathways showing significant alterations. These pathways were consistently associated with higher gene expression in CF epithelial cells compared to non-CF cells, suggesting that targeting these pathways could improve

clinical outcomes. The success of quantile discretization and Bayesian network analysis in the context of CF suggests that these approaches might be applicable to other contexts where exactly comparable data sets are hard to find.

Keywords: Cystic Fibrosis, Bayesian Network, Data.

[a] This work was supported by funding from the Cystic Fibrosis Foundation to B.A.S. (STANTO19G0, STANTO20P0, STANTO23R0 and STANTO19R0), the National Institutes of Health to B.A.S (P30-DK117469 and R01HL151385) and the Flatley Foundation.

1. Introduction

Worldwide initiatives are currently discussing the principles of acquiring, standardizing, storing, and making scientifically produced data accessible for reuse. However, one of the key difficulties is addressing the heterogeneity of the data, which is called batch effects. These batch effects occur when we compare multiple datasets obtained from different laboratories, platforms, or processed at different time points. These internal differences can lead to misinterpretations of the results and it is not only a common issue in omics data analysis but in many cross-study comparisons.[1,2] In recent years, there has been increasing consideration of batch effects in data analysis and several approaches have been proposed to address them.[3] The simplest way to handle batch effects is to include them in the statistical model during analysis. Other approaches involve estimating and creating a new dataset adjusted by batch effects, to perform the statistical analyses.[4] However, it is important to note that this technique can reduce statistical power, particularly when the batch-group is unbalanced, meaning that batch differences may be influenced by group differences. This correction can either diminish group differences or introduce new batch effects due to errors in batch effect estimation that may be inflated by false positives.[5]

Cystic fibrosis (CF) is a recessive genetic disorder characterized by alterations in electrolyte transport across polarized epithelia resulting from mutations in the CF transmembrane conductance regulator gene (*CFTR*).[6] Numerous studies on CF have identified similarities or specific gene signatures that are closely related.[7,8] However, the amount of available transcriptomic datasets for reanalysis and comparison is continually growing.[9] Integrating data from diverse sources can provide a more comprehensive understanding of underlying biological processes that may not be evident from individual studies alone, especially when dealing with multiple conditions and distinct variables.[10] The Meta-analysis instrument of individual microarray studies on CF can help assess the connections between respiratory disorders at the transcriptomic level and provide insights for pathway analysis, but deal with several conditions like: usage of antibiotics, type of mutations, infections by virus or bacteria.[10]

Meta-analysis is a statistical tool that allows the analysis of results from different scientific studies conducted in different locations or by using different methods.[11] In the late 1990s, network meta-analysis (NMA), also known as multiple-treatments or mixed-treatment comparison meta-analysis was introduced as an extension to standard meta-analysis[12]. NMA can compare multiple treatments simultaneously, even when direct comparisons are lacking in

existing studies.[12] One systematic review of NMA methods found that around two-thirds of NMA studies utilized a Bayesian approach.[13] The Bayesian network (BN) models are promising in the medical field because they represent the relationships between variables based on real-world, making them more contextually meaningful than purely numeric associations[14] It has been used in various areas of medical science and can include different types of variables, such as clinical, diagnosis, prognosis, and symptoms.[15] This versatility allows researchers to integrate prior beliefs with sample data and BN analysis has recently been utilized in epidemiology, public health, and medicine.[13,16] On the other hand, there is limited knowledge about BN meta-analysis, which may be attributed to researchers' lack of understanding or familiarity with Bayesian methods. Nevertheless, there is significant potential for the application of BN meta-analysis in medicine.[12]

Standard meta-analysis only allows for comparing two interventions at a time, whereas BN Meta-analysis enables the inclusion of evidence from both direct and indirect comparisons in a single analysis.[12] However, BN analysis interpretations still require specific assumptions for accuracy of the algorithm learning and interpretation of network structure, making it a challenging task.[17] To address these issues inherent in Meta analysis, our study proposes a novel approach to pairing multiple transcriptomic datasets by quantile discretization and integrating metadata variables in a new BN Meta Transcriptomic analysis. This approach aims to provide new and valuable insights into understanding the complexities of a multifactorial disease like CF.

2. Methods

2.1. *Data Selection*

We accessed datasets available in the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) by searching the keyword "cystic fibrosis". A total of 17 datasets were returned by this query, which was performed in November 2022. Nine datasets were excluded from further analysis due to methodological incompatibility or insufficient metadata, which involved the use of different cell tissues or experimental designs and did not measure the same patients variables. We retrieved metadata for these three studies. Three of these studies measured gene expression in airway epithelial cells. The first dataset (PRJEB9292), published by Balloy et al.,[18] included both non CF and CF epithelial cells infected with Pseudomonas aeruginosa for different time points. The second dataset (GSE139078) [19] involved epithelial cells from CF patients infected by Rhinovirus or control and treated with Ivacaftor or Lumacaftor/ivacaftor, modulator drugs used to enhance the functional of CFTR. The third study[20] included two datasets: a pilot dataset with 13 samples and a validation dataset contained 35 samples. All datasets provided patient genotype, modulator information, and infection status with either *Pseudomonas aeruginosa* or Rhinovirus.

