Published online 3 March 2023

NAR Genomics and Bioinformatics, 2023, Vol. 5, No. 1 **1**
https://doi.org/10.1093/nargab/lqad020

TiSA: TimeSeriesAnalysis—a pipeline for the analysis of longitudinal transcriptomics data

Yohan Lefol[1,2], Tom Korfage[3], Robin Mjelle[4], Christian Prebensen[1,5], Torben Lüders[1,6], Bruno Müller[7], Hans Krokan[4], Antonio Sarno[4], Lene Alsøe[1,2], CONSORTIUM LEMONAID, Jan-Erik Berdal[1,8], Pål Sætrom[4,9,10,11], Hilde Nilsen[1,2,*] and Diana Domanska[2,12]

[1]Institute of Clinical Medicine, University of Oslo, PO Box 1171, Blindern 0318, Norway, [2]Department of Microbiology, University of Oslo, Rikshospitalet, Oslo 0424, Norway, [3]Cytura Therapeutics BV, Kloosterstraat 9, Oss 5349AB, The Netherlands, [4]Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Erling Skjalgsons gate 1, Trondheim 7491, Norway, [5]Department of Infectious Diseases, Oslo University Hospital, Oslo 0424, Norway, [6]Department of Clinical Molecular Biology, Akershus University Hotspital, Lørenskog 1478, Norway, [7]Microsynth AG, Schützenstrasse 15, Balgach CH-9436, Switzerland, [8]Department of Infectious Diseases, Akershus University Hotspital, Lørenskog 1478, Norway, [9]Department of Computer and Information Science, Norwegian University of Science and Technology, Sem Sælandsvei 9 Gløshaugen, Trondheim 7491, Norway, [10]Bioinformatics Core Facility-BioCore, Norwegian University of Science and Technology, Erling Skjalgsons gate 1, Trondheim 7491, Norway, [11]K.G. Jebsen Center for Genetic Epidemiology, Norwegian University of Science and Technology, Håkon Jarls gate 11, Trondheim 7491, Norway and [12]Department of Pathology, Oslo University Hospital-Rikshospitalet, Sognsvannsveien 20, Oslo 0372, Norway

Received September 27, 2022; Revised January 12, 2023; Editorial Decision February 06, 2023; Accepted February 24, 2023

ABSTRACT

Improved transcriptomic sequencing technologies now make it possible to perform longitudinal experiments, thus generating a large amount of data. Currently, there are no dedicated or comprehensive methods for the analysis of these experiments. In this article, we describe our TimeSeries Analysis pipeline (TiSA) which combines differential gene expression, clustering based on recursive thresholding, and a functional enrichment analysis. Differential gene expression is performed for both the temporal and conditional axes. Clustering is performed on the identified differentially expressed genes, with each cluster being evaluated using a functional enrichment analysis. We show that TiSA can be used to analyse longitudinal transcriptomic data from both microarrays and RNA-seq, as well as small, large, and/or datasets with missing data points. The tested datasets ranged in complexity, some originating from cell lines while another was from a longitudinal experiment of severity in COVID-19 patients. We have also included custom figures to aid with the biological interpretation of the data, these plots include Principal Component Analyses, Multi Dimensional Scaling plots, functional enrichment dotplots, trajectory plots, and complex heatmaps showing the broad overview of results. To date, TiSA is the first pipeline to provide an easy solution to the analysis of longitudinal transcriptomics experiments.

INTRODUCTION

Transcriptomic analyses are used in a wide range of fields to observe the effect of genes on cellular activity. Commonly, two methods are used for these analyses, microarray and RNA sequencing (RNAseq) (1). RNAseq has been shown to be the preferred method due to lower false positives and a higher reproducibility (2,3). Microarrays remain in use due to their relative low cost and easy handling (4).

With the diminishing cost and complexity of both RNAseq and microarrays, an increasing number of experiments introduce the dimension of time. A time series experiment or longitudinal study involves the use of two or more time points in the experimental design. Longitudinal studies allow for the identification or estimation of onset times, time-varying factors, as well as the measurement of genetic trajectories (5).

Few tools exist to analyse longitudinal transcriptomic data, even though none of them have been specifically designed to do so. Commonly, a differential gene expression analysis of time series data will require the design of a specific matrix using one of the differential gene expression

*To whom correspondence should be addressed. Tel: +47 93246618; Email: hilde.nilsen@medisin.uio.no

pipelines, such as DESeq2 (6) or limma (7). An alternative method is to perform a permutational multivariate analysis of variants (PERMANOVA) (8). Both methods are implemented in a time series visualization pipeline (9) for the analysis of RNAseq data. However, these methods yield a single list of differentially expressed genes. This limits the exploration of individual time points and can lead to loss of valuable information. The PERMANOVA method has the added caveat of requiring numerous replicates to yield significant results. In pilot studies or animal model research, small amounts of replicates are often used due to their high cost. This information reveals that the current methods for the analysis of time series data are not adequate as they can be complicated to implement and may often be unreliable for experiments with lower numbers of replicates.

TiSA aims to solve the above caveats while providing a user friendly solution for the analysis of time series data of both RNAseq and microarray origin. We eliminate some of the complexity of other existing pipelines, and enable the investigation of individual time points without the need for extra analyses. TiSA also provides solutions to the interpretation of the identified differentially expressed genes (DEGs) by using a novel clustering method followed by a functional enrichment analysis.

To test and validate our TiSA's ability to analyse large, small, and unevenly sampled transcriptomic data, three separate datasets were used. The first is a in house RNAseq dataset created to evaluate Activation Induced Cytidine Deaminase (AICDA/AID) stimulation cocktails in peripheral blood mononuclear cells (PBMC). The over-expression of AID has been associated to several types of cancer such as B cell lymphoma (10–12). The over-expression has also been linked to disease progression, relapse, resistance to salvage therapy and a poor overall remission rate particularly for B cell lymphoma patients (13,14). Since this dataset results from a cell line, we expect very low variability between replicates of the same time points.

The second is a murine dataset with a similar experimental design to the first one. We use this dataset to test TiSA's capabilities with data from an animal organism as well as data with only one replicate per time point and group.

The third dataset is a microarray based analysis performed to identify differences between non-critical and critical patients of SARS-CoV2 infection. SARS-CoV-2 is a virus which emerged in late 2019 causing a world-wide pandemic. The coronavirus disease 2019 (COVID-19) is primarily a respiratory disease. As of March 20th, 2022, the World Health Organization has stated that the number of global cases is nearing 600 million with over 6 million deaths globally. This disease has been the primary focus of thousands of researchers worldwide, however there are still many unknown elements which affect disease progression, severity, and persistence (15,16). The intent of the third dataset is to test TiSA's ability to handle microarray data as well as data with uneven sampling, that is to say an uneven amount of samples per time point/group. The third dataset will also test TiSA's ability to identify relevant results despite of the natural variability found in human gene expression.

MATERIALS AND METHODS

Production of cell lines

Peripheral blood mononuclear cells (PBMC) were ordered and B cells were isolated using the Magnetic-activated cell sorting (MACS), specifically the human Miltenyi B-cell isolation kit II. Cells were then seeded. Controls were stimulated with LPS while non-controls were stimulated using one of the following stimulation cocktails: IgM.acD40.IL4.IL21, TGFb.acD40.IL4. Cells were harvested at days 1, 3 and 9 following seeding. The time course experiment was performed twice in parallel resulting in two replicates for every condition at all three time points.

