Computational analysis and interpretation of multi-omics data

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

Dr. rer. nat.

der Fakultät für

Biologie

an der

Universität Duisburg-Essen

vorgelegt von

Yingying Cao

aus Henan (Shangqiu), China

Februar 2021

Essen, Deutschland

Die der vorliegenden Arbeit zugrunde liegenden Experimente wurden in den Abteilungen für Bioinformatik des Zentrums für Medizinische Biotechnologie (ZMB) der Universität Duisburg-Essen durchgeführt.

1. Gutachter: Prof. Dr. Daniel Hoffmann
2. Gutachter: Prof. Dr. Andrea Vortkamp
3. Gutachter:

Vorsitzender des Prüfungsausschusses: PD Dr. Marc Seifert
Tag der mündlichen Prüfung: 07.05.2021

DOI: 10.17185/duepublico/74422
URN: urn:nbn:de:hbz:464-20210528-135617-7

Contents

Summary

In the last decade, with huge advances in high throughput sequencing (HT-seq) technologies and rapidly decreasing costs, cell and molecular biology are becoming increasingly a heavily "data-driven" science. HT-seq has transformed the scientific landscape in biology by allowing researchers to answer important biological questions in multiple biological layers with multi-omics data. In this thesis, I will introduce my work on computational analysis, interpretation and application of multi-omics data on bulk as well on single cell levels.

RNA-seq, the next-generation sequencing of RNAs is a powerful method to characterize genome-wide differential gene expression between different conditions. ChIP-seq, the high throughput chromatin immuno-precipitation sequencing technology has been a powerful tool to identify genome-wide profiles of histone modifications which have been identified to be the key epigenetic mechanisms in the regulation of gene expression. More and more studies start analyzing simultaneously the combination of RNA-seq data and ChIP-seq data of different histone modifications across different conditions. The integrative analysis of these corresponding data sets, in principle, becomes a desirable option to study gene regulation in the complex and dynamic biological processes for example in organ development and disease progression. However, computational tools for such analyses are still technically in their infancy. In the first part of this thesis, I introduce intePareto, a novel method to prioritize genes with consistent changes in RNA-seq and ChIP-seq data of different histone modifications between different conditions using Pareto optimization.

In addition to the rapid development and applications in bulk sequencing of pooled cell populations discussed above, the past decade has witnessed tremendous progress in single cell RNA sequencing (scRNA-seq) technologies which have further revolutionized our understanding of the fundamental biological and physiological phenomena at the single cell resolution. The scRNA-seq technology allows unprecedented detailed characterizations of heterogeneity of cell populations previously believed to be homogeneous, or identification of a continuous spectrum cell trajectory previously hidden in pooled cell populations. However, scRNA-seq also brings computational challenges due to the small amount of material available in each single cell for sequencing, resulting in high sparsity of the data with abundance of observed zeros also known as "dropout" or zero-inflation in scRNA-seq counts. The high proportion of zeros observed in many genes poses a big challenge for

further downstream data analysis and interpretation, and is therefore a major research focus. Some believe the abundant zeros are attributed to technical artifacts and should be corrected with non-zeros, thus different imputation methods and tools have been designed to explicitly correct the zeros, i.e. to impute the "dropout" with appropriate values to hopefully better represent the true expression values. Zero-inflated models are therefore widely used to model the scRNA-seq data, and zero-inflation is even treated as an inherent property of scRNA-seq data. However, this "dropout" or zero-inflation problem is far from being fully understood. It is necessary to understand the source of observed zeros before imputation method or zero-inflated model is designed and adopted. In the second part of this thesis, we provide convincing empirical evidence showing that the dichotomy of zero-inflation in scRNA-seq data is between read counts and UMI counts, and not between droplet-based and plate-based platforms, and that large number of "unexpected zeros" (zero-inflation) in read counts are due to amplification bias, and should not be blindly imputed or modeled by zero-inflation models.

From the end of 2019, there was an unprecedented COVID-19 pandemic caused by SARS-CoV-2. COVID-19 in severe form is a systemic disease leading to multi-organ dysfunction. The current research on SARS-CoV-2/COVID-19 with respect to virus entry routes and innate immune responses is still in a paradoxical state: the rapid accumulation of data frequently also increases the confusion about what we actually know. One reason for this paradox could be that the bulk of the data comes from many small studies from which general conclusions are drawn overhastily. In this situation, a meta-study that analyzes larger clusters of comparable data from several studies could bring more clarity. In the third part of this thesis, I introduce comprehensive comparative analyses with RNA-seq data sets of different cells infected with SARS-CoV, MERS-CoV and SARS-CoV-2, as well as RNA-seq data from COVID-19 patients. In addition, the dynamics of neutrophils and neutrophil extracellular traps are also examined in the progression of COVID-19. We have presented evidence for multiple SARS-CoV-2 entry mechanisms. We have also dissolved apparent conflicts on cellular innate immune responses to SARS-CoV-2 infection. Our results emphasize the complex interactions between host cells and SARS-CoV-2, offering new insights into the pathogenesis of SARS-CoV-2, and can further inform the development of antiviral drugs.

In brief, in this thesis I have examined various topics in regard to computational integration, interpretation of high-throughput sequencing data in bulk and single cell levels, as well as the application of large scale sequencing data analysis and interpretation to gain insights into the pathogenesis of SARS-CoV-2 to help combat COVID-19 pandemic.

Zusammenfassung

In den letzten zehn Jahren hat sich die Zell- und Molekularbiologie mit enormen Fortschritten bei der Hochdurchsatz-Sequenzierung (HT-seq) und schnell sinkenden Kosten zunehmend zu einer stark datengetriebenen Wissenschaft entwickelt. In dieser Dissertation stelle ich meine Arbeit zur rechnergestützten Analyse und Interpretation von Multi-Omics-Daten sowohl von Zellgemischen als auch auf Einzelzellebene vor.

RNA-seq, die HT-seq von RNAs, ist eine leistungsstarke Methode zur Charakterisierung der genomweiten differentiellen Genexpression zwischen verschiedenen Bedingungen. ChIP-seq ist eine HT-seq-Technik, zur Identifizierung genomweiter Profile von Histonmodifikationen – ein wichtiger epigenetischer Mechanismus zur Regulation der Genexpression. Immer mehr Studien analysieren Kombinationen von RNA-seq-Daten und ChIP-seq-Daten verschiedener Histonmodifikationen unter verschiedenen Bedingungen. Die integrative Analyse der entsprechenden Datensätze kann ein Licht auf Genregulation in komplexen Prozessen werfen, beispielsweise in der Organentwicklung und in Krankheitsverläufen. Berechnungswerkzeuge für solche Analysen stecken jedoch noch in den Kinderschuhen. Im ersten Teil dieser Arbeit stelle ich inctePareto vor, eine neuartige Methode zur Priorisierung von Genen mit konsistenten Änderungen der RNA-seq- und ChIP-seq-Daten verschiedener Histonmodifikationen zwischen verschiedenen Bedingungen unter Verwendung der Pareto-Optimierung.

Zusätzlich zu der oben diskutierten raschen Entwicklung und Anwendung bei der Massensequenzierung gepoolter Zellpopulationen wurden in den letzten zehn Jahren enorme Fortschritte bei der Einzelzell-RNA-Sequenzierung (scRNA-seq) erzielt, die unser Verständnis grundlegender biologischer und physiologischer Phänomene revolutioniert. Die scRNA-seq-Technologie ermöglicht beispiellos detaillierte Charakterisierungen der Heterogenität von Zellpopulationen, von denen zuvor angenommen wurde, dass sie homogen sind, oder die Identifizierung kontinuierlicher Trajektorien zwischen Zellzuständen. ScRNA-seq bringt jedoch aufgrund der geringen Menge an Material, die in jeder einzelnen Zelle für die Sequenzierung verfügbar ist, große interpretatorische Probleme mit sich. Das ist zum einen die Spärlichkeit von Daten, zum anderen die Fülle von beobachteten Null-Expressionen von Genen, die auch als “Dropout” oder Null-Inflation bezeichnet wird. Einige glauben, dass die häufig vorkommenden Nullen technischen Artefakten zugeschrieben werden und mit Nicht-Nullen korrigiert werden sollten. Daher wurden verschiedene Imputationsmethoden entwickelt, um die Nullen durch vermeintlich geeignetere Werte zu er-

setzen. Null-Inflations-Modelle werden häufig zur Modellierung dieser scRNA-seq-Daten verwendet, und Null-Inflation wird oft als inhärente Eigenschaft von scRNA-seq-Daten behandelt. Dieses "Dropout" - oder Null-Inflations-Problem ist jedoch weit davon entfernt, vollständig verstanden zu werden. Es ist notwendig, die Quelle der beobachteten Nullen zu verstehen, bevor Imputationsmethoden oder Null-Inflations-Modelle angewendet werden

Im zweiten Teil dieser Arbeit liefern wir überzeugende empirische Beweise dafür, dass das Auftreten von Null-Inflation in scRNA-seq-Daten nichts zu tun hat mit prinzipiellen Unterschieden zwischen tröpfchen- und plattenbasierten scRNA-seq Plattformen. Das Problem der Null-Inflation ist eher eines der Analyse von Reads im Gegensatz zu UMIs, weil erstere verzerrt werden durch Amplifikationsmethoden wie PCR, letzere aber nicht. Vom blinden Anwenden von Imputationsmethoden zur Korrektur von Null-Inflation ist also abzuraten.

Seit Ende 2019 grassiert COVID-19, eine beispiellose Pandemie, verursacht durch SARS-CoV-2. COVID-19 in schwerer Form ist eine systemische Erkrankung, die zum Versagen mehrerer Organe führen kann. Die aktuelle Forschung zu SARS-CoV-2 / COVID-19 in Bezug auf Viruseintrittswege und angeborene Immunantworten befindet sich in einem paradoxen Zustand: Einerseits nimmt die Datenmenge rapide zu, andererseits auch die Verwirrung darüber, was wir tatsächlich wissen. Ein Grund für dieses Paradox könnte sein, dass der Großteil der Daten aus vielen kleinen Studien stammt, aus denen überstürzt allgemeine Schlussfolgerungen gezogen werden. In dieser Situation könnte eine Metastudie, die größere Cluster vergleichbarer Daten aus mehreren Studien analysiert, mehr Klarheit bringen.

Im dritten Teil dieser Arbeit stelle ich umfassende vergleichende Analysen mit RNA-seq-Datensätzen verschiedener mit SARS-CoV, MERS-CoV und SARS-CoV-2 infizierter Zellen sowie RNA-seq-Daten von COVID-19-Patienten vor. Darüber hinaus wird die Dynamik von Neutrophilen und extrazellulären Neutrophilenfallen im Verlauf von COVID-19 untersucht. Wir legen Beweise für mehrere SARS-CoV-2-Eintrittsmechanismen vor, und wir lösen vermeintliche Konflikte bezüglich zellulärer angeborener Immunantworten auf SARS-CoV-2-Infektionen. Unsere Ergebnisse betonen die komplexen Wechselwirkungen zwischen Wirtszellen und SARS-CoV-2, bieten neue Einblicke in die Pathogenese von SARS-CoV-2, und können die Entwicklung antiviraler Medikamente fördern.

Chapter 1

Introduction

1.1 Gene regulation and histone modifications

1.1.1 Structure of DNA

A genome is a complete set of genetic instructions in an organism. It consists of DNA. DNA is formed as a double helix by two complementary strands of nucleotides that are covalently linked together. There are four types of nucleotides, each with a sugar-phosphate backbone attached by either of four chemical bases including adenine (A), guanine (G), cytosine (C), and thymine (T) (Alberts et al., 2002). A always pairs with T, and C always pairs with G, to form units called base pairs. The ends of the DNA strands indicate the polarities of the two strands by referring to one end as the 3' end and the other as the 5' end. The genetic information is stored as the order or the sequence of these four base pairs in DNA. (figure 1.1)

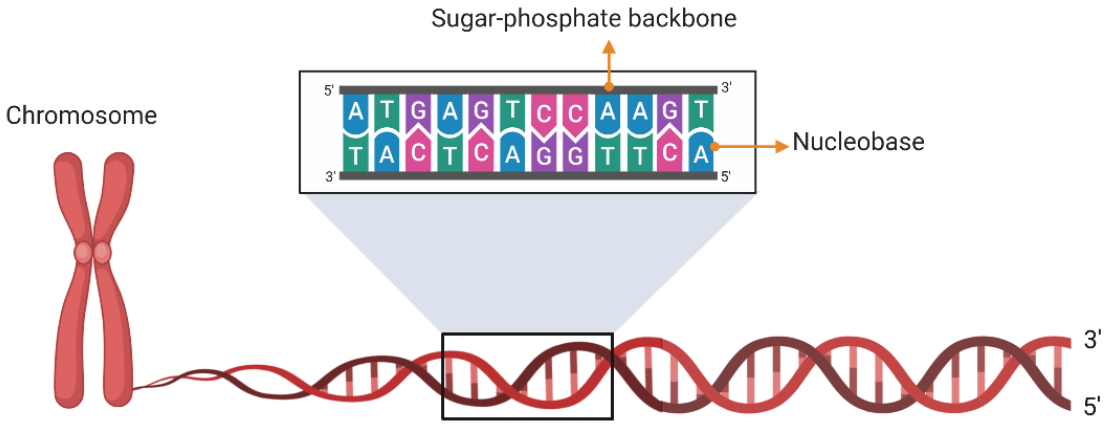


Figure 1.1: **DNA and its building blocks.**

1.1.2 Gene transcription

Gene refers to the basic unit of DNA carrying the genetic information that encodes the synthesis of a gene product, either protein or RNA. Transcription is the first step of gene expression. It is a process of RNA molecule synthesis in which a particular segment of DNA is transcribed into RNA. Different types of RNAs can be produced, including mNRA (messenger RNA). These synthesized RNA molecules are called transcripts. Different from DNA, RNA molecules are single-stranded. RNA nucleotides also have A, G and C, but instead of T they have another pyrimidine base called U (uracil). In eukaryotic cells, a precursor mRNA (pre-mRNA) synthesized by transcription of a gene's DNA template, usually undergoes several major processing events before the mature mRNA can get exported out of the nucleus. These processing events include acquisition of a 5' cap structure (Mizumoto and Kaziro, 1987), splicing of introns (Will and Lührmann, 1997), and the formation of a 3' end, usually modified by adding a poly-A tail (Manley and Di Giammartino, 2013). (figure 1.2)

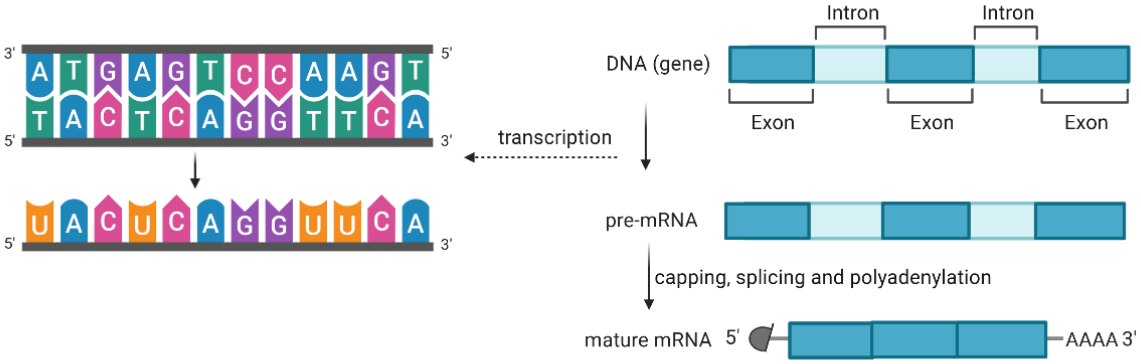


Figure 1.2: **Gene transcription.**

1.1.3 Structure of Chromatin

In eukaryotic cells, DNA in the nucleus is packed into a highly condensed structure called chromatin. The fundamental units of chromatin are nucleosomes, which are compactly arranged like "beads on a string", where the "string" is DNA and "beads" are nucleosomes. A single nucleosome core particle is assembled in an octameric DNA-protein complex with about 145-147 bp of DNA sequence tightly wrapped around a histone octamer core consisting of two copies of four histones H2A, H2B, H3 and H4 (Luger et al., 1997). The amino(N)-terminal "tails" of histones are the subject of numerous post-translational modifications (PTMs) such as acetylation, methylation, phosphorylation, ubiquitylation and sumoylation (Peterson and Laniel, 2004; Shah et al., 2020) (figure 1.3).

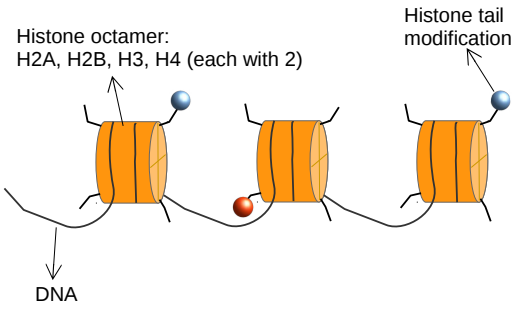


Figure 1.3: **"beads on a string" nucleosome array.**

1.1.4 Gene regulation

Even though nearly all cells in an organism contain essentially the same DNA, the cell types and cell functions differ because of the differences in their gene expression patterns. For instance, even closely related proliferating chondrocytes and hypertrophic chondrocytes can show distinct gene expression patterns (Wuelling et al., 2020). How does a cell know which sets of genes should be expressed in it? How does a gene know when it should be expressed? These questions can be answered by the study of gene regulation which is at the heart of all the complex biological processes during development, cell differentiation and disease progression.

Gene regulation is often primarily controlled at the level of transcription, i.e. which part of the DNA is transcribed to make an RNA molecule and the amount of RNA molecules are transcribed for a specific gene. Gene regulation mechanisms are very complex and dynamic processes involving numerous steps. Histone modifications are thought to play a crucial role in gene regulation by altering the extent of which DNA is wrapped around histones, the availability of genes in the DNA for transcription, as well as by recruiting proteins and complexes, for example like RNA polymerase II (RNAPII), to catalyze the gene transcription (Bannister and Kouzarides, 2011; Gibney and Nolan, 2010; Kouzarides, 2007b; Sims III et al., 2004) (figure 1.4).

1.1.5 Histone modifications

Different histone modifications are associated with different functions, which can occur combinatorially to form a "histone code" that is read by other proteins to switch transcription on or off, leading to activation or repression of target genes (Jenuwein and Allis, 2001; Strahl and Allis, 2000). Histone acetylation and histone methylation are well-explored PTMs (Barski et al., 2007; Benevolenskaya, 2007; Koch et al., 2007). Histone acetylations (e.g. H3K9ac and H3K27ac) are mainly linked to active transcription (Creyghton et al., 2010; Wang et al., 2008). Histone methylations like H3K9me3 and H3K27me3 are known as repressive marks and are associated with gene repression (Barski et al., 2007; Wang et al., 2008); H3K36me3 (Bannister et al., 2005), H3K4me1 (Benevolenskaya, 2007) and

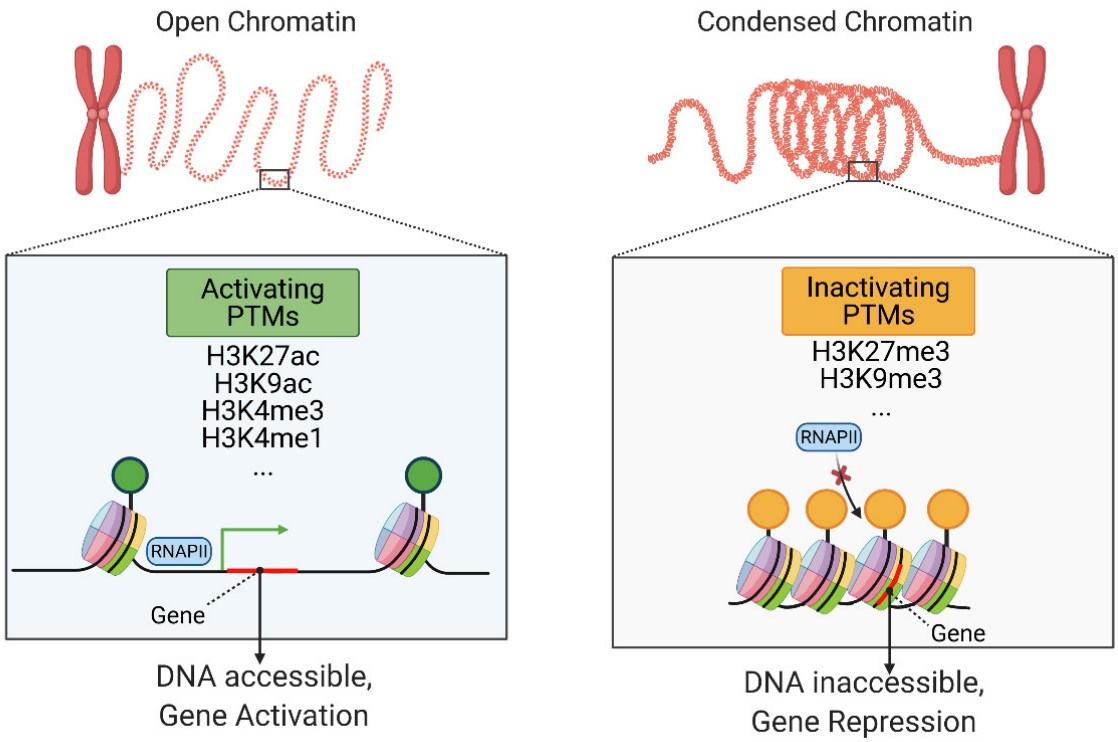


Figure 1.4: **Histone modification and gene expression regulation.**

H3K4me3 (Koch et al., 2007) are known as active marks and are often associated with gene activation (figure 1.4). Combination of different histone modification marks indicates different chromatin states. Consequently, aberrant histone modifications have been suggested to be involved in disease pathology by eliciting pathological gene expression programs (Mirabella et al., 2016).

1.1.6 Applications of RNA-seq and ChIP-seq technology

In recent years, high-throughput RNA sequencing (RNA-seq) has become a powerful technology for genome wide gene expression profiling, enabling extensively quantification of differences in gene expression between different conditions including different tissues and disease states. In combination with RNA-seq, chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq) has enabled researchers to investigate the relationships between histone modification marks and the regulation of gene transcription on a genome-wide scale. Genome-wide mapping of histone modification marks allows one to systematically catalogue the patterns of histone modifications, which are essential for full understanding of gene expression regulation at the level of transcription. ChIP-seq offers higher resolution, less noise, and greater coverage than its array-based predecessor ChIP-chip. With the decreasing cost of sequencing, ChIP-seq has become an indispensable tool for studying gene regulation and epigenetic mechanisms.

1.1.7 Integration of RNA-seq and ChIP-seq data

Computational methods that integrate RNA-seq data and ChIP-seq data of different histone modification marks are highly desirable to gain insights into the epigenetic regulation of transcription during development and disease progression. There are two important aspects that have to be considered. One is quantitative matching of RNA-seq and ChIP-seq data on the gene level, the other one is the choice of effective approach for the specific integration of high dimensional genomic data analysis.

Quantification of ChIP-seq data still remains a challenge. First, the evaluation of enriched peaks is difficult due to the lack of ground truth annotation (Nakato and Shirahige, 2017). In addition, many important histone modification marks like H3K9me3 and H3K27me3 do not occur in narrow well-defined peaks, but show broadly diffusing patterns (Beisel and Paro, 2011; Kouzarides, 2007a) with low signal-to-noise ratios at effective modification regions. Consequently, many false positives and false negative peaks are usually generated during peak calling analysis in ChIP-seq data of such histone modification marks. Another challenge for peak calling is the difficulty of handling the low reproducibility of peaks across replicates (Chen et al., 2012). Second, a simple fold ratio of the signal for the ChIP sample relative to that of the control sample around the peak provides important information for quantification, but it is still not adequate. For example, a fold ratio of 2 estimated from 20 and 10 tags (ChIP/control) has a different statistical significance from the same ratio estimated from 200 and 100 tags (Park, 2009). Some studies (Karlić et al., 2010) take the sum of tag counts surrounding transcription start site (TSS), which needs further normalization by the number of TSSs taken into consideration for each gene. This strategy fails to capture signals of several modifications that have greater enrichment in gene bodies, such as H3K36me3 (Barski et al., 2007). Recent studies prefer "best-bin" strategy (Dong et al., 2012; Singh et al., 2016), which searches for the bins showing the best correlation between chromatin feature signal and the expression level. However, the window selection is still a problem to maximize useful information content and minimize the incorporation of noise (Hoang et al., 2011).

Recently multiple computational models have been proposed to use histone modification marks to predict gene expression (Singh et al., 2016; Zeng et al., 2020). However, such integration analyses only focus on RNA-seq and ChIP-seq in one condition. One important task in biology is to understand how cellular function changes between different conditions, for example, different stages of disease progression or different developmental stages. One usual way of understanding such functional difference is to characterize differential gene expression. It is even more attractive to combine evidence from both measurements of gene expression in RNA-seq and evidence from measurements of various histone modifications in ChIP-seq, which allows assessment of activation or suppression state of genes. This combination of information from different biological layers gives a clearer picture of the cellular function than using any one of the data types alone.

We have therefore developed the R package named inte Pareto that allows such an

integrative analysis of different types of sequencing data. The intePareto workflow starts with RNA-seq and ChIP-seq data for two different cell types or conditions. The ChIP-seq data will in general comprise information of several histone modifications with activating or repressing function. The end product of intePareto is a list of genes prioritized according to congruence of changes in gene expression and histone modifications. We have applied this method to study epigenetic gene regulation mechanisms mediating cell state transitions from proliferating chondrocytes into hypertrophic chondrocytes during endochondral ossification with biological meaningful outputs (Wuelling et al., 2020).

1.2 Single cell RNA sequencing technology

1.2.1 Applications of scRNA-seq

In the last decade, breakthroughs in single cell RNA sequencing (scRNA-seq) technologies have further revolutionized our understanding of biological systems at the finest resolution – single cell level (Anchang et al., 2016; Sandberg, 2014). Instead of profiling cell populations, it is now possible to profile transcript abundance of an individual cell, which provides new opportunities for studying cellular heterogeneity and dynamic processes.

With the decreasing cost and unprecedented opportunities provided by scRNA-seq technologies, scRNA-seq has become routine in recent biological and medical research, for example (1) to examine cell-to-cell heterogeneity or cell type diversity in immune cell heterogeneity (Papalexi and Satija, 2018), cancer heterogeneity (González-Silva et al., 2020), and cell classification in neuroscience (Tasic, 2018); (2) to infer the order of cells along developmental trajectories (Saelens et al., 2019; Trapnell et al., 2014) during development and differentiation processes; etc.

1.2.2 Lab protocols for scRNA-seq

Single cell RNA sequencing lab protocols can be generally categorized as "full-length" or "tag-based" protocols. The "full-length" protocols including Smart-seq2 (Picelli et al., 2014) and Smart-seq3 (Hagemann-Jensen et al., 2020) allow full-length coverage across transcripts, enabling quantification of isoforms, analysis of alternative splicing or detection of single nucleotide variants. Tag-based protocols including the widely used 10X (10X Genomics Chromium, 10X Genomics, Pleasanton, CA) protocols (Zheng et al., 2017b) only capture either the 5'- or 3'-end of each RNA with the limitations of reduction in mappability and causing difficulty in quantification of isoforms or detection of single nucleotide variants. Previously, the main advantage of tag-based protocols is that they can be combined with UMI (Unique Molecular Identifier) by ligating a random nucleotide sequence onto each strand of DNA fragment before PCR (polymerase chain reaction) or IVT (in vitro transcription) amplification (Hagemann-Jensen et al., 2020; Hashimshony et al., 2016; Kivioja et al., 2012). Sequenced reads with the same UMI can be easily

identified as duplicates which makes it possible to computationally eliminate the effects of amplification bias (Islam et al., 2014). This is particularly important for quantification where many amplification cycles are required for single cell RNA sequencing with extremely small quantity of starting materials available in individual cells. However, the very recently developed Smart-seq3 (Hagemann-Jensen et al., 2020) protocol combines the advantages of full-length sequencing as well as the combination of UMI, making it a powerful and promising technology in the future.