2.2. *Data Harmonization and Analysis*

The metadata description included means and standard deviation for numeric variables and frequencies and percentages for categorical data. RNAseq datasets were individually

normalized by library size and log CPM (count per million) transformation and differential expression analyses were performed individually for each dataset using deseq2.[21] In the Balloy dataset, we compared CF vs. non-CF infected or not infected with *Pseudomonas aeruginosa*; in the De Jong dataset, CF epithelial cells infected with virus or not infected with virus; and in the Salas dataset, epithelial cells of CF patients compared to non-CF subjects. In this exploratory design, the DEGs were used to filter the large number of targets, and they were determined by applying specific criteria: genes with a P-value less than 0.05 and a log2 expression fold change greater than 1 or less than -1 were considered as differentially expressed. These criteria were chosen to serve as a filter and help reduce processing time. Each study was normalized individually, and each gene was discretized according to sample distributions. The count table with filtered genes were discretized into quartiles (1st - Minimal to 25%, 2nd - 25% to 50%, 3rd - 50% to 75%, and 4th - 75% to maximum values by sample distribution) using Hartemink's algorithm, which is available in the bnlearn package.[22,23] Afterward, all the transformed transcriptomic datasets were merged into a single discretized dataset, to which metadata was added. The learning algorithm used to establish the Bayesian network structure was based on the heuristic Hill climb method.[24,25] Bayesian network learning was used to visualize conditional dependencies between multiple clinical and transcriptome variables.[26] The dependencies are represented qualitatively by a directed acyclic graph where each node corresponds to a variable and a direct arc between nodes represents a direct influence. Robustness of the arcs was scored by a non-parametric bootstrap test (100×replicates).[27] For functional analysis of genes related to CF, virus infection, bacterial infection, and use of modulators, enrichment pathway analysis was performed using the clusterProfiler package and REACTOME geneset.[28,29] For the Pathway meta-analysis we use the qusage package.[30] All analyses were performed in R version 4.0.2[24] and the Bayesian network and discretization scripts are available in github (https://github.com/FfKB/BNCF). Figure 1. presents a summary of the study selection process and experimental design.

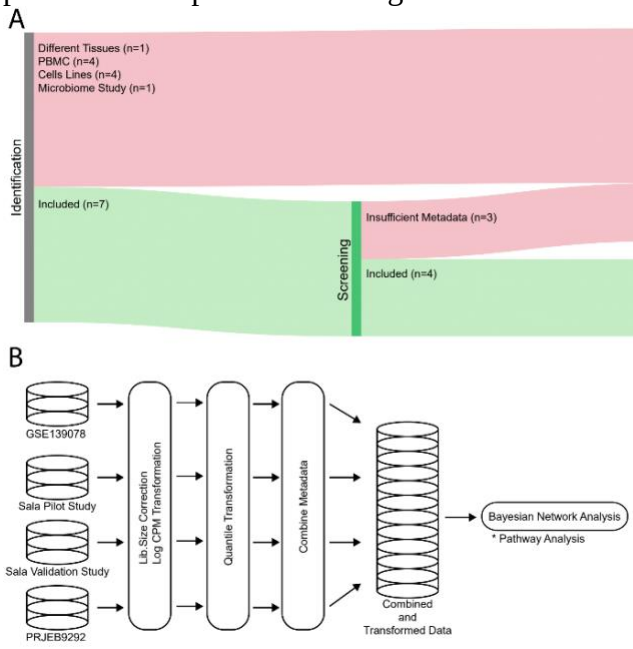

Figure 1. Experimental Design. A) Diagram illustrating the study selection process using a Sankey diagram. The excluded datasets are highlighted in red, while the eligible datasets are highlighted in green. B) Flowchart depicting the data processing steps in the study.

3. Results

3.1. *Study descriptions*

A total of three studies comprising four datasets were considered for analysis: GSE139078, Sala Study, and PRJEB9292. The GSE139078 dataset consists of CF patients who were infected with rhinovirus (RHV). The PRJEB9292 dataset includes four patients divided into four time points, enabling a comparison between gene expression in non CF subjects and CF patients infected with *Pseudomonas aeruginosa*. The Sala study included two datasets: the pilot study and the validation study, which involved a comparison of gene expression profiles between CF patients and non CF subjects. The analysis also includes the assessment of modulator use (Lumacaftor and Ivacftor; and Ivacftor alone) in three datasets (GSE139078, Sala Pilot, and Sala Validation). All CF patients included in these studies have the F508del/F508del genotype, a common genetic mutation (~50%) associated with CF. However, sex and age data were not available for all the datasets, thus, that metadata was not included in the Bayesian Network Analysis. These carefully selected datasets provide comprehensive insights into gene expression patterns related to CF, considering factors such as viral and bacterial infections and the influence of modulators (Table 1).

Table 1. The characteristics of subjects from the selected datasets.

	GSE139078	Sala Pilot	Sala Validation	PRJEB9292
Male sex, n(%)	48 (84.2)	-	-	-
Age, mean (SD)	3.4 (1.4)	35.3 (5.3)	34.1 (8.2)	-
Infection by virus, n(%)	38 (66.7)	-	-	-
Infection by *P. aeruginosa*, n(%)	-	-	-	32 (100)
Cystic Fibrosis, n(%)	57 (100)	7 (53.8)	24 (68.6)	4 (50)*
Modulators (Luma/Iva), n(%)	10 (17.5)	2 (15.4)	10 (28.6)	-
Modulators (Ivacaftor), n(%)	9 (15.8)	0 (0)	2 (5.7)	-
Genotypes F508del, n(%)	57 (100)	7 (53.8)	24 (68.6)	16 (50)

* = 4 Patients in 4 different timepoints (0, 2, 4 and 6).

