data

MDPI

Data Descriptor

Preprocessing of Public RNA-Sequencing Datasets to Facilitate Downstream Analyses of Human Diseases

Naomi Rapier-Sharman, John Krapohl, Ethan J. Beausoleil, Kennedy T. L. Gifford, Benjamin R. Hinatsu, Curtis S. Hoffmann, Makayla Komer, Tiana M. Scott and Brett E. Pickett *

Department of Microbiology and Molecular Biology, Brigham Young University, Provo, UT 84602, USA; naomi.rapier.sharman@gmail.com (N.R.-S.); jlkrapohl@gmail.com (J.K.); ejbeausoleil@gmail.com (E.J.B.); kennedylincoln@icloud.com (K.T.L.G.); hinatsu.ben@gmail.com (B.R.H.); hoffmac@byu.edu (C.S.H.); mkomer13@gmail.com (M.K.); tianarunner25@gmail.com (T.M.S.)
* Correspondence: brett_pickett@byu.edu

Abstract: Publicly available RNA-sequencing (RNA-seq) data are a rich resource for elucidating the mechanisms of human disease; however, preprocessing these data requires considerable bioinformatic expertise and computational infrastructure. Analyzing multiple datasets with a consistent computational workflow increases the accuracy of downstream meta-analyses. This collection of datasets represents the human intracellular transcriptional response to disorders and diseases such as acute lymphoblastic leukemia (ALL), B-cell lymphomas, chronic obstructive pulmonary disease (COPD), colorectal cancer, lupus erythematosus; as well as infection with pathogens including *Borrelia burgdorferi*, hantavirus, influenza A virus, Middle East respiratory syndrome coronavirus (MERS-CoV), *Streptococcus pneumoniae*, respiratory syncytial virus (RSV), severe acute respiratory syndrome coronavirus (SARS-CoV), and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). We calculated the statistically significant differentially expressed genes and Gene Ontology terms for all datasets. In addition, a subset of the datasets also includes results from splice variant analyses, intracellular signaling pathway enrichments as well as read mapping and quantification. All analyses were performed using well-established algorithms and are provided to facilitate future data mining activities, wet lab studies, and to accelerate collaboration and discovery.

Dataset: https://zenodo.org/record/4757764; DOI:10.5281/zenodo.4757764.

Dataset License: CC-BY.

Keywords: transcriptomics; RNA-sequencing; autoimmune diseases; cancer; pathogens; bacteria; viruses; data preprocessing

Citation: Rapier-Sharman, N.; Krapohl, J.; Beausoleil, E.J.; Gifford, K.T.L.; Hinatsu, B.R.; Hoffmann, C.S.; Komer, M.; Scott, T.M.; Pickett, B.E. Preprocessing of Public RNA-Sequencing Datasets to Facilitate Downstream Analyses of Human Diseases. *Data* **2021**, *6*, 75. https://doi.org/10.3390/data6070075

Academic Editor: Ralph A. Tripp

Received: 8 June 2021
Accepted: 12 July 2021
Published: 15 July 2021

1. Summary

The number of publicly available RNA-sequencing (RNA-seq) datasets is increasing, and we expect this momentum to continue. However, comprehensive results from statistical analyses such as differential gene expression are not consistently available in public transcriptomics repositories such as the Gene Expression Omnibus (GEO). Additionally, in the subset of cases where multiple differentially expressed gene (DEG) lists from different experiments are available, directly comparing them is difficult due to the differing parameters, assumptions, and biases present within each of the preprocessing algorithms (e.g., trimming, mapping, quantification; see sampling of pipelines and methods) [1–5]. A survey of the literature confirms that transcriptomic preprocessing pipelines utilize a variety of underlying statistical models, further complicating comparison between two datasets processed by different pipelines.

Our motivation for publishing these preprocessed public datasets was to make the results from these computational methods accessible to facilitate hypothesis generation, as

well as for subsequent analysis and interpretation by researchers. In particular, to aid those who may not have the necessary computational infrastructure or expertise to perform this work. The specialized bioinformatic expertise required to successfully preprocess and analyze RNA-seq data can pose a significant barrier for some research groups. To complete an RNA-seq analysis, researchers must find relevant studies, collect the necessary metadata, quantify read mapping, calculate the DEGs, perform Gene Ontology (GO) enrichment, and compute significant signaling pathways. In addition, it is not uncommon to read and write over one terabyte of data while completing a large meta-analysis, which can exceed the computational capacity available to some researchers. The results from preprocessing and analyzing datasets such as those presented in the current study can save researchers both time and resources by easily retrieving genes, biological functions, pathways, and/or splice variants that are significantly affected during a given disease or condition. These biological entities could be further evaluated to identify potential biomarkers and disease mechanisms that can be exploited to improve diagnosis or treatment of disease, and to increase the speed of research in these fields.

Although the scope of the current study is to report the results of various computational methods on public datasets, we are unable to accurately interpret the data in all of the pathologies for which we have preprocessed data. This makes it imperative that the research community reviews the genes, functions, and pathways that were identified.