1.2.3 Challenges in scRNA-seq data analysis

Imputation or not?

Single cell RNA sequencing brings computational challenges alongside as well due to the limited amount of material obtained from individual cells for sequencing, leading to high level of sparsity i.e. large amount of observed zeros in the resulting data which pose challenges for further downstream data analysis and interpretation. With publicly available scRNA-seq data from Smart-seq2 protocol and the corresponding bulk RNA-seq data from the same study (Zheng et al., 2017a), we show that the degree of sparsity or the proportion of zeros depends on the sequencing depth and the expression level of the gene (figure 1.5). Very recent research also highlights that sequencing depth explains most (95%) of the variation in the number of observed zeros per cell with the UMI count data (Choi et al., 2020). The term of "dropout" is often widely used to denote the observed high frequency of zeros in scRNA-seq data and usually modeled with zero-inflation models (Kharchenko et al., 2014; Lopez et al., 2018; Pierson and Yau, 2015). During the last few years, different imputation methods have been proposed to impute the "dropout" i.e. to correct the zeros with non-zeros in the data, to predict the expression level of a gene had there been no "dropout" or zero-inflation (Gong et al., 2018; Huang et al., 2018; Li and Li, 2018; Van Dijk et al., 2018).

Recent studies (Andrews and Hemberg, 2018; Choi et al., 2020; Hou et al., 2020) start questioning the practice of imputation, implying imputation can mislead the downstream data analysis and interpretation by introducing bias in estimation of gene expression levels and mask biologically cellar heterogeneity and transcriptional stochasticity. A benchmark study points out that false signals or irreproducible identification of cell-type specific markers can be introduced with imputed scRNA-seq data (Andrews and Hemberg, 2018). Another benchmark study shows that, in comparison with original data, imputation doesn't improve performance in downstream data analyses (Hou et al., 2020). Even so, the "dropout" problem has not yet been fully examined and understood, further investigations are needed to better understand the source of zeros before the adoption and application of imputation tools and zero-inflation models.

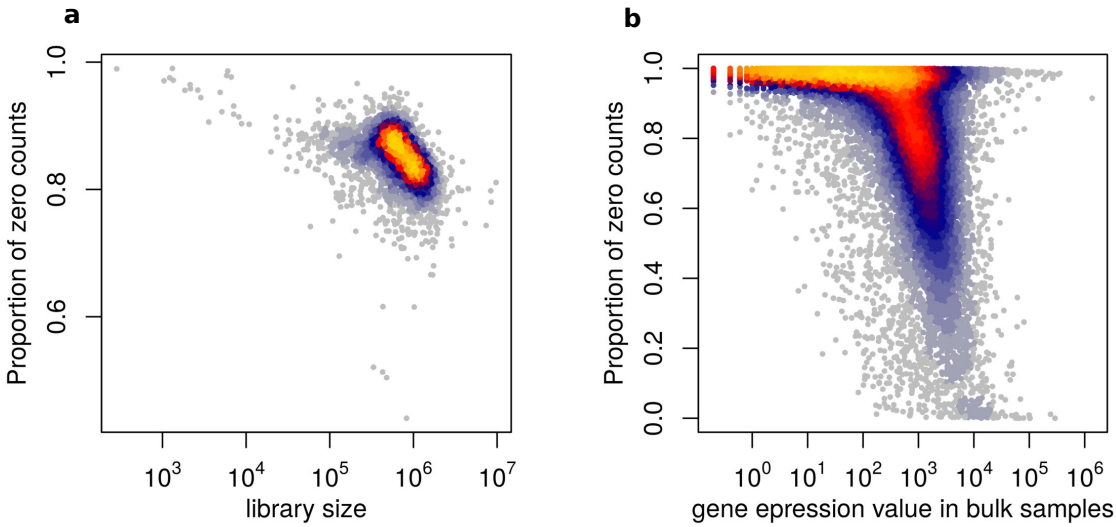


Figure 1.5: **Examination of characteristics of observed zeros in scRNA-seq data.** (a) Each dot represents a cell, library size is the sum of all the read counts for each cell, proportion of zero counts is the proportion of genes with zero counts in a cell. (b) Each dot represents a gene, x-axis represents the mean of each gene's expression values measured in read counts in corresponding bulk samples, y-axis represents the proportion of zero counts for each gene in the corresponding single cells. Data used here are from GEO (Gene Expression Omnibus) with accession number GSE98638 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE98638`).

Zero-inflation or not?

Recently, a paper written by Svensson (Svensson, 2020) highlights the potential of previous misunderstanding of "zero-inflation" phenomenon in scRNA-seq data and points out that the high throughput droplet-based data are not zero-inflated and can be sufficiently modeled using a negative binomial distribution, whereas plate-based data need "zero-inflation" to be accurately fitted (Svensson, 2020). One of the possible reasons he speculated was that uneven sampling of fragments from gene bodies in plate-based methods introduced additional layer of count noise (Svensson, 2020).

Actually, both droplet-based and plate-based protocols can incorporate UMI (Hagemann-Jensen et al., 2020; Hashimshony et al., 2016; Kivioja et al., 2012). For these UMI-based protocols, we usually only use the UMI counts for further analysis. The read counts before collapsing to UMIs are also available (figure 1.6) for analysis which has always been ignored in the previous research. Leveraging the fact that one can count reads in data sets that use UMIs, we show that, in the same data set, the read counts exhibit zero-inflation, while the UMI counts do not, which is very convincing empirical evidence to point out that the dichotomy of zero-inflation in scRNA-seq data is between read counts and UMI counts, and not between droplet-based and plate-based platforms (Cao et al., 2021a).

Development of new methods and protocols for scRNA-seq is currently a very dynamic

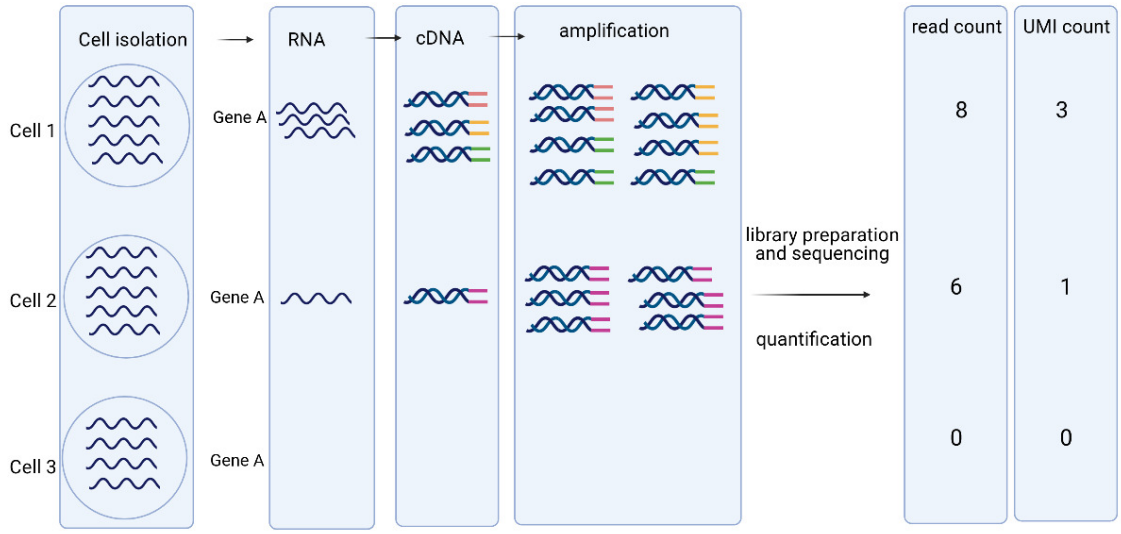


Figure 1.6: **Influence of UMI on quantification.**

and active area of research. It is very important to emphasize that "zero-inflation" is not protocol specific, i.e. whether droplet-based or plated-based, but actually depends on the way we quantify the gene expression values, i.e. measured with UMI counts or read counts. Plate-based methods can also incorporate UMIs, for example CEL-seq2 (Hashimshony et al., 2016) as well as the most recently developed Smart-seq3 (Hagemann-Jensen et al., 2020) technology. The reason why we should emphasize this is that the original article (Svensson, 2020) may mislead further technology development or data analysis, leading people to have an impression that "zero-inflation" is protocol depended (Amezquita et al., 2019) or assume there is no "zero-inflation" in droplet-based protocols (Galfre and Morandin, 2020; Gomes et al., 2019). In fact, the "zero-inflation" in the read counts data using droplet-based protocols can provide us an excellent opportunity to examine the characteristics of amplification bias by taking advantage of the read counts and corresponding UMI counts in the same data set.

1.3 Coronavirus

1.3.1 Coronavirus structure

Coronavirus is a family of enveloped, non-segmented, positive-sense and single-stranded RNA viruses with a genome of approximately 30 kilobases (Fehr and Perlman, 2015). The coronavirus genome consists of a 5' cap and a 3' poly (A) tail, allowing it to be used as an mRNA template for the translation of replicase polyproteins. The first two thirds of the virus genome encodes 16 non-structural proteins (nsps). The remaining third of the virus genome encodes accessory proteins and four major structural proteins including the spike (S) protein, envelope (E) protein, membrane (M) protein, and nucleocapsid (N) protein

(Fehr and Perlman, 2015) (figure 1.7).

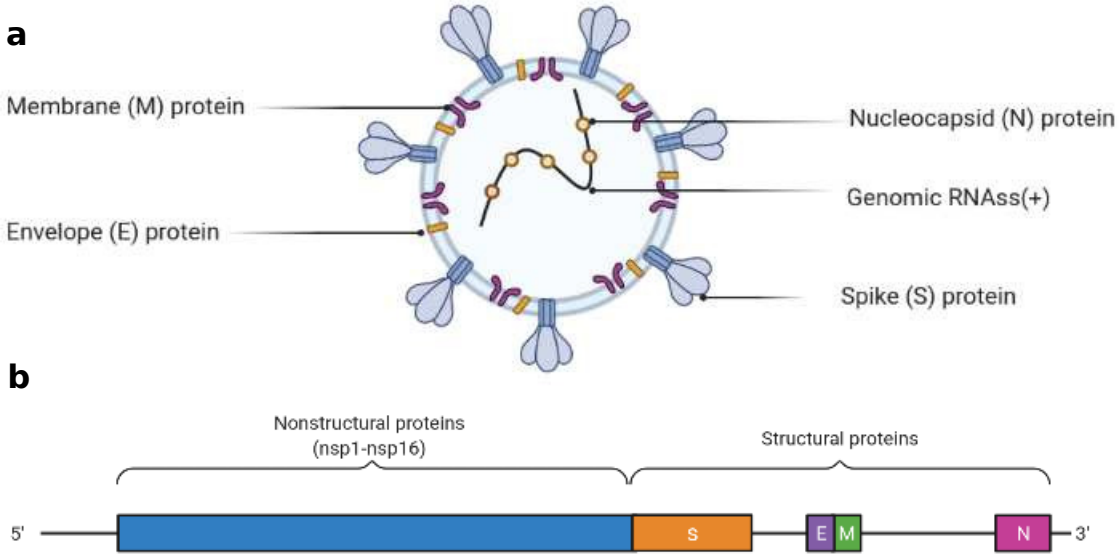


Figure 1.7: **Coronavirus.** (a) Coronavirus structure. (b) Coronavirus genome structure.

1.3.2 Coronavirus epidemics or pandemics

Over the past two decades, there have already been three epidemics or pandemics caused by three closely related coronaviruses – SARS-CoV (severe acute respiratory syndrome coronavirus) (Kuiken et al., 2003), MERS-CoV (Middle East respiratory syndrome coronavirus) (Zaki et al., 2012) and SARS-CoV-2 (Wu et al., 2020). Emerging in late 2019, SARS-CoV-2 still continues to be a major burden on the physical and mental health of the population, cause direct global economic losses, and make negatively influence on the stability of the world (Keni et al., 2020; Saladino et al., 2020). SARS-CoV, MERS-CoV or SARS-CoV-2 infections in severe form emerge as a systemic disease with multiple factors leading to multi-organ dysfunction with similar symptoms, including fever or feeling feverish/chills, headaches, sore throat, cough, runny or stuffy noise, and shortness of breath, etc.

Until now, there are no specific effective anti-SARS-CoV, anti-MERS-CoV or anti-SARS-CoV-2 therapeutics available for human use. SARS-CoV-2 is a new emerging coronavirus, however, its similarity to SARS-CoV (Xu et al., 2020) and MERS-CoV suggests lessons learned from SARS-CoV and MERS-CoV would provide invaluable information to help understand the pathogenesis of COVID-19, therefore guide the development of the potential anti-CoV therapeutics. There are several points of attack for potential anti-SARS-CoV/MERS-CoV/SARS-CoV-2 treatment strategies, two important of which are intervention on cell entry mechanisms or acting on the host immune systems.

1.3.3 Virus entry and innate immune response

Virus entry into host cells is the earliest step of the viral life cycle as the virus delivers the viral genome into the host cell. Virus entry is an important determinant of virus infectivity and pathogenesis (Belouzard et al., 2012; Lou et al., 2014), and also constitutes an important antiviral target (Teissier et al., 2011). SARS-CoV-2 uses similar virus entry mechanism as SARS-CoV (Mahmoud et al., 2020), requiring the S protein of SARS-CoV-2 to bind to ACE2 (angiotensin converting enzyme 2) through their receptor-binding domain (RBD) and using TMPRSS2 as an activating protease (Hoffmann et al., 2020) (figure 1.7).

Innate immune response is an essential component and the first line of the host defense against invading virus. Drug-like molecules that regulate innate immune responses can be introduced as promising antiviral option (Lou et al., 2014). The activation of antiviral innate immune response depends on initiation of specific signaling pathways, including sensing of molecular structures of the invading virus by several pattern recognition receptors (PRRs) such as cytoplasmic retinoic acid-inducible gene I (RIG-I) like receptors (RLRs) and Toll-like receptors (TLRs) (Bowie and Haga, 2005; Loo and Gale Jr, 2011). The RLRs family functions as cytoplasmic sensors of viral RNA and encompasses three members identified to date: RIG-I, MDA5 (melanoma differentiation-associated gene 5), and LGP2 (Laboratory of Genetics and Physiology 2) (Loo and Gale Jr, 2011). The TLRs family comprises ten members (TLR1-TLR10) in humans and plays a major role in sensing of viral infection (Bowie and Haga, 2005; Kawasaki and Kawai, 2014). These pattern recognition events trigger several signaling pathways leading to activation of downstream transcription factors such as interferon regulator factors (IRFs) and nuclear factor κB (NF-κB). The activation of transcription factors will subsequently stimulate the transcription of interferons (IFNs). IFNs bind and induce signaling through their corresponding IFN receptors and subsequently induce transcription of IFN-simulated genes (ISGs) and pro-inflammatory chemokines to establish an antiviral state to control viral replication and dissemination (Chiang and Gack, 2017; Park and Iwasaki, 2020) (figure 1.8).

Neutrophils are also increasingly recognized to play an important role in host innate immune responses to virus infections though multiple mechanisms, for example, by directly recognizing and harboring viruses through multiple receptors on their surface membrane, or though neutrophil extracellular traps; or by interacting with other immune cells, secreting cytokines, or modulating the adaptive immune responses to elicit antiviral responses (Naumenko et al., 2018).

However, the latest studies on virus cell entry routes and cellular innate immune responses to SARS-CoV-2 have produced contradictory claims. Recently, several clinical studies have found that SARS-CoV-2 can infect several organs where ACE2 expression could not be detected in healthy individuals (Hikmet et al., 2020; Ren et al., 2021; Zou et al., 2020), which is in contrast to the classic ACE2-dependent entry routes. Several studies on host immune responses point out that robust IFN responses and markedly

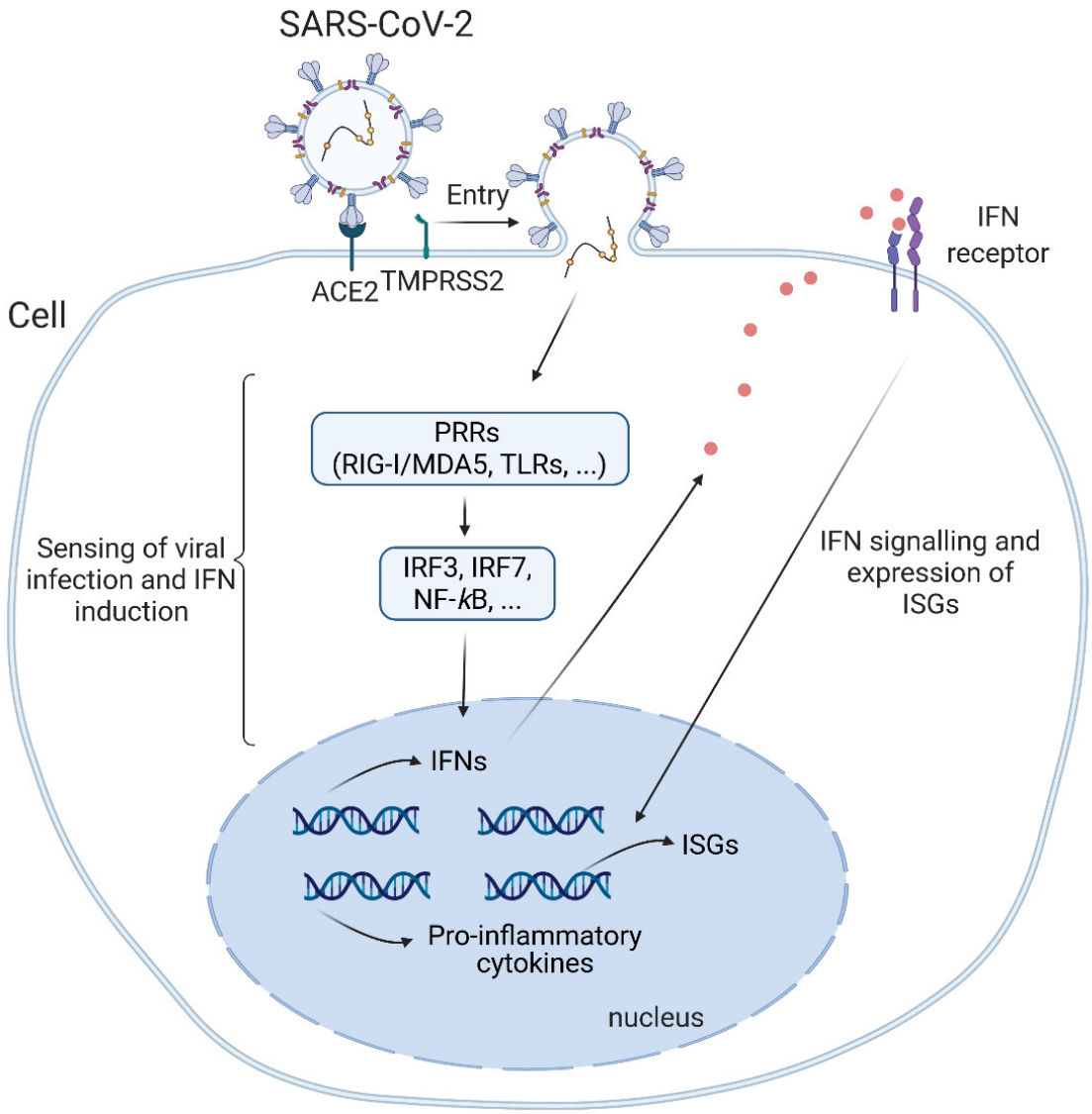


Figure 1.8: **Virus entry and innate immune response.**

elevated expression of ISGs are observed in SARS-CoV-2 infections of different cells and in patient samples (Broggi et al., 2020; Ren et al., 2021; Wei et al., 2020; Zhang et al., 2020; Zhou et al., 2020). In contrast, the study by (Blanco-Melo et al., 2020) states that weak IFN response and moderate ISG expression are characteristic for SARS-CoV-2 infection. It becomes imperative to clarify these matters for further therapy development. In the third part of my thesis, we have performed comprehensive comparative analyses with RNA-seq data sets of different cells infected with SARS-CoV-2, SARS-CoV and MERS-CoV, as well as RNA-seq data from COVID-19 patients (Cao et al., 2021b). We have presented evidence for multiple SARS-CoV-2 entry mechanisms. We have also dissolved apparent conflicts on cellular innate immune responses to SARS-CoV-2 infection. In addition, with the help of analysis of RNA-seq data of lung and bronchoalveolar lavage fluid samples of COVID-19 patients, dynamics of neutrophils and neutrophil extracellular traps have also been examined in the progression of COVID-19 (Wang et al., 2020). Our

results emphasize the complex interactions between host cells and SARS-CoV-2, offering new insights into the pathogenesis of SARS-CoV-2, and can further inform development of antiviral drugs.

Chapter 2

Methods

Computational tools for HT-seq data analysis are written in different computational languages such as Java, Perl, Python, and R, etc. Most of them are executed in Linux system, Mac terminal, or other similar environments. In this thesis, all computational tasks are conducted in Linux systems with a lot of command line tools and R programming environment with numerous packages.

Most of raw sequencing data are very big and the data processing tasks also require many computing resources. Therefore, most of the analyses especially the downloading and preprocessing of the raw sequencing data in this thesis were conducted with the remote servers. SSH (secure shell) is used to connect to the remote server with the IP address. Slurm is used to schedule and manage different tasks.

R is the main language used to do most of the downstream analyses after preprocessing of data on remote server. R is a well-developed programming language. Its strengths lie in statistical (modeling, statistical tests, ...) computing and the ease in producing of well-designed high-quality plots. There are numerous R packages in CRAN (`https://cran.r-project.org/`) and Bioconductor (`https://www.bioconductor.org/`) for processing and analyzing sequencing data with detailed documents and vignettes.

Git was used as version control to keep track of different changes in one project and to collaborate. Some schematic diagrams used in this thesis were created with BioRender (`https://biorender.com/`). The thesis was written in LaTeX(`https://www.latex-project.org/`).

In the following, I will further illustrate the general data analysis methods and algorithms used in this thesis.

2.1 RNA-seq data analysis

RNA-seq is the application of HT-seq technologies to cDNA molecules obtained from reverse transcription from RNA in order to get information about the RNA content of a sample. The basic experimental and analysis procedures include (figure 2.1):

1. mRNA isolation either with polyA+ selection protocol or rRNA depletion protocol

(Zhao et al., 2018);

2. library preparation which included a number of steps (RNA fragmentation, cDNA synthesis, adapter ligation, and amplification);

3. sequencing process including single-end or paired-end sequencing, this step generates millions of reads with associated quality scores as FASTQ files;

4. quantification and differential gene expression analysis.

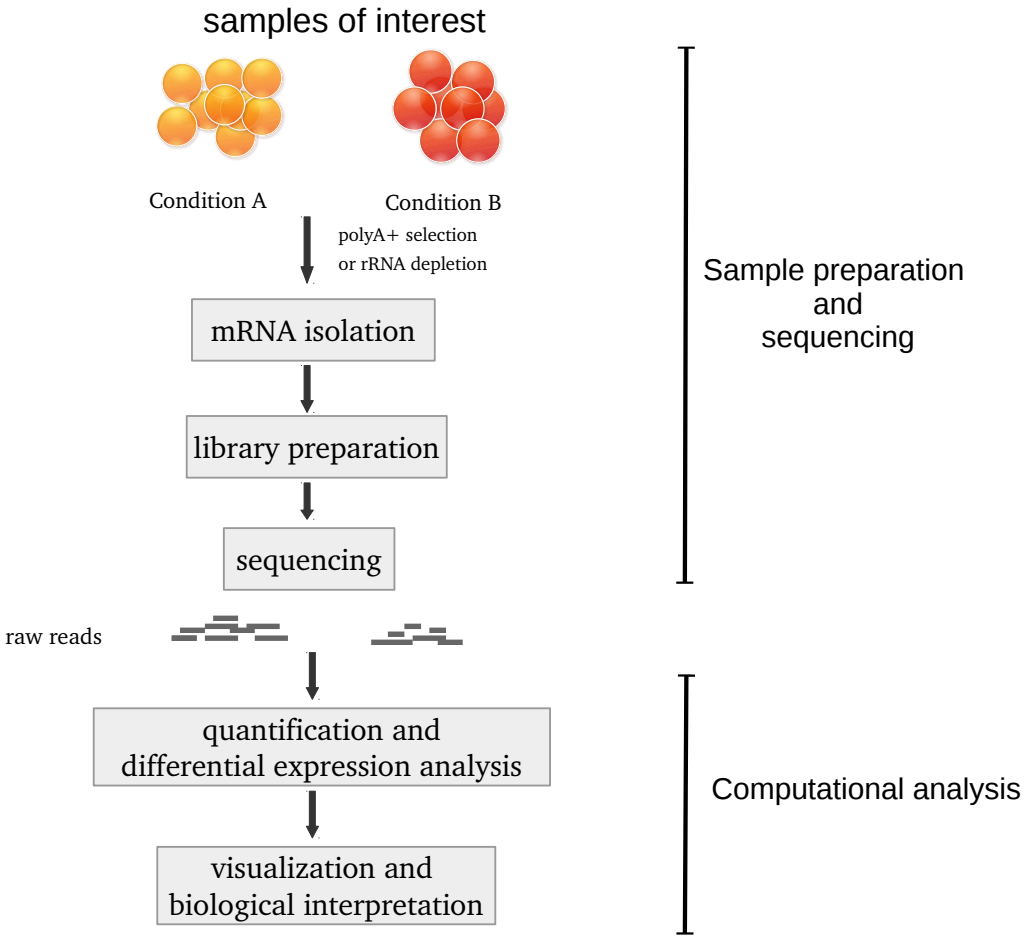


Figure 2.1: **Overview of RNA-seq data analysis.**

2.1.1 Assignment of reads in RNA-seq data

RNA-seq technologies (including bulk and single cell) usually generate millions of reads. A critical step for quantification of RNA-seq data is the efficient and accurate assignment of sequencing reads to transcripts that the reads are originated from to further infer gene expression levels (Bray et al., 2016; Dobin et al., 2013; Kim et al., 2015; Liao et al., 2019; Patro et al., 2017). The tools used for assignment of reads to transcripts can be generally categorized as "alignment-based" or "alignment-free".

The alignment-based tools rely on accurate base-to-base sequence alignments to a reference genome or transcriptome. Afterwards, gene expression levels can be estimated from the alignments at the annotated gene loci (Dobin et al., 2013; Kim et al., 2013, 2015). The benefits of alignment-based methods are that the output alignment files can also be used to further detect single nucleotide variations or conduct RNA-editing analysis (Sahraeian et al., 2017). However, they are usually computationally intensive and time consuming (Kim et al., 2013), regardless of the development of several fast aligners (Dobin et al., 2013; Kim et al., 2015).

The alignment-free methods avoid the base-to-base alignment. They work by assigning reads directly to transcripts using k-mer-based counting algorithms enabling fast assignments of reads to pre-indexed reference transcripts (Bray et al., 2016; Patro et al., 2017). Therefore, alignment-free tools rely on less computational resources and significantly increase the speed of RNA-seq data analysis.

Different assignment strategies can perform differently in quantification of genes, with the tools using similar strategies are reported to perform more similarly (Sahraeian et al., 2017; Srivastava et al., 2020; Wu et al., 2018).