RNA isolation from cell pellets of cell lines

Cells were pelleted by centrifugation and flash frozen with liquid nitrogen. Cell pellets were then shipped on dry ice. A human cell pellet from a blood sample was used as a positive control. Samples in 96-well plate were placed on dry ice and 3 times 150 ul of RNA shield (DNA/RNA Shield by Zymo, cat. no. R1100-250) was added. The pellets were resuspended in RNA shield and the plate was transferred to room temperature and incubated at room temperature (RT) for 30 min. The plate was centrifuged for 4 min at 5400 × g. Supernatant was transferred to a new plate and remaining pellets were frozen. Qiagen's RNeasy Plus 96 Kit (cat. no. 74192) was used for RNA isolation. RLT buffer was supplemented with β-mercaptoethanol (β-ME) before use. 1 volume of RLT plus buffer was added to samples, and samples were mixed by vigorous shaking. The following steps were according to the RNeasy® Plus 96 Handbook, as available from https://www.qiagen.com/de/resources/download.aspx?id=4842d50e-a987-477a-a819-98a017445ccd&lang=en. RNA was eluted with 45 ul of RNAse free water. Next, samples were quantified with RiboGreen RNA Reagent, RediPlate™96 RiboGreen™RNA Quantitation Kit, cat. no. R11490.

Sequencing and processing of cell lines

Libraries were prepared using an RNA Kit from Takara (Takara Bio USA, Inc. SMARTer® Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian) according to the manufacturer's instructions. Libraries were then pooled and analyzed by Agilent 5200 Fragment Analyzer system (Agilent Technologies) for quantity and size distribution. Sequencing was done on the Illumina NovaSeq platform using an S2 flow cell with 2 × 50 bp reads. 8379591470 passed filter reads were produced with a mean Q of 36, equaling 63 Mio read pairs per sample on average. The produced double-end reads which passed Illumina's chastity filter were subject to de-multiplexing and trimming of Illumina adaptor residuals using Illumina's bcl2fastq software version 2.20.0.422 (no further refinement or selection). Quality of the reads in fastq format was checked with the software FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (version 0.11.8). Raw reads having average Q-values below 20 or incorporating uncalled 'N' bases were filtered using

the BBTools (https://jgi.doe.gov/data-and-tools/software-tools/bbtools/) software suite (version 38.86). The splice aware RNA mapping software STAR(17) (version 2.7.7a) was used to map the surviving reads to the reference genome UCSC hg38 provided by IGenomes. To count the uniquely mapped reads to annotated genes, the software htseq-count (18) (HTSeq version 0.13.5) was used. DNA extraction, library construction, sequencing and data analysis described in this section were performed by Microsynth AG (Balgach, Switzerland).

Murine sequencing and data processing

SmallRNA-seq and mRNA-seq was performed on primary B-cells isolated from mice. 10 samples were collected at the following time points after stimulation with LPS, IL-4 and TGF-beta-1: 0 h, 15 min, 30 min, 1 h, 2 h, 3 h, 6 h, 12 h, 24 h, 48 h. The cells were analysed by flow after 4 days to confirm that they had undergone class switch recombination. For each stimulated sample there was also one unstimulated control sample. The RNA for small RNA and mRNA-seq were isolated from the same tube of cells. The TruSeq protocols (small RNA and poly-A) were used for both sequencing experiments. The raw sequencing data was aligned to the mouse genome (GCF_000001635.23_GRCm38.p3) using STAR-aligner, chimSegmentMin was set to 30, outFilterMultimapNmax to 20, alignSJoverhangMin to 8, alignSJDBoverhangmin to 1, outFilterMismatchNmax to 10, outFilterMismatchNoverLmax to 0.04, alignIntronMin to 20, and alignIntronMax to 1 000 000. Genes were then counted using htseq-count with the following GFF file: GCF_000001635.23_GRCm38.p3_genomic.gff along with the following parameters: stranded (-s) is set to no, idattr (-i) is set to gene, and type (-t) is set to exon.

Data normalization

RNAseq count files were split into their respective groups and attributed defining conditions, such as 'control' and 'experiment'. The counts for the respective conditions are merged, genes which have no values for every sample are removed. A sample defining file containing the sample names along with their associated condition is also created, this file contains the relevant information for the samples within the merged count file. The values are then normalized using the DESeq2 (6) normalization method. The subsequent DESeq2 object is stored in the time series object. If the data originates from microarrays, TiSA expects that the data be inputted as an Elist with the data being already normalized to the users specifications.

Differential gene expression—conditional and temporal

Differential gene expression is performed with either DESeq2 (6), version 1.32.0 or limma (7), version 3.48.1. Differential gene expression is performed on the conditional and temporal axes. The conditional axis consists of performing a differential gene expression at every time point while using the conditions (experiment vs control). The temporal axis is analysed by comparing each time point with the next time point in the series while using the later time point as the experiment, and the earlier time point as the control. Using three time points, two temporal analyses would be performed, one with time point 2 versus time point 1 and another with time point 3 versus time point 2. Each differential gene expression experiment is saved to the time series object using a unique key for the analysis.

A list of differentially expressed genes is obtained by retrieving all significant differentially expressed genes above a user-specified $\log_2$foldchange (or foldChange if using limma) threshold. Significance is established using a false discovery rate (FDR) threshold of 0.05. TiSA enables the use of *P*-value instead of FDR as a measure of significance, but retains the FDR threshold as the default.

Large heatmaps summarizing each dimension of the differential gene expression were designed using the ComplexHeatmap package (19), version 2.8.0.

PART clustering

Clustering is performed using the PART method from the clusterGenomics R package (20), version 1.0. The PART function was set using a minimum cluster size of 50, and a recursion of 100. The subsequent hclust function's distance parameter was set to 'euclidean' and clustering method was set to 'hclust'. For the purpose of reproducibility, a custom seed was set to '123456'.

Gene scaling for plotting

Trajectories of gene clusters are obtained by first calculating the value of each gene per time point and group (experiment or control). This is done by calculating the mean expression of the replicates for a time point and group. The genes are then scaled by dividing each gene by the sum of all the values for that gene. The illustration method for the trajectory plots were inspired by (9).

Gprofiler2—GO ancestor method, semantic similarity for upstream clustering

The gprofiler2 R package (21), version 0.2.1 was used to run each individual cluster through the gprofiler tool. In addition to the standard gprofiler figures, TiSA also provides alternative plotting solutions. Ancestor queries search gprofiler results of each cluster for GO terms which are affiliated to the requested ancestor GO IDs. This is done by using the GO.db package (22), version 3.13.0, which provides a list of children terms for all known GO terms. The children found for the queried ancestors are extracted and plotted.