Combining multiple individual datasets in a meta-analysis increases the statistical power of the derived results by increasing the signal-to-noise ratio. Multiple previous studies have shown that removing background noise makes it easier to gain additional insight on the underlying biological mechanisms that play a role in any given system [6–8]. Specific examples of prior meta-analyses that revealed novel results from existing data include Kori and Arga identifying 18 previously unknown cervical cancer receptors [9], Patel et al. discovering that the accepted transcriptional profile of Alzheimer's disease only applies to the temporal lobe, with distinct gene expression patterns appearing in other areas of the Alzheimer's-diseased brain [10], and Zhang et al. finding evidence that CD-38, LAG-3, and interferon-1 stimulated genes are linked to the progression to AIDS after HIV infection [11].

The goal of the current study was to produce lists of DEGs and GO terms for each target dataset together with statistically significant splice variants, and signaling pathways, and quantification results for subsequent analysis and interpretation by the research community. We used the ARMOR automated analytical workflow as well as custom scripts to preprocess and analyze samples across various targeted infectious diseases as well as other human diseases and conditions. For infectious diseases, we queried for samples quantifying the host response to *Borrelia burgdorferi*, hantavirus, influenza A virus, MERS, respiratory syncytial virus, *Streptococcus pneumoniae*, SARS-CoV, and SARS-CoV-2. We also targeted non-infectious diseases and conditions that include acute lymphoblastic leukemia, B-cell lymphomas, chronic obstructive pulmonary disease, colorectal cancer, and lupus erythematosus. Our primary focus was on datasets from humans or human-derived cell lines, with a small number of datasets from mice or murine cell lines. Providing consistently preprocessed RNA-seq datasets for secondary analysis and research minimizes the bioinformatics barrier, enabling researchers to perform preliminary and/or in depth in silico analyses.

The public availability of these analytical results supports the findable, accessible, interoperable, and reusable (FAIR) guidelines [12]. Our study preprocessed the raw data from 31 public datasets and makes the results from each publicly available. We are not aware of any prior report that contributes the results from such a large number of preprocessed RNA-seq datasets in a single study. Unfortunately, some meta-analyses that have been performed only report cherry-picked genes and pathways from the DEG list and do not consistently publish the complete lists of genes, functions, and/or pathways that were generated during the meta-analysis. Though we do not provide a comprehensive interpretation of these preprocessed results in this study, the files that were generated

by this study contain lists from which additional downstream work can be done. In an attempt to provide some validation to our work, we have included several examples from published research that support the results from our preprocessing workflow. We expect that our efforts to compile and preprocess these datasets will fuel future experiments to develop novel targeted diagnostics and/or treatments.

2. Data Description

The initial component of this data processing workflow required the manual searching and curation of appropriate metadata associated with each study to better inform our analytical design. Although this metadata review process was labor intensive, we believe it augments the value of the results. As such, we greatly appreciate the existing minimal information standards for Minimum Information about a Next-Generation Sequencing Experiment (MINSEQE) and Minimum Information for Biological and Biomedical Investigations (MIBBI) [13]. We highly recommend continued improvement and adherence to such standards.

Overall, we preprocessed and analyzed 31 datasets consisting of over 1250 samples (Table 1). The results from these analyses are reported in over 200 files that contain significant changes in gene expression (EdgeR files), Gene Ontology terms (Camera files), splice variants (DRIMSeq files), intracellular signaling pathways (SPIA files), and/or read mapping/quantification (Salmon files). We estimate that this work required over 25 terabytes of data being read and written throughout the workflow, hundreds of person-hours, and thousands of CPU-hours to complete.

We have provided a few examples of genes identified as significant by our analysis which have been reported in prior studies. We include these to validate the accuracy of our approach and to reiterate the value that these data contain.

For the DEG analysis of pre-treatment *Borrelia burgdorferi* infection vs. healthy controls, several of the top 10 significant DEGs identified in the current study have been previously associated with Lyme disease. Specifically, CoQ10 was shown to be an effective supplemental treatment for chronic Lyme Disease patients' fatigue [14], suggesting that the blockage of CoQ10A, a gene which had a $\log_2$ fold change (logFC) of -1.8 and a false-discovery rate-corrected *p*-value (FDR) of 1.22×10^{-12} in our results, could be a pathogenic mechanism of *B. burgdorferi*. Though no direct connection has been made between *B. burgdorferi* and LEMD3, a rheumatology review considered both LEMD3 mutation and *B. burgdorferi* infection as sources of painful, scleroderma-like disorders [15]. This indicates that the classic Lyme disease symptom of joint pains may originate from the downregulation of LEMD3 (logFC = -1.68, FDR = 1.31×10^{-12}) during infection. Mutations in C19orf12 result in hereditary neuropathies of paraspasticity and Silver Syndrome [16], suggesting that the neuropathic symptoms of *B. burgdorferi* infection may result at least partially from the downregulation of C19orf12 (logFC = -2.6, FDR = 1.20×10^{-11}).