3.2. *Filtering gene expression data for use in the model*

We began by selecting significant genes through a conventional RNAseq comparison within each dataset. In the Sala Pilot and Validation studies, we compared patients with CF against non-CF individuals to identify genes associated with CF in these datasets. The De Jong datasets exclusively included CF samples, so we compared the presence or absence of virus infection. Lastly, the Balloy dataset consisted of different time points of infection by *Pseudomonas aeruginosa*, with an uninfected control established as point zero for comparison. In all of the studies, we observed changes in gene expression across various comparisons, such

as CF versus non-CF, presence or absence of virus, and infection by *Pseudomonas aeruginosa*. It gave us an idea about which genes should be integrated in our Bayesian Network Model. In the De Jong study, we identified 280 genes (220 up-regulated and 60 down-regulated) (Figure 2A). In the Balloy study, we identified 350 genes (221 up-regulated and 129 down-regulated) (Figure 2B). In the Sala pilot study, we identified 789 genes (639 up-regulated and 150 down-regulated) (Figure 2C), and in the Sala validation study, we identified 2716 genes (2114 up-regulated and 602 down-regulated) (Figure 2D). The differences between all the comparisons can be accessed for both up-regulated genes (Figure 2E) and down-regulated genes (Figure 2F).

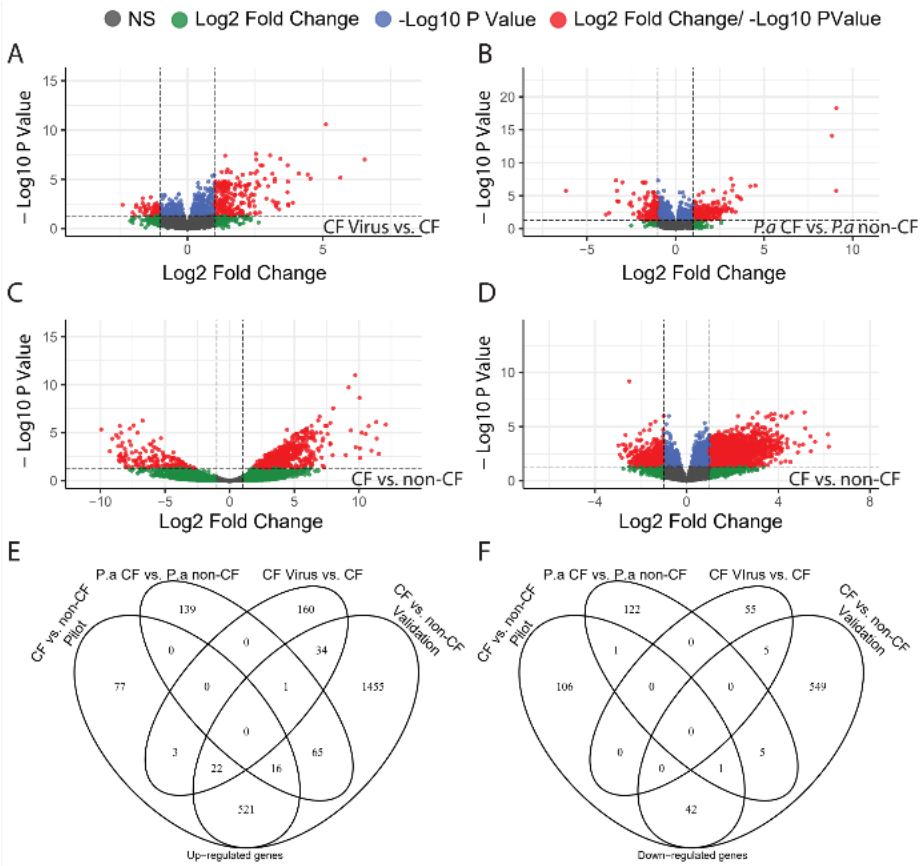

Figure 2. Differential gene expression analysis of epithelial cell datasets. A) GSE139078 shows gene expression changes in cystic fibrosis (CF) patients infected with a virus compared to non-infected CF patients. B) PRJEB9292 compares gene expression in CF patients and controls infected with *Pseudomonas aeruginosa* (P.a). C and D) The Sala Cohort dataset compares gene expression between CF patients and non CF subjects in a pilot study (C) and validation study (D). Red dots represent significant genes with fold changes above or below ±0.5, blue dots represent significant genes without fold change variation, and green dots represent non-significant genes with fold change variation. E and F) Venn diagrams represent the overlap and exclusivity of differentially expressed genes (DEGs) in each comparison, using the upregulated (>1 fold change and p-value <0.05) and downregulated (<-1 fold change and p-value <0.05) DEGs.

3.3. *The Bayesian network is capable of identifying genes associated with all conditions and covariates.*

To circumvent experimental design limitations and to measure the relationship between all conditions and covariates present, we discretized the log CPM table and retrieved all the significant genes obtained from all comparisons of each dataset combined with its respective metadata (infection type (viral or bacterial), CF, modulators (Luma/Iva or Ivacaftor) and genotype (F508del or non CF controls) to create a new dataset. In total we included 1976 genesin the Bayesian network model. As a result, the Bayesian network reveals which genes have a direct relationship with the presence of bacteria, virus, usage of modulators, CF, and the genotype (F508del). Each condition has its own network community despite the genotype, and it is associated with the presence of CF (Figure 3A). Genes present in each network community were used for functional analysis. The functional analysis revealed an Interferon signaling (alpha/beta and gamma) associated with CF, virus, and bacterial network communities. However, IL-9, IL-21, and IL-6 signaling were exclusively related to CF. Virus exposure was exclusively associated with the TGF-beta pathway, and the bacterial exposure did not have any exclusive pathway. Modulator treatment was associated with the response of EIF2AK1 to heme deficiency, late endosomal microautophagy, and IL-1 signaling (Figure 3B).