2.1.2 Quantification units for RNA-seq data

The output raw counts of reads assigned to each gene (or transcript) can not be directly compared and should be converted into informative gene expression units. Several different factors may affect the number of read counts assigned to each gene (Oshlack and Wakefield, 2009; Risso et al., 2011; Robinson and Oshlack, 2010). For instance, the observed raw counts are not directly comparable across different samples because of the different library sizes (or sequencing depths, the total number of assigned reads) (Robinson and Oshlack, 2010) – samples with the larger library size will certainly increase the observed counts for genes than the samples with smaller library sizes. Thus, a library size normalization is usually needed to obtain equal library size for all samples through rescaling or resampling the observed counts. After accounting for library size, CPM (counts per million) or RPM (reads per million) value for each gene or transcript can be calculated:

$$\text{CPM/RPM} = \frac{\text{Number of reads assigned to a gene}}{\text{Total number of assigned reads}} \times 10^6$$

Another important factor that needs to be taken into consideration is the gene length (or transcript length) (Oshlack and Wakefield, 2009). In comparison with the shorter genes or transcripts with the same expression levels, the longer ones usually have more reads assigned to them. After accounting for library size and gene length (or transcript length), TPM (transcripts per million) or RPKM/FPKM (reads/fragments per kilobase per million reads mapped) value for each gene or transcript can be computed (Zhao et al., 2020):

$$\text{TPM} = \frac{\text{X}}{\sum(\text{X})} \times 10^6$$

$$\text{Where X} = \frac{\text{Number of reads assigned to a gene} \times 10^3}{\text{gene length in base pair}}$$

$$\text{RPKM/FPKM} = \frac{\text{Number of reads or fragments assigned to a gene} \times 10^3 \times 10^6}{\text{Total number of assigned reads} \times \text{gene length in base pair}}$$

By definition, TPM and RPKM/FPKM are both relative measurement units of gene or transcript expression levels, i.e. they only represent the proportion of reads assigned to a gene or transcript among a pool of sequenced genes or transcripts in a sample or a cell, and should be interpreted with caution (Zhao et al., 2020). FPKM and RPKM are analogous, and FPKM is especially used in paired-end sequencing experiments. TPM is a slight modification of RPKM and they are closely related (Zhao et al., 2020).

2.1.3 Normalization of count data

Although TPM and RPKM/FPKM are meaningful quantification units, they are not the input for further differential expression (DE) analysis. The input for different DE analysis tools is usually a count matrix with genes along the rows and samples along the columns. The raw counts in the matrix needs normalization to make sure they are adjusted to account for the factors that prevent direct comparisons, allowing accurate estimation and detection of DE genes.

The factors that affects the number of read counts assigned to a gene can be can be divided into two categories: "within-sample" factors and "between-sample" factors. The "within-sample" factors refer to factors that only influence the comparisons for different genes within the same sample, like gene length or GC-content (Evans et al., 2018; Oshlack and Wakefield, 2009; Risso et al., 2011). The "between-sample" factors refer to factors that affect the comparison of same gene in different samples (Evans et al., 2018). The "within-sample" factors are usually not taken into consideration for normalization for DE analysis, since it only concerns the expression difference for the same gene between different conditions. So, the normalization before DE analysis mostly refers to "between-sample" normalization (Evans et al., 2018).

Different tools implement different normalization strategies to normalize these counts, including TMM (Trimmed Mean of M-values) used in edgeR (Robinson and Oshlack, 2010), and median-of-ratios method used in DESeq and DESeq2 (Anders and Huber, 2010; Love et al., 2014).

2.1.4 Modeling of count data

Poisson distribution

Count data are discrete, non-negative integer values. Count data generated from HT-seq technology can be modeled with Poisson distribution $X \sim \text{Poisson}(\lambda)$:

$$\text{Poisson}(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0, 1, 2, \ldots .$$

The parameter λ is equal to the expected mean of X and also equal to its variance σ^2. The figure 2.2 shows the Poisson probability mass function for different values of λ. We can see that the width of the distribution increases with the an increase of λ, indicating the uncertainty in measurement increases with an increase of the value of x.

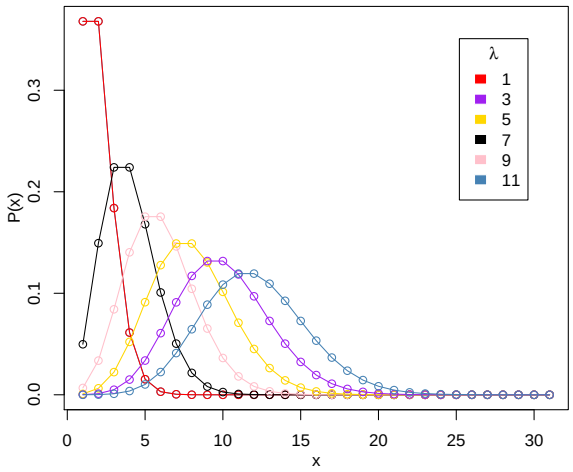


Figure 2.2: **Poisson distribution.**

Although Poisson distribution is often used for statistical modeling for count data, it only fit if the frequency of counts for a given gene is uniform across different samples or different cells, and variation from sample-to-sample or cell-to-cell is due to independent statistical sampling. These assumptions, however, may not fit for high-throughput count data in bulk or single cell RNA-seq data, which usually have additional unknown biological variation and other sources of biological or technical stochasticity, i.e. overdispersion (Anders and Huber, 2010).

Negative binomial distribution

Due to overdispersion discussed above, negative binomial distribution is usually widely used for bulk RNA-seq count or single cell RNA-seq UMI count data (Anders and Huber, 2010; Chen et al., 2018; Love et al., 2014).

Let's say $X \sim \text{NegBinomial2}(\lambda, \phi)$:

$$\text{NegBinomial2}(x|\lambda,\phi) = \binom{x+\phi-1}{x}\left(\frac{\lambda}{\lambda+\phi}\right)^{x}\left(\frac{\phi}{\lambda+\phi}\right)^{\phi} \qquad x = 0, 1, 2, \ldots .$$

This parameterization directly uses a mean parameter λ and dispersion parameter ϕ that controls overdispersion relative to the square of the mean:

$$\text{Var} = \lambda + \frac{\lambda^2}{\phi}\,.$$

The term $\frac{1}{\phi}$ represents the degree of overdispersion, and $\frac{\lambda^2}{\phi}$ represents the additional variance to the Poisson distribution. The negative binomial distribution goes closer to a Poisson distribution with the increasing of ϕ. (figure 2.3)

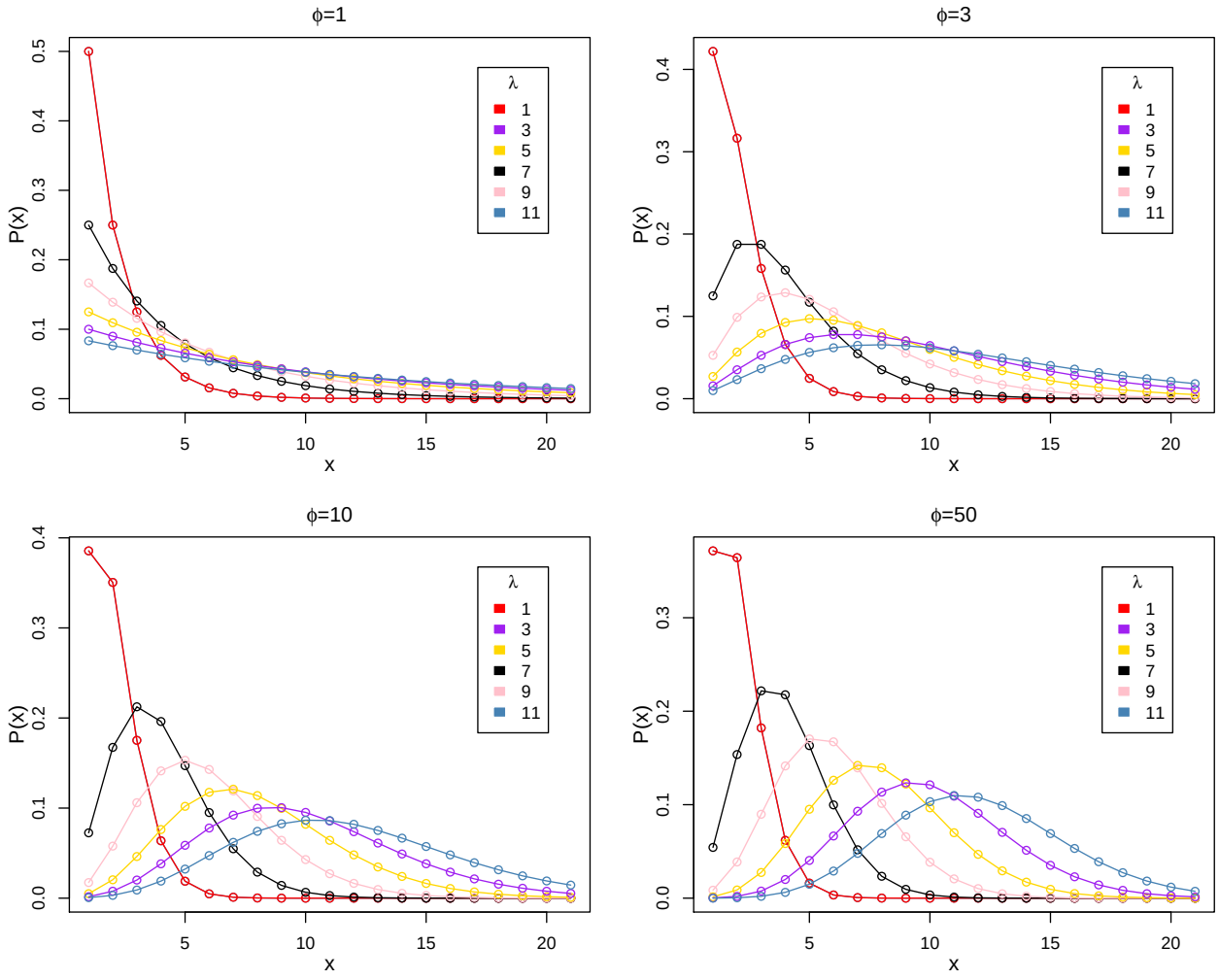


Figure 2.3: **Negative binomial distribuion.**

2.1.5 Test for differential expression

After normalization, DE analysis tools will perform statistical test analysis to discover quantitative changes in expression levels between different biological groups. For example, a statistical test is conducted to decide whether, for a given gene, an observed difference in read counts is greater or lower than what would be expected just due to natural random variation. The likelihood ration test (LRT) and Wald test are two widely used statistical tests for DE analysis by different tools (Love et al., 2014; Pimentel et al., 2017).

For LRT, two models are estimated for each gene, goodness of fit is then compared based on the ratio of their likelihoods:

$$\mathrm{LR} = -2\log\left(\frac{L(m_1)}{L(m_2)}\right) = 2(\mathrm{loglik}(m_2) - \mathrm{loglik}(m_1))\,.$$

m_1 is the reduced model, i.e. gene expression level unaffected by the treatment. m_2 is the full model, i.e. gene expression level affected by the treatment. $L(m_*)$ denotes the likelihood of the respective model. Because the full model is more complex, it will improve the goodness of fit. The question is whether the simpler reduced model can explain the data, and how much better the full model is than the reduced model. LRT compares the log likelihoods of the two models to test whether the observed difference in model fit is statistically significant.

The Wald test approximates the LRT, but it only estimates one model per gene and evaluates the null hypothesis which is $logFC = 0$. If the test fails to reject the null hypothesis, it suggests that removing the variables from the model will not substantially harm the fit of the model.

Wald test allows for comparison between two groups. LRT allows for comparisons of two groups or more groups, as well as time course analyses. When choosing an appropriate statistical test method, it is important to consider the experimental design of the data.

Keep in mind that these statistically tests have the multiple testing problem that are needed to be corrected. Some multiple test correction approaches include Bonferroni (Dunn, 1961) and FDR/Benjamini-Hochberg (Benjamini and Hochberg, 1995; Chen et al., 2010).

2.2 ChIP-seq data analysis

2.2.1 Alignment and Visualization

The alignment of the sequenced reads from ChIP-seq can be conducted with tools lile BWA (Li and Durbin, 2009), Bowtie (Langmead et al., 2009) and Bowtie2 (Langmead and Salzberg, 2012), etc. The common output file formats are Sequence Alignment Map (SAM) or its binary version BAM. Duplicated reads should be then marked and removed based on their alignment locations. A suite of tools developed for analyzing and visualizing HT-seq data named deepTools (`https://deeptools.readthedocs.io/en/develop/`)

can be used to visualize the distribution of the reads and check the quality of the data. For example, "plotHeatmap" can be used to visualize the distribution of ChIP-seq signals; "plotFingerprint" can be used to determine how well the ChIP-seq signal is separated from the background noise in the control sample; "plotCorrelation" can be used to evaluate the similary between different samples to determine whether different sample types are well separated, etc.

2.2.2 Peak calling

The peak-calling step can be then conducted to identify the stacks of aligned reads. Some of the stacks represent the signal of interest, i.e. binding of a transcription or histone modification. A lot of the stacks of reads are experimental or molecular noise. Peak calling comprises two main tasks: one is to identify candidate peaks, the other is to evaluate the significance of enriched peaks. There are a lot of tools available for peak calling (Thomas et al., 2017), including MACS2 (Zhang et al., 2008), MUSIC (Harmanci et al., 2014), BCP (Xing et al., 2012), etc. (figure 2.4)

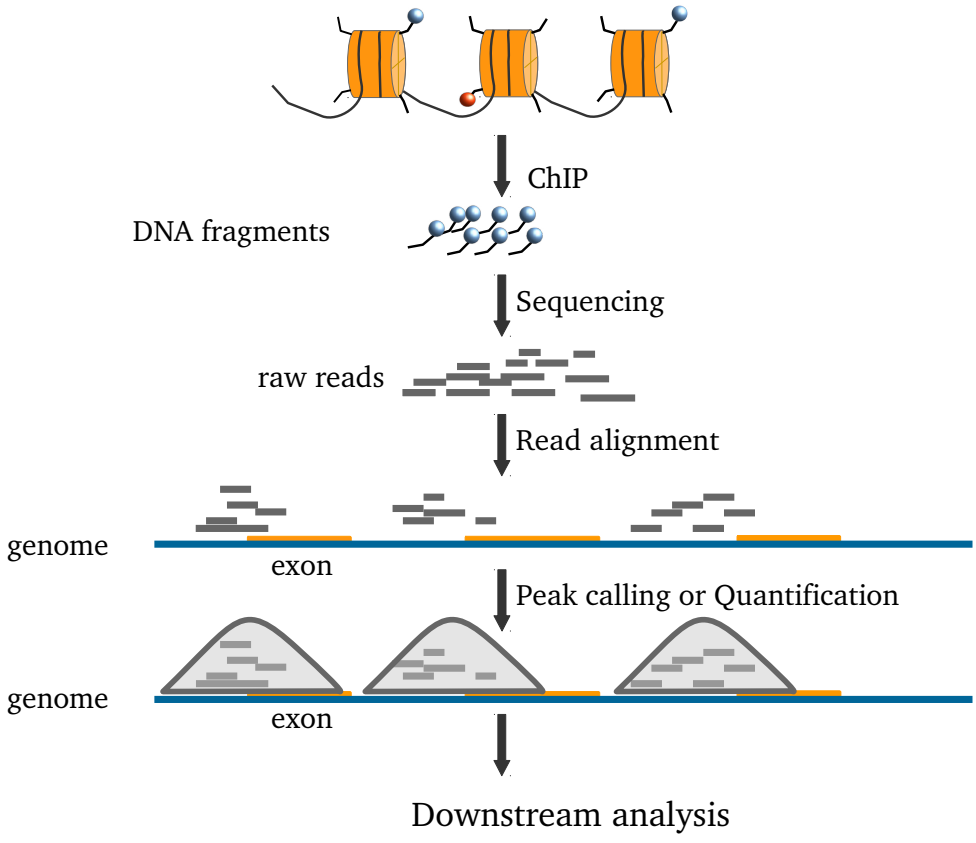


Figure 2.4: **Overview of ChIP-seq data analysis.**

2.2.3 Quantitative comparison

Besides detection of peaks, quantitative comparison between different groups is also a very important question. However, quantitative comparison is usually very complicated, and confounded by intrinsic noisiness and the variability caused by the complex steps in sample preparation (Steinhauser et al., 2016). Binary comparisons that identify common or unique peaks is a possible option for quantitative comparison, although some false positive and false negative peaks are usually obtained.

When the signal/noise ratios in different samples are similar, or proper normalization is conducted, the quantitative comparison can be conducted with the statistical methods similar with differential gene expression analysis, for example, the peak-based method DiffBind (Stark et al., 2011) or window-based method csaw (Lun and Smyth, 2016). DiffBind (Stark et al., 2011) works with consensus preaksets with start and end positions, which are derived from other peak callers. The reads in each interval are counted to get a count matrix. The differential analysis is then conducted by default using DESeq2. The csaw (Lun and Smyth, 2016) package uses a sliding window approach to count reads across the genome. Each window is then tested for significant differences using the methods in edgeR.

2.3 Multi-objective optimization

Optimization involves minimizing or maximizing one or more functions, also called the objective functions. Optimization problems that have two or more objective functions to be optimized at the same time are called multi-objective optimization problems. In a multi-objective problem, the objectives will often be opposite with each other, i.e. when maximizing two competing objectives, an increase in one objective would cause a decrease in the other which is known as "tradeoff". Instead of a single solution, the answer to multi-objective optimization problems is a set of solutions that define the best tradeoff between competing objectives.

2.3.1 Dominance

In a single-objective optimization problem, the superiority of a solution over other solutions can be easily determined by direct comparison of their objective function values. In a multi-objective optimization problem, the goodness of a solution is determined by the dominance. When solution S_1 is strictly better than S_2 in at least one objective, and solution S_1 is no worse than S_2 in all objectives, then solution S_1 dominates S_2.

2.3.2 Pareto optimization

Pareto optimization algorithm is one of the solutions to multi-objective optimization problems, and can be understood as finding a set of Pareto optimal solutions that define

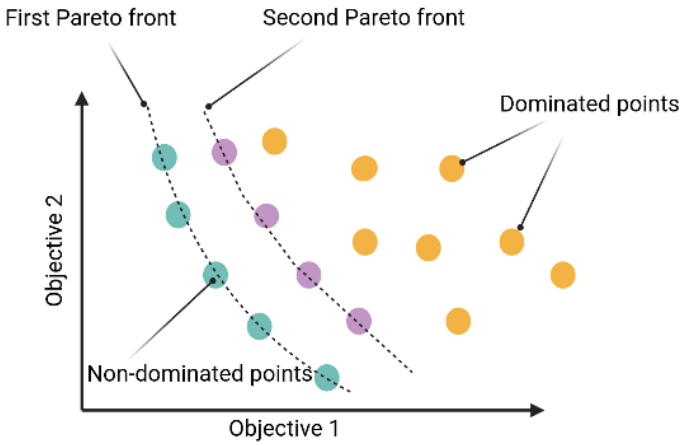


Figure 2.5: **Pareto optimization solutions.**

the best trade-offs between different competing objectives. Likewise, the superiority of a solution over the other can be determined by the Pareto dominance, i.e. a vector for a feasible solution is said to Pareto dominate another when it can not be improved in any of the objectives without degrading at least one of the other objectives. The Pareto dominance allows for comparison of two objective vectors in a precise sense. Pareto front is defined as the boundary mapped by the set of points representing the set of mutually non-dominated solutions from Pareto optimization. The first Pareto front is determined by the points that are not dominated by others. In the same way, the second Pareto front is determined after removal of the first Pareto front, so on and so forth. In the end, all the points can be ranked by different Pareto front levels. (figure 2.5).

Chapter 3

Contributed Articles

3.1 intePareto: an R package for integrative analyses of RNA-seq and ChIP-seq data

This section is based on the following publication:

Yingying Cao, Simo Kitanovski, and Daniel Hoffmann. **intePareto: an R package for integrative analyses of RNA-seq and ChIP-seq data.** BMC Genomics 21, 802 (2020).
https://doi.org/10.1186/s12864-020-07205-6

Cao *et al. BMC Genomics* 2020, **21**(Suppl 11):802
https://doi.org/10.1186/s12864-020-07205-6

BMC Genomics

SOFTWARE Open Access

intePareto: an R package for integrative analyses of RNA-Seq and ChIP-Seq data

Yingying Cao*, Simo Kitanovski and Daniel Hoffmann

From The International Conference on Intelligent Biology and Medicine (ICIBM) 2020
Virtual. 9–10 August 2020

Abstract

Background: RNA-Seq, the high-throughput sequencing (HT-Seq) of mRNAs, has become an essential tool for characterizing gene expression differences between different cell types and conditions. Gene expression is regulated by several mechanisms, including epigenetically by post-translational histone modifications which can be assessed by ChIP-Seq (Chromatin Immuno-Precipitation Sequencing). As more and more biological samples are analyzed by the combination of ChIP-Seq and RNA-Seq, the integrated analysis of the corresponding data sets becomes, theoretically, a unique option to study gene regulation. However, technically such analyses are still in their infancy.

Results: Here we introduce *intePareto*, a computational tool for the integrative analysis of RNA-Seq and ChIP-Seq data. With *intePareto* we match RNA-Seq and ChIP-Seq data at the level of genes, perform differential expression analysis between biological conditions, and prioritize genes with consistent changes in RNA-Seq and ChIP-Seq data using Pareto optimization.

Conclusion: *intePareto* facilitates comprehensive understanding of high dimensional transcriptomic and epigenomic data. Its superiority to a naive differential gene expression analysis with RNA-Seq and available integrative approach is demonstrated by analyzing a public dataset.

Keywords: RNA-Seq, ChIP-Seq, Integrative analysis

Background

With increasing accessibility and application of high-throughput sequencing (HT-Seq), it has become possible, in principle, to combine and integrate complex transcriptomic (RNA-Seq, [1]) and epigenomic data as a multi-omics approach to understand mechanisms of gene regulation [2]. One of the most important epigenetic regulators of gene expression are histone modifications [3]. Several types of histone modifications can change the state of the chromatin in different ways and increase or decrease gene expression.

There are many interesting applications of integrative analysis of RNA-Seq and ChIP-Seq data. For instance, the consistent co-occurrence of histone modification patterns and up- or down-regulated gene expression can improve our understanding of the "histone code" [4]; or, the comparison of histone modification states with quantitative gene expression can lead to the discovery of new enhancer regions [5]; or, expression and simultaneous occurrence of different modifications at a gene can reveal gene regulation dynamics along a developmental trajectory [6]. Separate analyses of RNA-Seq or ChIP-Seq data alone can not fully explain the complex mechanisms underlying the regulation of gene expression. Efforts to quantitatively integrate available RNA-Seq and ChIP-Seq data of histone modifications in various conditions are crucial for

*Correspondence: yingying.cao@uni-due.de
Bioinformatics and Computational Biophysics, Faculty of Biology and Center for Medical Biotechnology (ZMB), University of Duisburg-Essen, Universitätsstr.2, 45141 Essen, Germany

improving our understanding of the role of epigenetics in gene regulation.

Several computational methods have been proposed to use histone modifications for predicting gene expression [7, 8]. However, these methods generally focus on the prediction of gene expression with ChIP-Seq data of different histone modifications in one cell type or state. An important task for quantitative integration of RNA-Seq and ChIP-Seq data is the identification of genes of important biological function that are differentially expressed and therefore define cell types or states. Integration could answer questions like these: For which genes do we see *consistent* changes in expression and in histone modifications as we compare different cell types or conditions? Which genes show increased expression in combination with acquisition of activating histone modifications, or decreased expression in combination with more suppressive histone modifications?

Such genes with consistent transcriptomic and epigenomic changes are more likely to point to essential functional differences and to play an important role in cell differentiation or the development of disease.

Although identification of such genes is obviously highly attractive, and matched data sets of RNA-Seq and ChIP-Seq are increasingly available, promising technical implementations are still rare and not readily available [9]. One reason may be the sheer complexity of the data, consider e.g. that there are numerous histone marks with similar but probably not identical function, such as activating marks H3K4me3, H3K4me1, H3K36me3, H3K27ac, or repressive marks H3K9me3 and H3K27me3.

There are a few methods developed to detect genes with congruent changes in RNA-Seq and ChIP-Seq between two experimental conditions. For example, Klein et al., 2014 [10] and Schäfer et al., 2017 [11] developed approaches based on Bayesian inference of mixture models [10] and hierarchical models and clustering [11]. These early methods are a great step forward towards integrative analysis, but they still suffer from limitations, e.g. with respect to the number of genomic variables that may be analyzed, or because of the danger of losing important information in the aggregating of data. Further more, their integration [11] is based on transcript level, from a biological perspective, data integration on gene level is easier to interpret than at the transcript level.

Here we present a quantitative method for the integrative analysis of RNA-Seq and ChIP-Seq data for several different histone modifications. We frame integrative analysis as multi-objective optimization problem that we solve by Pareto optimization [12]. Multi-objective optimization has significant advantages compared to single-objective optimization, e.g., in classification, system optimization, and inverse problems [13]. With our new R package *intePareto* we provide a first solution of Pareto optimization to the integration of RNA- and ChIP-Seq data sets. Specifically, *intePareto* is a flexible and user-friendly tool (1) to match these data sets on gene level, (2) to integrate them in a quantitative fashion, (3) to examine abundance correlations of histone modifications and gene expression, and (4) to prioritize genes based on the consistence of changes between conditions in both RNA-Seq and ChIP-Seq using Pareto optimization. The result of the last step is an informative rank-ordered gene list.

We demonstrate that integration of RNA-Seq data and ChIP-Seq data by Pareto Optimization outperforms a clustering method based on Bayesian inference of a hierarchical model [11], and the analysis of RNA-Seq alone.

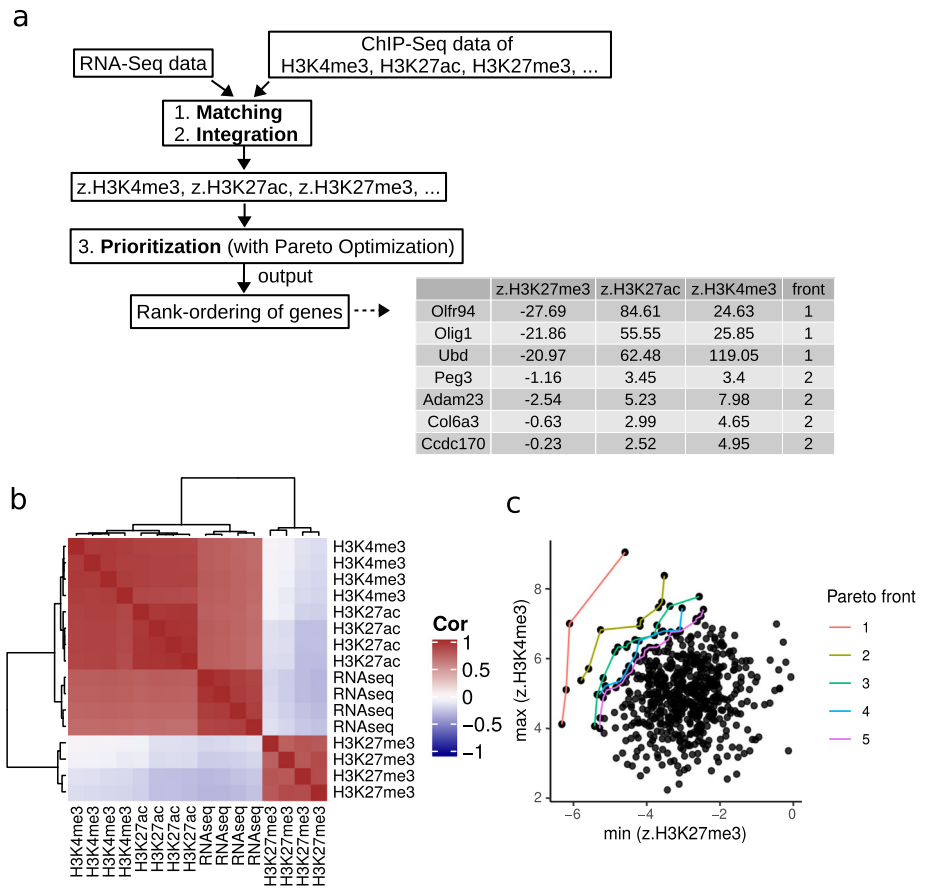
Implementation

intePareto is implemented as an R package that provides an easy-to-use workflow to quantitatively integrate RNA-Seq and ChIP-Seq data of one or more different histone modifications. A typical application, as presented here with 4 RNA-Seq samples and 28 ChIP-Seq samples (case study in Additional file 1), runs in less than one hour on a standard personal computer. In this section, we describe the implementation of *intePareto* in detail. The pipeline takes as first input RNA-Seq data, preprocessed by RNA-Seq quantification software, for instance estimated read counts from Kallisto [14], or other suitable quantities [15–17]. Kallisto performs well in terms of speed and quantification, so we use as input file format the output format of Kallisto. Other quantification inputs [15–17] are also accepted if structured in the same input file format. Second, the pipeline takes ChIP-Seq reads, aligned to the reference genome with tools like BWA [18], and then processed further with Samtools [19]. The workflow then comprises three main steps, 1. "Matching", 2. "Integration", 3. "Prioritization" sections (Fig. 1a).