Multi-dimensional scaling (MDS) plots are made available by exploiting the semantic distance. The semantic similarity is obtained by using the godata function of the GOSemSim R package (23), version 2.18.1. For a standard MDS plot, the semantic similarity between each GO is calculated using the mgoSim function using Wang's measure method (24), the combination method is set to 'NULL'. The semantic similarity between each go is then inputted in the cmdscale function to obtain the multi-dimensional values. In addition, a nearest ancestor approach is implemented where each GO term is brought up to it's nearest common ancestor. This illustration method is inspired by

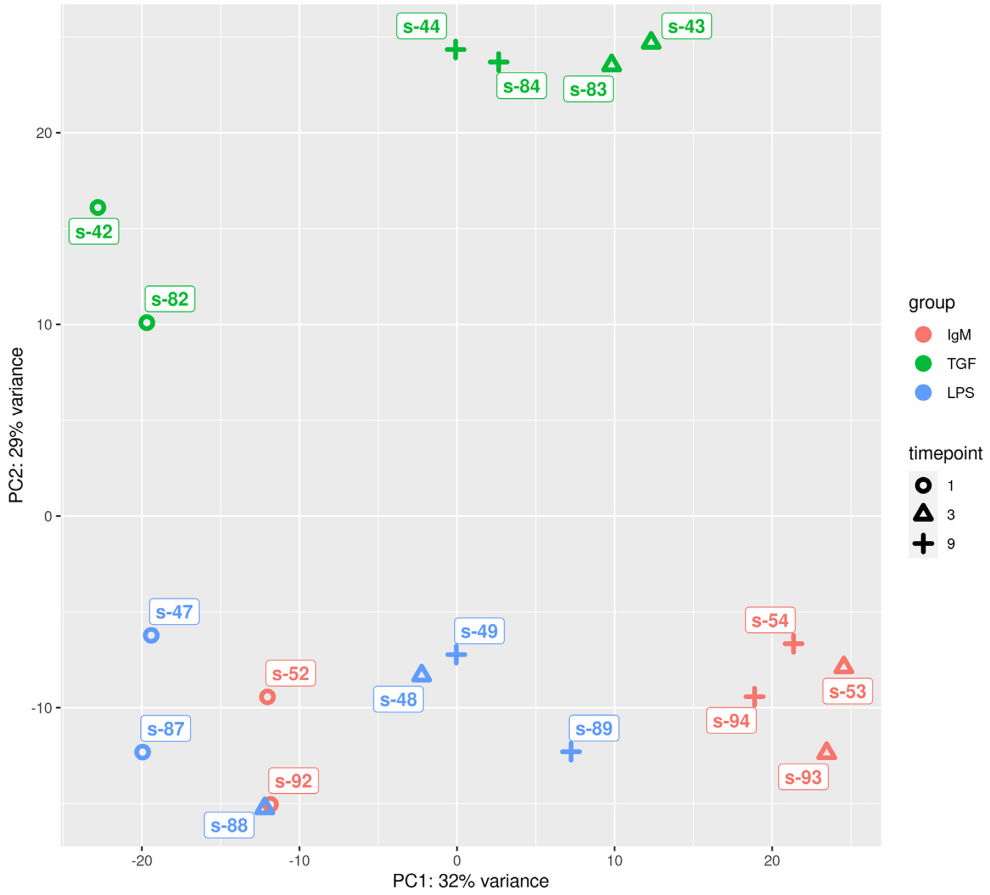

Figure 1. Samples indicate temporal trend. PCA plot showing each sample of all three groups contained within the analysis. Groups are distinguished by color, with IgM in red, TGFb in green, and LPS in blue. The shapes of each sample distinguishes the time points, with circle being the first time point, triangles the second, and crosses the third. In this PCA, time points are days 1, 3 and 9 respectively.

(25). The semantic similarity between terms is calculated the same way as previously described, a clustering is performed using the 'ward.D2' aggregation method with a minimum cluster size of two. To represent the found clusters (or ancestors), their semantic similarity must be calculated. Since ancestors represent two or more GOs, the semantic similarities of the GOs within an ancestor must be combined. The mgosim function is used along with Wang's measurement method and the best mean average (BMA) combining method.

RESULTS

Analysis of AID stimulation cocktails in PBMCs

Two AID stimulation cocktails were compared to an LPS control. The first contained anti-IgM, anti-CD40, IL-4 and IL-21 (IgM.acD40.IL4.IL21) from here on named the IgM cocktail. The second cocktail contained tumor growth factor beta (TGFb), anti-CD40 and IL-4 (TGFb.acD40.IL4) from here on named the TgFb cocktail. LPS was used as a control to ensure the activation and proliferation of the PBMC cells (26). To evaluate the performance of both stimulation cocktails, each were compared with the LPS stimulation cocktail. A time series PCA illustrating the different groups as well as the time points for each sample was created (Figure 1). The PCA showed that the stimulation cocktails are distinguishable from one another. With the IgM and LPS groups being relatively similar and the TGFb group separated by the second principal component (PC2). PC1 appears to explain variability over time, while PC2 picked up on differences between groups. This indicates that variation over time is stronger than variation between groups.

For both cocktails of interest, differential gene expression was performed both conditionally and temporally, with significant genes defined as having a FDR below 0.05 and an absolute log2FoldChange greater or equal to 2. Conditional differential gene expression analysis consist of comparing the two groups at each time point separately. The temporal analysis compares each subsequent time point irrelevant of grouping. This method allowed for the extraction of DEGs of significance within all dimensions of the dataset. The number of DEGs found for each experiment is seen in table (Table 1).

PART clustering was then performed using all significant differentially expressed genes. This resulted in 14 clusters for the IgM group comparison and 12 for the TGFb group comparison. The overview of the clustering can be viewed in a heatmap format, as seen in (Figure 2). To further explore the differences between the clusters, Gprofiler was utilised as it has the ability to query many databases (27). First, the

Table 1. Number of significant differentially expressed genes. Summarisation of the number of DEGs found in the conditional and temporal differential gene expression experiments for all stimulation cocktails tested. The number of unique genes after merger of differential gene expression experiments are shown per conditional and temporal respectively. The total number of unique genes (conditional and temporal merged) is shown in the right-most column

	Conditional			Temporal		**Total**
	TP 1	TP 2	TP 3	TP 2 – TP 1	TP 3 – TP 2	
IgM – LPS	591	590	621	557	17	1691
Total	1253			567		
TGFb – LPS	301	417	701	187	6	1194
Total	1104			192		

REACTOME database was used to obtain an overview of cellular processes (28) involved in the different clusters. This revealed that both stimulation cocktails induced strong cell cycle activity, the most significant activity being seen in the IgM group. Both cocktails also indicated immune activation, however the REACTOME results did not distinguish which immune activity was being activated by the stimulation cocktails.

To further explore the immune activation induced by these cocktails, a dotplot was designed to illustrate the results from a query of any terms which are children for the following parent immune biological processes: regulation of immune system process (GO:0002682), immune response (GO:0006955), and immune effector process (GO:0002252) (Figure 3A). The dotplot showed that most immune related pathways were more significant in the IgM group. Additionally, the 'adaptive immunity' GO term was highly significant in the IgM group, while it was present in the TGFb group, it was at a much lower significance. This indicated that B cells are more strongly activated in the IgM group leading to class-switch recombination and somatic hyper mutation, the two activities in which the AICDA gene participates.