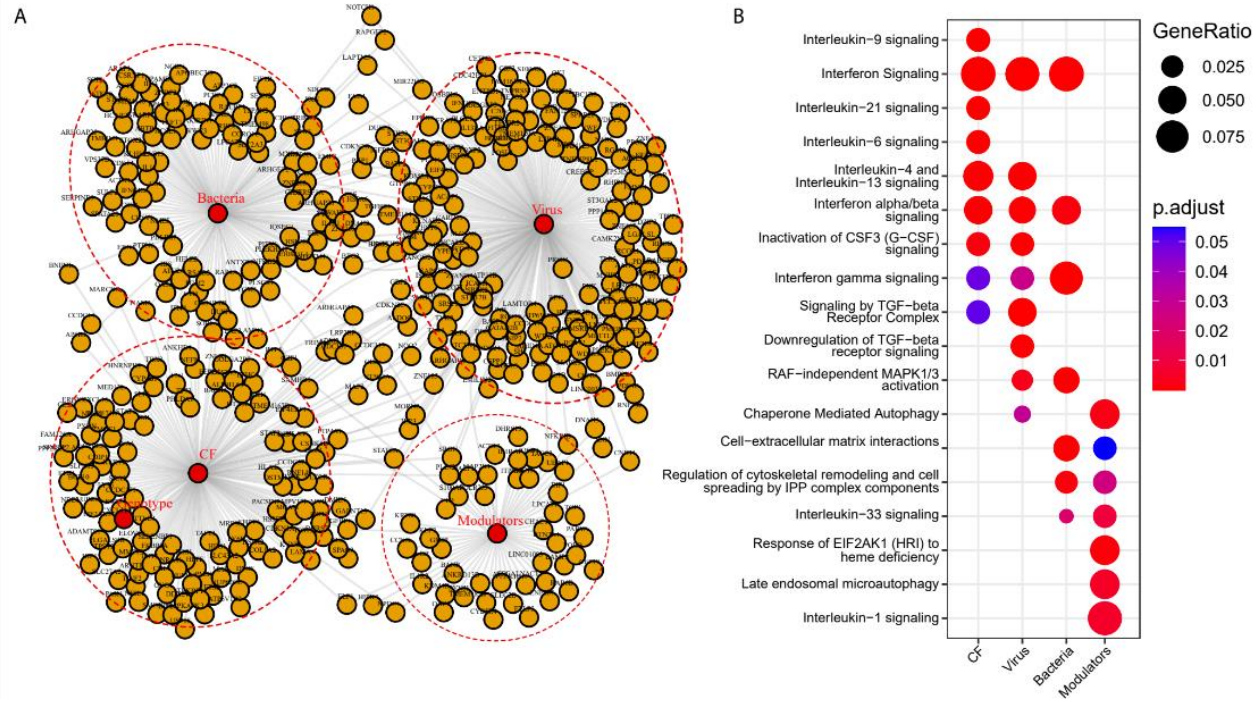

Figure 3. Bayesian Network signatures associated with cystic fibrosis (CF), infection, and mediators. Associations were extracted using Bayesian Network analysis and reconstructed using the "igraph" package in R. A) The main variables (CF, mutations, mediators, and infection) are represented by red nodes and clusters are depicted with red dotted lines. B) Genes presented in each cluster were used for over-represented pathway analysis.

3.4. *The CF Bayesian signature pathway is consistent across all datasets and shows higher expression levels when compared to non-CF epithelial cells.*

The pathways that were discovered in the Bayesian Network Analysis, related to CF were subjected to qusage pathway meta-analysis to measure their activation levels in each study individually, as well as their combination across all studies. As a result, the Interferon signaling, interferon gamma signaling, interleukin 4 and 13 signaling, interleukin 6 signaling, interleukin 21 signaling, and Inactivation of CSF3 G-CSF signaling pathways exhibited an overall alteration across all studies with significant p-values, while the pathways Interleukin 9 signaling and Signaling of TBF-g receptor complex were not significant (Figure 4). We investigated the gene composition of these significant pathways in CF and non-CF to understand their expression. Across all significant pathways investigated (Figure 5). A) Interferon signaling, B) Interferon gamma signaling, C) Interleukin 4 and interleukin 13 signaling, D) Interleukin 6 signaling, E) Inactivation of CSF3 G CSF signaling, and F) Interleukin 21 signaling. The analysis revealed a considerable proportion of epithelial cells derived from CF patients displayed a heightened expression of these genes present in the upper quartile (+75%), in comparison with non-CF. These genes were poorly expressed in all samples in the quantile transformed integrated dataset (Figure 5).