Matching

Our first problem is to link histone modification data with the corresponding gene expression data. Hence, the first step is to match quantitative histone modification data from ChIP-Seq to the biologically corresponding gene expression data as measured by RNA-Seq, or in other words: to find the target genes for histone modifications.

This matching of RNA-Seq and ChIP-Seq data is complicated by the fact that one gene usually has multiple transcripts, and multiple transcript starting sites (TSSs), which means that there are multiple promoters that can drive gene expression [20]. Another more challenging task is that the link between enhancers and genes is much more difficult to determine. Contrary to promoters that reside approximately 3 kilobases (kb) upstream from the transcription start site (TSS) of a gene, enhancers are often found dozens of kb away from the genes they influence.



	z.H3K27me3	z.H3K27ac	z.H3K4me3	front
Olfr94	-27.69	84.61	24.63	1
Olig1	-21.86	55.55	25.85	1
Ubd	-20.97	62.48	119.05	1
Peg3	-1.16	3.45	3.4	2
Adam23	-2.54	5.23	7.98	2
Col6a3	-0.63	2.99	4.65	2
Ccdc170	-0.23	2.52	4.95	2

Fig. 1 intePareto workflow. **a** The general pipeline of intePareto. **b** Heatmap of correlation matrix of RNA-Seq and ChIP-Seq for different histone marks after genewise matching. The color represents Spearman's rank correlation across all samples. **c** Top Pareto fronts

Moreover, enhancers are tissue- and cell type-specific and highly variable [21–23].

Several methods for predicting target genes for histone modifications have been published [24–26]. However, the lack of agreement between them discouraged us to include them in our pipeline [27].

For ChIP-Seq data of histone modification marks that are enriched in promoter regions, like H3K4me3 and H3K27me3, *intePareto* offers two matching strategies: (1) *highest* – choose the promoter with maximum ChIP-Seq abundance value among all the promoters as a representative of the ChIP-Seq signal for this gene; (2) *weighted.mean* – calculate the abundance weighted mean of all the promoters to represent the ChIP-Seq signal for this gene. In this study the promoter region was defined as 5 kb stretch with the TSS at the center; we found that this value safely included all relevant ChIP-Seq signals. This definition can be adapted if necessary.

More matching strategies will be offered in future versions with increasing availability of validated annotated enhancers and of studies that examine the relationship between the density of ChIP-Seq and expression level of RNA-Seq. After the genewise match of RNA-Seq and ChIP-Seq data, the correlation of RNA-Seq and ChIP-Seq can be examined for each histone mark (Fig. 1b)

Integration

After the genewise matching of RNA-Seq and ChIP-Seq, these two data types are integrated by calculation of *log* fold changes (FC) between conditions, as implemented in DESeq2 [28]. For that purpose we propose to use DESeq2 because it works well for both RNA-Seq and ChIP-Seq data [29]. Another benefit is that *apeglm* algorithm [30] is used to shrink the logFC values to zero when the counts are low, dispersion is high, or the number of biological replicates is small. To normalize the data for sequencing depth and RNA composition, the median of ratios method is implemented [28]. *intePareto* determines the Z scores for each gene g and each histone modification type h, defined as:

$$Z_{g,h} = \frac{logFC_g^{(\mathrm{RNA})}}{\mathrm{sd}\left(logFC_g^{(\mathrm{RNA})}\right)} \cdot \frac{logFC_{g,h}^{(\mathrm{ChIP})}}{\mathrm{sd}\left(logFC_{g,h}^{(\mathrm{ChIP})}\right)}.$$

A combination of gene and histone mark has a high, positive Z score if between the compared conditions or cell populations gene expression and histone modification change strongly and in the same direction, i.e. both up or both down.

Prioritization

intePareto takes the Z scores for different, user-selected histone modifications as input, so that for each gene we have several Z scores.

To this end, we can collect all Z scores in an objective function, namely the vector of the n Z scores (one for each histone modification), i.e. $(\alpha_1 Z_1, \alpha_2 Z_2, \ldots, \alpha_n Z_n)$, where $\alpha_i \in \{-1, 1\}$, depending on whether the histone mark is repressive or activating.

We can then interpret the identification of genes that show strong and consistent changes across histone marks as a multi-objective optimization problem, and we solve this problem by a Pareto optimization algorithm [12, 31].

The result is a ranking of genes in Pareto fronts. Using marks H3K27me3 and H3K4me3 as an example, genes in the first Pareto front could minimize Z scores for the repressive mark H3K27me3, and simultaneously maximize the Z scores for the activating mark H3K4me3. This simultaneous optimization is understood in the sense that genes in the first Pareto front are not *dominated* by other genes, i.e. no genes outside the first Pareto front have a lower H3K27me3 Z score and simultaneously a higher H3K4me3 Z score. The second Pareto front is determined in the same way after removal of the first Pareto front, etc. Fig. 1c shows an example of the resulting rank ordering. The Additional file 1 gives more details and an example application of *intePareto*.

Results

Evaluation of *intePareto* using publicly available data

RNA-Seq and ChIP-Seq data

We evaluate *intePareto* based on publicly available RNA-Seq and ChIP-Seq data from a study of Tet methylcytosine dioxygenase 2 (Tet2) knockout mouse embryonic stem cells (mESCs) that are compared to wild type mESCs [32]. With Tet2 assumed to be involved in the regulation of DNA methylation at enhancers, we expected to find congruent changes between the epigenomes and transcriptomes of Tet2 knockout and wildtype mESCs. For each cell type, the data consists of biologically replicated RNA-Seq data and ChIP-Seq data for histone marks H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3 and H3K36me3 (see Additional file 1 for details).

Data processing

The raw RNA-Seq data in FASTQ format was aligned and quantified with Kallisto (version 0.43.1) [14] against a reference transcriptome downloaded from the ENSEMBL database [33]. The outputs of this step are estimated counts of reads and TPM values for each gene of a given cell condition. The raw ChIP-Seq data in FASTQ format was aligned with BWA (0.7.17) [18] also against a reference genome from ENSEMBL. The resulting files were sorted and the corresponding index files were built with Samtools (version 0.1.19) [19].

Analysis with intePareto

We know that the histone marks H3K4me3, H3K27me3, and H3K9me3 are enriched at gene promoter regions [34, 35]. Other marks such as H3K4me1 and H3K27ac are often associated with gene enhancers as well as active promoter regions, while H3K36me3 is associated with the gene body [34–36]. To define the epigenetic signal for marks that are prevalent at gene promoters we counted the number of ChIP-Seq reads falling into the promoter region of specific genes. For H3K36me3 we counted the total number of reads that fall into the genomic body.

Matching of RNA-Seq and ChIP-Seq data was performed with *highest* strategy as described in "Implementation" section (also see Additional file 1). We demonstrate that our matching strategy captures meaningful epigenetic and transcriptomic signals, by showing that the gene expression is positively correlated with the signal of active marks, and negatively correlated with the signal of repressive marks (Fig. 2) [37, 38]. The matched data was integrated (*doIntegration* function), followed by a prioritization (*doPareto* function) based on Pareto optimization. The optimization task was devised such that it prioritizes genes having high positive Z-scores for active histone marks (H3K4me1, H3K4me3, H3K27ac, H3K36me3) and low negative Z-scores for repressive histone marks (H3K9me3, H3K27me3). The resulting list of genes were sorted according to ascending fronts (Additional file 2).

Downstream analysis of the output of *intePareto*

Gene Ontology (GO) enrichment analysis [39] of the top genes resulting from Pareto optimization by *intePareto* shows (Fig. 3a) that all enriched GO terms are known functional characteristics of Tet2 according to the data source [32] and other research. Specifically, Tet2 can influence the cell differentiation and proliferation of ESCs through altering of the methylation status of DNA, especially in neurogenic differentiation [32, 40], and the development of the heart [41, 42] and other organs [43]. Figure 3b is the heatmap of the 14 genes in the first Pareto front. There are distinct patterns between the up-regulated and down-regulated genes. The clustering dendrogram at the top of the heatmap hints at the functional similarity of H3K27me3 and H3K9me3, and the functional similarity of H3K4me1, H3K4me3, H3K27ac, and H3K36me3. This is in line with previous reports about the function of these histone marks [37, 38]. It is worth noting

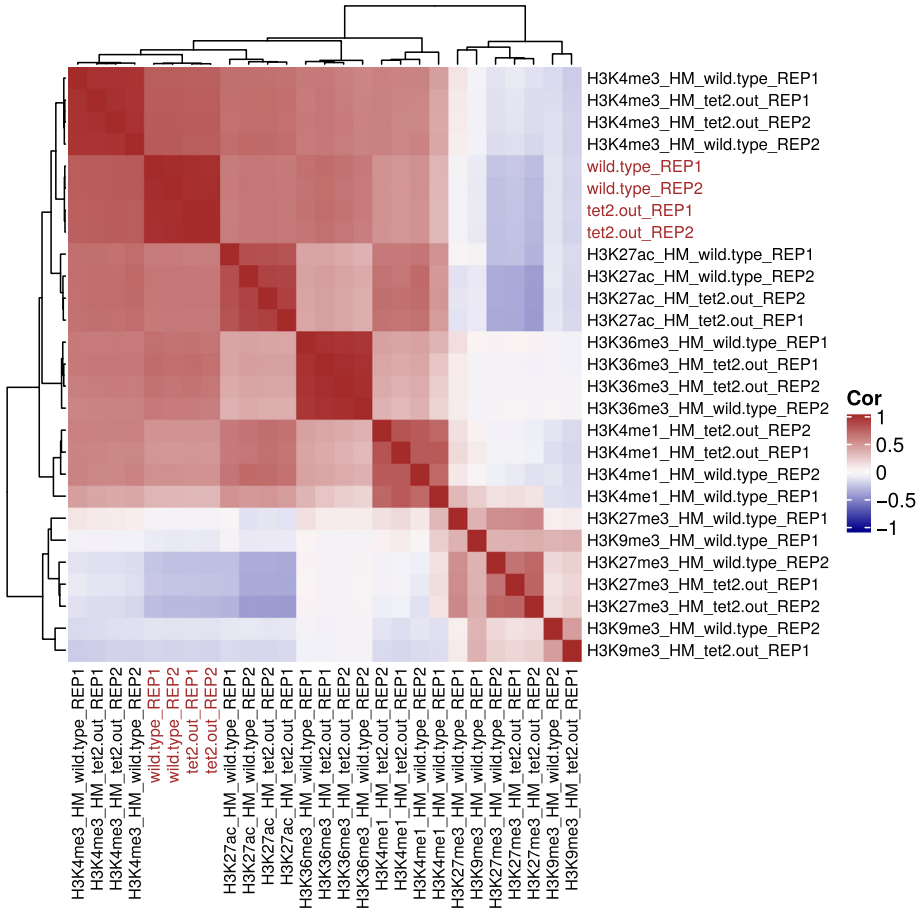


Fig. 2 Heatmap of correlation matrix. Heatmap shows correlation matrix of all RNA-Seq and ChIP-Seq samples. Sample names marked in red are RNA-Seq samples. Sample names marked in black are ChIP-Seq samples

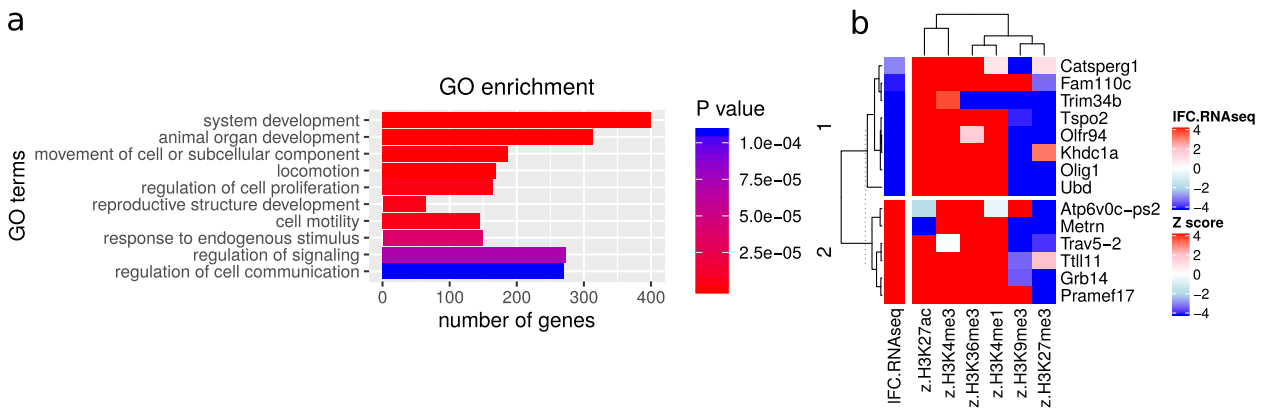


Fig. 3 Top genes detected by *intePareto*. **a** Top 10 GO terms that are enriched for genes detected by *intePareto* in the first 5 Pareto fronts. **b** Genes in the first Pareto front. The logFoldChange of RNA-Seq (IFC.RNA-Seq) is calculated by DESeq2 in tet2 knockout condition over wild type. The Z score is calculated by *intePareto*

that the gene Eif2s3y, which was recently confirmed as strongly down-regulated [44] in Tet2 knockdown mESC, was not significantly down-regulated in RNA-Seq data alone. However, it popped up in the top two Pareto fronts of our integrative analysis. This also highlights the benefits of integrative analysis of both data types, which can reduce false negatives or false positives from analyses based on a single sample or data type.

Comparison with existing approach

To evaluate the performance of *intePareto*, we compared our results to those of an integrative analysis with a recently published hierarchical Bayesian model-based clustering approach ("model-based approach") [11], and to the analysis of RNA-Seq alone (Additional file 3). As quality metric for the comparisons, we used the enrichment score of interesting GO terms. For a set of genes (G; e.g. high-priority genes assigned to Pareto front 1), we define the enrichment score for GO term i as the fraction $f_i = |G \cap GO_i|/|G|$, with GO_i the set of all genes annotated with GO term i.

The GO terms of interest were those confirmed in previous research such as "neurogenesis" [32, 40], "cardiac chamber development" [41, 42], "mammary gland formation" [43, 45], and "limb morphogenesis" [46]. Both our integrative approach and the model-based approach found that the genes in the top-ranked genes were enriched in "neurogenesis" (Fig. 4a) and "limb morphogenesis" (Fig. 4d). Analysis based on RNA-Seq alone did not find this enrichment. *intePareto* also found that the top-ranked genes are more enriched in "cardiac chamber development" (Fig. 4b) and "mammary gland formation" (Fig. 4c) as they should be. These functions were not identified by RNA-Seq analysis alone or the model-based approach. An alternative to GO enrichment, that yields complementary information, is pathway enrichment.

Discussion and conclusions

Integrative methods such as those implemented in *intePareto* can collect more evidence from the increasing amount of HT-Seq data of different modalities, such as RNA-Seq and ChIP-Seq data. This will hopefully allow

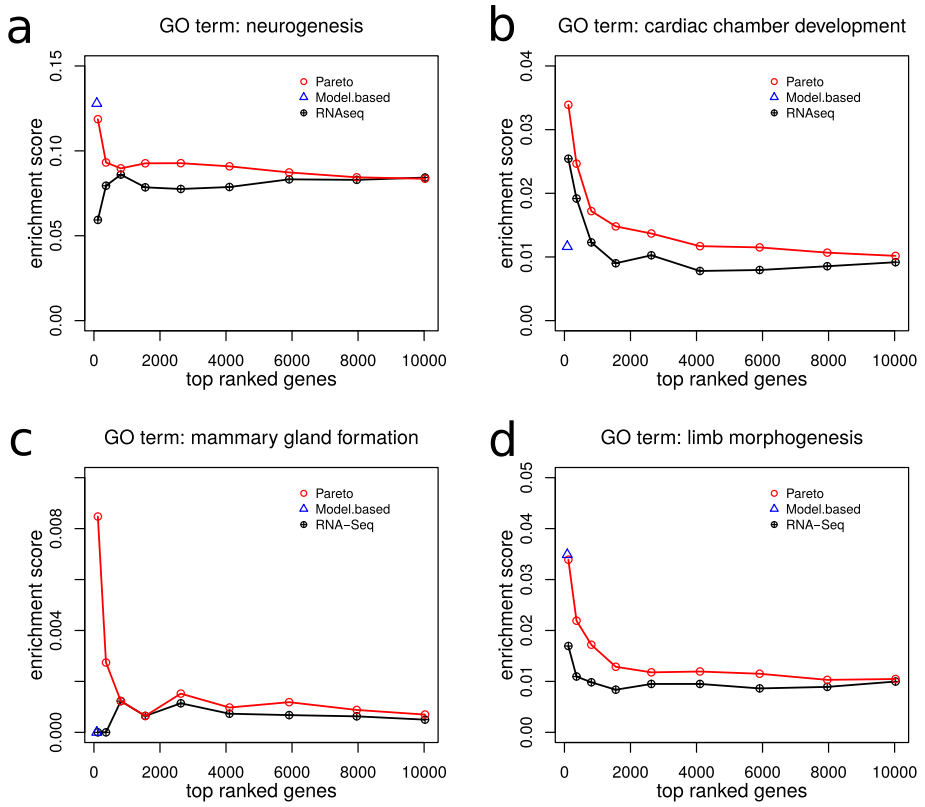


Fig. 4 Comparison of *intePareto* with a model-based clustering approach and analysis of RNA-Seq alone. (a-d) In each of the four panels, the first point from the left on the red line marks the number of genes (x-axis) in the first two (Since there are only 14 genes in the first Pareto front shown in Fig. 3b) Pareto fronts together with the enrichment score (y-axis) of the respective GO term in that Pareto front. Accordingly, the second point refers to the genes in the first three Pareto fronts, etc. Assume that the first i Pareto fronts comprise a total n_i genes, then the corresponding point on the black line takes the first n_i genes, ranked by q-value obtained from the differential gene expression analysis based on RNA-Seq data alone. Note that the red line from the intePareto analysis always lies above the black line, indicating a stronger enrichment of the relevant GO terms in the integrated data compared to RNA-Seq data alone. The blue triangles mark the corresponding values of the existing integrative analysis method

deeper insight into molecular mechanisms underlying processes like cell differentiation or disease progression. The approach chosen here can be generalized to further HT-Seq data types, e.g. from DNA methylation or chromatin accessibility.

Another use of *intePareto* lies in quality control. Specifically, the correlation matrix (Fig. 2) that is generated in the analysis procedure can be used to check ChIP-Seq data quality, which is still not straightforward [47–49]. Such quality checks prior to detailed data analysis and interpretation can avoid errors caused by low-quality ChIP-seq data, and point to possible reasons of failure.

As mentioned above, our approach can be extended in several directions. For instance, improvements are possible if the relationship between distal (even transchromosmal) regulatory elements like enhancers, and their target genes are clarified.

However, it is also true that our approach has inherent limitations. Gene regulation is of such a complexity [50–52] that it probably cannot be completely mapped on a simple approach as that proposed here. We would have to jointly consider the multitude of effects of chromatin remodelers [53, 54], transcription factor co-occupancy [55, 56], different combination of histone modification marks [4, 57], DNA methylation [58], and even RNA modifications [59, 60], which are laborious to capture and profile simultaneously [61]. Nevertheless, we think that a robust, easy-to-use approach such as *intePareto* that exploits subsets of these genomic modalities is a valuable addition to the toolbox of basic and applied genomics.

Availability and requirements

Project name: intePareto
Project home package: https://cran.r-project.org/web/packages/intePareto
https://github.com/yingstat/intePareto (development version)
Operation system(s): Platform independent
Programming language: R (≥3.6.0)
License: GPL (≥2)
Restrictions to use by non-academics: None

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-020-07205-6.

Additional file 1: CaseStudy. Codes and details of an example application with *intePareto*. The data used in the case study are the public data we analyzed in this paper.

Additional file 2: Results_of_intePareto. Full list of the results of integrative analysis using *intePareto*.

Additional file 3: Results_of_RNASeq_data_analysis. Full list of the results of differential gene analysis with RNA-Seq data.

Abbreviations

HT-Seq: High-throughput sequencing; RNA-Seq: RNA-sequencing; ChIP-Seq: Chromatin immuno-precipitation sequencing; TSS: Transcript starting sites; kb: Kilobases; TPM: Transcripts per million; Tet2: Tet methylcytosine dioxygenase 2; mESC: Mouse embryonic stem cell; ESC: Embryonic stem cell; GO: Gene ontology

Acknowledgements

Not applicable.

About this supplement

This article has been published as part of BMC Genomics Volume 21 Supplement 11 2020: Bioinformatics methods for biomedical data science. The full contents of the supplement are available at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-21-supplement-11.

Authors' contributions

YC, SK, and DH conceived the study. YC performed data analysis and drafted the manuscript. SK helped analysis and revision of the manuscript. DH directed analysis and revision of the manuscript. The author(s) read and approved the final manuscript.

Funding

This work was supported by Deutsche Forschungsgemeinschaft [HO 1582/12-1]. Publication costs are funded by the University of Duisburg-Essen. The funding body was not involved and had no role in the study.

Availability of data and materials

All original data are available from NCBI Gene Expression Omnibus (GEO) under accession number GSE48519.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 21 October 2020 Accepted: 29 October 2020
Published: 29 December 2020

References

1. Wang Z, Gerstein M, Snyder M. Rna-seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.
2. Stricker SH, Köferle A, Beck S. From profiles to function in epigenomics. Nat Rev Genet. 2017;18(1):51.
3. Kouzarides T. Chromatin modifications and their function. Cell. 2007;128(4):693–705.
4. Strahl BD, Allis CD. The language of covalent histone modifications. Nature. 2000;403(6765):41–45.
5. Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why? Mol Cell. 2013;49(5):825–37.
6. Ziller MJ, Edri R, Yaffe Y, Donaghey J, Pop R, Mallard W, Issner R, Gifford CA, Goren A, Xing J, et al. Dissecting neural differentiation regulatory networks through epigenetic footprinting. Nature. 2015;518(7539):355–9.
7. Singh R, Lanchantin J, Robins G, Qi Y. Deepchrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics. 2016;32(17):i639–48.
8. Zeng W, Wang Y, Jiang R. Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network. Bioinformatics. 2020;36(2):496–503.
9. Ickstadt K, Schäfer M, Zucknick M. Toward integrative bayesian analysis in molecular biology. Ann Rev Stat Appl. 2018;5:141–67.
10. Klein H-U, Schäfer M, Porse BT, Hasemann MS, Ickstadt K, Dugas M. Integrative analysis of histone ChIP-Seq and transcription data using bayesian mixture models. Bioinformatics. 2014;30(8):1154–62.
11. Schäfer M, Klein H-U, Schwender H. Integrative analysis of multiple genomic variables using a hierarchical bayesian model. Bioinformatics. 2017;33(20):3220–7.