Our results for Respiratory Syncytial Virus (RSV) identified a handful of gene products that are suspected to be critical to the patient response during RSV infection. Specifically, González-Sanz et al. demonstrated that interferon-stimulated gene 15 (ISG15; logFC = 4.2, FDR = 3.20×10^{-41}) has a strong anti-viral effect in vitro and suggest that the same effects may be part of the human innate immune response in vivo [17]. IFIT1 (logFC = 5.04, FDR = 6.54×10^{-40}), IFIT2 (logFC = 4.78, FDR = 5.85×10^{-39}), and IFIT3 (logFC = 4.87, FDR = 4.92×10^{-39}) are all proteins that have an anti-viral effect on RSV [18], indicating that their upregulation during infection is likely a protective measure against the virus. During viral infections, PARP9 (logFC = 2.99, FDR = 1.45×10^{-38}) and DTX3L (logFC = 2.21, FDR = 9.92×10^{-27}) form a complex to induce interferon hyper-responsiveness without toxicity [19].

Our B-cell lymphoma preprocessed dataset also yielded DEGs that have been identified in previous wet-lab experiments. CXCL9 (logFC = 11, FDR = 4.31×10^{-141}) has been shown to promote the progression of diffuse large B-cell lymphoma by starting a cascade that upregulates oncogenes such as CCND1 (logFC = 2.23, FDR = 1.08×10^{-22}) [20]. Upregulated VCAM1 (logFC = 7.85, FDR = 2.29×10^{-120}) is associated with a poor prognosis for

patients with non-Hodgkin's lymphomas and is under investigation as a serum biomarker for disease progression assessment [21].

Due to the varying origins of the RNA-sequencing data we preprocessed, our results may contain background noise that potentially reflects laboratory artifacts, differences in protocols, or other biases. Although human error is also a possibility, meta-analyses generally reduce the statistical "noise" of outlier samples by "drowning them out" by including large numbers of samples in the process. We also performed quality control on the sample data before any statistical analysis was started. Additionally, our chosen bioinformatic workflow implements a false-discovery rate (FDR) multiple hypothesis correction on all initial *p*-values, effectively reducing the occurrence of false-positives. Overall, we feel that the impact of any noise or error on our statistical analyses has been minimized.

The statistically significant findings from each of these datasets could be further analyzed by performing Boolean comparisons of DEGs, GO terms, and pathways. Such an analysis would identify entities that are unique to a given dataset or shared between multiple datasets. The results from such meta-analyses could then be used to generate testable hypotheses and design robust validation experiments in the wet lab. The data generated in this work can facilitate more in-depth data mining activities that enable biomarker identification, improving understanding of disease, and the repurposing of existing drugs. We anticipate that making these preprocessed RNA-seq datasets publicly available will ensure that scientific data remains findable, accessible, interoperable, and reusable (FAIR), while simultaneously fueling collaboration, innovation, and discovery.

Table 1. Summary of preprocessed datasets.

Disease/Disorder	Organism	Tissue Type	Sample Type	# of Studies	# of Samples	GEO Identifier	EdgeR, Camera, DRIMseq	Salmon	SPIA
Acute Lymphoblastic Leukemia (ALL)	Homo sapiens	Blood and bone marrow	Total RNA	1	10	GSE162894 [22]	Yes	Yes	Yes
B-cell Lymphomas	Homo sapiens	B-cells	mRNA	7	322	GSE153437 [23] GSE130751 [24] GSE110219 [25] GSE95013 [26] GSE62241 [27] GSE50514 [28] GSE45982 [29]	Yes	Yes	Yes
Borrelia burgdorferi	Homo sapiens	PBMC	mRNA	1	97	GSE63085 [30]	Yes	Yes	No
Chronic Obstructive Pulmonary Disease (COPD)	Homo sapiens	Lung tissue	mRNA	1	189	GSE57148 [31]	Yes	No	No
Colorectal cancer	Homo sapiens	Colorectal tissue	lncRNA	3	44	GSE104836 [32] GSE124526 [33] GSE155457 [34]	Yes	Partial	Yes
Hantavirus	Homo sapiens	PBMC, HUVEC	Total RNA	2	36	GSE133751 [35] GSE158712 [36]	Yes	No	Yes
Influenza A	Homo sapiens	A549	mRNA	1	4	GSE147507 [37]	Yes	No	Yes
Lupus Erythematosus	Homo sapiens	B-cells	mRNA	3	335	GSE92387 [38] GSE118254 [39] GSE110999 [40]	Yes	Yes	Yes
Middle East Respiratory Syndrome Coronavirus (MERS-CoV)	Homo sapiens	Calu-3	mRNA	3	31	GSE139516 [41] GSE122876 [42] GSE56192 [1]	Yes	Partial	Yes
Streptococcus pneumoniae	Homo sapiens, Mus Musculus	Nasal samples, nasal lavage, polymorphonuclear leukocytes, A549	Total RNA, mRNA	5	104	GSE150811 [1] GSE79595 [43] GSE116604 [44] GSE117580 [45] GSE124949 [46]	Yes	Yes	Yes

