ERJ open research
THE BEST IN OPEN ACCESS BASIC, TRANSLATIONAL & CLINICAL RESPIRATORY RESEARCH

Early View

Research letter

Comparison of genome-wide gene expression profiling by RNA-Seq *versus* microarray in bronchial biopsies of COPD patients before and after ICS treatment. Does it provide new insights?

Benedikt Ditz, Jeunard G. Boekhoudt, Hananeh Aliee, Fabian J. Theis, Martijn Nawijn, Corry-A Brandsma, Pieter S. Hiemstra, Wim Timens, Gaik W. Tew, Michele A. Grimbaldeston, Margaret Neighbors, Victor Guryev, Maarten van den Berge, Alen Faiz

Please cite this article as: Ditz B, Boekhoudt JG, Aliee H, *et al*. Comparison of genome-wide gene expression profiling by RNA-Seq *versus* microarray in bronchial biopsies of COPD patients before and after ICS treatment. Does it provide new insights?. *ERJ Open Res* 2021; in press (https://doi.org/10.1183/23120541.00104-2021).

Comparison of genome-wide gene expression profiling by RNA-Seq versus microarray in bronchial biopsies of COPD patients before and after ICS treatment. Does it provide new insights?

Authors: Benedikt Ditz[1,2]*, Jeunard G Boekhoudt[2,3] *, Hananeh Aliee[4], Fabian J. Theis[4,5], Martijn Nawijn[2,3], Corry-A Brandsma[2,3], Pieter S. Hiemstra[6], Wim Timens[2,3], Gaik W Tew[7], Michele A. Grimbaldeston[7], Margaret Neighbors[7], Victor Guryev[8], Maarten van den Berge[1,2#] and Alen Faiz[1,2,9#]

* shared first author

shared last author

[1]*University of Groningen, University Medical Center Groningen, Department of Pulmonary Diseases, Groningen, The Netherlands*

[2]*University of Groningen, University Medical Center Groningen, GRIAC (Groningen Research Institute for Asthma and COPD), Groningen, The Netherlands*

[3]University of Groningen, University Medical Center Groningen, Department of Pathology & Medical Biology, *section Medical Biology,* Groningen, The Netherlands

[4]Institute of Computational Biology, Helmholtz Centre, Munich, Germany

[5]Technical University of Munich, Department of Mathematics, Munich, Germany

[6]Leiden University Medical Center, Department of Pulmonology, Leiden, the Netherlands

[7]OMNI Biomarker Development, Genentech Inc, South San Francisco. USA

[8]European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

[9]University of Technology Sydney, Faculty of Science, Ultimo NSW 2007, Australia

Correspondence: A. Faiz Ph.D., University of Technology Sydney, Respiratory Bioinformatics and Molecular Biology (RBMB), School of Life Sciences, Sydney, Australia

To the editor:

In the era of "Big Data", microarray technology has provided researchers with the ability to measure the expression of thousands of genes in a single experiment[1]. However, array technology is limited, as it can only measure transcripts present in medium to high abundance and can only quantify genes for which oligonucleotide probes are specifically designed. RNA-Seq, the direct sequencing of RNA, is rapidly becoming more popular in analyzing gene expression. RNA-Seq performs better with respect to the detection of low abundance transcripts, identifying genetic variants, and detecting more differentially expressed genes with higher fold-change[2,3]. Bulk tissue cell-type deconvolution represents a recently developed computational method to interrogate the proportions of cell-types in a sample using cell-type-specific gene expression references[4]. This method is mainly based on RNA-seq data, however, little has been done to determine whether this technique can be utilised for microarray technology. We sought to investigate whether gene expression profiling in COPD bronchial biopsies, using RNA seq, provides additional insight into the transcriptional effects before and after ICS, compared to microarrays. Furthermore, we aimed to determine whether cellular deconvolution techniques can be conducted on microarray data by using two current methods Non-negative least squares (NNLS) and support vector regression (SVR) and comparing to RNA-Seq data. To this end, we analyzed the steroid response before and after 6 months of inhaled corticosteroids (ICS) treatment in participants with COPD. Therefore, we utilized gene expression data from bronchial biopsies, which were measured using both microarray (Affymetrix Hugene_ST1.0 array) and RNA-Seq (Illumina HiSeq 2500 platform). The bronchial biopsies were obtained from the Groningen and Leiden Universities study of Corticosteroids in Obstructive Lung Disease (GLUCOLD) study[5]. The methods of microarray sequencing in GLUCOLD have been previously described[6]. With respect to RNA-seq, the Ribozero GOLD libraries were sequenced using 50bp single-read sequencing. The FastQC program version 0.11.5 (https://github.com/s-andrews/FastQC) was utilized to perform quality control checks on the raw sequence data, the sequences were then trimmed using the java program trimmomatic 0.33[7]. The RNA-Seq mapping was conducted using the Spliced Transcripts Alignment to a Reference (STAR) version 2.5.3a[8]. Principal component analysis was performed (using R) to detect extreme outliers. After these quality checks, all samples were found to be of sufficient quality.