12. Ngatchou P, Zarei A, El-Sharkawi A. Pareto multi objective optimization. In: Proceedings of the 13th International Conference On Intelligent Systems Application to Power Systems. New York: IEEE; 2005. p. 84–91.
13. Handl J, Kell DB, Knowles J. Multiobjective optimization in bioinformatics and computational biology. IEEE/ACM Trans Comput Biol Bioinforma. 2007;4(2):279–92.
14. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic rna-seq quantification. Nat Biotechnol. 2016;34(5):525–7.
15. Liao Y, Smyth GK, Shi W. The r package rsubread is easier, faster, cheaper and better for alignment and quantification of rna sequencing reads. Nucleic Acids Res. 2019;47(8):47–47.
16. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417–9.
17. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from rna-seq reads using lightweight algorithms. Nat Biotechnol. 2014;32(5):462–4.
18. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics. 2009;25(14):1754–60.
19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and samtools. Bioinformatics. 2009;25(16):2078–9.
20. Strausberg RL, Levy S. Promoting transcriptome diversity. Genome Res. 2007;17(7):965–8.
21. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489(7414):109–13.
22. Kulaeva OI, Nizovtseva EV, Polikanov YS, Ulianov SV, Studitsky VM. Distant activation of transcription: mechanisms of enhancer action. Mol Cell Biol. 2012;32(24):4892–7.
23. Rubtsov MA, Polikanov YS, Bondarenko VA, Wang Y-H, Studitsky VM. Chromatin structure can strongly facilitate enhancer action over a distance. Proc Natl Acad Sci. 2006;103(47):17690–5.
24. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M, et al. Genehancer: genome-wide integration of enhancers and target genes in genecards. Database. 2017;2017. https://doi.org/10.1093/database/bax028.
25. Wang S, Sun H, Ma J, Zang C, Wang C, Wang J, Tang Q, Meyer CA, Zhang Y, Liu XS. Target analysis by integration of transcriptome and ChIP-Seq data with beta. Nat Protoc. 2013;8(12):2502–15.
26. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. Vista enhancer browser–a database of tissue-specific human enhancers. Nucleic Acids Res. 2007;35(suppl_1):88–92.
27. Benton ML, Talipineni SC, Kostka D, Capra JA. Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. BMC Genomics. 2019;20(1):511.
28. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq data with deseq2. Genome Biol. 2014;15(12):550.
29. Stark R, Brown G, et al. Diffbind: differential binding analysis of ChIP-Seq peak data. R Packag Version. 2011;100:4–3.
30. Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics. 2019;35(12):2084–92.
31. Roocks P. Computing Pareto Frontiers and Database Preferences with the rPref Package. The R Journal. 2016;8(2):393–404. https://doi.org/10.32614/RJ-2016-054.
32. Hon GC, Song C-X, Du T, Jin F, Selvaraj S, Lee AY, Yen C-a, Ye Z, Mao S-Q, Wang B-A, et al. 5mc oxidation by tet2 modulates enhancer activity and timing of transcriptome reprogramming during differentiation. Mol Cell. 2014;56(2):286–97.
33. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, et al. Ensembl 2019. Nucleic Acids Res. 2019;47(D1):745–51.
34. Koch CM, Andrews RM, Flicek P, Dillon SC, Karaöz U, Clelland GK, Wilcox S, Beare DM, Fowler JC, Couttet P, et al. The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Res. 2007;17(6):691–707.
35. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. High-resolution profiling of histone methylations in the human genome. Cell. 2007;129(4):823–37.
36. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh T-Y, Peng W, Zhang MQ, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet. 2008;40(7):897–903.
37. Karlić R, Chung H-R, Lasserre J, Vlahoviček K, Vingron M. Histone modification levels are predictive for gene expression. Proc Natl Acad Sci. 2010;107(7):2926–31.
38. Dong X, Weng Z. The correlation between histone modifications and gene expression. Epigenomics. 2013;5(2):113–6.
39. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, Vilo J. g: Profiler–a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. 2016;44(W1):83–89.
40. Li X, Yao B, Chen L, Kang Y, Li Y, Cheng Y, Li L, Lin L, Wang Z, Wang M, et al. Ten-eleven translocation 2 interacts with forkhead box o3 and regulates adult neurogenesis. Nat Commun. 2017;8(1):1–14.
41. Greco CM, Kunderfranco P, Rubino M, Larcher V, Carullo P, Anselmo A, Kurz K, Carell T, Angius A, Latronico MV, et al. Dna hydroxymethylation controls cardiomyocyte gene expression in development and hypertrophy. Nat Commun. 2016;7(1):1–15.
42. Fuster JJ, MacLauchlan S, Zuriaga MA, Polackal MN, Ostriker AC, Chakraborty R, Wu C-L, Sano S, Muralidharan S, Rius C, et al. Clonal hematopoiesis associated with tet2 deficiency accelerates atherosclerosis development in mice. Science. 2017;355(6327):842–7.
43. Cakouros D, Hemming S, Gronthos K, Liu R, Zannettino A, Shi S, Gronthos S. Specific functions of tet1 and tet2 in regulating mesenchymal cell lineage determination. Epigenetics Chromatin. 2019;12(1):1–20.
44. Huang Y, Chavez L, Chang X, Wang X, Pastor WA, Kang J, Zepeda-Martínez JA, Pape UJ, Jacobsen SE, Peters B, et al. Distinct roles of the methylcytosine oxidases tet1 and tet2 in mouse embryonic stem cells. Proc Natl Acad Sci. 2014;111(4):1361–6.
45. Holliday H, Baker LA, Junankar SR, Clark SJ, Swarbrick A. Epigenomics of mammary gland development. Breast Cancer Res. 2018;20(1):100.
46. Li R, Zhou Y, Cao Z, Liu L, Wang J, Chen Z, Xing W, Chen S, Bai J, Yuan W, et al. Tet2 loss dysregulates the behavior of bone marrow mesenchymal stromal cells and accelerates tet2-/–driven myeloid malignancy progression. Stem Cell Rep. 2018;10(1):166–79.
47. Carroll TS, Liang Z, Salama R, Stark R, de Santiago I. Impact of artifact removal on chip quality metrics in ChIP-Seq and chip-exo data. Front Genet. 2014;5:75.
48. Mendoza-Parra MA, Gronemeyer H. Assessing quality standards for ChIP-Seq and related massive parallel sequencing-generated datasets: When rating goes beyond avoiding the crisis. Genomics data. 2014;2:268–73.
49. Nakato R, Shirahige K. Recent advances in ChIP-Seq analysis: from quality management to whole-genome annotation. Brief Bioinform. 2017;18(2):279–90.
50. Wu L, Belasco JG. Let me count the ways: mechanisms of gene regulation by mirnas and sirnas. Mol Cell. 2008;29(1):1–7.
51. Reddy TE, Pauli F, Sprouse RO, Neff NF, Newberry KM, Garabedian MJ, Myers RM. Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. Genome Res. 2009;19(12):2163–71.
52. Theunissen TW, Jaenisch R. Mechanisms of gene regulation in human embryos and pluripotent stem cells. Development. 2017;144(24):4496–509.
53. Pray-Grant MG, Daniel JA, Schieltz D, Yates JR, Grant PA. Chd1 chromodomain links histone h3 methylation with saga-and slik-dependent acetylation. Nature. 2005;433(7024):434–8.
54. Wysocka J, Swigut T, Xiao H, Milne TA, Kwon SY, Landry J, Kauer M, Tackett AJ, Chait BT, Badenhorst P, et al. A phd finger of nurf couples histone h3 lysine 4 trimethylation with chromatin remodelling. Nature. 2006;442(7098):86–90.
55. Liu L, Jin G, Zhou X. Modeling the relationship of epigenetic modifications to transcription factor binding. Nucleic Acids Res. 2015;43(8):3873–85.
56. Slattery M, Zhou T, Yang L, Machado ACD, Gordân R, Rohs R. Absence of a simple code: how transcription factors read the genome. Trends Biochem Sci. 2014;39(9):381–99.
57. Berger SL. Histone modifications in transcriptional regulation. Curr Opin Genet Dev. 2002;12(2):142–8.
58. Suzuki MM, Bird A. Dna methylation landscapes: provocative insights from epigenomics. Nat Rev Genet. 2008;9(6):465–76.
59. Roundtree IA, Evans ME, Pan T, He C. Dynamic rna modifications in gene expression regulation. Cell. 2017;169(7):1187–200.

60. Fu Y, Dominissini D, Rechavi G, He C. Gene expression regulation mediated through reversible m 6 a rna methylation. Nat Rev Genet. 2014;15(5):293.
61. Atkinson TJ, Halfon MS. Regulation of gene expression in the genomic context. Comput Struct Biotechnol J. 2014;9(13):e201401001.

Publisher's Note

Supplementary for intePareto: An R package for integrative analysis of RNA-Seq and ChIP-Seq data

Yingying Cao, Simo Kitanovski, Daniel Hoffmann

A frequent question in biology is: How do the functions of different cell types differ? E.g. we may be interested in what the effect of a mutation or gene knockout is in terms of functional differences between wild type and mutant/knockout, or how cellular function changes between two developmental stages of a cell type. One way of understanding such functional difference is to characterize them at the level of differences in repertoires in active genes or suppressed genes. The characterization of differential gene expression is helpful in this respect, but even more expressive is the combination of evidence from different experiments, namely measurements of gene expression (RNA-Seq) and measurements of various histone modifications (ChIP-Seq) that allow assessment of activation or suppression state of genes. In our experience, this combination of information gives a clearer picture of the cellular function at the molecular level than using any of the information types alone.

We have therefore developed the R package intePareto that allows such a combination of different types of sequencing data. The intePareto workflow starts with RNA-Seq and ChIP-Seq data for two different cell types or conditions. The ChIP-Seq data will in general comprise information on several histone modifications with activating or repressing function. The end product of intePareto is a list of genes prioritized according to congruence of changes of gene expression and histone modifications.

In the following we demonstrate the technical workflow with the published dataset GSE48519 where a Tet2 knockout cell line is compared to the wild type [4]. The raw data were downloaded from Gene Expression Omnibus. This set contains 4 RNA-Seq samples, 31 ChIP-Seq samples with histone modification mark of H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, H3K36me3 and control for Tet2 knockout and wild type mouse embryonic stem cells (mESCs) separately.

Preprocessing of RNA-Seq data and ChIP-Seq data

The quality of the fastq data was examined with FastQC [1]. The RNA-Seq data was then preprocessed by Kallisto [2], and ChIP-Seq data was preprocessed by BWA [5], Samtools [6]. The meta data below gives an overview of the preprocessed files that are the input files for intePareto.

```
library(intePareto)
```

```
rna_meta
```

```
##         SRR condition                                                files
## 1 SRR925874 wild.type ./data/tet2/kallisto/outputSRR925874/abundance.tsv
## 2 SRR925875 wild.type ./data/tet2/kallisto/outputSRR925875/abundance.tsv
## 5 SRR925878  tet2.out ./data/tet2/kallisto/outputSRR925878/abundance.tsv
## 6 SRR925879  tet2.out ./data/tet2/kallisto/outputSRR925879/abundance.tsv
```

```
chip_meta[1:2,]
```

```
##         SRR    mark condition                                      files
## 1 SRR925639 H3K4me1 wild.type ./data/tet2/output/SRR925639_sorted.bam
## 2 SRR925640 H3K4me3 wild.type ./data/tet2/output/SRR925640_sorted.bam
```

1. match: Match RNA-Seq and ChIP-Seq data on the gene level

Take the meta data of the preprocessed RNA-Seq and ChIP-Seq data as input. The first step of intePareto is to match the RNA-Seq data and ChIP-Seq data on the gene level. There are two strategies available now to do the matching step: (1) highest - choose the maximum promoter abundance value among all the promoters as a representative of the ChIP-Seq signal for this gene. (2) weighted.mean - calculate the weighted mean of the abundance value of all the promoters to represent the ChIP-Seq signal for this gene. Here we choose "highest":

```
library(intePareto)
chip_meta_noH3K36me3 <- chip_meta[!chip_meta$mark%in%"H3K36me3",]
res.1 <- doMatch(rnaMeta = rna_meta, # metadata of RNA-Seq
                 chipMeta = chip_meta_noH3K36me3, # metadata or ChIP-Seq
                 region = "promoter", # specify the region
                 method = "highest", # specify the strategy to do the match
                 ensemblDataset = "mmusculus_gene_ensembl"
                 # specify the dataset of corresponing species
)
chip_meta_H3K36me3 <- chip_meta[chip_meta$mark%in%"H3K36me3",]

res.2 <- doMatch(rnaMeta = rna_meta, # metadata of RNA-Seq
                 chipMeta = chip_meta_H3K36me3, # metadata or ChIP-Seq
                 method = "highest", # we don't need this parameter if we choose
                 # genebody, but it doesn't matter if we choose
                 region = "genebody", # specify the region
                 ensemblDataset = "mmusculus_gene_ensembl"
                 # specify the dataset of corresponing species
)
res.1$matched.data <- merge(res.1$matched.data,
                            res.2$matched.data)
res.1$res.chip <- merge(res.1$res.chip,
                        res.2$res.chip)
res <- res.1
```

Figure 1 shows the correlation matrix, the color represents the value of correlation coefficients of Spearman's rank correlation of all samples. From this figure we can see that RNA-Seq (wild.type_REP1, wild.type_REP2, tet2.out_REP1, tet2.out_REP2) positively correlate with active histone modification markers (H3K4me3, H3K27ac, H3K4me1, and H3K36me3), and negatively correlate with repressive histone modification markers (H3K27me3 and H3K9me3). This can confirm our match strategy works well for the match of RNA-Seq and ChIP-Seq data on the gene level.

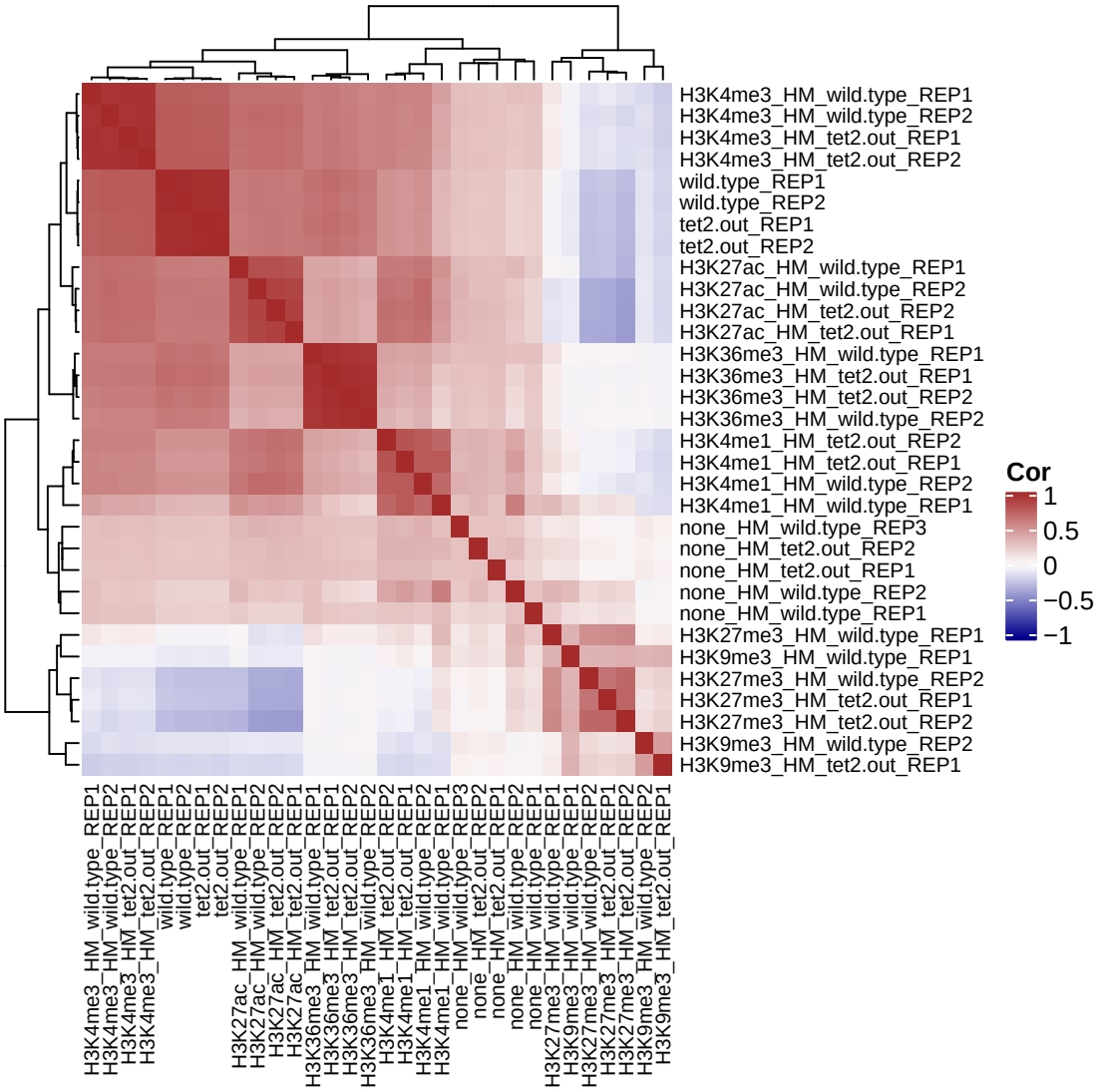


Figure 1: Correlation of RNA-Seq and ChIP-Seq

2. integration: Integrate RNA-Seq and ChIP-Seq data by calculating logFoldChange and Z scores

After the match of RNA-Seq and ChIP-Seq at the gene level, the integration of these two types of data is conducted through the **doIntegration** function by calculating logFoldChange of RNA-Seq and ChIP-Seq and then calculate Z scores for each marker. The input of the function is a list result from **doMatch** function.

```
df_final <- doIntegration(res = res, # result list from "doMatch" function
                          type = "apeglm", # shrinkage estimator, default is "apeglm"
                          ref = "wild.type", # specifying the reference level
                          apeAdapt = FALSE)
```

The result of the second step from **doIntegration** function contains the logFoldChange of RNA-Seq and ChIP-Seq as well as the Z score of each mark for each gene (shown as the table below).

Table 1: integration results

	z.H3K9me3	z.none	z.H3K27me3	z.H3K27ac	z.H3K4me3	z.H3K4me1	z.H3K36me3
0610009B22Rik	1.5096313	-0.4556445	1.1294972	-0.0600427	0.1922781	-0.3990826	0.1116033
0610010F05Rik	0.0727513	-0.3159894	0.1848376	1.2379816	-0.3030021	0.1750952	-0.1077852
0610010K14Rik	0.2818030	0.5665779	0.3684958	0.1205920	0.0840617	0.0560468	0.0950096
0610012G03Rik	-0.0402419	-0.0577281	-0.0344298	0.0053385	0.0248794	0.0119307	-0.0709771
0610030E20Rik	-0.2843424	-0.0069848	-0.0345165	0.5311879	-0.1859840	-0.0157660	-0.1489075
0610040J01Rik	0.0263985	-0.6136573	0.0738724	1.2450397	0.0459432	-0.0293661	0.3395275
1110002E22Rik	-0.1117175	-0.6161363	0.3220724	-0.0374538	0.0682505	-0.4427649	0.4736603
1110004F10Rik	-0.0075126	0.0382641	-0.0531325	-0.0302648	0.0218445	0.0215032	0.0285630
1110008P14Rik	0.5878211	0.1613085	-0.0158789	-0.0216326	-0.1131650	0.1088885	0.2975963
1110012L19Rik	-0.7605611	-0.2186961	-0.4329753	-0.3474599	-0.1035559	-0.2430795	0.3229549
1110017D15Rik	-0.5635032	-0.0215326	0.1116994	-0.2396299	0.2999805	0.4402020	-0.2901453
1110032A03Rik	-0.0308223	1.3990756	-0.3721691	0.5582615	-0.7094221	-0.9638899	0.4072242
1110032F04Rik	-0.0619083	0.2234994	-0.0579622	0.1148733	0.0706187	0.1086061	0.1350982
1110051M20Rik	0.0778253	-0.0865843	-0.2830442	-0.1500254	0.1005083	-0.2821450	0.0665700
1110059E24Rik	-0.0749205	-0.1335756	0.0250346	0.2520036	-0.1724477	-0.2854025	-0.1534097

3. prioritization: prioritization of genes based on Z scores with Pareto optimization

Take the Z scores of several different histone modifications as input, the prioritization of genes based on Z scores can be formulated as multiobjective optimization problem and solved with Pareto optimization [8]. The aim of Pareto optimization method is to find the Pareto optimal trade-off (Pareto front) between conflicting objectives (such as minimizing Z score of H3K27me3 and maximizing z-score of H3K4me3 for each gene). The results of Pareto optimization method is a rank-ordering of the genes by the level of the congruent changes in RNA-Seq and ChIP-Seq (shown as table below).

```
nr.fronts <- 50 # choose a large number to include all the fronts
objective <- data.frame(mark = c("z.H3K27ac", "z.H3K4me3", "z.H3K4me1",
                                 "z.H3K36me3", "z.H3K9me3", "z.H3K27me3"),
                        obj=c("max","max","max","max","min","min"),
                        stringsAsFactors=FALSE)
res_final <- doPareto(df_final = df_final,
                      objective = objective,
                      nr.fronts = nr.fronts)
```

Table 2: prioritization results

	z.H3K9me3	z.none	z.H3K27me3	z.H3K27ac	z.H3K4me3	z.H3K4me1	z.H3K36me3	front
Atp6v0c-ps2	16.589667	3.2706193	-20.013204	-1.725424	11.8417083	-0.5113998	10.440533	1
Catsperg1	-8.061640	-1.5505929	1.034593	10.164148	12.7109471	0.8538197	12.002940	1
Fam110c	7.090283	4.5737313	-3.192932	38.702950	7.4691643	4.8392381	12.702844	1
Grb14	-3.354402	2.4001542	-14.829860	27.547845	10.8787232	4.9987109	15.616601	1
Khdc1a	-9.654657	-5.3475747	3.041933	17.776863	11.8376748	20.2619271	14.673243	1
Metrn	-8.142343	-6.7852712	-5.716880	-10.044830	8.9894024	11.5985664	12.760880	1
Olfr94	-5.706979	0.4598344	-36.221443	99.403578	21.5264205	20.2764413	1.578383	1
Olig1	-6.158681	11.2312845	-20.694139	34.781308	20.0243582	8.6956798	4.906146	1
Pramef17	5.581220	8.4221983	-6.412906	11.671704	6.1742990	21.5723380	21.153664	1
Trav5-2	-8.224873	2.4279795	-3.729377	23.326949	-0.1153585	13.0210337	18.820450	1
Trim34b	-10.368544	-0.9556500	-14.980885	5.767722	3.5631355	-11.3481747	-6.496720	1
Tspo2	-3.858320	-5.9325064	-4.143173	13.234274	30.3344625	12.5968673	38.445580	1
Ttll11	-3.243813	0.0535220	1.956712	98.616938	7.3652294	7.1142097	18.991588	1
Ubd	-26.955101	8.3544039	-12.420790	73.875377	115.7006521	42.1527121	11.752621	1
Zfp786	-6.054174	2.8041174	-0.747937	6.618457	3.1202473	7.2548820	5.319367	2

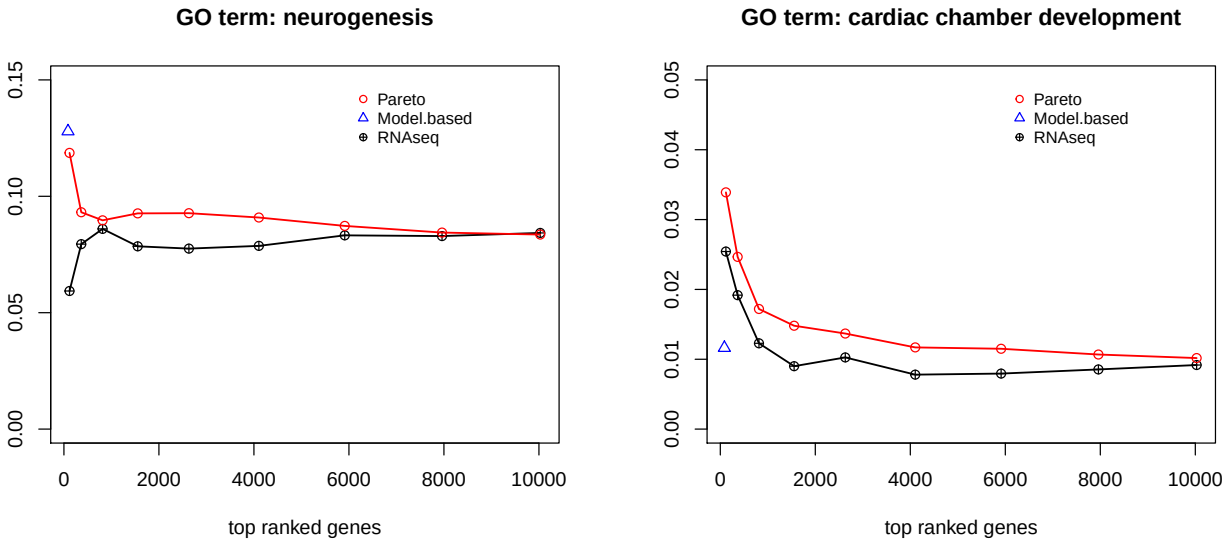


Figure 2: Compare inte Pareto with another approach and RNAseq alone

Compare our integrative results with other approach and RNA-Seq alone

To evaluate the performance of our method we did the integrative analysis with a recent published the hierarchical Bayesian model-based clustering approach (model-based approach) [9] and analysis of RNA-Seq alone, a functional quality metric, enrichment score of interesting terms for each data set is used to do comparison research between our integrative approach, the model-based approach and the analysis of RNA-Seq data alone. The enrichment score is defined as $N(G_i \cap G_r) \mid N(G_r)$, in which G_i stands for the genes in the interesting GO terms, G_r stands for the genes in the result of analysis.

Both our integrative approach and the model-based approach found that the genes in the final result or the top rank gene (inte Pareto) enriched in "limb morphogenesis" GO term [Figure 2], which is consist with results in a recently published research [7]. Our approach also found genes in the top-ranked list enriched in "mammary gland formation" which is discussed in this review [3]. This result will be hidden when RNA-Seq data alone was analyzed or the model-based approach was used. When compared to the model-based approach[9], our method outperformed it by having comparable results at the top genes and offering more informative results by providing the rank-ordered list of the remaining genes.

```
sessionInfo()
```

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04.6 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/libblas/libblas.so.3.6.0
## LAPACK: /usr/lib/lapack/liblapack.so.3.6.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=de_DE.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=de_DE.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=de_DE.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=de_DE.UTF-8 LC_IDENTIFICATION=C
```

```
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
## [1] knitr_1.24           circlize_0.4.8       ComplexHeatmap_2.0.0
## [4] intePareto_0.1.1     apeglm_1.6.0
##
## loaded via a namespace (and not attached):
##   [1] bitops_1.0-6                matrixStats_0.55.0
##   [3] bit64_0.9-7                 httr_1.4.1
##   [5] RColorBrewer_1.1-2          progress_1.2.2
##   [7] GenomeInfoDb_1.20.0         numDeriv_2016.8-1.1
##   [9] tools_3.6.1                 backports_1.1.4
##  [11] R6_2.4.0                    rpart_4.1-15
##  [13] Hmisc_4.2-0                 DBI_1.1.0
##  [15] lazyeval_0.2.2              BiocGenerics_0.30.0
##  [17] colorspace_1.4-1            GetoptLong_0.1.7
##  [19] nnet_7.3-12                 prettyunits_1.0.2
##  [21] tidyselect_1.0.0            gridExtra_2.3
##  [23] DESeq2_1.24.0               bit_1.1-14
##  [25] compiler_3.6.1              Biobase_2.44.0
##  [27] htmlTable_1.13.1            DelayedArray_0.10.0
##  [29] scales_1.0.0                checkmate_1.9.4
##  [31] genefilter_1.66.0           stringr_1.4.0
##  [33] digest_0.6.20               Rsamtools_2.0.0
##  [35] foreign_0.8-72              rmarkdown_1.15
##  [37] XVector_0.24.0              base64enc_0.1-3
##  [39] pkgconfig_2.0.2             htmltools_0.3.6
##  [41] highr_0.8                   bbmle_1.0.20
##  [43] GlobalOptions_0.1.0         htmlwidgets_1.3
##  [45] rlang_0.4.5                 rstudioapi_0.10
##  [47] RSQLite_2.1.2               rPref_1.3
##  [49] shape_1.4.4                 BiocParallel_1.18.1
##  [51] acepack_1.4.1               dplyr_0.8.3
##  [53] RCurl_1.95-4.12             magrittr_1.5
##  [55] GenomeInfoDbData_1.2.1      Formula_1.2-3
##  [57] Matrix_1.2-17               Rcpp_1.0.2
##  [59] munsell_0.5.0               S4Vectors_0.22.1
##  [61] stringi_1.4.3               yaml_2.2.0
##  [63] MASS_7.3-51.4               SummarizedExperiment_1.14.1
##  [65] zlibbioc_1.30.0             plyr_1.8.4
##  [67] blob_1.2.0                  parallel_3.6.1
##  [69] crayon_1.3.4                lattice_0.20-38
##  [71] Biostrings_2.52.0           splines_3.6.1
##  [73] annotate_1.62.0             hms_0.5.1
##  [75] locfit_1.5-9.1              pillar_1.4.2
##  [77] igraph_1.2.4.1              GenomicRanges_1.36.0
##  [79] rjson_0.2.20                geneplotter_1.62.0
##  [81] biomaRt_2.40.5              stats4_3.6.1
##  [83] XML_3.98-1.20               glue_1.3.1
##  [85] evaluate_0.14               latticeExtra_0.6-28
##  [87] data.table_1.12.2           RcppParallel_4.4.3
```