Table 1. *Cont.*

Disease/Disorder	Organism	Tissue Type	Sample Type	# of Studies	# of Samples	GEO Identifier	EdgeR, Camera, DRIMseq	Salmon	SPIA
Respiratory syncytial virus (RSV)	Homo sapiens	A549	mRNA	1	4	GSE147507 [37]	Yes	No	Yes
Severe acute respiratory syndrome coronavirus (SARS-CoV)	Homo sapiens	MRC5	Total RNA	1	15	GSE56192 [1]	Yes	No	Yes
Severe acute respiratory syndrome coronavirus 2 (SARS-CoV2)	Homo sapiens	A549, NHBE, Calu-3, RUES2-derived lung cells, M1 + M2 macrophages	mRNA, scRNA	4	38	GSE147507 [37] GSE149312 [47] GSE150708 [48] GSE153970 [49]	Yes	Yes	Yes

[1] No published study is currently associated with this dataset on NCBI's GEO.

3. Methods

The raw data for these experiments have been previously released by the primary authors and conform to the appropriate ethical oversight to protect patient autonomy and patient identity. Thirty of the 33 primary RNA-sequencing datasets from which we gathered samples for meta-analysis have been published in the peer-reviewed literature, increasing overall confidence that each dataset has acceptable quality.

GEO queries were used to identify all of the relevant publicly available RNA-seq experiments data from NCBI for each targeted condition. Samples involving drug experiments, treatments, xenografts, irrelevant tissue type, irrelevant disease, or otherwise unrelated to our disease vs. healthy comparisons were excluded. All samples that had one or more of these disqualifying attributes were excluded from the dataset prior to our analysis, meaning that only a subset of the samples from an individual experiment were represented in our meta-analyses. Healthy control samples were obtained from the same RNA sequencing projects as the disease samples.

Fastq sequencing files were downloaded from the Sequence Read Archive (SRA) using sratools. The fastq files, the associated metadata, and a configuration file for each dataset were then used as input to the Automated Reproducible MOdular Workflow for Preprocessing and Differential Analysis of RNA-seq Data (ARMOR) workflow [50]. This workflow uses a configuration file to appropriately set up each python-based snakemake workflow [51]. Specifically, this workflow trims reads with TrimGalore! [52], calculates quality control metrics with FastQC [53], maps and quantifies reads to the human GRCh38 transcriptome with Salmon [54], generates DEG lists with edgeR [55], performs GO enrichment with Camera [56], and calculates significant splice variants with DRIMseq [57]. Together, TrimGalore! and FastQC ensure that only the high-quality regions of sequences are considered in the statistical analyses performed by downstream modules, and that the quality of the included regions can be manually confirmed by the researcher at any point during the analysis. The DEGs from the ARMOR workflow were then used as input to an R script that implements the signaling pathway impact analysis (SPIA) algorithm to identify intracellular signaling pathways that were significantly represented by the DEGs [58].

Author Contributions: Conceptualization, B.E.P.; methodology, B.E.P.; formal analysis, N.R.-S., J.K., E.J.B., K.T.L.G., B.R.H., C.S.H., M.K., T.M.S. and B.E.P.; investigation, N.R.-S., J.K., E.J.B., K.T.L.G., B.R.H., C.S.H., M.K., T.M.S. and B.E.P.; writing—original draft preparation, N.R.-S.; writing—review and editing, N.R.-S., J.K. and B.E.P.; supervision, B.E.P. All authors have read and agreed to the published version of the manuscript.

Funding: We thank the BYU College of Life Sciences for providing the resources necessary to complete this work. This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study, due to the public availability of included datasets.

Informed Consent Statement: Patient consent was waived due to the datasets already being publicly available.

Data Availability Statement: The data we have announced in this publication is available for download online at Zenodo: https://zenodo.org/record/4757764, DOI:10.5281/zenodo.4757764.