Additionally, the trajectory of the genes within the two largest immune clusters was measured. Cluster 14 of the IgM stimulation cocktail and cluster 11 of the TGFb stimulation cocktail contained the majority of immune related GO terms. The values of these 262 genes were scaled and their trajectory measured for both stimulation cocktails and the LPS control (Figure 3B). Some genes only appeared in one of the two clusters, therefore the genes were merged into a single plot to better represent the pathways of interest. This figure shows that the TGFb group had some genes with a marginal increase compared to the control while all of the cluster's genes in the IgM stimulation cocktail were much higher in expression when compared to both the LPS control and the TGFb group.

Analysis of AID stimulation cocktail in a murine model

A similar AID stimulation cocktail experiment was performed with a murine model. This experiment had two mice, one control and one treated, blood was harvested and sequenced at 10 different time points. Due to having single replicates, each subsequent time point was merged in order to create time points with two replicates per group. A PCA plot showing all samples at their respective time points was created to validate this approach (Supplementary Figure S1). The PCA showed that the time points were close in proximity to each other, which validated our approach to the analysis of this dataset.

Using the same approach as the previous dataset, a search for GO children of various immune GO ancestors was performed and a multi-dimensional scaling (MDS) plot was generated (Figure 4A). The MDS plot isolates the various GO children found for the queried ancestors and colors them based on their associated ancestor. In addition, it indicates to which cluster the children were associated. This MDS plot shows that the immune ancestors group quite well, but more importantly it reveals a child of the 'Immune response' ancestor, found in cluster 22. This child represents adaptive immunity linked to somatic recombination, which served as an indication for CSR activity. The trajectory of cluster 22 showed that the mouse treated with the stimulation cocktail had increased CSR activity at later time points (Figure 4B). Earlier time points showed similar expression for both treated and untreated mice, however the beginning of an increase in expression for the treated mouse was observed at 180 min (3 h), the expression continued to increase and appeared to slowly fall off during the last two time points (24 and 48 h).

These results indicated that the stimulation cocktail activated adaptive immunity. Flow cytometry experiments confirmed class switch recombination for the treated mouse (Supplementary Figure S2).

Identification of biological pathways relevant to SARS-CoV-2 infection severity

A longitudinal covid study using microarray was performed by Prebensen et al. in a currently unpublished study. This longitudinal covid analysis utilized three time points. All patients were able to give a sample at the first time point, however some patients did not return for sampling at time point two. The drop-out was more pronounced at the third time point, a common feature of clinical datasets. Conditional differential gene expression was performed and a summary heatmap drawn (Figure 5), it illustrates TiSA's capability of adapting to missing samples.

868 differentially expressed genes were identified and clustered using the PART method, 9 clusters were found. A REACTOME analysis of these 9 clusters was performed, with enriched terms found in seven of the clusters, as seen in (Figure 6A), five of these clusters were of immunological interest (Figure 6B).

Cluster 9 showed very strong similarity with both the non-critical and critical group, while the other four clusters showed either different trajectories or expression levels. Cluster 2 seemed to be primarily defined by innate immune processes and these were found to be activated in the critical patient group. Cluster 6 results seemed to indicate a diversity of immune pathways, in order to get a better view of the cluster the top biological processes were further analysed. This revealed that the cluster mainly maps to adaptive/humoral immune responses, phagocytosis, and B-cell activation. This cluster's trajectory suggests that levels of adaptive immunity drop between time point 2 (3 days) and time point 3 (8 days) in the case of severe patients, specifically B-cell related immunity. Cluster 7 showed trans-

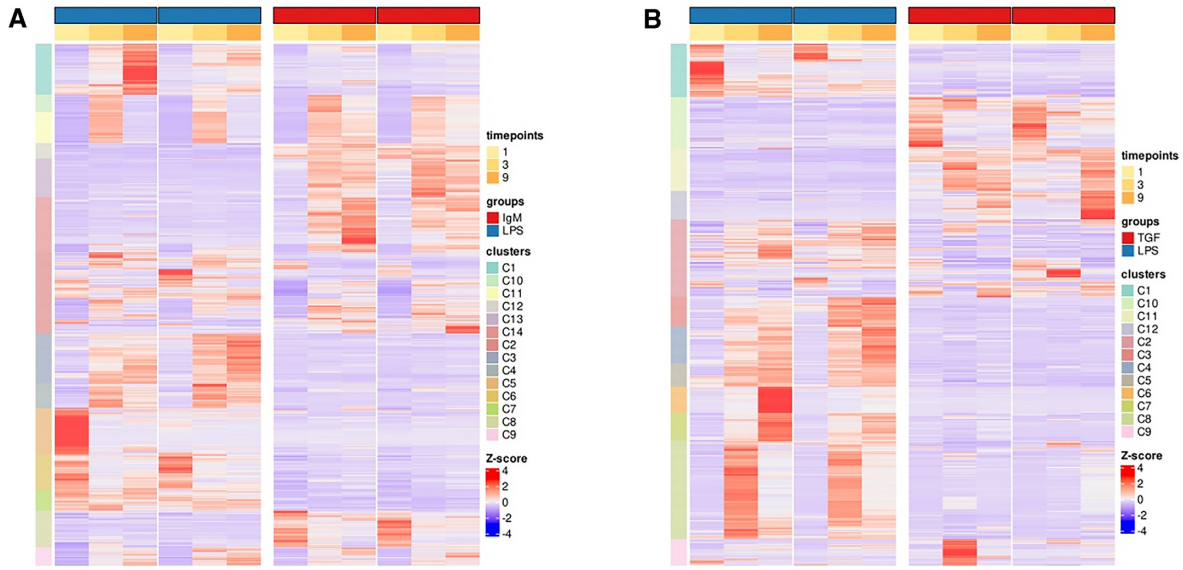

Figure 2. Summarisation of the PART clustering results for both AID stimulation cocktails. The heatmap for the IgM versus LPS comparison (**A**) and the TGFb versus LPS comparison (**B**). Each heatmap shows the PART clustering results of the indicated stimulation cocktail comparison. The genes shown are significant genes with a log2foldchange greater than 2 or below –2. There are 1691 genes in (A) and 1194 in (B). Clusters are indicated with the vertical colored bar on the left-hand side of the heatmap. Groups are indicated by the color at the top of the heatmap, with blue being the control (LPS) and red being the stimulation cocktail. Time points are shown with the yellow/orange bar below the groups. Illustrated within the heatmap itself is the z-score.

lation related pathways, along with a higher expression and an elevated trajectory in critical patients. Cluster 8 showed one immune related pathways, though on it's own it is not very informative. Some biological processes of this cluster indicated inflammation, however most processes pointed towards general immune related pathways.

DISCUSSION

We report here our transcriptomic time series analysis pipeline, TiSA for short, which can be used with both microarray and RNAseq data. Our objective was to provide an easy to use pipeline which can analyse longitudinal transcriptomic data from any annotated organism while catering to both clinical and biological researchers. Both of these fields often have difficulties obtaining large and/or complete datasets. This was the motivation to develop TiSA in a way where it only requires two replicates per time point. Provided there are two replicates per time point, TiSA will adjust for any uneven sampling, such as nine samples at the first time point and only four at the second. To test the TiSA's capability of handling various datasets and still extract valuable biological information, we tested it with three datasets.