```
##  [89] png_0.1-7                  vctrs_0.2.4
##  [91] gtable_0.3.0               purrr_0.3.2
##  [93] clue_0.3-57                assertthat_0.2.1
##  [95] ggplot2_3.2.1              emdbook_1.3.11
##  [97] xfun_0.9                   xtable_1.8-4
##  [99] coda_0.19-3                survival_2.44-1.1
## [101] tibble_2.1.3               GenomicAlignments_1.20.1
## [103] AnnotationDbi_1.46.1       memoise_1.1.0
## [105] IRanges_2.18.3             cluster_2.1.0
```

References

[1] Simon Andrews et al. *FastQC: a quality control tool for high throughput sequence data*. 2010.

[2] Nicolas L Bray et al. "Near-optimal probabilistic RNA-seq quantification". In: *Nature biotechnology* 34.5 (2016), p. 525.

[3] Holly Holliday et al. "Epigenomics of mammary gland development". In: *Breast Cancer Research* 20.1 (2018), pp. 1–11.

[4] Gary C Hon et al. "5mC oxidation by Tet2 modulates enhancer activity and timing of transcriptome reprogramming during differentiation". In: *Molecular cell* 56.2 (2014), pp. 286–297.

[5] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *bioinformatics* 25.14 (2009), pp. 1754–1760.

[6] Heng Li et al. "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[7] Rong Li et al. "TET2 Loss Dysregulates the Behavior of Bone Marrow Mesenchymal Stromal Cells and Accelerates Tet2-/–Driven Myeloid Malignancy Progression". In: *Stem cell reports* 10.1 (2018), pp. 166–179.

[8] Patrick Ngatchou, Anahita Zarei, and A El-Sharkawi. "Pareto multi objective optimization". In: *Proceedings of the 13th International Conference on, Intelligent Systems Application to Power Systems*. IEEE. 2005, pp. 84–91.

[9] Martin Schäfer, Hans-Ulrich Klein, and Holger Schwender. "Integrative analysis of multiple genomic variables using a hierarchical Bayesian model". In: *Bioinformatics* 33.20 (2017), pp. 3220–3227.

3.2 UMI or not UMI, that is the question for scRNA-seq zero-inflation

This section is based on the following publication:

Yingying Cao, Simo Kitanovski, Ralf Küppers, and Daniel Hoffmann. **UMI or not UMI, that is the question for scRNA-seq zero-inflation.** Nature Biotechnology (2021): 1-2. Publisher - Springer Nature.
`https://doi.org/10.1038/s41587-020-00810-6`

nature biotechnology

MATTERS ARISING

https://doi.org/10.1038/s41587-020-00810-6

Check for updates

UMI or not UMI, that is the question for scRNA-seq zero-inflation

Yingying Cao[1], Simo Kitanovski[1], Ralf Küppers[2] and Daniel Hoffmann[1]✉

ARISING FROM V. Svensson *Nature Biotechnology* https://doi.org/10.1038/s41587-019-0379-5 (2020)

In your February 2020 issue, Svensson[1] addressed the problem of zero-inflation in single-cell RNA-sequencing (scRNA-seq) data—that is, the observation that more genes in more cells than expected appear to have zero expression. Using examples, the Correspondence demonstrates that droplet-based methods that make use of unique molecular identifier (UMI)[2,3] counts to quantify gene expression are adequately modeled with negative binomial distributions without zero-inflation. We agree with this, and we also share the concern that there is confusion about the validity of zero-inflation and the necessity of computational methods to eliminate it.

Even so, we find Svensson's subsequent discussion of plate-based scRNA-seq methods misleading because a reader not deeply immersed in the subject may be tempted to draw the conclusion that zero-inflation is a matter of the technical platform; specifically, that droplet-based scRNA-seq data are not zero-inflated, whereas plate-based scRNA-seq data are zero-inflated. Such a conclusion would be misguided and potentially could damage prospects for important technological developments and applications in the highly dynamic scRNA-seq field. We therefore felt the need for a clarifying response.

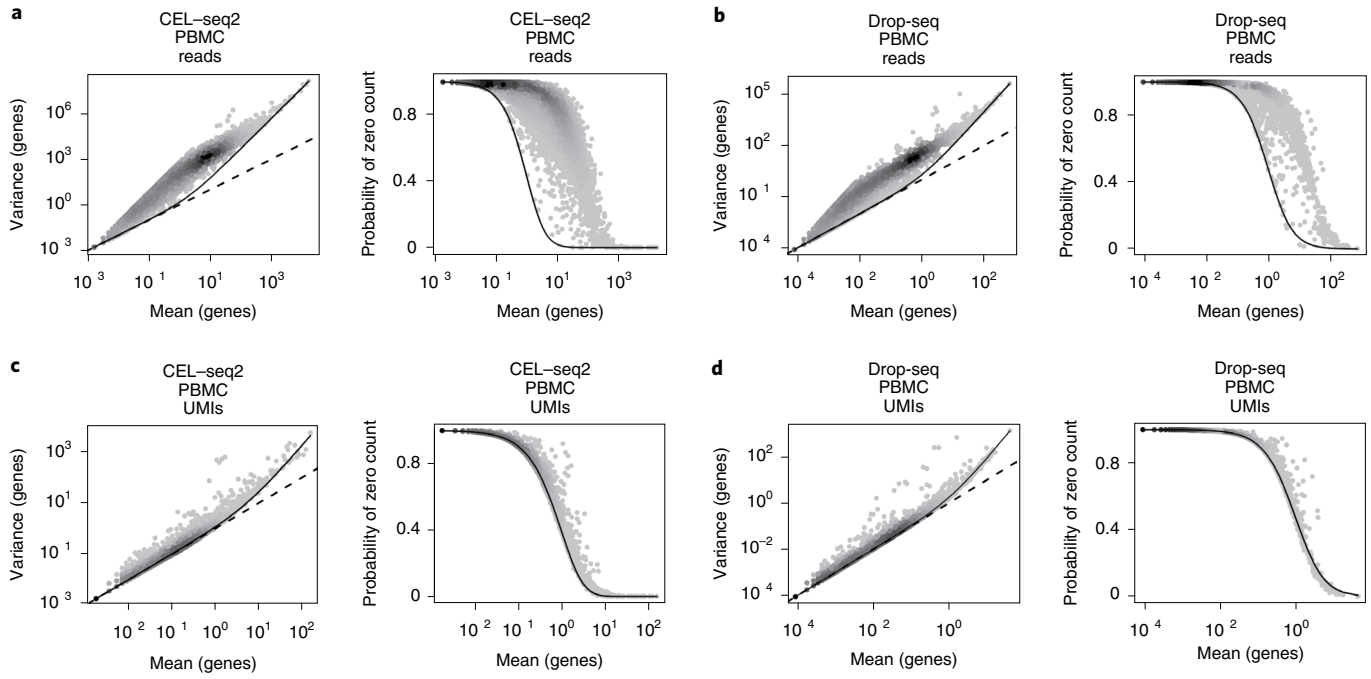


Fig. 1 | A comparison of the effect of UMI use and the sequencing platform on zero-inflation. a–**d**, Read counts for plate-based cell expression by linear amplification and sequencing 2 (CEL-seq2; **a**) and droplet-based Drop-seq (**b**) versus UMI counts for CEL-seq2 (**c**) and Drop-seq (**d**) on a sample of heterogeneous cells (peripheral blood monocytes; PBMCs)[4]. In the left-hand plot of each panel, the solid curve is a least-squares fit ($\mathrm{var} = \mu + \varphi\mu^2$, valid for a negative binomial distribution with mean μ and dispersion φ, as in the Correspondence[1]) used to determine φ. Dashed lines are diagonals with intercept = 0 and slope = 1. In the right-hand plot of each panel, the solid curve is the predicted fraction of zeros with that φ. The density of the actual scRNA-seq data is represented from low (light gray) to high (black). The data are clearly zero-inflated for both plate-based and droplet-based scRNA-seq with read count quantification (**a**,**b**), whereas for UMI count quantification zero-inflation is suppressed for both plate-based and droplet-based scRNA-seq (**c**,**d**).

[1]Bioinformatics and Computational Biophysics, Faculty of Biology and Center for Medical Biotechnology, University of Duisburg-Essen, Essen, Germany. [2]Institute of Cell Biology (Cancer Research), University Hospital Essen, Essen, Germany. ✉e-mail: daniel.hoffmann@uni-due.de

Our response can be stated crisply as follows: what matters most for zero-inflation with current scRNA-seq is not the technical platform (droplet versus plate), but whether gene expression is measured in terms of UMI counts or read counts—suppressed zero-inflation with UMIs, stronger zero-inflation with read counts. For UMI count experiments, we typically have the raw read counts as well, so this point can be made by direct comparison within the same experiment. Figure 1 shows exemplary cases from published data[4] for all four combinations of plate-based (Fig. 1a,c) and droplet-based (Fig. 1b,d) scRNA-seq, with read counts (Fig. 1a,b) and UMI counts (Fig. 1c,d), demonstrating that even for heterogeneous samples not the platform but the use of UMIs makes the difference for zero-inflation. Supplementary Table 1 shows this for more datasets in a form similar to Table 1 of the Correspondence[1].

The point that we are making here has been made already by others for data across technical platforms[5] and also specifically for droplet-based data[6], but it has apparently not received the necessary attention. Curiously, the Correspondence[1] itself mentions possible reasons for zero-inflation, including that the use of UMI counts deflates amplification bias, although the Correspondence[1] ignores that the use of UMI counts is not limited to droplet-based methods[7]. The main reason for suppressed zero-inflation with UMI counts is likely that UMI counts collapse multiple reads from the same original RNA molecule to a single read, thus also collapsing for many weakly expressed genes the gap between zero and non-zero expression that had been artificially widened by amplification. After this collapsing, the non-zero-inflated negative binomial is again the appropriate distribution.

Although measuring UMI counts is a good way to avoid zero-inflation problems, UMI counting is not a panacea. For instance, if accurate mapping of reads or detection of isoforms is a major objective of an scRNA-seq study, a tag-based protocol with UMIs could be less useful than a full-length sequencing protocol without UMIs.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-020-00810-6.

Received: 19 February 2020; Accepted: 23 December 2020;
Published online: 01 February 2021

References

1. Svensson, V. Droplet scRNA-seq is not zero-inflated *Nat. Biotechnol.* **38**, 147–150 (2020).
2. Kivioja, T. et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2012).
3. Islam, S. et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).
4. Ding, J. et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).
5. Chen, W. et al. UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biol.* **19**, 70 (2018).
6. Townes, F. W. et al. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **20**, 295 (2019).
7. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).

Acknowledgements
We thank Deutsche Forschungsgemeinschaft for funding (grants KU1315/14-1 and HO1582/12-1).

Author contributions
Y.C. analyzed data and wrote the first draft, S.K. and R.K. reviewed work and revised the text, D.H. supervised work and wrote the final text.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-020-00810-6.

Correspondence and requests for materials should be addressed to D.H.

Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

nature research

https://doi.org/10.1038/s41587-020-00810-6

Supplementary information

UMI or not UMI, that is the question for scRNA-seq zero-inflation

In the format provided by the authors and unedited

Supplementary Information for "UMI or not UMI, that is the question for scRNAseq zero-inflation"

Yingying Cao[a], Simo Kitanovski[a], Ralf Küppers[b], Daniel Hoffmann[a]

Table 1. Overview of analyzed data sets [1]

Data	Method	Cells	Genes	Measurements	percentage of genes over x percentage points away from expected fraction			
					1 percentage point	5 percentage points	10 percentage points	20 percentage points
Cortex	10xChromium	5720	23313	reads	72.17%	59.31%	52.78%	43.95%
				UMIs	9.62%	2.69%	1.25%	0.24%
	DroNc-seq	3087	22864	reads	72.07%	57.72%	49.07%	36.09%
				UMIs	15.82%	1.89%	0.50%	0.06%
	Sci-RNA-seq	5792	23493	reads	74.05%	59.61%	51.80%	41.49%
				UMIs	8.19%	1.60%	0.37%	0.08%
PBMC	Seq-Well	5676	22978	reads	54.24%	37.97%	28.26%	15.51%
				UMIs	5.17%	1.58%	0.91%	0.54%
	Drop-seq	11052	25015	reads	57.97%	44.28%	38.06%	29.96%
				UMIs	7.88%	0.52%	0.18%	0.05%
	inDrops	11350	21514	reads	63.53%	50.24%	43.25%	32.57%
				UMIs	3.29%	0.67%	0.21%	0.06%
	CEL-Seq2	560	22730	reads	86.03%	75.22%	69.34%	61.77%
				UMIs	17.70%	2.26%	0.73%	0.17%
	10xChromiumv3	4027	22499	reads	62.61%	48.07%	41.13%	31.68%
				UMIs	13.01%	3.26%	1.28%	0.45%
	10xChromiumv2	11768	24968	reads	57.83%	44.52%	37.94%	29.11%
				UMIs	5.03%	1.18%	0.44%	0.18%

Table shows results for data sets [1] for heterogeneous cells from cortex (top) and PMBCs (bottom), including scRNAseq method, cells and genes analyzed, and percentage of genes over x 1, , 10, 20 percentage points away from expected fraction for a negative binomial model without zero-inflation. Percentages are reported for both read counts and UMI counts. 10xChromium: Chromium Single Cell 3' Reagent; DroNc-seq: single nucleus RNA-seq with Drop-seq; Sci-RNA-seq: single-cell combinatorial-indexing RNA-seq; Seq-Well; Drop-seq: Droplet-Sequencing; inDrops: indexing droplets; CEL-seq2: improved version of CEL-seq; 10xChromiumv3 / v2: Chromium Single Cell 3' Reagent (version 3 / version 2, respectively).

References

1. Ding, J. *et al. BioRxiv* 632216 (2019).

[a]Bioinformatics and Computational Biophysics, Faculty of Biology and Center for Medical Biotechnology, University of Duisburg-Essen, Essen, Germany
[b]Institute of Cell Biology (Cancer Research), University Hospital Essen, Essen, Germany

3.3 Excessive Neutrophils and Neutrophil Extracellular Traps in COVID-19

This section is based on the following publication:

Jun Wang, Qian Li, Yongmei Yin, Yingying Zhang, Yingying Cao, Xiaoming Lin, Lihua Huang, Daniel Hoffmann, Mengji Lu, and Yuanwang Qiu. **Excessive neutrophils and neutrophil extracellular traps in COVID-19.** Frontiers in immunology 11 (2020): 2063.
https://doi.org/10.3389/fimmu.2020.02063

frontiers in Immunology

ORIGINAL RESEARCH
published: 18 August 2020
doi: 10.3389/fimmu.2020.02063

Excessive Neutrophils and Neutrophil Extracellular Traps in COVID-19

Jun Wang[1,2,3†], Qian Li[3,4†], Yongmei Yin[5†], Yingying Zhang[1], Yingying Cao[2], Xiaoming Lin[5], Lihua Huang[1,6], Daniel Hoffmann[2], Mengji Lu[3*] and Yuanwang Qiu[1,5,6*]*

[1] Center of Clinical Laboratory, The Fifth People's Hospital of Wuxi, Jiangnan University, Wuxi, China, [2] Bioinformatics and Computational Biophysics, University of Duisburg-Essen, Essen, Germany, [3] Institute of Virology, University Hospital of Essen, University of Duisburg-Essen, Essen, Germany, [4] Department of Laboratory Medicine, Maternal and Child Health Hospital of Hubei Province, Wuhan, China, [5] Radiology Department, The Fifth People's Hospital of Wuxi, Jiangnan University, Wuxi, China, [6] Department of Infectious Diseases, The Fifth People's Hospital of Wuxi, Jiangnan University, Wuxi, China

Background: Cases of excessive neutrophil counts in the blood in severe coronavirus disease (COVID-19) patients have drawn significant attention. Neutrophil infiltration was also noted on the pathological findings from autopsies. It is urgent to clarify the pathogenesis of neutrophils leading to severe pneumonia in COVID-19.

Methods: A retrospective analysis was performed on 55 COVID-19 patients classified as mild ($n = 22$), moderate ($n = 25$), and severe ($n = 8$) according to the Guidelines released by the National Health Commission of China. Trends relating leukocyte counts and lungs examined by chest CT scan were quantified by Bayesian inference. Transcriptional signatures of host immune cells of four COVID19 patients were analyzed by RNA sequencing of lung specimens and BALF.

Results: Neutrophilia occurred in 6 of 8 severe patients at 7–19 days after symptom onset, coinciding with lesion progression. Increasing neutrophil counts paralleled lesion CT values (slope: 0.8 and 0.3–1.2), reflecting neutrophilia-induced lung injury in severe patients. Transcriptome analysis revealed that neutrophil activation was correlated with 17 neutrophil extracellular trap (NET)-associated genes in COVID-19 patients, which was related to innate immunity and interacted with T/NK/B cells, as supported by a protein–protein interaction network analysis.

Conclusion: Excessive neutrophils and associated NETs could explain the pathogenesis of lung injury in COVID-19 pneumonia.

Keywords: coronavirus, COVID-19, neutrophil extracellular trap, pneumonia, neutrophilia, lymphopenia

OPEN ACCESS

Edited by:
Cees Van Kooten, Leiden University, Netherlands

Reviewed by:
Claudio Costantini, University of Perugia, Italy
Payel Sil, National Institute of Environmental Health Sciences (NIEHS), United States

***Correspondence:**
Daniel Hoffmann
Daniel.Hoffmann@uni-due.de
Mengji Lu
mengji.lu@uni-due.de
Yuanwang Qiu
qywang839@126.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Innate Immunity, a section of the journal Frontiers in Immunology

Received: *05 June 2020*
Accepted: *29 July 2020*
Published: *18 August 2020*

Citation:
Wang J, Li Q, Yin Y, Zhang Y, Cao Y, Lin X, Huang L, Hoffmann D, Lu M and Qiu Y (2020) Excessive Neutrophils and Neutrophil Extracellular Traps in COVID-19. Front. Immunol. 11:2063. doi: 10.3389/fimmu.2020.02063

INTRODUCTION

As of early May 2020, more than 3 million cases of coronavirus disease 2019 (COVID-19) have been confirmed worldwide, resulting in hundreds of thousands of deaths (1). According to the Guidelines of the Diagnosis and Treatment of New Coronavirus Pneumonia (version 7) published by the National Health Commission of China, COVID-19 patients can be classified as mild, moderate, and severe cases. Severe patients easily develop acute respiratory distress syndrome (ARDS) or multiple organ failure, with a 4–15% death rate (2, 3)

It is not well-understood what drives the exacerbated host response involving a cytokine storm in severe COVID-19 (4). Specifically, it is unclear what initiates and propagates the

cytokine storm. Neutrophil infiltration was noted in three recent reports on the pathological findings from autopsied COVID-19 patients (5–7). Neutrophil infiltration in pulmonary capillaries, acute capillaritis with fibrin deposition, extravasation of neutrophils into the alveolar space, and neutrophilic mucositis were observed. Similarly, increased neutrophil counts were reported to occur simultaneously in the peripheral blood of severe and non-surviving COVID-19 patients (3, 8). Neutrophilia predicts poor outcomes in patients with COVID-19, and our previous research also indicated the neutrophil-to-lymphocyte ratio (NLR) is an independent risk factor for severe disease (8, 9).

Recently, two serum markers of neutrophil extracellular traps (NETs), myeloperoxidase (MPO)-DNA, and citrullinated histone H3 (Cit-H3) levels were found to be elevated in the serum of COVID-19 patients (10). This suggested that neutrophilia and excessive NETs may contribute to cytokine release and respiratory failure. As a contributor to pathological inflammation of pneumonia, excessive neutrophils lead to tissue injury by oxidative burst, phagocytosis, and the formation of neutrophil NETs, known as NETosis. NETs are composed of extracellular webs of DNA, histones, microbicidal proteins, and oxidative enzymes that are released by neutrophils to corral infections (11–15). The ability of NETs to damage tissues is well-documented in infection and sterile disease. NETs directly kill epithelial and endothelial cells (16, 17), and excessive NETosis damages the epithelium in pulmonary fungal infection (18) and the endothelium in transfusion-related acute lung injury (19).

In the present study, first, the dynamics of neutrophil counts in COVID-19 patients ($n = 23$) during hospitalization were examined, together with the corresponding lung injury, to clinically define the relationship between lung injury and leukocyte counts. Second, transcriptional signatures of host immune cells from COVID-19 patients ($n = 4$) were analyzed by RNA sequencing of lung specimens or bronchoalveolar lavage fluids (BALF). Immune cell frequency was analyzed by MCPcouter. We used average expression of genes enriched in neutrophil degranulation and activation to screen highly correlated genes and further identified NET associated genes in the correlated gene list to construct an interactive network from the STRING database.

METHODS

Participants and Study Design

The study was approved by the Ethics Committee of the Fifth People's Hospital, Wuxi (No. 2020-006-1). The 55 confirmed COVID-19 patients were enrolled in this retrospective study from January 23 to March 15, 2020. Written informed consent was obtained from all patients from the Fifth People's Hospital, Wuxi, China.

The clinical handling of COVID-19 patients was performed according to the Guidelines of the Diagnosis and Treatment of New Coronavirus Pneumonia (version 7) published by the National Health Commission of China. Mild, moderate, and severe cases were defined by the following conditions: (1) epidemiological history, (2) fever or other respiratory symptoms, (3) frequency of typical CT image abnormalities of viral pneumonia, and (4) positive RT-PCR result for SARS-CoV-2 RNA. In addition, mild cases were diagnosed if no typical CT image abnormality of viral pneumonia (#3 above) was seen and severe patients also met at least one of the following conditions: (1) shortness of breath, respiratory rate (RR) $\geq$30 times/min, (2) oxygen saturation (resting state) $\leq$93%, or (3) PaO_2/FiO_2 $\leq$300 mm Hg.

Data Collection

All medical records including epidemiological, demographic, clinical manifestation, laboratory data, radiological characteristics, treatment, and outcome data were reviewed and collected. Laboratory confirmation of SARS-CoV-2 infection was performed by real-time RT-PCR (Bojie Ltd, 119 Shanghai, China) according to Chinese CDC approval. Five sets of RNA-seq data from BALF of two COVID-19 patients were acquired from BIG Data Center (accession number CRA002390), and corresponding data of three healthy controls were from the NCBI SRA database (accession numbers SRR10571724, SRR10571730, and SRR10571732). Four RNA-seq data from lung specimens of two COVID-19 patients and two healthy controls were acquired from the GEO database (accession numbers GSM4462416, GSM4462415, GSM4462414, and GSM4462413).

Chest CT Protocols