In 21 GLUCOLD participants, both microarrays and RNA-seq data in bronchial biopsies were available before and after 6 months treatment with fluticasone (ICS), with or without added salmeterol. Differential expression and cell-type composition analyses were performed to compare individual gene expression as well as single-cell (sc)RNA-Seq expression signatures. The differential expression analysis was conducted in R using the "limma" package (*limma_3.30.13)* for both microarray and RNA-Seq datasets while correcting for age and smoking status[9]. Differentially expressed genes (DEGs) were defined as having a fold-change (FC) >±|1.5| and a False Discovery Rate (FDR)-adjusted p-value less than 0.05[10]. scRNA-Seq signatures for basal, rare, ciliated, and mucus-secretory cells (club and goblet cells) were utilized from our previously-published data to determine differences in cell-type composition, using mRNA expression levels. scRNA-Seq data from bronchial biopsies genes were selected which represented the unique profiles of each cell type, as previously explained[11]. Due to similar expression profiles, club cell and goblet cell scRNA-Seq signatures were combined to generate a uniform scRNA-Seq signature of mucus-secretory cells. For deconvolution, we first performed AutoGeneS to select informative genes and used two different regression methods to infer cell type proportions: NNLS and SVR[12].

By comparing genome-wide gene expression profiling in the RNA-Seq and microarray dataset, the differential expression analysis showed a stronger signal (more significant genes and higher fold change) in the RNA-seq dataset (Figure 1A). Our analysis of the RNA-Seq data identified 4 increased DEGs before and after 6 months of ICS treatment, while 56 DEGs were decreased (Figure 1C). In contrast, the microarray analysis only identified 1 DEG increased by ICS treatment while 7 DEGs were decreased. An overlap of these two analyses showed that 87.5 % of microarray DEGs were identified with RNA-Seq (Figure 1B).

Fold-changes between the two datasets (Figure 1D), using genes measured with both techniques, showed a high level of correlation (Pearson: r= 0.6615, p-value < 2.2e-16). Importantly, the magnitude of FC was overall higher in the RNA-Seq compared to the microarray dataset. As an example, gene *RGS13,* which encodes a regulator of G protein signaling, was found to be downregulated after ICS treatment in the RNA-Seq dataset (logFC: -1.01; FDR: 0.017), but not in the microarray dataset (logFC: -0.34; FDR: 0.08)[13]. Subsequently, we utilized g:profiler to perform functional profiling on the top 50 most significantly decreased DEGs uniquely identified in RNA-Seq[14]. Several pathways that were enriched among the highest down-regulated DEGs belonged to immune system pathways, such as immune response, lymphocyte activation, or regulation of leukocyte activation. This indicates that RNA-Seq

captures differences in transcriptional biological processes, measured in bronchial biopsies from COPD participants, before and after 6 months of ICS treatment, which are missed by microarrays. Cellular deconvolution found a significant Pearson correlation between microarray and RNA-Seq using the SVR for the three cell types secretary (goblet and club), basal and ciliated ($p<0.05$, Figure 1E), however, this was not found for rare cells, which cellular deconvolution techniques usually have problems with. Interestingly, no correlation was observed for the NNLS, indicating that this method gave different results depending on the platform used. NNLS result is likely due to how this program deals with 0 values which are not present in microarray data. SVR stands for Support Vector Regression, which tries to fit the regression within a certain threshold. NNLS stands for the nonnegative least squares method. Additionally, we have included references providing benchmarking of the two methods[12,15].Spearman correlations were then conducted to determine the relationship between cellular deconvolution conducted on microarray and RNA-Seq data.