With the first dataset we sought to determine which AID stimulation cocktail best simulates AID over-expression in PBMCs. To test and validate the most appropriate AID stimulation cocktail, we designed a time series experiment for two different stimulation cocktails along with two different controls. The time series format was selected to properly identify the time point where AID expression is at its peak following activation.

Of the two cocktails used, (IgM.acD40.IL4.IL21 and TGFb.acD40.IL4), we expect the IgM-containing cocktail to have the best performance. This is based on the findings of Van Belle et al. (29). We expect anti-IgM to mimic the binding of an antigen via the internalization of the B-cell receptor (BCR), the anti-CD40 to mimic the ligation of a T-cell and thus induce the expression of AID (30). IL-4 and IL-21 are expected to further stimulate T-cell assistance, specifically the T cell activation step which induces B cell proliferation and the terminal B-cell differentiation (30,31) respectively. Interestingly, the TGFb component is expected to inhibit B-cell differentiation (32) and circumvent the need for BCR activation. However, the TGFb cocktail has previously been observed to enhance AID expression (33) as well as induce class switch recombination (CSR) (34). We expect the IgM cocktail to generate a stronger AID induction as well as a stronger CSR activity.

Our analysis was able to identify AID activation using both stimulation cocktails, however only the IgM-containing cocktail was observed to have significantly activated CSR indicating that the IgM cocktail was the better of the two cocktails in regards to the activation of AID. These were the expected results and indicate that our ancestor querying approach (Figures 3 and 4) proved to be useful in the evaluation of CSR activation. It also suggests that this approach could be used in other query based approaches, for example, the evaluation of a treatments effect on neuronal related biological processes.

In line with our exploration of the ideal AID stimulation cocktail, we also tested TiSA's effectiveness with a murine dataset which uses the LPS.CD40L.TGFb-1 stimulation cocktail (34). With this dataset we sought to determine TiSA's ability to process datasets with few replicates but with many time points. It is common for pilot biological studies to utilize single replicates when employing animal models due to the complexity and cost of developing and maintaining these models, it is therefore of great importance that a time series analysis pipeline be able to analyse data with the minimum number of replicates.

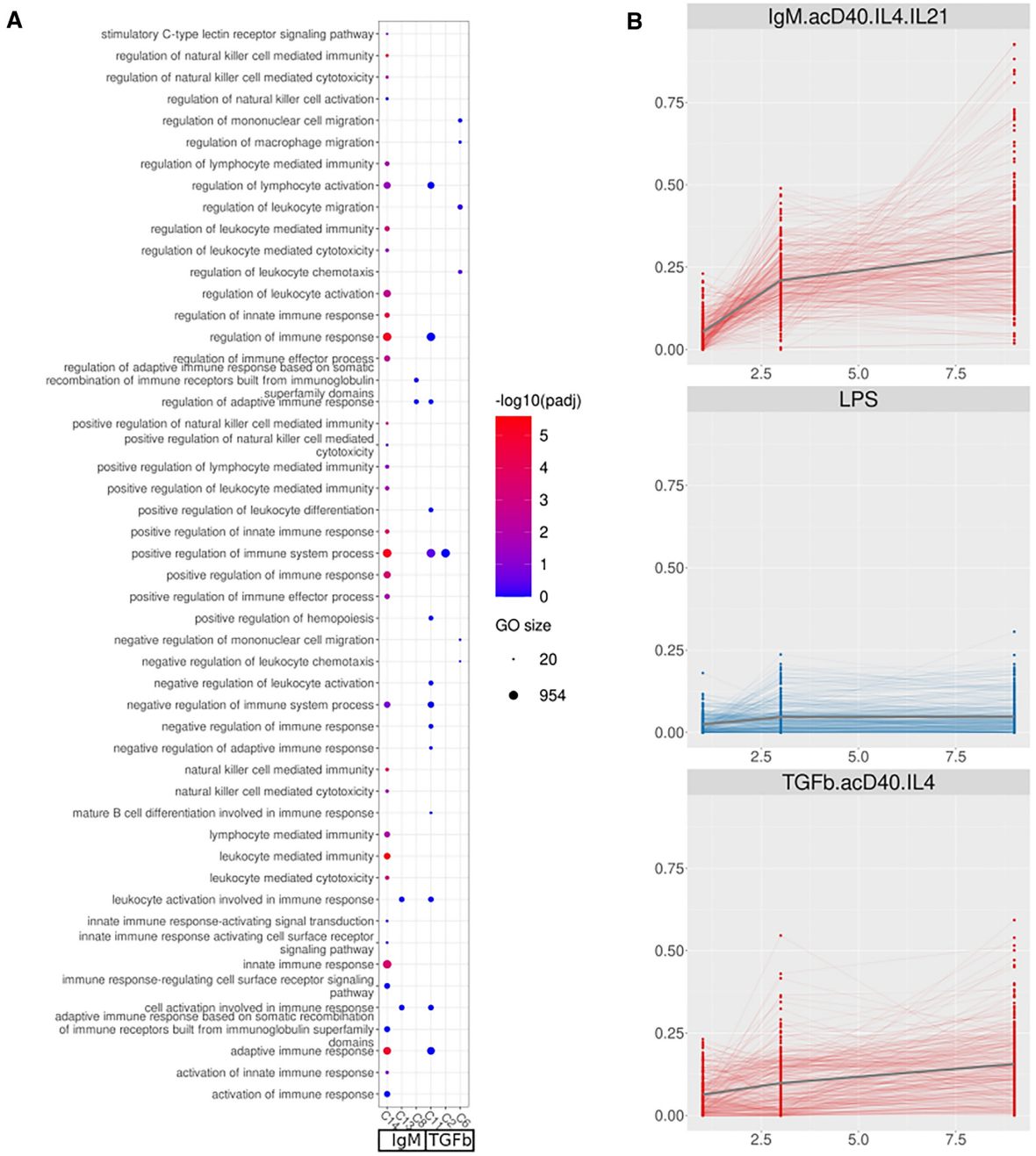

Figure 3. Immune related results found for both AID stimulation cocktails. A dotplot of all GO terms with a adjusted *P*-value <0.05 and affiliated to three immune related GO IDs (GO:0002682, GO:0006955, GO:0002252) (**A**). GO terms found are split based on the cluster in which they were found, with the first three clusters (C14, C13 and C8) originating from the IgM group and the last three (C11, C2 and C8) from the TGFb group. Clusters which map to adaptive immune response (C14 and C11) have been merged, the scaled trajectory of the genes (262) for both stimulation cocktails and the LPS control is shown (**B**). In (A) the color represents the negative log transformation of the adjusted pvalue, with red being most significant and blue less significant. Term size (amount of genes contained in the GO) is indicated by size of dot while term names are indicated on the y axis. (B) shows time on the x axis and scaled value on the y axis.

For this analysis, we utilized a log2foldchange threshold of 1 for the DEGs to be inputted to the PART clustering method. This resulted in 2851 unique significant differentially expressed genes across the conditional and/or temporal analysis. Twenty two clusters were identified by the PART method.