Acknowledgments: We thank the high-performance computing resources provided by the BYU Research Computing Center. We also gratefully acknowledge those who generated, provided, and submitted the original data.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Ma, T.; Huo, Z.; Kuo, A.; Zhu, L.; Fang, Z.; Zeng, X.; Lin, C.-W.; Liu, S.; Wang, L.; Liu, P.; et al. MetaOmics: Analysis Pipeline and Browser-Based Software Suite for Transcriptomic Meta-Analysis. *Bioinformatics* **2019**, *35*, 1597–1599. [CrossRef]
2. Wang, H.-Q.; Zheng, C.-H.; Zhao, X.-M. JNMFMA: A Joint Non-Negative Matrix Factorization Meta-Analysis of Transcriptomics Data. *Bioinformatics* **2015**, *31*, 572–580. [CrossRef]
3. Menon, R.; Garg, G.; Gasser, R.B.; Ranganathan, S. TranSeqAnnotator: Large-Scale Analysis of Transcriptomic Data. *BMC Bioinform.* **2012**, *13* (Suppl. 17), S24. [CrossRef]
4. Jin, H.; Liu, Z. A Benchmark for RNA-Seq Deconvolution Analysis under Dynamic Testing Environments. *Genome Biol.* **2021**, *22*, 102. [CrossRef]
5. Medina, I.; Carbonell, J.; Pulido, L.; Madeira, S.C.; Goetz, S.; Conesa, A.; Tárraga, J.; Pascual-Montano, A.; Nogales-Cadenas, R.; Santoyo, J.; et al. Babelomics: An Integrative Platform for the Analysis of Transcriptomics, Proteomics and Genomic Data with Advanced Functional Profiling. *Nucleic Acids Res.* **2010**, *38*, W210–W213. [CrossRef]
6. Matikas, A.; Zerdes, I.; Lövrot, J.; Richard, F.; Sotiriou, C.; Bergh, J.; Valachis, A.; Foukakis, T. Prognostic Implications of PD-L1 Expression in Breast Cancer: Systematic Review and Meta-Analysis of Immunohistochemistry and Pooled Analysis of Transcriptomic Data. *Clin. Cancer Res.* **2019**, *25*, 5717–5726. [CrossRef]
7. Haas Bueno, R.; Recamonde-Mendoza, M. Meta-Analysis of Transcriptomic Data Reveals Pathophysiological Modules Involved with Atrial Fibrillation. *Mol. Diagn. Ther.* **2020**, *24*, 737–751. [CrossRef] [PubMed]
8. Aevermann, B.D.; Pickett, B.E.; Kumar, S.; Klem, E.B.; Agnihothram, S.; Askovich, P.S.; Bankhead, A.; Bolles, M.; Carter, V.; Chang, J.; et al. A Comprehensive Collection of Systems Biology Data Characterizing the Host Response to Viral Infection. *Sci. Data* **2014**, *1*, 140033. [CrossRef] [PubMed]
9. Kori, M.; Yalcin Arga, K. Potential Biomarkers and Therapeutic Targets in Cervical Cancer: Insights from the Meta-Analysis of Transcriptomics Data within Network Biomedicine Perspective. *PLoS ONE* **2018**, *13*, e0200717. [CrossRef] [PubMed]
10. Patel, H.; Dobson, R.J.B.; Newhouse, S.J. A Meta-Analysis of Alzheimer's Disease Brain Transcriptomic Data. *J. Alzheimers Dis.* **2019**, *68*, 1635–1656. [CrossRef] [PubMed]
11. Zhang, L.-L.; Zhang, Z.-N.; Wu, X.; Jiang, Y.-J.; Fu, Y.-J.; Shang, H. Transcriptomic Meta-Analysis Identifies Gene Expression Characteristics in Various Samples of HIV-Infected Patients with Nonprogressive Disease. *J. Transl. Med.* **2017**, *15*, 191. [CrossRef] [PubMed]
12. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef]
13. Taylor, C.F.; Field, D.; Sansone, S.-A.; Aerts, J.; Apweiler, R.; Ashburner, M.; Ball, C.A.; Binz, P.-A.; Bogue, M.; Booth, T.; et al. Promoting Coherent Minimum Reporting Guidelines for Biological and Biomedical Investigations: The MIBBI Project. *Nat. Biotechnol.* **2008**, *26*, 889–896. [CrossRef] [PubMed]
14. Nicolson, G.L.; Settineri, R.; Ellithorpe, R. Glycophospholipid Formulation with NADH and CoQ10 Significantly Reduces Intractable Fatigue in Western Blot-Positive 'Chronic Lyme Disease' Patients: Preliminary Report. *Funct. Foods Health Dis.* **2012**, *2*, 35–47. [CrossRef]
15. Czirjak, L.; Varju, C. Clinical Features of Scleroderma-Like Disorders: A Challenge for the Rheumatologist. *Curr. Rheumatol. Rev.* **2006**, *2*, 369–379. [CrossRef]
16. Finsterer, J.; Löscher, W.N.; Wanschitz, J.; Iglseder, S. Orphan Peripheral Neuropathies. *J. Neuromuscul. Dis.* **2021**, *8*, 1–23. [CrossRef] [PubMed]
17. González-Sanz, R.; Mata, M.; Bermejo-Martín, J.; Álvarez, A.; Cortijo, J.; Melero, J.A.; Martínez, I. ISG15 Is Upregulated in Respiratory Syncytial Virus Infection and Reduces Virus Growth through Protein ISGylation. *J. Virol.* **2016**, *90*, 3428–3438. [CrossRef]
18. Drori, Y.; Jacob-Hirsch, J.; Pando, R.; Glatman-Freedman, A.; Friedman, N.; Mendelson, E.; Mandelboim, M. Influenza A Virus Inhibits RSV Infection via a Two-Wave Expression of IFIT Proteins. *Viruses* **2020**, *12*, 1171. [CrossRef]
19. Zhang, Y.; Mao, D.; Roswit, W.T.; Jin, X.; Patel, A.C.; Patel, D.A.; Agapov, E.; Wang, Z.; Tidwell, R.M.; Atkinson, J.J.; et al. PARP9-DTX3L Ubiquitin Ligase Targets Host Histone H2BJ and Viral 3C Protease to Enhance Interferon Signaling and Control Viral Infection. *Nat. Immunol.* **2015**, *16*, 1215–1227. [CrossRef]
20. Ruiduo, C.; Ying, D.; Qiwei, W. CXCL9 Promotes the Progression of Diffuse Large B-Cell Lymphoma through up-Regulating β-Catenin. *Biomed. Pharmacother.* **2018**, *107*, 689–695. [CrossRef]
21. Shah, N.; Cabanillas, F.; McIntyre, B.; Feng, L.; McLaughlin, P.; Rodriguez, M.A.; Romaguera, J.; Younes, A.; Hagemeister, F.B.; Kwak, L.; et al. Prognostic Value of Serum CD44, Intercellular Adhesion Molecule-1 and Vascular Cell Adhesion Molecule-1 Levels in Patients with Indolent Non-Hodgkin Lymphomas. *Leuk. Lymphoma* **2012**, *53*, 50–56. [CrossRef]
22. Pullarkat, V.A.; Lacayo, N.J.; Jabbour, E.; Rubnitz, J.E.; Bajel, A.; Laetsch, T.W.; Leonard, J.; Colace, S.I.; Khaw, S.L.; Fleming, S.A.; et al. Venetoclax and Navitoclax in Combination with Chemotherapy in Patients with Relapsed or Refractory Acute Lymphoblastic Leukemia and Lymphoblastic Lymphoma. *Cancer Discov.* **2021**, *11*, 1440–1453. [CrossRef]
23. Faramand, R.; Jain, M.; Staedtke, V.; Kotani, H.; Bai, R.; Reid, K.; Lee, S.B.; Spitler, K.; Wang, X.; Cao, B.; et al. Tumor Microenvironment Composition and Severe Cytokine Release Syndrome (CRS) Influence Toxicity in Patients with Large B-Cell Lymphoma Treated with Axicabtagene Ciloleucel. *Clin. Cancer Res.* **2020**, *26*, 4823–4831. [CrossRef]