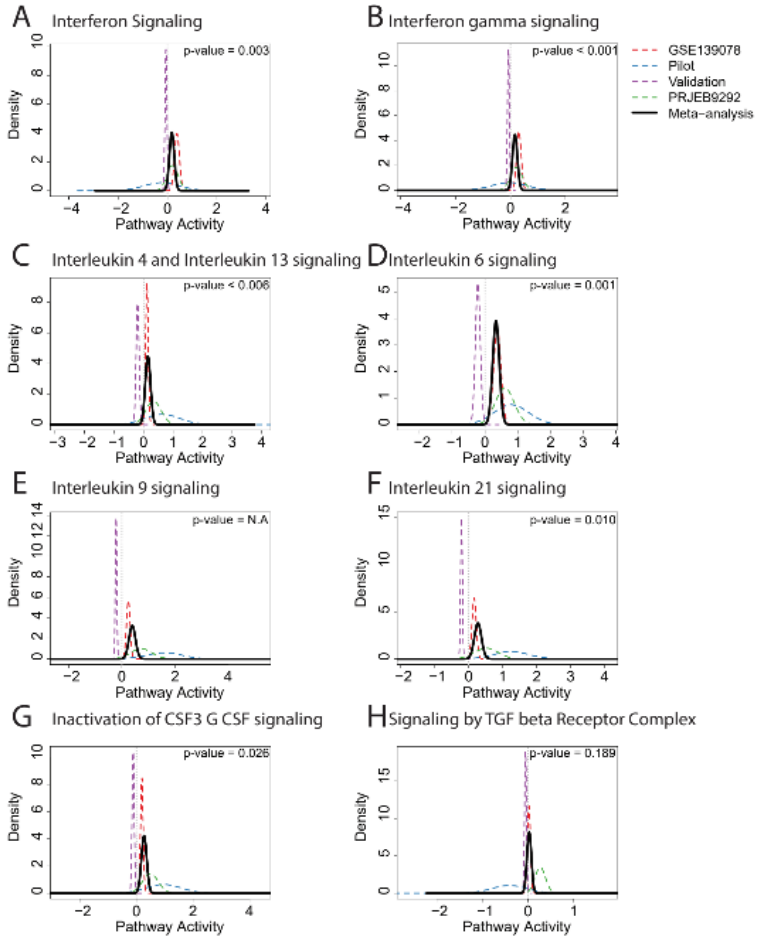

Figure 4. Meta-analysis of pathway enrichment across datasets. The accumulated pathway analysis between all studies was conducted using the pipeline available in the qusage package.

Dotted lines separate studies by color: red for GSE139078, blue for Sala pilot study, purple for Sala validation study, and green for PJREB292. Significant pathways increased related to cystic fibrosis (CF) were identified, including A) Interferon signaling, B) Interferon gamma signaling, C) Interleukin 4 signaling, D) Interleukin 6 signaling, E) Interleukin 9 signaling, F) Interleukin 21 signaling. Pathways decreased in CF include: G) Inactivation of CSF3 and G-CSF signaling, and H) Signaling by TGF-beta receptor complex.

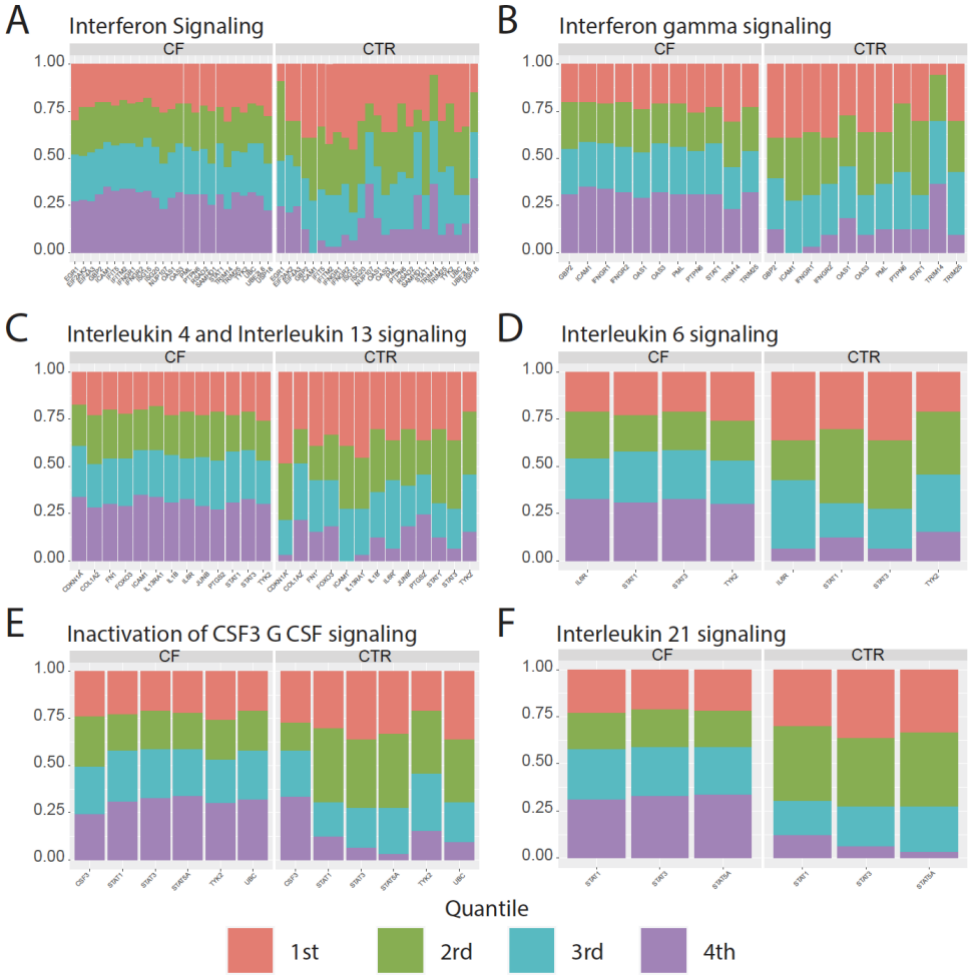

Figure 5. Quantile distribution of expressed genes in each significant pathway related to cystic fibrosis (CF). A) Interferon signaling, B) Interferon gamma signaling, C) Interleukin 4 and Interleukin 13 signaling, D) Interleukin 6 signaling, E) Inactivation of CSF3 and G-CSF signaling, and F) Interleukin 21 signaling.