In conclusion, the SVR method allows cellular deconvolution to be conducted in microarrays samples which reflects RNA-Seq. With respect to differential expression analysis, more DEGs were detected by RNA-seq than microarrays, which were associated with immunological pathways, with greater fold changes. While the fold change of 1.5 or 2 traditionally used for microarray cutoffs may have been too stringent. Therefore, re-sequencing samples, previously measured by microarray, may provide valuable new insights that may otherwise be overlooked.

Figure

Figure 1: Gene expression profiling in participants with COPD, before and after ICS treatment
A) Heatmaps visualizing the significant changes in gene expression after 6 months of ICS treatment in the RNA-Seq dataset in comparison to the microarray dataset **B)** A Venn diagram showing the overlap between DEGs from the (left circle) RNA-Seq dataset and from the (right circle) microarray. **C)** Volcano plot showing the differential expression analysis results for the RNA-Seq dataset. Up-regulated DEGs are given the color red and down-regulated DEGs are given the color blue. **D)** Comparison of log2 fold changes from RNA-Seq and Microarray. **E)** Heatmaps visualizing the correlation between cellular deconvolution results using microarray and RNAseq data. The deconvolution was applied on selected genes using AutoGeneS and inferred cellular proportions using two different regression methods: SVR and NNLS. The legend next to the heatmap depicts the genes per cell type. NA indicates that the gene was not found in that particular, but in the other dataset *=p<0.05. Pearson correlations were used to test associations.

Acknowledgements

OMNI Biomarker Development Genentech (Margaret Neighbors, Michele A. Grimbaldeston and Gaik W Tew) NHLBI LungMAP Consortium (Hananeh Aliee, Fabian J. Theis and M.C. Nawijn)

REFERENCES

1. Brown MPS, Grundy WN, Lin D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A*. 2000;97(1):262-267.

2. Marioni JC, Mason CE, Mane SM, et al. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008;18(9):1509-1517.

3. Zhang W, Yu Y, Hertwig F, et al. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol*. 2015;16(1):1-12.doi:10.1186/s13059-015-0694-1

4. Avila Cobos F, Vandesompele J, Mestdagh P, De Preter K. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*. 2018 Jun 1;34(11):1969-1979.

5. Lapperre TS, Snoeck-Stroband JB, Gosman MM, et al. Effect of Fluticasone With and Without

Salmeterol on Pulmonary Outcomes in Chronic Obstructive Pulmonary Disease: a randomized trial. *Ann Intern Med*. 2009;151(8):517-527.

6. van den Berge M, Steiling K, Timens W, et al. Airway gene expression in COPD is dynamic with inhaled corticosteroid treatment and reflects biological pathways associated with disease activity. *Thorax*. 2014;69(1):14-23.

7. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120.

8. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21.

9. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47.

