ORIGINAL RESEARCH LETTER

Comparison of genome-wide gene expression profiling by RNA Sequencing *versus* microarray in bronchial biopsies of COPD patients before and after inhaled corticosteroid treatment: does it provide new insights?

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 analysing 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 inhaled corticosteroids (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), which tries to fit the regression within a certain threshold, and comparing them to RNA-Seq data. To this end, we analysed the steroid response before and after 6 months of ICS treatment in participants with COPD. Therefore, we utilised 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) [5]. The methods of microarray sequencing in GLUCOLD have been described previously [6]. With respect to RNA-Seq, the RiboZero GOLD libraries were sequenced using 50 bp single-read sequencing. The FastQC programme (version 0.11.5; https://github.com/s-andrews/FastQC) was utilised to perform quality control checks on the raw sequence data; the sequences were then trimmed using the java programme trimmomatic 0.33 [7]. The RNA-Seq mapping was conducted using 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.

@ERSpublications
More DEGs are detected by RNA-Seq than microarrays in COPD lung biopsies and are associated with immunological pathways. Performing bulk tissue cell-type deconvolution in microarray lung samples, using the SVR method, reflects RNA-Seq results. https://bit.ly/2N8sY3s

Cite this article as: Ditz B, Boekhoudt JG, Aliee H, *et al.* Comparison of genome-wide gene expression profiling by RNA Sequencing *versus* microarray in bronchial biopsies of COPD patients before and after inhaled corticosteroid treatment: does it provide new insights? *ERJ Open Res* 2021; 7: 00104-2021 [https://doi.org/10.1183/23120541.00104-2021].

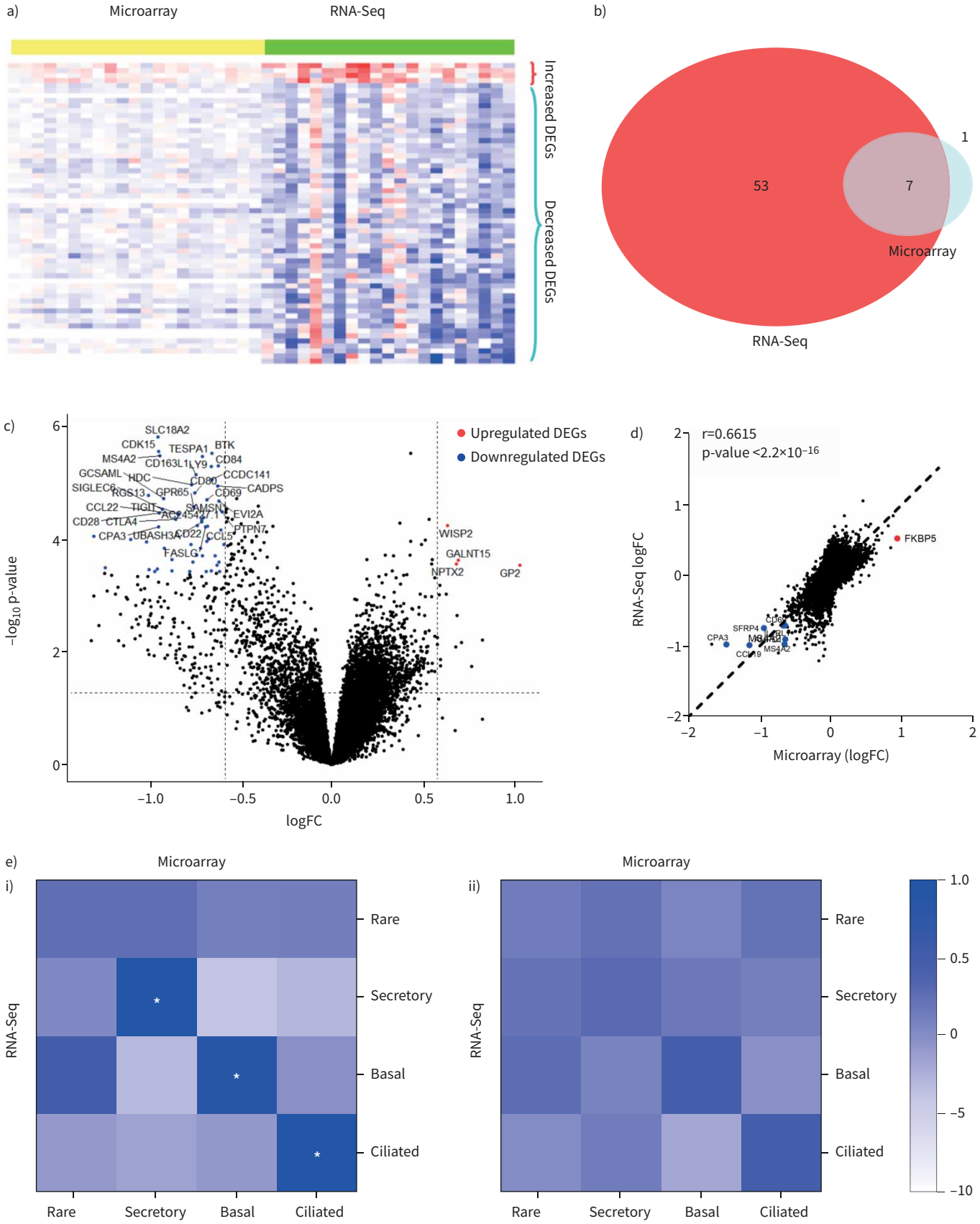

FIGURE 1 Gene expression profiling in participants with COPD, before and after inhaled corticosteroid (ICS) treatment. a) Heatmaps visualising the significant changes in gene expression after 6 months of ICS treatment in the RNA-Seq dataset in comparison to the microarray dataset; b) a Venn