All images were obtained on the CT system (Somatom Definition AS+, Siemens Healthineers, Germany) with patients in supine position. The main scanning parameters were as follows: tube voltage = 120 kV, automatic tube current modulation (about 95 mAs), pitch = 1.2 mm, slice thickness = 7 mm, field of view = 350 mm × 350 mm. All images were then reconstructed with a slice thickness of 0.6 mm with the same increment.

Image Analysis

Two professional radiologists (Y.M.Y. and X.M.L.), who were blinded to the laboratory test data, reported chest CT features and assessed the CT features by consensus. The lesion CT values were assessed using the Skyview pacs system. The region-of-interest was selected manually marking the area of highest intensity (most restricted area) of the lesion in CT images.

RNA-Seq Library Sequencing and Analysis

Kallisto was used to pseudoalign the RNA-seq reads and perform bootstrap analysis using an index based on the ENSEMBL GRCh38 *Homo sapiens* release 99 transcriptomes (20). Gene expression levels were then calculated as transcripts per million (TPM). Sleuth (version 0.30.0) (21) was used to perform differential gene expression (DEGs) analysis with the Wald test. Benjamini-Hochberg-adjusted false discovery rate ($q < 0.1$) was used to correct for multiple comparisons.

To compare lung and BALF samples of COVID-19 patients with healthy controls, differentially expressed genes were exhibited in a scaled heatmap using pheatmap (22). MCP-counter was used to characterize immune cell subpopulations (23). The MCP-counter scores obtained from the three underlying transcriptome platforms (Affymetrix Human Genome U133 Plus 2.0, Affymetrix 133A, and Illumina HiSeq) were used to estimate

the expression of each cell population. Functional enrichment analysis of the 29 upregulated marker genes of neutrophils was conducted with Metascape (http://metascape.org/) (24). Gene set enrichment analysis (GSEA) was performed in pre-ranked list mode with 1,000 permutations and weighted enrichment statistic (25). The gene interaction was analyzed by STRING (26). Gene interaction networks were visualized with eXamine (27).

Statistical Analyzes

Quantitative parameters are described as the median value followed by the inter-quartile range (IQR) in parentheses. Principal component analysis was performed with R package "FactoMineR" to identify those clinical parameters that contribute most to distinguishing severe, moderate, and mild cases of COVID-19 (28). Figures were produced with R package "ggplot2" (29). Logistic regression was conducted with R package "rstanarm" (30) to identify associations of laboratory parameters with severity of cases.

Severe cases were typed as severe and others (moderate and mild cases) as non-severe. The generalized linear model was then used to calculate coefficients (mean value with 5%, 95% confidence interval) of all parameters for severe. Finally, we used the function of exp [exp(x) = ex] for coefficients. The results were an odd's ratio (mean, 5–95% credible interval). Receiver operating characteristic curves (ROC) were calculated by R package "pROC." The area under the ROC curve (AUC) and cut-off values of selected parameters were used to distinguish mild and severe cases (31). Numerical Bayesian linear regression was carried out with Stan using Hamiltonian Monte Carlo (**Supplemental Materials**; **Supplementary Figure 1**) (32).

RESULTS

Characteristics of COVID-19 Patients

Fifty-five confirmed COVID-19 patients were hospitalized in The Fifth People's Hospital of Wuxi from Jan 23 to Mar 15, 2020. The median age of patients was 45 years (IQR 25–61), and 27 (49%) were male. Based on the previously described guidelines, 22 (40%), 25 (45%), and 8 (15%) of the 55 COVID-19 patients were classified as mild, moderate, and severe cases, respectively. There were five patients with diabetes (9%), 13 with hypertension (24%), eight with surgical history (15%), and two with co-infections (4%). The most common symptoms at onset were fever in 28 cases (51%), sputum production in 13 cases (24%), cough in 22 cases (40%), and fatigue in 17 cases (31%) (**Table 1**).

The clinical handling and relevant time-points of 33 patients including eight severe and 25 moderate cases are shown in **Figure 1**. The median time from the date of onset of symptoms to hospital admission, lymphopenia, ARDS, and neutrophilia was 3, 7, 8, and 9 d, respectively. Lymphopenia occurred in seven of eight severe patients and 11 of 25 moderate cases within 7 d, ARDS occurred in all eight severe patients within 8 d, and neutrophilia occurred in six of eight severe patients and one of 25 moderate cases within 9 d (**Figure 1**).

The laboratory test of each patient on the day of hospital admission showed that the median neutrophil count in severe COVID-19 patients (3.4, IQR: 1.8–6.7) was higher than in the moderate (3.0, 2.4–3.6) and mild (2.9, 2.3–3.5) groups. In contrast, lymphocyte and monocyte counts in severe COVID-19 patients were lower than in the other two groups (**Table 2**). By logistic regression, the following ORs for effects on having a severe case were obtained: neutrophil counts (1.5, 95% CI: 1.0–2.1), ratio of neutrophil to lymphocyte (NLR; 1.2, 95% CI: 1.1–1.4), C-reactive protein (CRP; log-scaled; 2.6, 95% CI: 1.6–4.7), Fibrinogen (FIB, 2.6, 95% CI: 1.5–4.9), and thrombin time (TT, 2.5, 95% CI: 1.4–5.0). These findings suggest that higher neutrophil counts, the NLR, and CRP, FIB, and TT levels as potential prognostic factors. The ORs of lymphocyte (0.28, 95% CI: 0.08–0.85) and monocyte (0.02, 95% CI: 0.00–1.16) counts suggest an association of lower lymphocyte and monocyte counts with severe pneumonia.

TABLE 1 | Demographic and clinical characteristics of 55 COVID-19 patients.

Variable	Value
Age (year)	45.0 (25.0–61.0)
Gender—no./(%)	
Male	27 (49.1)
Female	28 (50.9)
Clinical diagnosis—no./(%)	
Severe	8
Moderate	25
Mild	22
Initial symptoms—no./(%)	
Fever (>38°C)	28 (50.9)
Sputum production	13 (23.6)
Headache	5 (9.1)
Chill	7 (12.7)
Shivering	2 (3.6)
Nausea or vomiting	1 (1.8)
Diarrhea	13 (23.6)
Fatigue	17 (30.9)
Cough	22 (40)
Pharyngalgia	2 (3.6)
Rhinorrhea	6 (10.9)
Chest pain	1 (1.8)
Shortness of breath	5 (9.1)
Chest tightness	9 (16.4)
Chronic disease—no./(%)	
Diabetes	5 (9.1)
Hypertension	13 (23.6)
Thyroid disease	2 (3.6)
Malignant tumor	2 (3.6)
Gastritis	2 (3.6)
Coronary artery disease	1 (1.8)
Surgical history	8 (14.5)
Co-infection—no./(%)	
Initial	0
Progressive	2 (3.6)

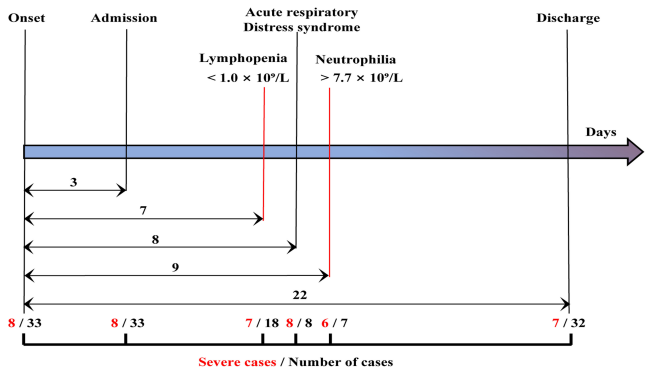


FIGURE 1 | Clinical courses of the study patients. The time lines showed the days of hospital admission, lymphopenia, acute respiratory distress syndrome, neutrophilia, and discharge from symptom onset for each case. The median time from onset of symptoms to hospital admission and discharge was 3, 22 days, respectively. Among the 33 patients (black), eight cases were severe cases (red). A total of 18 cases exhibited lymphopenia within 7 days, including seven severe patients. All eight patients presented with acute respiratory distress syndrome within 8 days. Seven cases presented with neutrophilia, including six severe cases within 9 days.

Principal Component Analysis and Dynamic Monitoring of Laboratory Parameters

Principal component analysis was performed to visualize the contribution of all mentioned clinical parameters on disease severity (**Figure 2A**). Nine variables contributed most strongly. Among them, higher CRP, FIB, neutrophil count, and NLR, and lower lymphocyte count were associated with increased disease severity. These parameters may therefore be used for prognosis. To assess the diagnostic value of the top two contributors, CRP and lymphocytes, the AUC and cut-off values from the ROC curves were calculated for the severe and mild cases, respectively (**Supplementary Figure 2B**). The cut-off values for severe patients were CRP (26.1) and lymphocytes (1.0), and for mild patients the values were CRP (2.2) and lymphocytes (1.4) (see dashed lines in **Figure 2B**).

Next, dynamic changes of neutrophil, lymphocyte, and monocyte counts in the peripheral blood of COVID-19 patients were monitored (**Figure 2C**). Dramatically increased neutrophil counts were found in severe COVID-19 patients in comparison to the other two groups. In contrast, lymphocyte counts persisted at lower values in severe COVID-19 patients. Monocyte counts were lower in severe cases, although the monocyte count fluctuated over a wide range. Timing of the occurrence of maximum neutrophil, minimum lymphocyte, and minimum monocyte counts, and the corresponding counts in COVID-19 patients, during hospitalization are shown in **Figure 2D**. From day 7 to day 9 after symptom onset, neutrophil counts erupted ($>7.7 \times 10^9$/L) and peaked in six of eight severe COVID-19 patients. In contrast, only one moderate (1/26) COVID-19 patient was found with neutrophilia. Lymphopenia occurred in seven of eight severe patients but only in four mild (4/22) COVID-19 patients. Monopenia ($<1 \times 10^8$/L) was found in three moderate (3/25) and four severe (4/8) COVID-19 patients. Overall, monitoring blood cell parameters revealed neutrophilia as a characteristic of severe COVID-19 patients.

Bayesian Linear Regression of CT Values and Changing Neutrophil and Lymphocyte Counts

Neutrophilia and lymphopenia obviously occurred in severe COVID-19 patients during hospitalization. Here was a case of severe patient. The CRP level remained low when neutrophilia occurred, and the D-dimer levels increased after neutrophilia. Series of chest CT images exhibited enlarged patches and ground-glass nodules in the sub-pleura area of both lungs during neutrophilia. Interestingly, all observed lesions were reduced or gradually absorbed along with the return of neutrophils to normal levels after neutrophilia (**Figures 3A,B**). The CT value of lesions, reflecting lung lesions, was further demonstrated to have the same trend with neutrophils but the opposite trend with lymphocytes (**Figure 3C**).

To estimate the overall correlation of CT value with neutrophil and lymphocyte counts across patients with a visual inspection of possible trends, linear models were fitted to summarize the dependency of z-values of CT value (CTz, see **Supplementary Information**) of neutrophil and lymphocyte counts. Thus, Bayesian linear regression was used to quantify the observed trends of CTz values as a function of parameters mentioned above. For log-transformed neutrophil counts, a slope for the moderate cases of 0.3 [−0.3, 0.9] (0.05 and 0.95 quantiles in square brackets) was obtained, i.e., with a slope that could be flat. For the severe cases, the mean slope was 0.8 [0.3, 1.2], i.e., clearly positive. Thus, no clear trend for moderate cases was visible, whereas an increase in CTz value with neutrophil counts was significantly correlated for severe cases. For CTz as a function of lymphocyte counts, the slope was −0.1 [−0.4, 0.6] for moderate cases and −0.3 [−0.5, 0.0] for the severe cases, supporting the trends in **Figure 3D**. Overall, the results showed that the CTz value has no average trend with changing neutrophil and lymphocyte counts for moderate cases (green). However, for the severe cases (red), there are clear trends for CTz value with changing cell counts; specifically, CTz value increased for increasing neutrophil counts, whereas CTz value decreased for increasing lymphocyte counts (**Figure 3D**).

Immune Cell Transcriptional Signatures of the Lung and BALF in COVID-19 Patients

Immune cell transcriptional signatures were established from RNA-seq data of BALF and lung specimens of COVID-19 patients and healthy controls. Marker genes of neutrophils, T cells, monocytes, and B cells were identified from Microenvironment Cell Populations-counter (MCP-counter). Their representation in the RNA-seq data were exhibited using a scaled heatmap by comparing both lung and BALF samples of COVID-19 patients to healthy controls (**Figure 4A**).

The results revealed that 112 marker genes represented four immune populations: neutrophils (46 genes), T cells (13 genes), monocytes (10 genes), and B cells (43 genes). For lung

TABLE 2 | Laboratory parameters of mild, moderate, and severe COVID-19 cases.

Baseline variables	Reference range	Severe cases ($n = 8$)	Moderate cases ($n = 25$)	Mild cases ($n = 22$)	*Odds ratio for severe (95% CI)
Age (year)		59 (50–73)	45 (30–60)	39.5 (22.3–52)	1.07 (1.02–1.12)
Female (%)		3 (37.5)	11 (44)	14 (64)	1.90 (0.57–7.34)
Blood routine					
White blood cell ($\times 10^9$/L)	3.5–9.5	5.4 (3.4–7.6)	4.8 (4.1–5.7)	5.3 (4.7–6.8)	1.23 (0.86–1.78)
Neutrophil ($\times 10^9$/L)	1.8–6.3	3.4 (1.8–6.7)	3.0 (2.4–3.6)	2.9 (2.3–3.4)	1.47 (1.05–2.14)
Lymphocyte ($\times 10^9$/L)	1.1–3.2	1.0 (0.7–1.6)	1.3 (0.9–1.5)	1.9 (1.1–2.8)	0.28 (0.08–0.85)
Monocyte ($\times 10^9$/L)	0.1–0.6	0.4 (0.2–0.6)	0.5 (0.4–0.6)	0.5 (0.4–0.6)	0.02 (0.00–1.16)
Platelet ($\times 10^9$/L)	125.0–350.0	154.0 (121.0–182.8)	191.0 (156.5–213.5)	194.5 (163.8–214.5)	0.98 (0.97–1.01)
PDW (CV %)	15.5–18.1	15.4 (11.4–16.6)	14.2 (13.7–15.9)	12.8 (11.1–14.0)	1.12 (0.83–1.50)
Red blood cell ($\times 10^{12}$/L)	4.30–5.80	4.3 (4.0–4.9)	4.8 (4.1–5.0)	4.4 (4.0–4.7)	0.48 (0.15–1.50)
RDW (CV %)	11.5–14.9	12.9 (12.1–13.9)	12.4 (11.7–13.6)	11.9 (11.6–12.3)	1.91 (1.19–3.41)
Ratio of neutrophils to lymphocytes		2.4 (1.4–16.2)	2.3 (1.7–2.9)	1.8 (0.9–2.8)	1.21 (1.06–1.42)
Ratio of monocytes to lymphocytes		0.3 (0.3–0.8)	0.4 (0.2–0.5)	0.3 (0.2–0.4)	2.86 (0.28–27.0)
C-reactive protein (mg/L)	0.0–10.0	41.1 (13.8–139.9)	6.2 (1.1–12.7)	2.1 (0.5–17.7)	2.64 (1.64–4.65)
Biochemical indicators					
ALT (U/L)	4.0–44.0	17.0 (14.0–60.0)	19.0 (16.0–35.3)	26.0 (14.0–43.3)	1.02 (0.99–1.04)
AST (U/L)	8.0–38.0	28.0 (23.0–49.0)	23.5 (20.8–31.3)	24.5 (19.0–31.0)	1.04 (0.99–1.09)
Total bilirubin (μmol/L)	2.0–21.0	7.0 (3.0–12.0)	5.0 (2.8–9.0)	6.5 (4.8–10.3)	1.01 (0.87–1.16)
Direct bilirubin (μmol/L)	2.0–7.0	0.1 (0.1–1.0)	0.1 (0.1–1.8)	0.1 (0–1.4)	0.99 (0.67–1.37)
Serum total protein (g/L)	67.0–83.0	68.0 (65.0–75.0)	69.5 (65.0–73.3)	69.0 (65.0–71.3)	0.98 (0.87–1.16)
Serum albumin (g/L)	35.0–50.0	39.0 (34.0–43.0)	43.5 (38.8–47.3)	41.5 (38.8–45.0)	0.85 (0.73–1.00)
Creatine kinase (U/L)	0.0–171.0	101.0 (54.0–151.0)	69.0 (53.8–106.8)	68.0 (46.8–102.0)	1.000 (0.99–1.01)
Creatine kinase MB (U/L)	0.0–12.0	11.0 (10.0–13.0)	10.0 (9.0–12.5)	10.0 (7.8–14.8)	0.97 (0.79–1.17)
Blood urea nitrogen (mmol/L)	3.1–8.0	5.9 (3.3–10.1)	4.2 (3.5–4.9)	4.0 (3.0–4.6)	1.60 (1.18–2.31)
Serum creatinine (μmol/L)	53.0–97.0	64.0 (38.0–88.0)	54.5 (43.5–64.5)	48.5 (39.3–58.5)	1.04 (1.01–1.08)
Serum potassium (mmol/L)	3.8–5.0	3.8 (3.2–4.2)	4.1 (3.8–4.2)	4.0 (3.9–5.0)	0.24 (0.06–0.79)
Serum sodium (mmol/L)	136.0–149.0	140.0 (13.9.0–141.0)	142.0 (141.0–143.0)	142.0 (140.0–143.0)	0.59 (0.38–0.89)
Serum chlorine (mmol/L)	98.0–106.0	105.0 (103.0–106.0)	104.0 (102.8–106.0)	105.0 (103.0–106.0)	0.72 (0.95–1.25)

(Continued)

TABLE 2 | Continued

Baseline variables	Reference range	Severe cases ($n = 8$)	Moderate cases ($n = 25$)	Mild cases ($n = 22$)	*Odds ratio for severe (95% CI)
Blood coagulation function					
D-dimer (mg/L)	0.0–0.5	0.6 (0.3–1.2)	0.3 (0.2–0.6)	0.3 (0.2–0.5)	1.314 (0.579–2.986)
PT (s)	11.5–15.5	13.2 (12.2–13.4)	13.2 (12.9–13.6)	13.2 (13.2–13.5)	0.22 (0.05–0.90)
APTT (s)	26.0–40.0	37.5 (35.5–42.3)	38.2 (36.3–42.9)	41.3 (37.6–44.3)	0.93 (0.79–1.09)
Fibrinogen (g/L)	2.0–4.0	4.9 (4.4–5.9)	3.6 (2.9–4.8)	3.6 (2.7–4.1)	2.61 (1.52–4.87)
TT (s)	14.0–21.0	17.1 (16.2–18.2)	16.2 (15.8–16.8)	16.2 (15.9–17.4)	2.46 (1.35–4.97)
Blood gas analysis					
$PaCO_2$ (mm Hg)	35.0–48.0	42.5 (39.3–44.0)	43.0 (40.5–47.0)	42.0 (40.3–45.0)	0.92 (0.76–1.09)
PaO_2 (mm Hg)	83.0–108.0	83.0 (64.5–00.5)	106.0 (93.5–134.0)	103.5 (93.3–124.3)	0.95 (0.91–0.98)
PaO_2/FiO_2 (mm Hg)	400.0–500.0	395.2 (300.0–478.6)	504.8 (445.2–632.6)	461.9 (395.6–591.7)	0.99 (0.98–1.00)
Lactic acid (mmol/L)	0.5–2.2	1.9 (1.3–3.4)	1.6 (1.3–1.9)	1.7 (1.1–2.3)	2.44 (1.07–5.93)

The Odd Ratio of log normalization.

tissue, the most up-regulated marker genes were enriched in neutrophils, second in monocytes, and only a small proportion were enriched in B cells. Marker genes of T cells were almost all lowly expressed. For BALF, the most upregulated marker genes were similarly enriched in neutrophils, but more up-regulated genes in monocytes and B cells were observed in COVID-19 patients compared to healthy controls, which is different from the lung samples.

Functional enrichment analysis of the 27 upregulated marker genes of neutrophils were further conducted with Metascape. The enrichment analysis revealed that five gene sets with lowest *q*-value were related to neutrophil degranulation and activation (**Figure 4B**) and there were 15 marker genes involved. Then, we calculated the average expression of these genes as an evaluating score for neutrophil activation (NAS).

To further assess the abundance of infiltrating immune cells of the lung and BALF in COVID-19 patients, the MCP-counter score was used to quantify the absolute abundance of immune cell subpopulations. Notably, the neutrophil scores were higher and T cell scores were lower in lung samples of COVID-19 patients. The higher abundance of cytotoxic T lymphocytes contributed for cell injury, not for anti-virus. Due to the marker genes for cytotoxic T lymphocytes was *KLRC1* (Killer Cell Lectin Like Receptor C1). For the BALF samples, the score of neutrophils, cytotoxic lymphocytes, B cells, monocytes, and dendritic cells were found to be higher in one of the COVID-19 patients compared to the three healthy controls (**Figure 4C**).

Neutrophil Activation Related Genes Enrichment Analysis

To explore the outcome of neutrophil activation in COVID-19, we further analyzed the correlation of NAS with 1,363 DEGs that overlapped in both the lung and BALF samples. The spearman correlation was used separately for COVID-19 patients and healthy controls. Then, the R value for every single gene was acquired for COVID-19 patients (R1) and healthy cases (R2). All DEGs were ranked based on ΔR (R1-R2). The "R value" of the top 84 genes (R1 > 0) in the two groups are displayed in **Figure 5A**. Of these 84 genes, 16 genes were NETs associated genes (**Figure 5B**; **Table 3**) (33–46) Of the 16 genes, *LGALS9, HCK, LCP1, CEACAM1* were involved in the cytokine-mediated signaling pathway. *S100A8, LGALS9,* and *CTSC* were involved in regulation of apoptotic signal by enrichment annotation from the Metascape tool (**Figure 5B**; **Table 3**) (33–46).

To further investigate the role of NETs in COVID-19, we generated a gene set termed "NET-associated genes" based on genes coding for proteins enriched in NETs released from human neutrophils with mass spectrometry (**Supplementary Table 1**). Pre-ranked GSEA by ΔR resulted in significant enriched gene sets of "NET-associated genes" (Enrichment Score = 0.80) and "Regulation of inflammatory response" (Enrichment Score = 0.72) (**Figure 5C**).

NETs Associated Genes From RNA-Seq Data in COVID-19 Patients

As known, the formation of NETs could induce direct lung injury (17). There were 16 NETs associated genes related with neutrophils activation in COVID-19 patients. To further illustrate the interaction between these NETs associated genes with other neutrophils activation related genes, we constructed a protein-to-protein interaction network from the STRING database (**Figure 6**). We found that the NETs interacted with *STAT1* induced Interferon stimulated genes by *IL2RG*, implying that NETs associated genes may be triggered by IFN signaling.

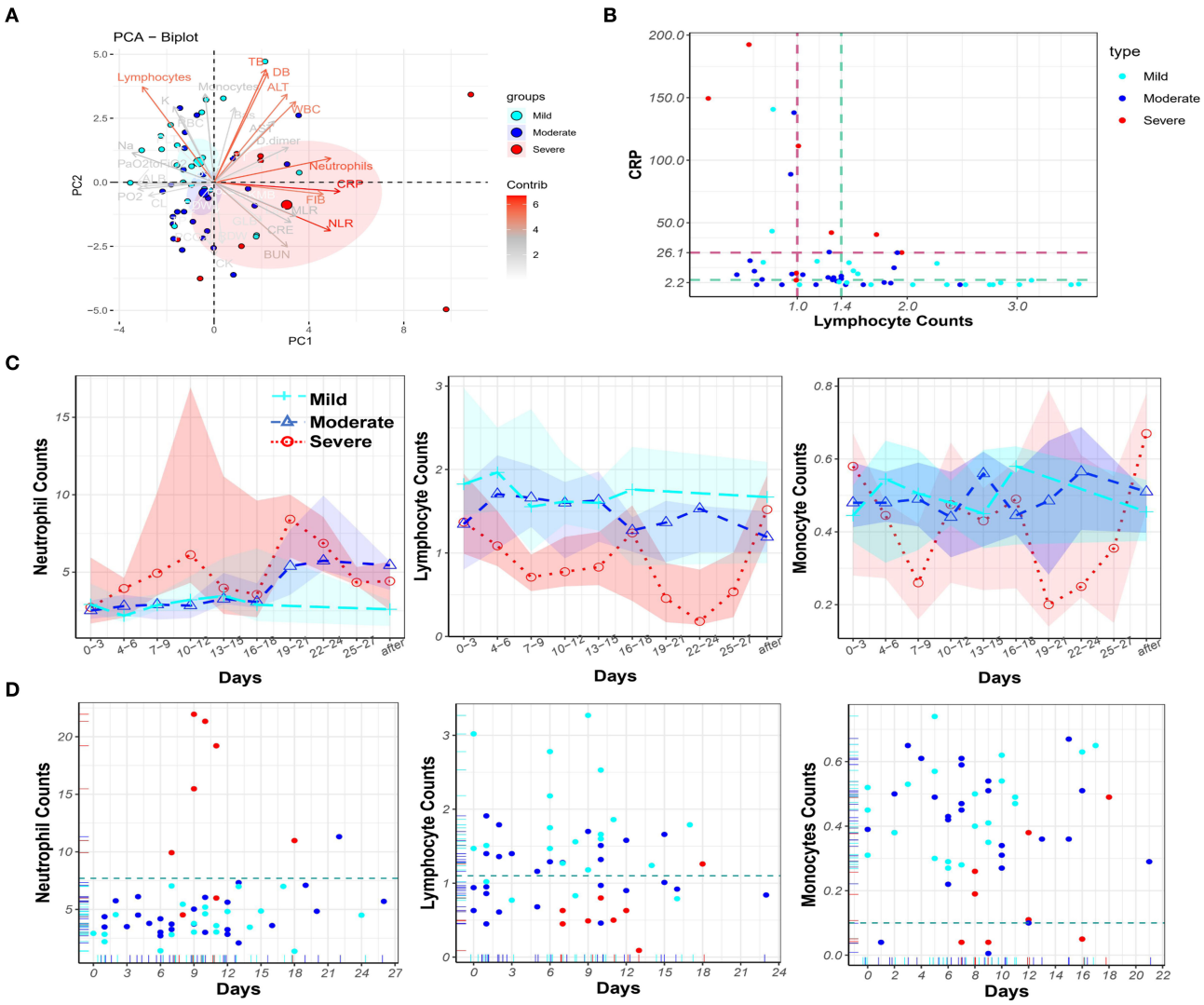


FIGURE 2 | Principal component analysis of laboratory parameters and dynamic monitoring of blood cells in the peripheral blood of COVID-19 patients. **(A)** Principal component analysis to identify variables for distinguishing the disease severity of COVID-19 patients. The nine variables that contributed mostly to distinguishing the disease severity were white blood cell counts (WBC), neutrophil counts, neutrophil-to-lymphocyte ratio (NLR), C-reactive protein (CRP), FIB, ALT, Total Bilirubin (TB), Direct Bilirubin (DB), and lymphocyte counts. **(B)** CRP levels and lymphocyte counts in 55 cases with the cut-off values CRP = 26.1 and lymphocyte = 1.0 for eight severe cases (brown) and the cut-off values CRP = 4.3 and lymphocyte = 1.4 for 22 mild cases (green). **(C)** The dynamic change of neutrophil, lymphocyte and monocyte counts over time in COVID-19 patients in mild (cyan), moderate (blue), and severe (red) groups, circled dots: mean value; colored background area: IQR (interval quartile range). The **(D)** Time points of maximum neutrophil, minimum lymphocyte, and minimum monocyte counts, and the corresponding counts in mild (cyan), moderate (blue), and severe (red) COVID-19 patients during hospitalization.