Between the results of cluster 22 seen in (Figure 4) and the flow cytometry results seen in (Supplementary Figure S2), we show that the cocktail used was indeed able to induce AID expression as well as CSR. In addition, we have demonstrated TiSA's ability to utilize a murine dataset as well as a dataset with a

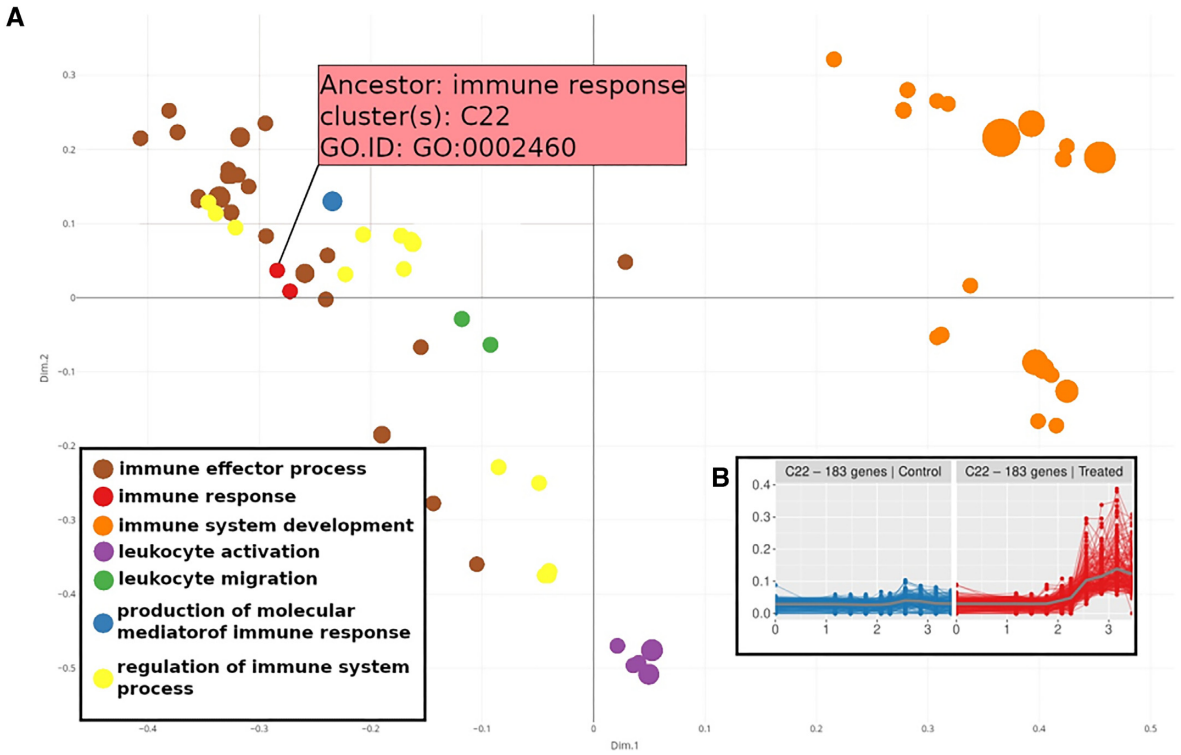

Figure 4. Presence of adaptive immunity and class switch recombination. A MDS plot showing the various immune GOs found in the dataset (**A**). MDS plots show the semantic similarity between GOs; a reflection of their similarity in regards to the genes which are associated to them. The figure also highlights one specific child from the 'Immune response' ancestor—'adaptive immune response based on somatic recombination of immune receptors build from immunoglobulin superfamily domains (GO:002460)'. This GO was found in cluster 22, whose trajectory in both groups (**B**). In the trajectory plot, the y axis is the scaled expression of the genes while the x axis is the log 10 transformed time points in minutes.

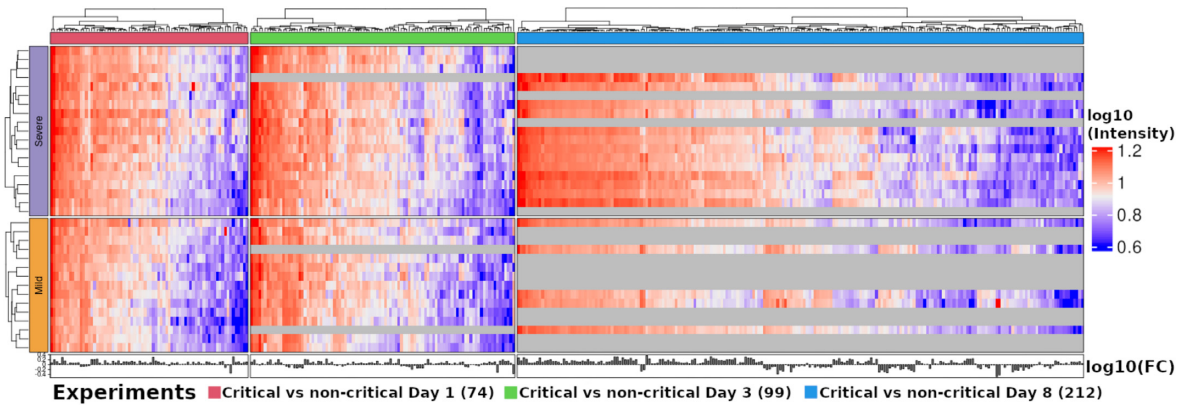

Figure 5. An overview of the SARS-CoV2 severity dataset. A heatmap summarizing the conditional differential gene expression analysis. Each row represents a individual patient and columns represent genes. Patients are grouped by condition (non-critical in orange and critical in mauve) while genes are grouped based on the differential gene expression analysis performed at various time points, starting from the left at time point 1 and progressing to time points 3. Illustrated in the heatmap itself is the log transformed intensity values. The histogram below the heatmap shows the log transformed fold change for the gene above it. Horizontal gray lines represent missing samples for those patients.

very small amount of replicates, something that is often the case in pilot studies within the molecular biology field.

The SARS-CoV-2 pandemic has been of scientific interest since it's beginning in 2019. Prebensen *et al.* (a currently unpublished study) obtained several micro array results for non-critical and critical patients at three different time points, where critical and non-critical is defined as patients being placed on mechanical ventilation or not. This project utilized TiSA as described in this article to identify the genomic differences between non-critical and critical patients. We used the dataset in this article to illustrate TiSA's ability to handle a microarray dataset as well as datasets with uneven sampling, that is datasets which may not have all time points available for each replicate.