24. Li, M.; Chiang, Y.-L.; Lyssiotis, C.A.; Teater, M.R.; Hong, J.Y.; Shen, H.; Wang, L.; Hu, J.; Jing, H.; Chen, Z.; et al. Non-Oncogene Addiction to SIRT3 Plays a Critical Role in Lymphomagenesis. *Cancer Cell* **2019**, *35*, 916–931.e9. [CrossRef] [PubMed]
25. Porpaczy, E.; Tripolt, S.; Hoelbl-Kovacic, A.; Gisslinger, B.; Bago-Horvath, Z.; Casanova-Hevia, E.; Clappier, E.; Decker, T.; Fajmann, S.; Fux, D.A.; et al. Aggressive B-Cell Lymphomas in Patients with Myelofibrosis Receiving JAK1/2 Inhibitor Therapy. *Blood* **2018**, *132*, 694–706. [CrossRef]
26. Teater, M.; Dominguez, P.M.; Redmond, D.; Chen, Z.; Ennishi, D.; Scott, D.W.; Cimmino, L.; Ghione, P.; Chaudhuri, J.; Gascoyne, R.D.; et al. AICDA Drives Epigenetic Heterogeneity and Accelerates Germinal Center-Derived Lymphomagenesis. *Nat. Commun.* **2018**, *9*, 222. [CrossRef]
27. Raju, S.; Kretzmer, L.Z.; Koues, O.I.; Payton, J.E.; Oltz, E.M.; Cashen, A.; Polic, B.; Schreiber, R.D.; Shaw, A.S.; Markiewicz, M.A. NKG2D-NKG2D Ligand Interaction Inhibits the outgrowth of Naturally Arising Low-Grade B Cell Lymphoma In Vivo. *J. Immunol.* **2016**, *196*, 4805–4813. [CrossRef] [PubMed]
28. Rouhigharabaei, L.; Finalet Ferreiro, J.; Tousseyn, T.; van der Krogt, J.-A.; Put, N.; Haralambieva, E.; Graux, C.; Maes, B.; Vicente, C.; Vandenberghe, P.; et al. Non-IG Aberrations of FOXP1 in B-Cell Malignancies Lead to an Aberrant Expression of N-Truncated Isoforms of FOXP1. *PLoS ONE* **2014**, *9*, e85851. [CrossRef]
29. Verma, A.; Jiang, Y.; Du, W.; Fairchild, L.; Melnick, A.; Elemento, O. Transcriptome Sequencing Reveals Thousands of Novel Long Non-Coding RNAs in B Cell Lymphoma. *Genome Med.* **2015**, *7*, 110. [CrossRef] [PubMed]
30. Bouquet, J.; Soloski, M.J.; Swei, A.; Cheadle, C.; Federman, S.; Billaud, J.-N.; Rebman, A.W.; Kabre, B.; Halpert, R.; Boorgula, M.; et al. Longitudinal Transcriptome Analysis Reveals a Sustained Differential Gene Expression Signature in Patients Treated for Acute Lyme Disease. *MBio* **2016**, *7*, e00100–e00116. [CrossRef]
31. Jeong, I.; Lim, J.-H.; Oh, D.K.; Kim, W.J.; Oh, Y.-M. Gene Expression Profile of Human Lung in a Relatively Early Stage of COPD with Emphysema. *Int. J. Chronic Obstr. Pulm. Dis.* **2018**, *13*, 2643–2655. [CrossRef]
32. Li, M.; Zhao, L.-M.; Li, S.-L.; Li, J.; Gao, B.; Wang, F.-F.; Wang, S.-P.; Hu, X.-H.; Cao, J.; Wang, G.-Y. Differentially Expressed LncRNAs and MRNAs Identified by NGS Analysis in Colorectal Cancer Patients. *Cancer Med.* **2018**, *7*, 4650–4664. [CrossRef]
33. Deng, X.; Li, S.; Kong, F.; Ruan, H.; Xu, X.; Zhang, X.; Wu, Z.; Zhang, L.; Xu, Y.; Yuan, H.; et al. Long Noncoding RNA PiHL Regulates P53 Protein Stability through GRWD1/RPL11/MDM2 Axis in Colorectal Cancer. *Theranostics* **2020**, *10*, 265–280. [CrossRef] [PubMed]
34. Lazar, S.B.; Pongor, L.; Li, X.L.; Grammatikakis, I.; Muys, B.R.; Dangelmaier, E.A.; Redon, C.E.; Jang, S.-M.; Walker, R.L.; Tang, W.; et al. Genome-Wide Analysis of the FOXA1 Transcriptional Network Identifies Novel Protein-Coding and Long Noncoding RNA Targets in Colorectal Cancer Cells. *Mol. Cell. Biol.* **2020**, *40*. [CrossRef] [PubMed]