4. Discussion

Integrating data from transcriptomics or other high-throughput systems, such as proteomics, metabolomics, and lipidomics, is expected to yield new insights. Unfortunately, it also introduces significant heterogeneity arising from various designs or methodologies, commonly known as batch effects. Batch effects are pervasive across all types of high-throughput biological platforms, including single measurement methods like PCR or ELISA.[31] When performing a meta-analysis, batch effects may create bias and reduce statistical power,

making it challenging to detect all relevant features, especially those with small effect sizes or in unbalanced samples.[4] On the other hand, integrating several smaller datasets theoretically improves statistical power, provided that technical heterogeneity, including batch effects, is effectively resolved.

Efforts to mitigate batch effects have been proposed, as they are known to interfere with downstream statistical analysis, potentially introducing false significance between groups that only exist between batches without biological meaning.[32,33] Batch effects can also lead to the loss of biological signals contained in the data.[34,35] The proposed quantile transform approach tends to be respectful of each dataset's characteristics, and by mapping each variable's probability in a probabilistic graphical model, it can handle variables present in the metadata, such as group allocation, clinical data, and dichotomous variables, which can be added and probabilistically related to each other.[36] To achieve this, we evaluated four distinct Cystic Fibrosis Datasets with CF genotype, modulator therapy, and different types of infection, incorporating gene expression with these variables, without applying any batch correction while respecting each dataset's individuality. This approach has demonstrated a high level of accuracy in classifying cancer types when applied to expression datasets. [37]

To reduce processing time, we filtered the genes by selecting those that were differentially expressed in all datasets. For the Baloy dataset, we identified 350 differentially expressed genes (221 upregulated and 129 downregulated genes). In their original publication,[18] the authors found a significantly higher number of upregulated genes than down regulated genes compared to noninfected control cells, although their comparisons were done at each time point. In our study, we bulked the controls and the *Pseudomonas aeruginosa* infection time point 0 as a control and compared to *Pseudomonas aeruginosa* infection. In De Jong's study,[19] the author separated the cells by classes and made two different comparisons: virus infection versus controls and virus infections plus modulator with either Ivacaftor or Ivacaftor/Lumacaftor. We compared all cells together against the controls and identified 195 upregulated genes and 60 downregulated genes. In the study by Sala et al.,[20] our comparisons were similar, with 639 and 2114 upregulated genes in the pilot and validation datasets, respectively, and 568 and 1834 downregulated genes, and 150 and 112 upregulated genes, and 320 and 403 downregulated genes in our analysis, respectively. Differences can be noticed between the studies not only in how the comparisons were done, but also in the methods used for comparisons. In our study, all the analyses were performed with the DESEQ2 package,[21] whereas Sala and De Jong's studies used edgeR.[38]

The pathway analysis performed by Balloy[18] and Sala[20] did not use the same geneset. In our study, we used the Reactome geneset [39], and only De Jong[19] used Reactome geneset as well. However, the inflammatory responses were similar in all studies. In Sala's study, they associated the chaperone pathway in CF, while in our study, it was associated with the modulators. Other pathways, such as Interleukin 6, 9, and 21, were exclusively associated with CF in our analysis. The role of IL-6 is controversial; however, it participates in proinflammatory responses with TNF-a and interleukin-1b. IL-6 is a regulator of the host inflammatory response and is negatively associated with pulmonary function in chronic infection in CF and during acute exacerbation of respiratory symptoms or during a period of apparent clinical stability. In bronchoalveolar lavage fluid, IL-6 was significantly elevated in

infants with CF.[40] Increased expression of IL-9 and IL-9R is responsible for the mucus-overproducing in the lung epithelium of patients with cystic fibrosis[41] and IL-21 is a multifunctional cytokine that acts on various immune cells.[42] Interestingly, in mice fibroblasts, IL-21R is expressed and upregulates matrix metalloproteinases in response to IL-21 by CD8+ T cells.[43]

When it comes to viral infection, we found that viruses have only one exclusive pathway associated with our analysis, which is related to TGF-b signaling. This pathway is involved in pulmonary fibrosis and other organ-related processes. Viruses utilize various mechanisms to modulate this pathway, including altering TGF-b protein expression and its receptors, as well as modulating the SMAD cascades, TGF-b lead to enhanced cell growth and induction of fibrosis.[44] On the other hand, bacterial infection does not influence any pathways in our analysis. As for the use of modulators, we identified three exclusive pathways: "Response of EIF2AK1 to heme deficiency," "late endosomal microautophagy," and "IL-1 signaling". The HRI kinase (or EIF2AK1) plays two main roles during development: it ensures a balanced synthesis of globin and heme and promotes the survival of erythroid precursors during iron deficiency.[45] Inhibitors of P-gp (P-Glycoprotein) such as fostamatinib[46] and Ivacaftor can be associated with various stress conditions, including oxidative stress, heme deficiency, osmotic shock, and heat shock.[47] In the context of CF, the usage of modulators is associated with an autophagy pathway, which compromises CFTR recycling to lysosomal degradation.[48] Moreover, in our study, the genes associated with modulators were linked to this pathway. In CF patients, CFTR modulators have been shown to increase airway nitric oxide (NO) by increasing the concentrations of IL-1α, IL-1β, and other Th17-associated cytokines in sputum, which is related to NO metabolism.[49]