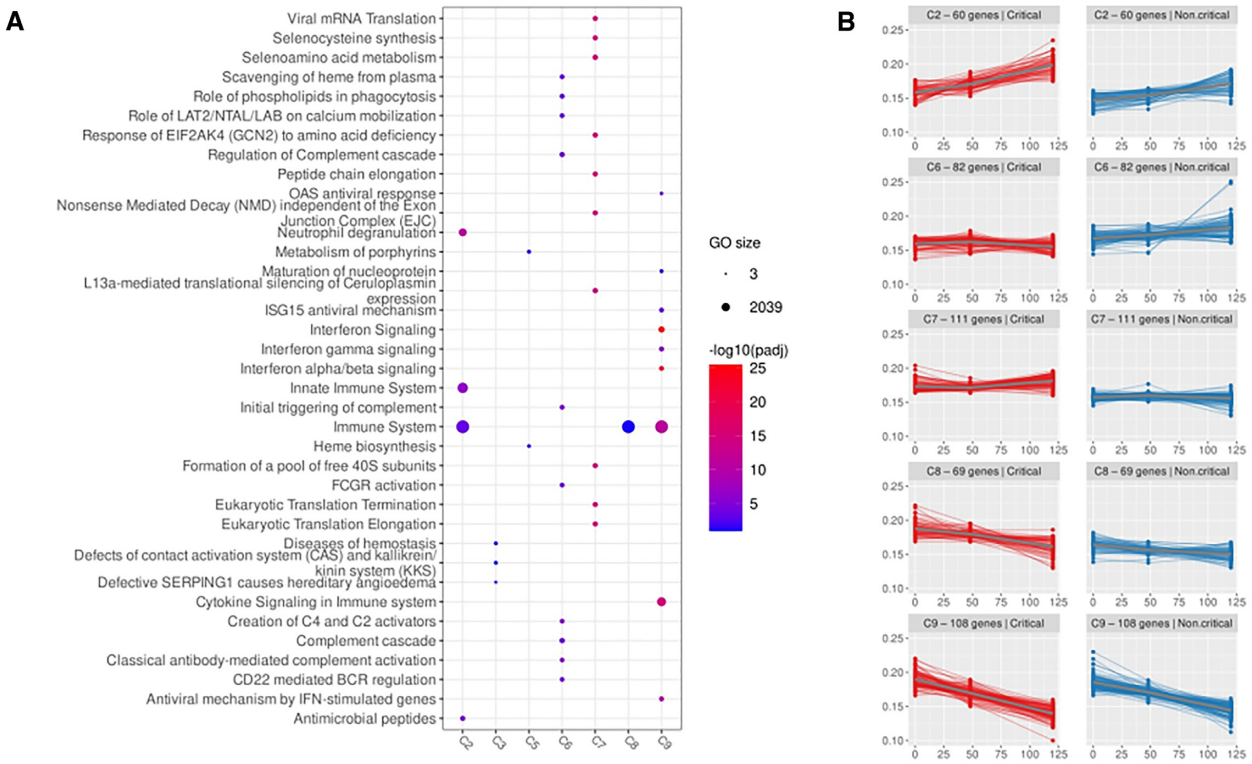

Figure 6. Clusters of importance in the SARS-CoV2 severity dataset. Dotplot of the top 10 REACTOME pathways for all clusters (**A**). The color represents the negative log transformation of the adjusted pvalue, with red being most significant and blue less significant. Term size is indicated by size of dot while term names are indicated on the y axis. The scaled trajectory of the five clusters (C2, C6, C7, C8 and C9) for which REACTOME pathways were found (**B**). The names and quantity of genes as well as grouping is indicated in the title of each subplot. Time points are indicated on the x axis and the scaled value on the y axis.

Several clusters were found to be of interest (Figure 6). Cluster 2 indicates that innate immune responses were higher in the critical group as opposed to the non-critical group. This observation follows what is already known of the disease, that neutrophil levels are elevated based on disease severity (35–37). Cluster 6 seems to show a trend found in naturally infected SARS-CoV2 patients, where adaptive immunity is stable in non-critical patients, but inconsistent in critical patients (38).

The information revealed in this analysis fits with what is already understood between the differences of critical and non-critical SARS-CoV2 patients. As such, we have demonstrated TiSA's ability to analyse microarray datasets as well as datasets which possess uneven sampling. In addition, we have shown that the TiSA is capable of extracting meaningful results from patient derived data, data which inherently carries much variability within each group.

Overall TiSA has demonstrated the ability to analyse a variety of datasets, however like any method some limitations exist. As seen with the first dataset, TiSA is limited to comparative analyses of two groups. This was by design as the intent was to preserve simplicity within the pipeline. In addition, TiSA is primarily designed to function on local computers in order to maximize it's availability to a diversity of users. This comes with the caveat that TiSA will be limited by the resources available. The main bottleneck in this pipeline will be the time required for PART clustering which will strongly vary based on the computer's resources and the number of genes inputted. As a result, we have implemented three separate parameters within the pipeline which can adjust the number of genes submitted to PART clustering or adjust the clustering parameters such as number of recursions and minimum cluster size. Adjusting both of these parameters will adjust the speed and which PART clustering is performed. More information on these parameters can be found within the pipeline itself.

CONCLUSION

We show that TiSA can be used for the analysis and interpretation of both microarray and RNAseq data. We also demonstrate TiSA's ability to solve certain challenges faced with biological data such as few replicates and uneven sampling within experimental groups. TiSA utilizes the PART clustering method which identifies small genomic clusters, with each cluster being analysed independently by gprofiler. Many of the plots designed for TiSA aid in the visualization and biological interpretation of the data.

TiSA has already been successfully used to evaluate the performance of two AID stimulation cocktails using PBMC cells. It has also been able to validate the activation of AID in a murine time series experiment with few replicates. Lastly, it has shown to be capable of analysing unevenly sampled and highly variable dataset all while providing meaningful biological results. This was seen through the

analysis and interpretation of a SARS-CoV-2 longitudinal microarray experiment.

Overall, we show TiSA's ability to analyse longitudinal transcriptomic data from both RNAseq and microarray sources. TiSA is capable of identifying meaningful biological pathways in difficult datasets, such as datasets with a low number of replicates or high variability. TiSA is also made accessible to users with minimal R knowledge as it comes equipped with a clear installation tutorial on our github page as well as several analysis tutorials using a Rmarkdown format.

DATA AVAILABILITY

TiSA along with the tested datasets can be found on github (https://github.com/Ylefol/TimeSeriesAnalysis).

The three datasets presented in this manuscript are also available at the gene expression omnibus (https://www.ncbi.nlm.nih.gov/geo/) using the accession identifiers below.

- PBMC AID stimulation cocktail - GSE213255
- Murine AID stimulation cocktail - GSE212955
- Longitudinal covid severity experiment - GSE213313

ETHICAL CONSIDERATIONS

The longitudinal SARS-CoV2 study performed by Prebensen *et al*. was approved by the Regional Committees for Medical Research Ethics South-East Norway (reference number 2020_39).

The generation of murine data was approved by FOTS; the Norwegian Food Safety Authority (Project ID 4095).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

FUNDING

Cancer Society [223314]; Eurostars [312005].
Conflict of interest statement. None declared.