35. Lu, S.; Zhu, N.; Guo, W.; Wang, X.; Li, K.; Yan, J.; Jiang, C.; Han, S.; Xiang, H.; Wu, X.; et al. RNA-Seq Revealed a Circular RNA-MicroRNA-MRNA Regulatory Network in Hantaan Virus Infection. *Front. Cell. Infect. Microbiol.* **2020**, *10*, 97. [CrossRef] [PubMed]
36. Li, Y.; Quan, C.; Xing, W.; Wang, P.; Gao, J.; Zhang, Z.; Jiang, X.; Ma, C.; Carr, M.J.; He, Q.; et al. Rapid Humoral Immune Responses Are Required for Recovery from Haemorrhagic Fever with Renal Syndrome Patients. *Emerg. Microbes Infect.* **2020**, *9*, 2303–2314. [CrossRef] [PubMed]
37. Blanco-Melo, D.; Nilsson-Payant, B.E.; Liu, W.-C.; Uhl, S.; Hoagland, D.; Møller, R.; Jordan, T.X.; Oishi, K.; Panis, M.; Sachs, D.; et al. Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell* **2020**, *181*, 1036–1045.e9. [CrossRef] [PubMed]
38. Jenks, S.A.; Cashman, K.S.; Zumaquero, E.; Marigorta, U.M.; Patel, A.V.; Wang, X.; Tomar, D.; Woodruff, M.C.; Simon, Z.; Bugrovsky, R.; et al. Distinct Effector B Cells Induced by Unregulated Toll-like Receptor 7 Contribute to Pathogenic Responses in Systemic Lupus Erythematosus. *Immunity* **2018**, *49*, 725–739.e6. [CrossRef]
39. Scharer, C.D.; Blalock, E.L.; Mi, T.; Barwick, B.G.; Jenks, S.A.; Deguchi, T.; Cashman, K.S.; Neary, B.E.; Patterson, D.G.; Hicks, S.L.; et al. Epigenetic Programming Underpins B Cell Dysfunction in Human SLE. *Nat. Immunol.* **2019**, *20*, 1071–1082. [CrossRef]
40. Wang, S.; Wang, J.; Kumar, V.; Karnell, J.L.; Naiman, B.; Gross, P.S.; Rahman, S.; Zerrouki, K.; Hanna, R.; Morehouse, C.; et al. IL-21 Drives Expansion and Plasma Cell Differentiation of Autoreactive CD11chiT-Bet+ B Cells in SLE. *Nat. Commun.* **2018**, *9*, 1758. [CrossRef]
41. Zhang, X.; Chu, H.; Wen, L.; Shuai, H.; Yang, D.; Wang, Y.; Hou, Y.; Zhu, Z.; Yuan, S.; Yin, F.; et al. Competing Endogenous RNA Network Profiling Reveals Novel Host Dependency Factors Required for MERS-CoV Propagation. *Emerg. Microbes Infect.* **2020**, *9*, 733–746. [CrossRef]
42. Yuan, S.; Chu, H.; Chan, J.F.-W.; Ye, Z.-W.; Wen, L.; Yan, B.; Lai, P.-M.; Tee, K.-M.; Huang, J.; Chen, D.; et al. SREBP-Dependent Lipidomic Reprogramming as a Broad-Spectrum Antiviral Target. *Nat. Commun.* **2019**, *10*, 120. [CrossRef]
43. Aprianto, R.; Slager, J.; Holsappel, S.; Veening, J.-W. Time-Resolved Dual RNA-Seq Reveals Extensive Rewiring of Lung Epithelial and Pneumococcal Transcriptomes during Early Infection. *Genome Biol.* **2016**, *17*, 198. [CrossRef] [PubMed]
44. Kuipers, K.; Lokken, K.L.; Zangari, T.; Boyer, M.A.; Shin, S.; Weiser, J.N. Age-Related Differences in IL-1 Signaling and Capsule Serotype Affect Persistence of Streptococcus Pneumoniae Colonization. *PLoS Pathog.* **2018**, *14*, e1007396. [CrossRef]
45. Jochems, S.P.; Marcon, F.; Carniel, B.F.; Holloway, M.; Mitsi, E.; Smith, E.; Gritzfeld, J.F.; Solórzano, C.; Reiné, J.; Pojar, S.; et al. Inflammation Induced by Influenza Virus Impairs Human Innate Immune Control of Pneumococcus. *Nat. Immunol.* **2018**, *19*, 1299–1308. [CrossRef] [PubMed]