The overall pathway activation in all studies discovered by the Bayesian network approach in CF confirms previous studies describing a hyperinflammatory state in CF, as well as the participation of other pathways such as interleukin 4, 6, 13, and 21. Notably, interleukin 4 and 13 were not exclusively associated with CF status. The roles of IL-4 and IL-13 in the epithelium of CF patients share several biological properties, including chloride secretion.[50] On the other hand, IL-4 inhibits antiviral immunity,[51] and neutralization of IL-13 reduces death and disease severity in COVID-19 without affecting viral load, indicating an immunopathogenic role for this cytokine.[52] Additionally, G-CSF and GM-CSF can induce elastase and MMP-9 release by neutrophils [53]. Interestingly, all the genes presented in the pathway analysis were in the last quantile of expression in our dataset. The main limitation of this study is that it serves as the initial proof of concept for quantile discretization in the integration of raw datasets. A comparison with different methods should be conducted. Additionally, clinical non-numeric data were included in a single analysis. Therefore, this analysis must be interpreted carefully and should serve as a guide for future models aiming to integrate all datasets and variables in a similar manner. Unfortunately, this study was limited to using only four CF datasets due to the considerable challenge of aligning complete metadata, which encompasses treatment, genotype mutation profiling, and infection status. It is uncommon to find metadata with all these features available, and new studies using this approach must be conducted to assess its efficacy. Despite these limitations, this study sheds light on various biological processes related to CF, particularly concerning viral and bacterial

infections, as well as the impact of modulators on epithelial cells within a single assessment, providing valuable insights into these complex.

5. Conclusion

The analysis of integrated data remains a powerful hypothesis generation tool among data scientists. However, dealing with the heterogeneity of multiple datasets poses real challenges. In this study, we proposed a novel approach to integrate several datasets while respecting the unique characteristics of each individual dataset. By applying quantile transformation to multiple datasets and integrating them, we obtained biologically meaningful results that align with existing literature and established associations with other variables such as modulators, virus, and bacterial infections, and included access to good quality metadata. Our analysis revealed an inflammatory signature in CF patients, with exclusive associations observed in interleukin 4, 6, 13, and 21 pathways. Furthermore, we identified potential links between virus infections and the TGF-b pathway, as well as associations between modulators and pathways such as "Response of EIF2AK1 to heme deficiency," "late endosomal microautophagy," and "IL-1 signaling." These findings contribute to a better understanding of the complex interactions in CF and highlight potential targets for further research and development of new integration protocols. Nonetheless, additional studies employing this methodology are imperative to determine the extent to which this innovative approach can uncover novel associations compared to traditional methods.

References

1. Lin, S. et al. Comparison of the transcriptional landscapes between human and mouse tissues. Proc. Natl. Acad. Sci. U. S. A. 111, 17224–17229 (2014).
2. Fei, T., Zhang, T., Shi, W. & Yu, T. Mitigating the adverse impact of batch effects in sample pattern detection. Bioinformatics 34, 2634–2641 (2018).
3. Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics 19, 562–578 (2018).
4. Nygaard, V., Rødland, E. A. & Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics 17, 29–39 (2016).
5. Buhule, O. D. et al. Stratified randomization controls better for batch effects in 450K methylation analysis: a cautionary tale. Front. Genet. 5, 354 (2014).
6. Carraro, G. et al. Transcriptional analysis of cystic fibrosis airways at single-cell resolution reveals altered epithelial cell states and composition. Nat. Med. 27, 806–814 (2021).
7. Clarke, L. A., Sousa, L., Barreto, C. & Amaral, M. D. Changes in transcriptome of native nasal epithelium expressing F508del-CFTR and intersecting data from comparable studies. Respir. Res. 14, 38 (2013).
8. Hampton, T. H. & Stanton, B. A. A novel approach to analyze gene expression data demonstrates that the DeltaF508 mutation in CFTR downregulates the antigen presentation pathway. Am. J. Physiol. Lung Cell. Mol. Physiol. 298, L473–82 (2010).

9. Brazma, A. Minimum Information About a Microarray Experiment (MIAME)--successes, failures, challenges. ScientificWorldJournal 9, 420–423 (2009).

10. Clarke, L. A., Botelho, H. M., Sousa, L., Falcao, A. O. & Amaral, M. D. Transcriptome meta-analysis reveals common differential and global gene expression profiles in cystic fibrosis and other respiratory disorders and identifies CFTR regulators. Genomics 106, 268–277 (2015).

11. Hackenberger, B. K. Bayesian meta-analysis now - let's do it. Croat. Med. J. 61, 564–568 (2020).

12. Liu, Y. et al. A Gentle Introduction to Bayesian Network Meta-Analysis Using an Automated R Package. Multivariate Behav. Res. 1–17 (2022).

13. Chambers, J. D. et al. An assessment of the methodological quality of published network meta-analyses: a systematic review. PLoS One 10, e0121715 (2015).

14. Huang, S. et al. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics Proteomics 15, 41–51 (2018).

15. McLachlan, S., Dube, K., Hitman, G. A., Fenton, N. E. & Kyrimi, E. Bayesian networks in healthcare: Distribution by medical condition. Artif. Intell. Med. 107, 101912 (2020).

16. Lee, A. W. Review of mixed treatment comparisons in published systematic reviews shows marked increase since 2009. J. Clin. Epidemiol. 67, 138–143 (2014).

17. Briganti, G., Scutari, M. & Linkowski, P. Network Structures of Symptoms From the Zung Depression Scale. Psychol. Rep. 124, 1897–1911 (2021).

18. Balloy, V. et al. Normal and Cystic Fibrosis Human Bronchial Epithelial Cells Infected with Pseudomonas aeruginosa Exhibit Distinct Gene Activation Patterns. PLoS One 10, e0140979 (2015).

19. De Jong, E. et al. Ivacaftor or lumacaftor/ivacaftor treatment does not alter the core CF airway epithelial gene response to rhinovirus. J. Cyst. Fibros. 20, 97–105 (2021).

20. Sala, M. A. et al. The proteostatic network chaperome is downregulated in F508del homozygote cystic fibrosis. J. Cyst. Fibros. 20, 356–363 (2021).

21. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014).

22. Hartemink, A. J., Gifford, D. K., Jaakkola, T. S. & Young, R. A. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. Pac. Symp. Biocomput. 422–433 (2001).

23. Scutari, M. Bayesian network constraint-based structure learning algorithms: Parallel and optimized implementations in the bnlearn R package. J. Stat. Softw. 77, (2017).

24. R. Core Team. An Introduction to R. (Samurai Media Limited, 2015).

25. Liu, Y., Wang, L. & Sun, M. Efficient Heuristics for Structure Learning of -Dependence Bayesian Classifier. Entropy 20, (2018).

26. Prada-Medina, C. A. et al. Systems Immunology of Diabetes-Tuberculosis Comorbidity Reveals Signatures of Disease Complications. Sci. Rep. 7, 1999 (2017).

27. Friedman, N., Goldszmidt, M. & Wyner, A. Data analysis with Bayesian networks: A bootstrap approach. (2013) doi:10.48550/ARXIV.1301.6695.

28. Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb) 2, 100141 (2021).

29. Yaari, G., Bolen, C. R., Thakar, J. & Kleinstein, S. H. Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. Nucleic Acids Res. 41, e170 (2013).

30. Meng, H., Yaari, G., Bolen, C. R., Avey, S. & Kleinstein, S. H. Gene set meta-analysis with Quantitative Set Analysis for Gene Expression (QuSAGE). PLoS Comput. Biol. 15, e1006899 (2019).

31. Goh, W. W. B., Wang, W. & Wong, L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. Trends Biotechnol. 35, 498–507 (2017).

32. Leek, J. T. et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. 11, 733–739 (2010).

33. Li, T., Zhang, Y., Patil, P. & Johnson, W. E. Overcoming the impacts of two-step batch effect correction on gene expression estimation and inference. Biostatistics 24, 635–652 (2023).

34. Nyamundanda, G., Poudel, P., Patil, Y. & Sadanandam, A. A Novel Statistical Method to Diagnose, Quantify and Correct Batch Effects in Genomic Studies. Sci. Rep. 7, 10849 (2017).

35. Cai, H. et al. Identifying differentially expressed genes from cross-site integrated data based on relative expression orderings. Int. J. Biol. Sci. 14, 892–900 (2018).

36. Guha, N., Baladandayuthapani, V. & Mallick, B. K. Quantile Graphical Models: Bayesian Approaches. J. Mach. Learn. Res. 21, 1–47 (2020).

37. Jung, S., Bi, Y. & Davuluri, R. V. Evaluation of data discretization methods to derive platform independent isoform expression signatures for multi-class tumor subtyping. *BMC Genomics* 16 Suppl 11, S3 (2015).

38. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140 (2010).

39. Cold Spring Harbor Laboratory. Reactome, a Knowledgebase of Biological Processes. (2003).

40. Nixon, L. S., Yung, B., Bell, S. C., Elborn, J. S. & Shale, D. J. Circulating immunoreactive interleukin-6 in cystic fibrosis. Am. J. Respir. Crit. Care Med. 157, 1764–1769 (1998).

41. Hauber, H.-P. et al. Increased expression of interleukin-9, interleukin-9 receptor, and the calcium-activated chloride channel hCLCA1 in the upper airways of patients with cystic fibrosis. Laryngoscope 113, 1037–1042 (2003).

42. Asao, H. Interleukin-21 in Viral Infections. Int. J. Mol. Sci. 22, (2021).

43. Brodeur, T. Y. et al. IL-21 Promotes Pulmonary Fibrosis through the Induction of Profibrotic CD8+ T Cells. J. Immunol. 195, 5251–5260 (2015).

44. Mirzaei, H. & Faghihloo, E. Viruses as key modulators of the TGF-β pathway; a double-edged sword involved in cancer. Rev. Med. Virol. 28, (2018).

45. Rios-Fuller, T. J. et al. Translation Regulation by eIF2α Phosphorylation and mTORC1 Signaling Pathways in Non-Communicable Diseases (NCDs). Int. J. Mol. Sci. 21, (2020).

46. Duran, G. E. & Sikic, B. I. The Syk inhibitor R406 is a modulator of P-glycoprotein (ABCB1)-mediated multidrug resistance. PLoS One 14, e0210879 (2019).

47. Rolf, M. G. et al. In vitro pharmacological profiling of R406 identifies molecular targets underlying the clinical effects of fostamatinib. Pharmacol Res Perspect 3, e00175 (2015).

48. Maiuri, L., Raia, V. & Kroemer, G. Strategies for the etiological therapy of cystic fibrosis. Cell Death Differ. 24, 1825–1844 (2017).

49. Nissen, G. et al. Interleukin-1 beta is a potential mediator of airway nitric oxide deficiency in cystic fibrosis. J. Cyst. Fibros. 21, 623–625 (2022).

50. Zünd, G., Madara, J. L., Dzus, A. L., Awtrey, C. S. & Colgan, S. P. Interleukin-4 and interleukin-13 differentially regulate epithelial chloride secretion. J. Biol. Chem. 271, 7460–7464 (1996).

51. Moran, T. M., Isobe, H., Fernandez-Sesma, A. & Schulman, J. L. Interleukin-4 causes delayed virus clearance in influenza virus-infected mice. J. Virol. 70, 5230–5235 (1996).

52. Donlan, A. N. et al. IL-13 is a driver of COVID-19 severity. JCI Insight 6, (2021).

53. Castellani, S. et al. G-CSF and GM-CSF Modify Neutrophil Functions at Concentrations found in Cystic Fibrosis. Sci. Rep. 9, 12937 (2019).