10. Benjamini Y, Drai D, Elmer G, et al. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res*. 2001;125:279-284.

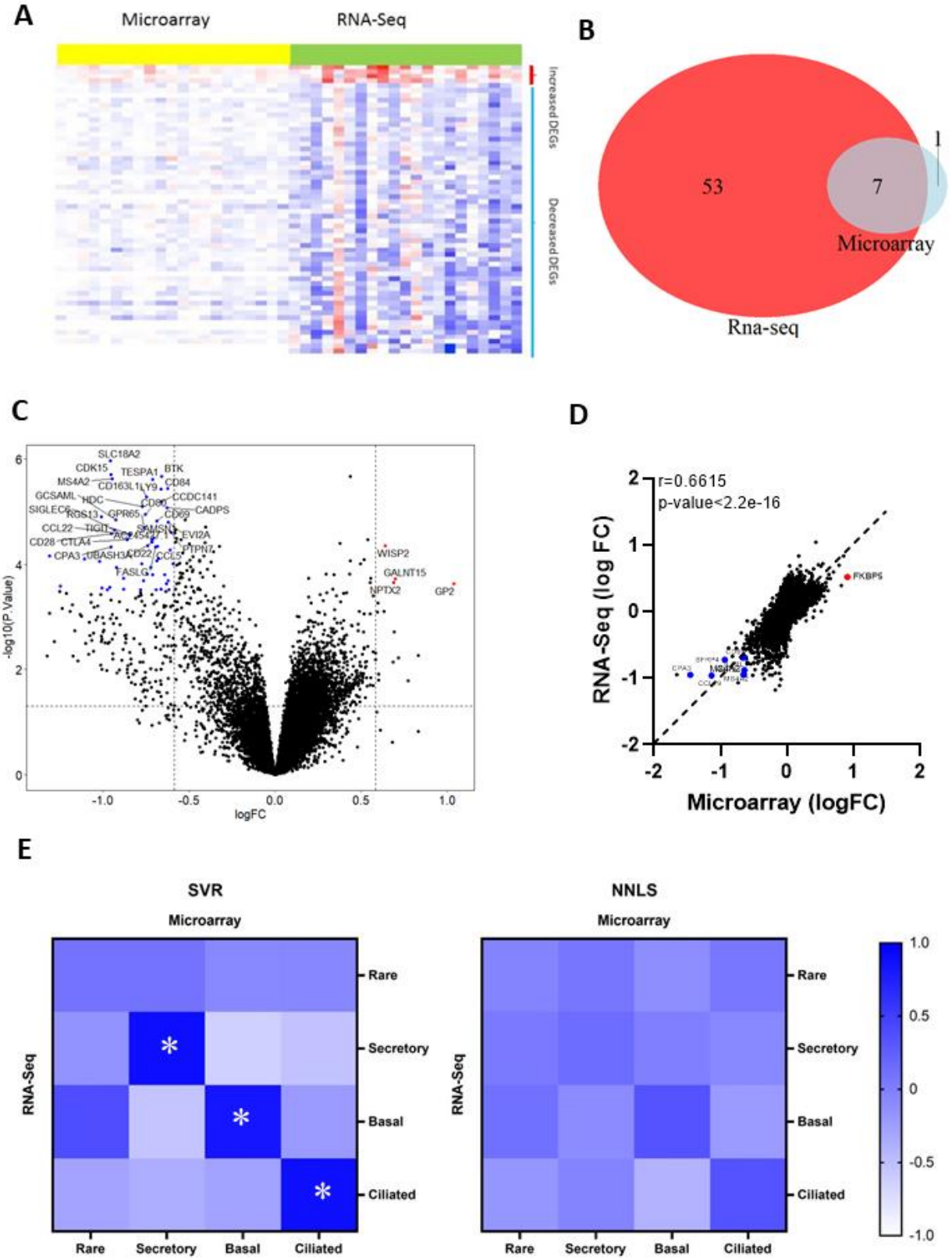
11. Vieira Braga FA, Kar G, Berg M, et al. A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat Med*. 2019;25(7):1153-1163.

12. Aliee H, Theis FJ. AutoGeneS: Automatic gene selection using multi-objective optimization for RNA-seq deconvolution. *bioRxiv*. 2020:2020.02.21.940650.

13. Bansal G, Xie Z, Rao S, et al. Suppression of immunoglobulin E-mediated allergic responses by regulator of G protein signaling 13. *Nat Immunol*. 2008;9(1):73-80.

14. Reimand J, Kull M, Peterson H, et al. G:Profiler-a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res*. 2007;35:193-200.

15. Avila Cobos F, Alquicira-Hernandez J, Powell JE, et al. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat Commun*. 2020;11(1).


A
Microarray
RNA-Seq
Increased DEGs
Decreased DEGs
B
53
7
1
Microarray
Rna-seq
C
SLC18A2
CDK15
TESPA1
BTK
MS4A2
CD84
CD163L1
LY9
GCSAML
HDC
CD80
CCDC141
SIGLEC8
RGS13
GPR65
CD69
CADPS
CCL22
TIGIT
SAMSN1
CD28
CTLA4
EVI2A
CPA3
UBASH3A
CD22
CCL5
FYB1
FASLG
WISP2
GALNT15
NPTX2
GP2
-log10(P.Value)
logFC
D
r=0.6615
p-value<2.2e-16
FKBP5
RNA-Seq (log FC)
Microarray (logFC)
E
SVR
NNLS
Microarray
RNA-Seq
Rare
Secretory
Basal
Ciliated
1.0
0.5
0
-0.5
-1.0