46. Weight, C.M.; Venturini, C.; Pojar, S.; Jochems, S.P.; Reiné, J.; Nikolaou, E.; Solórzano, C.; Noursadeghi, M.; Brown, J.S.; Ferreira, D.M.; et al. Microinvasion by Streptococcus Pneumoniae Induces Epithelial Innate Immunity during Colonisation at the Human Mucosal Surface. *Nat. Commun.* **2019**, *10*, 3060. [CrossRef] [PubMed]
47. Lamers, M.M.; Beumer, J.; van der Vaart, J.; Knoops, K.; Puschhof, J.; Breugem, T.I.; Ravelli, R.B.G.; Paul van Schayck, J.; Mykytyn, A.Z.; Duimel, H.Q.; et al. SARS-CoV-2 Productively Infects Human Gut Enterocytes. *Science* **2020**, *369*, 50–54. [CrossRef]
48. Duan, F.; Guo, L.; Yang, L.; Han, Y.; Thakur, A.; Nilsson-Payant, B.E.; Wang, P.; Zhang, Z.; Ma, C.Y.; Zhou, X.; et al. Modeling COVID-19 with Human Pluripotent Stem Cell-Derived Cells Reveals Synergistic Effects of Anti-Inflammatory Macrophages with ACE2 Inhibition Against SARS-CoV-2. *Res. Sq.* **2020**. [CrossRef]
49. Vanderheiden, A.; Ralfs, P.; Chirkova, T.; Upadhyay, A.A.; Zimmerman, M.G.; Bedoya, S.; Aoued, H.; Tharp, G.M.; Pellegrini, K.L.; Manfredi, C.; et al. Type I and Type III Interferons Restrict SARS-CoV-2 Infection of Human Airway Epithelial Cultures. *J. Virol.* **2020**, *94*. [CrossRef] [PubMed]
50. Orjuela, S.; Huang, R.; Hembach, K.M.; Robinson, M.D.; Soneson, C. ARMOR: An Automated Reproducible MOdular Workflow for Preprocessing and Differential Analysis of RNA-Seq Data. *G3 Genes Genomes Genet.* **2019**, *9*, 2089–2096. [CrossRef] [PubMed]
51. Köster, J.; Rahmann, S. Snakemake—A Scalable Bioinformatics Workflow Engine. *Bioinformatics* **2012**, *28*, 2520–2522. [CrossRef] [PubMed]
52. Babraham Bioinformatics—Trim Galore! Available online: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (accessed on 7 June 2021).
53. Babraham Bioinformatics—FastQC A Quality Control Tool for High Throughput Sequence Data. Available online: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 7 June 2021).
54. Patro, R.; Duggal, G.; Love, M.I.; Irizarry, R.A.; Kingsford, C. Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression. *Nat. Methods* **2017**, *14*, 417–419. [CrossRef] [PubMed]
55. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics* **2010**, *26*, 139–140. [CrossRef] [PubMed]
56. Wu, D.; Smyth, G.K. Camera: A Competitive Gene Set Test Accounting for Inter-Gene Correlation. *Nucleic Acids Res.* **2012**, *40*, e133. [CrossRef]
57. Nowicka, M.; Robinson, M.D. DRIMSeq: A Dirichlet-Multinomial Framework for Multivariate Count Outcomes in Genomics. *F1000Research* **2016**, *5*, 1356. [CrossRef] [PubMed]
58. Tarca, A.L.; Draghici, S.; Khatri, P.; Hassan, S.S.; Mittal, P.; Kim, J.-S.; Kim, C.J.; Kusanovic, J.P.; Romero, R. A Novel Signaling Pathway Impact Analysis. *Bioinformatics* **2009**, *25*, 75–82. [CrossRef]