REFERENCES

1. Kogenaru,S., Yan,Q., Guo,Y. and Wang,N. (2012) RNA-seq and microarray complement each other in transcriptome profiling. *BMC Genom.*, **13**, 629.
2. Richard,H., Schulz,M.H., Sultan,M., Nurnberger,A., Schrinner,S., Balzereit,D., Dagand,E., Rasche,A., Lehrach,H., Vingron,M. *et al.* (2010) Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Res.*, **38**, e112.
3. Sultan,M., Schulz,M.H., Richard,H., Magen,A., Klingenhoff,A., Scherf,M., Seifert,M., Borodina,T., Soldatov,A., Parkhomchuk,D. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.
4. Mantione,K.J., Kream,R.M., Kuzelova,H., Ptacek,R., Raboch,J., Samuel,J.M. and Stefano,G.B. (2014) Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. *Med. Sci. Mon. Basic Res.*, **20**, 138.
5. Kerner,B., North,K.E. and Fallin,M.D. (2009) Use of longitudinal data in genetic studies in the genome-wide association studies era: summary of Group 14. *Genetic Epidemiol.*, **33**(Suppl. 1), S93–S98.
6. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
7. Smyth,G.K. (2005) Limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, pp. 397–420.
8. Anderson,M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecol.*, **26**, 32–46.
9. Nguyen,L.H. (2021) TimeSeriesExperiment: Analysis for short time-series data. R package version 1.10.1.
10. Casellas,R., Basu,U., Yewdell,W.T., Chaudhuri,J., Robbiani,D.F. and Di Noia,J.M. (2016) Mutations, kataegis and translocations in B cells: understanding AID promiscuous activity. *Nat. Rev. Immun.*, **16**, 164–176.
11. Tubbs,A. and Nussenzweig,A. (2017) Endogenous DNA damage as a source of genomic instability in cancer. *Cell*, **168**, 644–656.
12. Shimizu,T., Marusawa,H., Endo,Y. and Chiba,T. (2012) Inflammation-mediated genomic instability: roles of activation-induced cytidine deaminase in carcinogenesis. *Cancer Sci.*, **103**, 1201–1206.
13. Kawamura,K., Wada,A., Wang,J.-Y., Li,Q., Ishii,A., Tsujimura,H., Takagi,T., Itami,M., Tada,Y., Tatsumi,K. *et al.* (2016) Expression of activation-induced cytidine deaminase is associated with a poor prognosis of diffuse large B cell lymphoma patients treated with CHOP-based chemotherapy. *J. Cancer Res. Clin. Oncol.*, **142**, 27–36.
14. Arima,H., Fujimoto,M., Nishikori,M., Kitano,T., Kishimoto,W., Hishizawa,M., Kondo,T., Yamashita,K., Hirata,M., Haga,H. *et al.* (2018) Prognostic impact of activation-induced cytidine deaminase expression for patients with diffuse large B-cell lymphoma. *Leuk. Lymphom.*, **59**, 2085–2095.
15. Bohn,M.K., Hall,A., Sepiashvili,L., Jung,B., Steele,S. and Adeli,K. (2020) Pathophysiology of COVID-19: mechanisms underlying disease severity and progression. *Physiology*, **35**, 288–301.
16. Liu,Y.-C., Kuo,R.-L. and Shih,S.-R. (2020) COVID-19: The first documented coronavirus pandemic in history. *Biom. J.*, **43**, 328–333.
17. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
18. Anders,S., Pyl,P.T. and Huber,W. (2015) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.
19. Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.
20. Nilsen,G., Borgan,Ø., LiestØl,K. and Lingjærde,O.C. (2013) Identifying clusters in genomics data by recursive partitioning. *Stat. Appl. Genet. Mol. Biol.*, **12**, 637–652.
21. Kolberg,L., Raudvere,U., Kuzmin,I., Vilo,J. and Peterson,H. (2020) gprofiler2–an R package for gene list functional enrichment analysis and namespace conversion toolset g: Profiler. *F1000Research*, **9**, 709–736.
22. Carlson,M. (2021) GO.db: A set of annotation maps describing the entire Gene Ontology. R package version 3.13.0.
23. Yu,G., Li,F., Qin,Y., Bo,X., Wu,Y. and Wang,S. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, **26**, 976–978.
24. Wang,J.Z., Du,Z., Payattakool,R., Yu,P.S. and Chen,C.-F. (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**, 1274–1281.
25. Brionne,A., Juanchich,A. and Hennequet-Antier,C. (2019) ViSEAGO: a Bioconductor package for clustering biological functions using Gene Ontology and semantic similarity. *BioData Min.*, **12**, 16.
26. Jansky,L., Reymanova,P. and Kopecky,J. (2003) Dynamics of cytokine production in human peripheral blood mononuclear cells stimulated by LPS, or infected by Borrelia. *Phys. Res.*, **52**, 593–598.
27. Raudvere,U., Kolberg,L., Kuzmin,I., Arak,T., Adler,P., Peterson,H. and Vilo,J. (2019) g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.*, **47**, W191–W198.
28. Jassal,B., Matthews,L., Viteri,G., Gong,C., Lorente,P., Fabregat,A., Sidiropoulos,K., Cook,J., Gillespie,M., Haw,R. *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.
29. Van Belle,K., Herman,J., Boon,L., Waer,M., Sprangers,B. and Louat,T. (2016) Comparative in vitro immune stimulation analysis of primary human B cells and B cell lines. *J. Immun. Res.*, **2016**, 5281823.

30. Guo,B. and Rothstein,T.L. (2013) IL-4 upregulates Igα and Igβ protein, resulting in augmented IgM maturation and B cell receptor–triggered B cell activation. *J. Immun.*, **191**, 670–677.
31. Kuchen,S., Robbins,R., Sims,G.P., Sheng,C., Phillips,T.M., Lipsky,P.E. and Ettinger,R. (2007) Essential role of IL-21 in B cell activation, expansion, and plasma cell generation during CD4+ T cell-B cell collaboration. *J. Immun.*, **179**, 5886–5896.
32. Tamayo,E., Alvarez,P. and Merino,R. (2018) TGFβ superfamily members as regulators of B cell development and function—implications for autoimmunity. *Int. J. Mol. Sci.*, **19**, 3928.
33. Muramatsu,M., Sankaranand,V., Anant,S., Sugai,M., Kinoshita,K., Davidson,N.O. and Honjo,T. (1999) Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells. *J. Biol. Chem.*, **274**, 18470–18476.
34. Nakamura,M., Kondo,S., Sugai,M., Nazarea,M., Imamura,S. and Honjo,T. (1996) High frequency class switching of an lgM+ B lymphoma clone CH12F3 to lgA+ cells. *Int. Immun.*, **8**, 193–201.
35. Aschenbrenner,A.C., Mouktaroudi,M., Krämer,B., Oestreich,M., Antonakos,N., Nuesch-Germano,M., Gkizeli,K., Bonaguro,L., Reusch,N., Baßler,K. *et al.* (2021) Disease severity-specific neutrophil signatures in blood transcriptomes stratify COVID-19 patients. *Genome Med.*, **13**, 7.
36. Kong,M., Zhang,H., Cao,X., Mao,X. and Lu,Z. (2020) Higher level of neutrophil-to-lymphocyte is associated with severe COVID-19. *Epidemiol. Inf.*, **148**, e139.
37. Li,X., Liu,C., Mao,Z., Xiao,M., Wang,L., Qi,S. and Zhou,F. (2020) Predictive values of neutrophil-to-lymphocyte ratio on disease severity and mortality in COVID-19 patients: a systematic review and meta-analysis. *Crit. Care*, **24**, 647.
38. Almendro-Vázquez,P., Laguna-Goya,R., Ruiz-Ruigomez,M., Utrero-Rico,A., Lalueza,A., Maestro de la Calle,G., Delgado,P., Perez-Ordoño,L., Muro,E., Vila,J. *et al.* (2021) Longitudinal dynamics of SARS-CoV-2-specific cellular and humoral immunity after natural infection or BNT162b2 vaccination. *PLoS Path.*, **17**, e1010211.