diagram showing the overlap between differentially expressed genes (DEGs) from the RNA-Seq dataset and from the microarray; c) volcano plot showing the differential expression analysis results for the RNA-Seq dataset; d) comparison of log_2 fold-changes (FC) from RNA-Seq and microarray; e) heatmaps visualising the correlation between cellular deconvolution results using microarray and RNA-Seq data. The deconvolution was applied on selected genes using AutoGeneS and inferred cellular proportions using two different regression methods: i) support vector regression and ii) non-negative least squares. The legend next to the heatmap depicts the genes per cell type. *: p<0.05. Pearson correlations were used to test associations.

In 21 GLUCOLD participants, both microarrays and RNA-Seq data in bronchial biopsies were available before and after 6 months of 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 <0.05 [10]. scRNA-Seq signatures for basal, rare, ciliated and mucus-secretory cells (club and goblet cells) were utilised from our previously published data to determine differences in cell-type composition, using mRNA expression levels. scRNA-Seq data from bronchial biopsy genes were selected, which represented the unique profiles of each cell type, as explained previously [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 four increased DEGs before and after 6 months of ICS treatment, while 56 DEGs were decreased (figure 1c). In contrast, the microarray analysis only identified one DEG increased by ICS treatment, while seven 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's r=0.6615, p-value $<2.2\times10^{-16}$). Importantly, the magnitude of fold-change was overall higher in the RNA-Seq compared to the microarray dataset. As an example, gene *RGS13*, which encodes a regulator of G-protein signalling, 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 utilised 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 most downregulated 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: secretory (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. The NNLS result is probably due to the way this programme deals with 0 values which are not present in microarray data. 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 microarray 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. The fold-change of 1.5 or 2 traditionally used for microarray cut-offs may have been too stringent; therefore, re-sequencing samples, previously measured by microarray, may provide valuable new insights that may otherwise be overlooked.

Benedikt Ditz[1,2,10], Jeunard G. Boekhoudt[2,3,10], Hananeh Aliee[4], Fabian J. Theis[4,5], Martijn Nawijn[2,3], Corry-Anke 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[1,2,11] and Alen Faiz[1,2,9,11]

[1]Dept of Pulmonary Diseases, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [2]University of Groningen, University Medical Center Groningen, GRIAC (Groningen

Research Institute for Asthma and COPD), Groningen, The Netherlands. [3]Dept of Pathology and Medical Biology, section Medical Biology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [4]Institute of Computational Biology, Helmholtz Centre, Munich, Germany. [5]Dept of Mathematics, Technical University of Munich, Munich, Germany. [6]Dept of Pulmonology, Leiden University Medical Center, Leiden, The Netherlands. [7]OMNI Biomarker Development, Genentech Inc, San Francisco, CA, USA. [8]European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [9]Faculty of Science, University of Technology Sydney, Ultimo, NSW, Australia. [10]Co-first authors. [11]Co-senior authors.

Correspondence: Alen Faiz, Building 4, Room 04.07.418, University of Technology Sydney, Thomas St, Ultimo, NSW 2007, Australia. E-mail: a.faiz@umcg.nl

Received: 11 Feb 2021 | Accepted: 2 March 2021

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

Conflict of interest: B. Ditz has nothing to disclose. J.G. Boekhoudt has nothing to disclose. H. Aliee has nothing to disclose. F.J. Theis has nothing to disclose. M. Nawijn reports grants from the European Commission (EU H2020 programme), GSK Ltd and Lung Foundation Netherlands during the conduct of the study. C-A. Brandsma has nothing to disclose. P.S. Hiemstra has nothing to disclose. W. Timens reports personal fees from Roche Diagnostics/Ventana, Merck Sharp Dohme, Bristol-Myers-Squibb and Diaceutics outside the submitted work. G.W. Tew is an employee of Genentech Inc., a member of the Roche Group. M.A. Grimbaldeston is an employee of Genentech Inc., a member of the Roche Group. M. Neighbors is a full-time employee of Genentech Inc., and holds stock and options in the Roche Group. V. Guryev has nothing to disclose. M. Van Den Berge has nothing to disclose. A. Faiz has nothing to disclose.

Support statement: This study was supported by Longfonds grant 4.2.16.132JO and the Ministerie van Economische Zaken en Klimaat. Funding information for this article has been deposited with the Crossref Funder Registry.

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 USA* 2000; 97: 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: 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: 133.
4 Avila Cobos F, Vandesompele J, Mestdagh P, *et al.* Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* 2018; 34: 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: 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: 14–23.
7 Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; 30: 2114–2120.
8 Dobin A, Davis CA, Schlesinger F, *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; 29: 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: 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: 1153–1163.
12 Aliee H, Theis FJ. AutoGeneS: automatic gene selection using multi-objective optimization for RNA-seq deconvolution. *bioRxiv* 2020; preprint [https://doi.org/10.1101/2020.02.21.940650].
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: 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: W193–W200.
15 Avila Cobos F, Alquicira-Hernandez J, Powell JE, *et al.* Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat Commun* 2020; 11: 5650.