Besides, NETs in turn may activate B cells via *TNFSF13B* and inhibit the function of T and NK cells via *LGAS9* and *CEACAM1*, which are negative regulators for T and NK cells. *LGAS9* is a possible promoter of protein-arginine deiminase type 4 (PAD4). *PAD4*, a key NETs associated gene, lies downstream of ROS and promotes chromatin decondensation (47, 48). Of note, we also observed ROS related genes including *HCK, RAC2*, and *NCF2* among NETs associated genes (**Figure 6**).

To annotate the function of NETs associated genes, they were categorized as metabolic enzymes (*RAC2, NCF2*), structural proteins (*LCP1*), anti-microbial related (*TREM1*), peroxisomal (*SH3BGRL3*), and others (*C1QC, LGALS9, SERPINA1, C1QB, CCL7, CCL8, CEACAM1, HCK*, and *CXCL16*) (**Table 3**). Thus, we speculate that NETs may be activated by innate immunity such as IFN signaling, in COVID-19 patients. NETs may negatively regulate the immune function of T cells and NK cells via *LGAS9* and *CEACAM1*, respectively, leading to insufficient anti-viral immunity and injuring the lung tissue directly.

DISCUSSION

In this study, a set of laboratory test parameters and the corresponding chest CT images of 55 COVID-19 patients were collected during hospitalization. Among these variables,

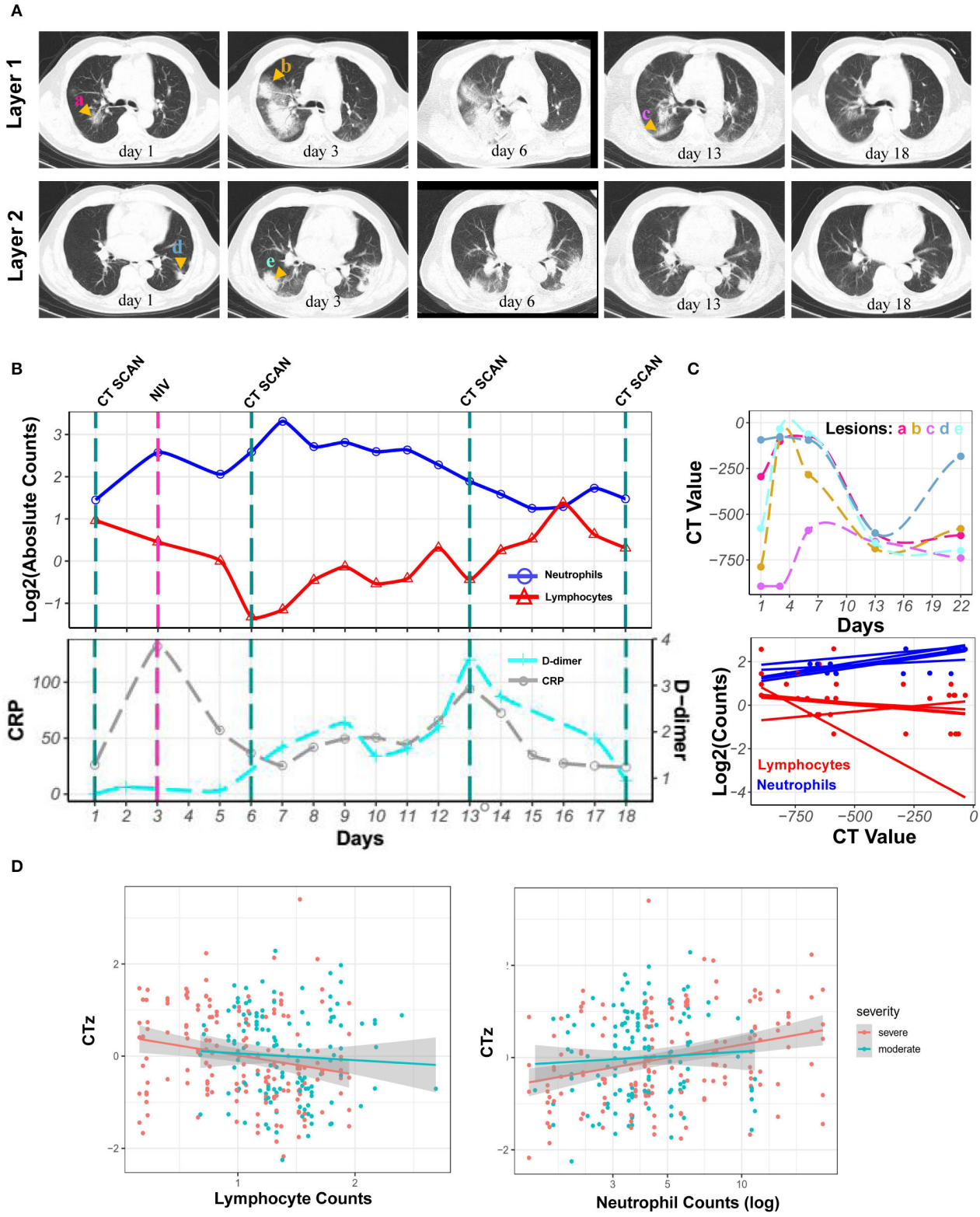
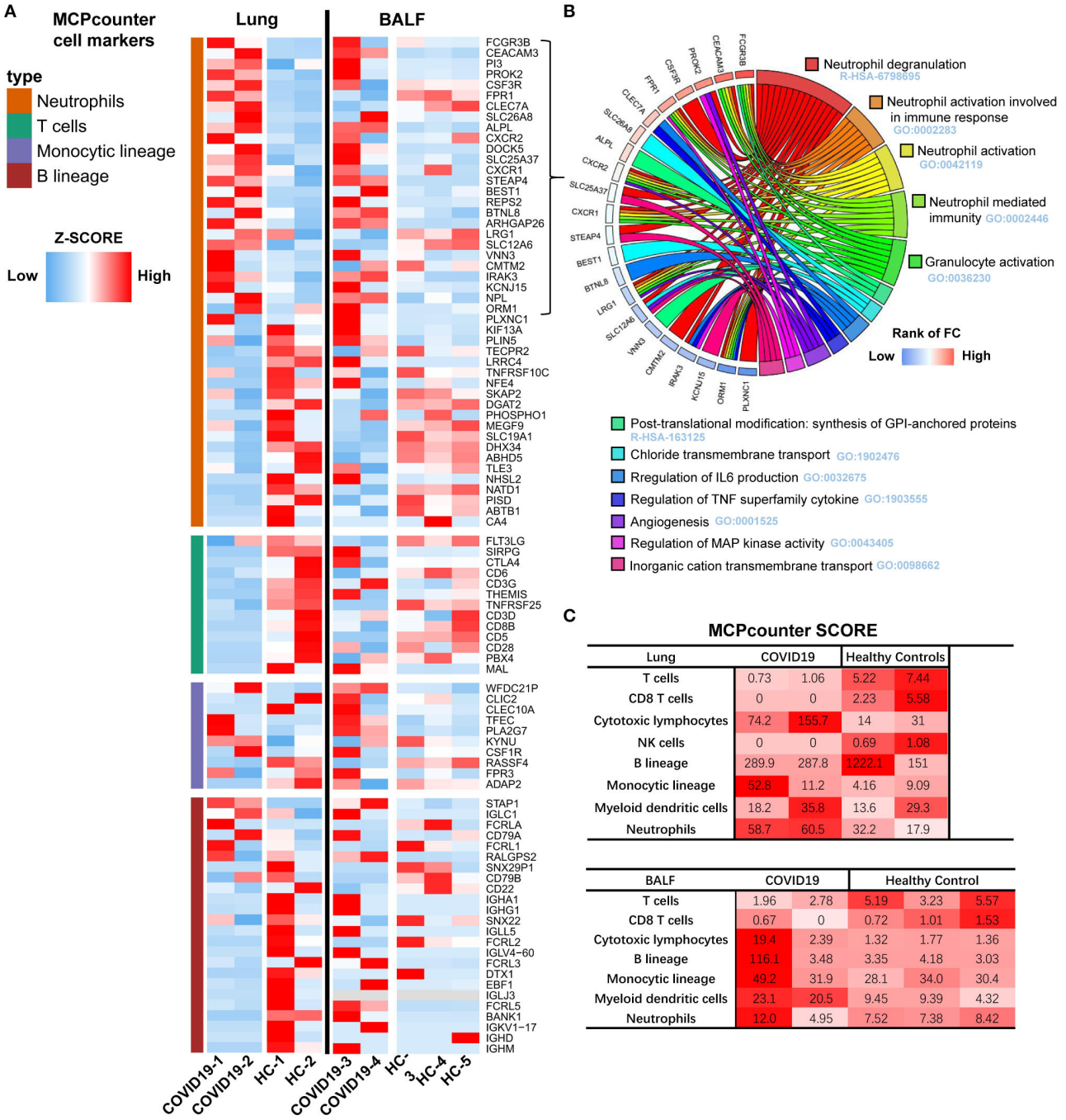


FIGURE 3 | Kinetics of laboratory parameters and serial chest CT images of severe COVID-19 patient with the development of neutrophilia. **(A)** Normal chest CT with axial planes at indicated time point. **(B)** The dynamics of neutrophil counts (blue line), lymphocyte counts (red line) with $\log_2$ scaling, and CRP (gray line) and D-dimer (cyan line) levels at indicated time point. **(C)** CT value of lesions and its correlation with $\log_2$ scaled neutrophil and lymphocyte counts at indicated time point. **(D)** Least-square fits of linear models to summarize the z-values of CT values as a function of log-transformed neutrophil counts for 23 patients. Points are pairs of CTz values (*z*-values of individual CT measurements) and log-neutrophil counts, colored according to severity of COVID-19. Colored lines are the corresponding least-square fits to the data form each severity group. Gray areas are 95% confidence intervals.



Lung	COVID19		Healthy Controls	
T cells	0.73	1.06	5.22	7.44
CD8 T cells	0	0	2.23	5.58
Cytotoxic lymphocytes	74.2	155.7	14	31
NK cells	0	0	0.69	1.08
B lineage	289.9	287.8	1222.1	151
Monocytic lineage	52.8	11.2	4.16	9.09
Myeloid dendritic cells	18.2	35.8	13.6	29.3
Neutrophils	58.7	60.5	32.2	17.9

BALF	COVID19		Healthy Control		
T cells	1.96	2.78	5.19	3.23	5.57
CD8 T cells	0.67	0	0.72	1.01	1.53
Cytotoxic lymphocytes	19.4	2.39	1.32	1.77	1.36
B lineage	116.1	3.48	3.35	4.18	3.03
Monocytic lineage	49.2	31.9	28.1	34.0	30.4
Myeloid dendritic cells	23.1	20.5	9.45	9.39	4.32
Neutrophils	12.0	4.95	7.52	7.38	8.42

FIGURE 4 | Transcriptome analysis of the lung and BALF in COVID-19 patient. **(A)** Marker genes from Microenvironment Cell Populations-counter (MCP-counter) were used to identify of Neutrophils, T cells, Monocytes, and B cells in both Lung and BALF samples of COVID-19 patients and healthy controls, respectively. The RNA-seq data TPM are shown in a scaled heatmap. **(B)** Circle plots for functional enrichment analysis of 29 marker genes of Neutrophils. **(C)** The absolute abundance of immune cell subpopulations as scores in COVID-19 patients and healthy cases.

excessive neutrophils were associated with disease severity, as shown by principal component analysis. Bayesian inference across patients quantified that the increased trend of pneumonia lung injury, as represented by CT values, was in accord with the increased trend in neutrophil counts. Transcriptome analysis of lung specimens and BALF from COVID-19 patients also indicated the most up-regulated marker genes were neutrophil related. Importantly, many neutrophil activation genes were categorized as NET-associated genes. These genes were further assessed to interact with T and NK cells via negative regulatory

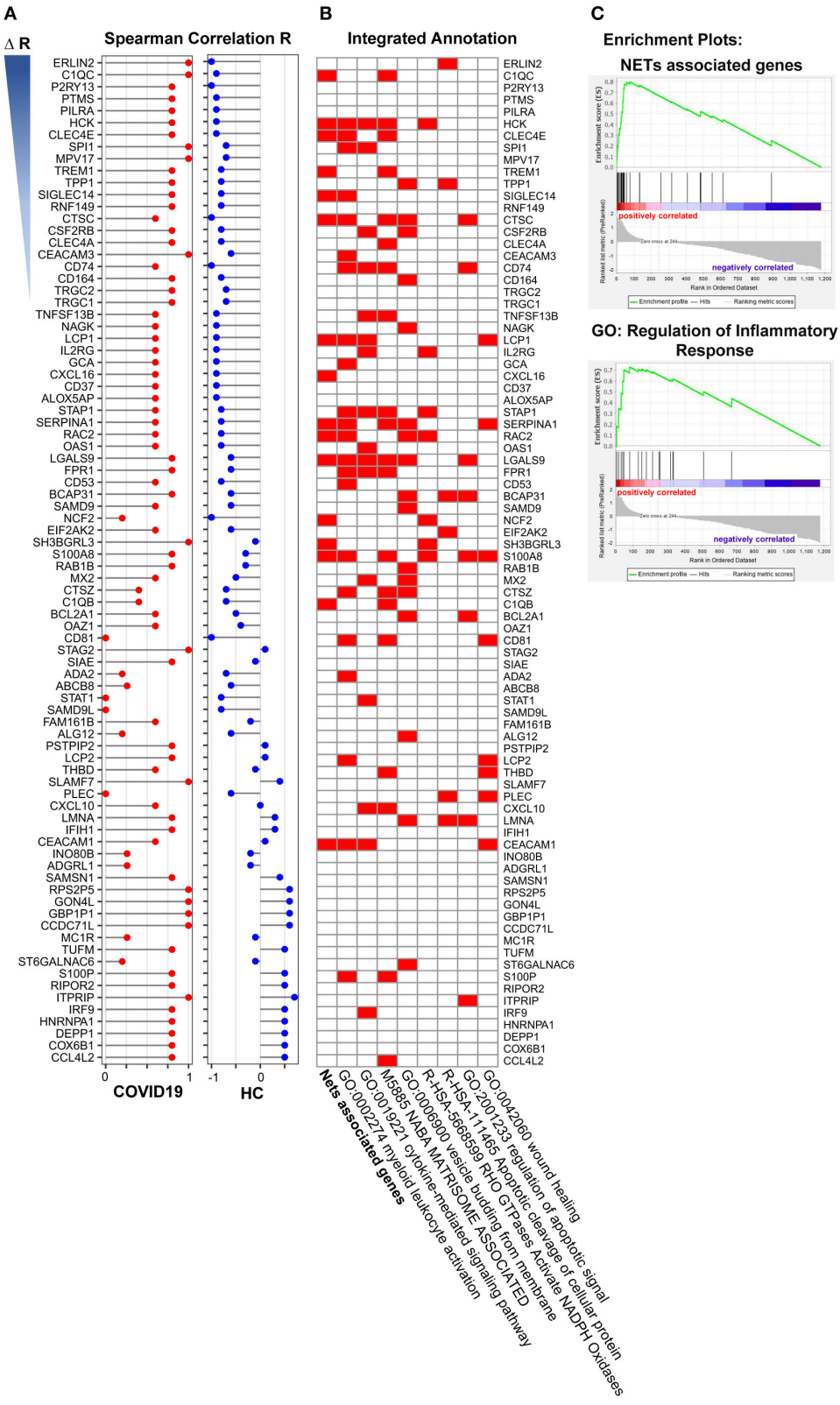


FIGURE 5 | Gene enrichment analysis of neutrophil activation related genes. **(A)** The 15 annotated genes of neutrophils activation were calculated the average expression of every single samples as neutrophils activation score, and the correlation of the score with overlapped 1,363 differently expressed genes both in COVID-19 and Healthy control were analyzed. The selected 84 genes were ranked based on ΔR (R1-R2, R1 from COVID-19 patients, R2 from healthy control). **(B)** Functional enrichment analysis of these 84 genes, of which 16 genes were NETs associated genes. **(C)** Nets associated genes set (Enrichment Score, 0.80) and the GO term of regulation of inflammatory (Enrichment Score, 0.72) by GSEA with DEGs from pre-ranked by ΔR.

molecules in COVID-19 patients leading to insufficient anti-viral response and lung injury (**Figure 6**).

Our previous study also found an increased neutrophil-to-lymphocyte ratio in the most severe disease cases (9). Recently, neutrophil infiltration was also noted in the lung tissue of autopsied COVID-19 patients (5–7). Since neutrophilia predicts poor outcomes in patients with COVID-19 (8), we propose that the change in neutrophil counts in peripheral blood or tissues may be closely associated with pathological injury in COVID-19 patients. We demonstrated here that the dynamics of neutrophil counts in COVID-19 patients during hospitalization exhibited the same trend as the corresponding lung injury.

TABLE 3 | Annotation of Nets associated genes.

Function	Gene name	References
Metabolic enzymes	RAC2; NCF2	(33, 34)
Structural proteins	LCP1	(35)
An-microbial related proteins	TREM1; S100A8; C1QB; C1QC	(35–37)
Peroxisomal enzyme	SH3BGRL3	(38)
Not classified	LGALS9; SERPINA1; CEACAM1; HCK; CXCL16; CLEC4E; CTSC; SIGLEC14	(39–46)

NETs, as confirmed contributors to pathological inflammation of pneumonia, can damage tissues by killing epithelial and endothelial cells (16, 17) of pulmonary tissue in infection and sterile disease. Recently, two elevated NETs markers have been observed in serum from COVID-19 patients, which suggests that neutrophilia and excessive NETs may contribute to cytokine release and respiratory failure in COVID19 patients (10). However, evidence is still lacking regarding NETosis in lungs. We analyzed the differentially expressed genes in lung tissue and BALF samples from COVID-19 patient in comparison to healthy controls. Among all up-regulated genes in neutrophil modules in COVID-19 patients, we found 17 genes derived from the neutrophil activation pathway were NETs associated genes. Thus, NETs may be activated in the lung of COVID-19 patients. It is also poorly understood how NETosis induces the cytokine storm or modulates the host immune response. Our STRING analysis suggests that NETs associated genes could interact with T, NK, and B cells through regulation of *LGALS9, CEACAM1*, and *TNFSF13B* expressions, respectively. We suspect that the progression of lesions in COVID-19 patients may be induced by NETs as well as NETs-T/NK/B cell interactions.

In conclusion, the clear trend of lung injury in accord with the trend of increasing neutrophils was quantified by Bayesian inference analysis in COVID-19 patients. The transcriptome signature of immune cells also indicated elevated neutrophil markers in the lung and BALF samples of COVID-19 patients.

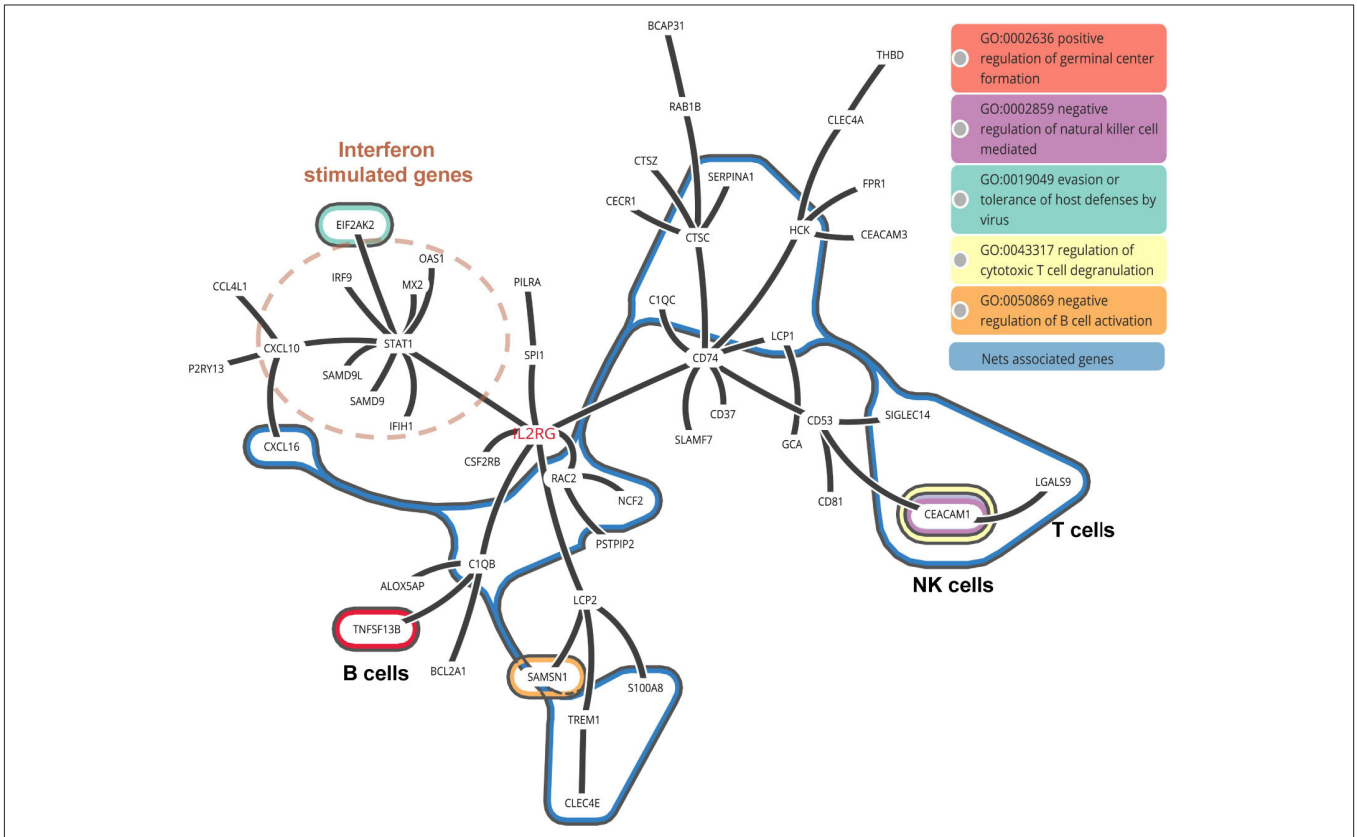


FIGURE 6 | PPI network of NETs associated genes in COVID-19 patients. The interaction between NETs associated genes with other neutrophil activated genes.

Importantly, among the excessive neutrophil activated genes, 17 were NETs associated genes and these genes interacted with T cells and NK cells through negative regulation. Therefore, we posit that NETosis in lung tissue leads to an insufficient anti-viral response in COVID-19 patients. We hope that future studies will investigate the predictive power of circulating NETs in well-phenotyped longitudinal cohorts.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the Fifth People's Hospital, Wuxi (No. 2020-006-1). The patients provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JW, YQ, and QL conceived and designed the experiments. QL, JW, DH, and ML drafted and revised the manuscript. YY, YZ, and XL carried out the data collection. JW, DH, and YC carried out the data analysis and interpretation. DH, YQ, ML, and LH contributed reagents, materials, and analysis tools. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the foundation of Wuxi Medical Development Discipline for Infectious Disease (FZXK006) and Wuxi Young Medical Talents (QNRC072), Health and Science Bureau of Wuxi (MS201731, CSE31N1712, Q201743). The funding source was not involved in the study design; in the collection, analysis, and interpretation of data, in the writing of the report, and in the decision to submit the paper for publication. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

ACKNOWLEDGMENTS

We are grateful to the doctors, nurses, disease control workers, and researchers for their fight against COVID-19 under extreme conditions. Some of them have lost their lives in this fight.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2020.02063/full#supplementary-material

REFERENCES

1. WHO. *Coronavirus Disease (COVID-2019) Situation Reports*. Available online at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (accessed March 30, 2020).
2. WHO. *Clinical Management of Severe Acute Respiratory Infection When Novel Coronavirus (nCoV) Infection Is Suspected*. Geneva: World Health Organization (2020). Available online at: https://apps.who.int/iris/handle/10665/330893 (accessed January 28, 2020).
3. Wang D, Hu B, Hu C. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *JAMA*. (2020) 323: 1061–9. doi: 10.1001/jama.2020.1585
4. Mehta P, McAuley DF, Brown M, Sanchez E, Tattersall RS, Manson JJ, et al. COVID-19: consider cytokine storm syndromes and immunosuppression. *Lancet*. (2020) 395:1033–4. doi: 10.1016/S0140-6736(20)30628-0
5. Fox SE, Akmatbekov A, Harbert JL, Li G, Brown JQ, Vander Heide RS. Pulmonary and cardiac pathology in Covid-19: the first autopsy series from New Orleans. *Lancet Respir Med*. (2020) 8:681–6. doi: 10.1016/S2213-2600(20)30243-5
6. Yao XH, Li TY, He ZC, Ping YF, Liu HW, Yu SC, et al. A pathological report of three COVID-19 cases by minimally invasive autopsies. *Zhonghua Bing Li Xue Za Zhi*. (2020) 49:411–7. doi: 10.3760/cma.j.cn112151-20200312-00193
7. Barnes BJ, Adrover JM, Baxter-Stoltzfus A, Borczuk A, Cools-Lartigue J, Crawford JM, et al. Targeting potential drivers of COVID-19: neutrophil extracellular traps. *J Exp Med*. (2020) 217:e20200652. doi: 10.1084/jem.20200652
8. Du RH, Liang LR, Yang CQ, Wang W, Cao TZ, Li M, et al. Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: a prospective cohort study. *Eur Respir J*. (2020) 55:e2000524. doi: 10.1183/13993003.00524-2020
9. Liu J, Li S, Liu J, Liang B, Wang X, Wang H, et al. Longitudinal characteristics of lymphocyte responses and cytokine profiles in the peripheral blood of SARS-CoV-2 infected patients. *EBioMedicine*. (2020) 55:102763. doi: 10.1016/j.ebiom.2020.102763
10. Zuo Y, Yalavarthi S, Shi H, Gockman K, Zuo M, Madison JA, et al. Neutrophil extracellular traps (NETs) as markers of disease severity in COVID-19. *JCI Insight*. (2020) 5:e138999. doi: 10.1101/2020.04.09.20059626
11. Liu S, Su X, Pan P, Zhang L, Hu Y, Tan H, et al. Neutrophil extracellular traps are indirectly triggered by lipopolysaccharide and contribute to acute lung injury. *Sci Rep*. (2016) 6:37252. doi: 10.1038/srep37252
12. Cheng OZ, Palaniyar N. NET balancing: a problem in inflammatory lung diseases. *Front Immunol*. (2013) 4:1. doi: 10.3389/fimmu.2013.00001
13. Papayannopoulos V. Neutrophil extracellular traps in immunity and disease. *Nat Rev Immunol*. (2018) 18:134–47. doi: 10.1038/nri.2017.105
14. Mikacenic C, Moore R, Dmyterko V, West TE, Altemeier WA, Liles WC, et al. Neutrophil extracellular traps (NETs) are increased in the alveolar spaces of patients with ventilator-associated pneumonia. *Crit Care*. (2018) 22:358. doi: 10.1186/s13054-018-2290-8
15. Maruchi Y, Tsuda M, Mori H, Takenaka N, Gocho T, Huq MA, et al. Plasma myeloperoxidase-conjugated DNA level predicts outcomes and organ dysfunction in patients with septic shock. *Crit Care*. (2018) 22:176. doi: 10.1186/s13054-018-2109-7
16. Saffarzadeh M, Juenemann C, Queisser MA, Lochnit G, Barreto G, Galuska SP, et al. Neutrophil extracellular traps directly induce epithelial and endothelial cell death: a predominant role of

histones. *PLoS ONE*. (2012) 7:e32366. doi: 10.1371/journal.pone.0032366
17. Villanueva E, Yalavarthi S, Berthier CC, Hodgin JB, Khandpur R, Lin AM, et al. Netting neutrophils induce endothelial damage, infiltrate tissues, and expose immunostimulatory molecules in systemic lupus erythematosus. *J Immunol*. (2011) 187:538–52. doi: 10.4049/jimmunol.1100450
18. Branzk N, Lubojemska A, Hardison SE, Wang Q, Gutierrez MG, Brown GD, et al. Neutrophils sense microbe size and selectively release neutrophil extracellular traps in response to large pathogens. *Nat Immunol*. (2014) 15:1017–25. doi: 10.1038/ni.2987
19. Thomas GM, Carbo C, Curtis BR, Martinod K, Mazo IB, Schatzberg D, et al. Extracellular DNA traps are associated with the pathogenesis of TRALI in humans and mice. *Blood*. (2012) 119:6335–43. doi: 10.1182/blood-2012-01-405183
20. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. (2016) 34:525–7. doi: 10.1038/nbt.3519
21. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods*. (2017) 14:687–90. doi: 10.1038/nmeth.4324
22. Kolde, R. *Pheatmap: Pretty Heatmaps*. R package v. 16. R Foundation for Statistical Computing (2012).
23. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol*. (2016) 17:218. doi: 10.1186/s13059-016-1113-y
24. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun*. (2019) 10:1523. doi: 10.1038/s41467-019-09234-6
25. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*. (2005) 102:15545–50. doi: 10.1073/pnas.0506580102
26. Damian S, Annika L, David L, Alexander J, Stefan W, Jaime H, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucl Acids Res*. (2019) 47:607–13. doi: 10.1093/nar/gky1131
27. Dinkla K, El-Kebir M, Bucur CI, Siderius M, Smit MJ, Westenberg MA, et al. eXamine: exploring annotated modules in networks. *BMC Bioinformatics*. (2014) 15:201. doi: 10.1186/1471-2105-15-201
28. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. *J Stat Softw*. (2008) 25:1–18. doi: 10.18637/jss.v025.i01
29. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York (2016). Available online at: https://ggplot2.tidyverse.org.
30. Goodrich B, Gabry J, Ali I, Brilleman S, Novik JB, Wolfe R. *rstanarm: Bayesian Applied Regression Modeling via Stan. R Package Version 2.17.4*. Available online at: http://mc-stan.org/.
31. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. (2011) 12:77. doi: 10.1186/1471-2105-12-77
32. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: a probabilistic programming language. *J Stat Softw*. (2017) 76:1–27. doi: 10.18637/jss.v076.i01
33. Lim MB, Kuiper JW, Katchky A, Goldberg H, Glogauer M. Rac2 is required for the formation of neutrophil extracellular traps. *J Leukoc Biol*. (2011) 90:771–6. doi: 10.1189/jlb.1010549
34. Jacob CO, Yu N, Yoo DG, Perez-Zapata LJ, Barbu EA, Kaplan MJ, et al. Haploinsufficiency of NADPH oxidase subunit neutrophil cytosolic factor 2 is sufficient to accelerate full-blown lupus in NZM 2328 mice. *Arthritis Rheumatol*. (2017) 69:1647–60. doi: 10.1002/art.40141
35. Urban CF, Ermert D, Schmid M, Abu-Abed U, Goosmann C, Nacken W, et al. Neutrophil extracellular traps contain calprotectin, a cytosolic protein complex involved in host defense against *Candida albicans*. *PLoS Pathog*. (2009) 5:e1000639. doi: 10.1371/journal.ppat.1000639
36. Lin YT, Tseng KY, Yeh YC, Yang FC, Fung CP, Chen NJ, et al. TREM-1 promotes survival during *Klebsiella pneumoniae* liver abscess in mice. *Infect Immun*. (2014) 82:1335–42. doi: 10.1128/IAI.01347-13
37. Leffler J, Martin M, Gullstrand B, Tydén H, Lood C, Truedsson L, et al. Neutrophil extracellular traps that are not degraded in systemic lupus erythematosus activate complement exacerbating the disease. *J Immunol*. (2012) 188:3522–31. doi: 10.4049/jimmunol.1102404
38. Bruschi M, Petretto A, Santucci L, Vaglio A, Pratesi F, Migliorini P, et al. Neutrophil extracellular traps protein composition is specific for patients with *Lupus nephritis* and includes methyl-oxidized αenolase (methionine sulfoxide 93). *Sci Rep*. (2019) 9:7934. doi: 10.1038/s41598-019-44379-w
39. Wiersma VR, Clarke A, Pouwels SD, Perry E, Abdullah TM, Kelly C, et al. Galectin-9 is a possible promoter of immunopathology in rheumatoid arthritis by activation of peptidyl arginine deiminase 4 (PAD-4) in granulocytes. *Int J Mol Sci*. (2019) 19:20. doi: 10.3390/ijms20164046
40. Agarwal S, Loder SJ, Cholok D, Li J, Bian G, Yalavarthi S, et al. Disruption of neutrophil extracellular traps (NETs) links mechanical strain to post-traumatic inflammation. *Front Immunol*. (2019) 10:2148. doi: 10.3389/fimmu.2019.02148
41. Dwyer M, Shan Q, D'Ortona S, Maurer R, Mitchell R, Olesen H, et al. Cystic fibrosis sputum DNA has NETosis characteristics and neutrophil extracellular trap release is regulated by macrophage migration-inhibitory factor. *J Innate Immun*. (2014) 6:765–79. doi: 10.1159/000363242
42. Roberts H, White P, Dias I, McKaig S, Veeramachaneni R, Thakker N, et al. Characterization of neutrophil function in Papillon-Lefèvre syndrome. *J Leukoc Biol*. (2016) 100:433–44. doi: 10.1189/jlb.5A1015-489R
43. Sara N, Laura F, Uma S, Lynn K, Patrizia S, Giorgio B, et al. Src family kinases and Syk are required for neutrophil extracellular trap formation in response to β-glucan particles. *J Innate Immun*. (2014) 7:59–73. doi: 10.1159/000365249
44. Sharma A, Steichen AL, Jondle CN, Mishra BB, Sharma J. Protective role of mincle in bacterial pneumonia by regulation of neutrophil mediated phagocytosis and extracellular trap formation. *J Infect Dis*. (2014) 209:1837–46. doi: 10.1093/infdis/jit820
45. Rayes RF, Vourtzoumis P, Bou Rjeily M, Seth R, Bourdeau F, Giannias B, et al. Neutrophil extracellular trap-associated CEACAM1 as a putative therapeutic target to prevent metastatic progression of colon carcinoma. *J Immunol*. (2020) 204:2285–94. doi: 10.4049/jimmunol.1900240
46. Syed R, Jerry JF, Aaron FC, Tamara DB, Rebecka L, Takashi A, et al. Siglec-5 and Siglec-14 are polymorphic paired receptors that modulate neutrophil and amnion signaling responses to group B Streptococcus. *J Exp Med*. (2014) 211:1231–42. doi: 10.1084/jem.20131853
47. Rohrbach AS, Slade DJ, Thompson PR, Mowen KA. Activation of PAD4 in NET formation. *Front Immunol*. (2012) 3:360. doi: 10.3389/fimmu.2012.00360
48. Wang Y, Li M, Stadler S, Correll S, Li P, Wang D, et al. Histone hypercitrullination mediates chromatin decondensation and neutrophil extracellular trap formation. *J Cell Biol*. (2009) 184:205–13. doi: 10.1083/jcb.200806072

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

3.4 Comprehensive comparison of transcriptomes in SARS-CoV-2 infection: alternative entry routes and innate immune responses.

This section is based on the following publication:

Yingying Cao, Xintian Xu, Simo Kitanovski, Lina Song, Jun Wang, Pei Hao, and Daniel Hoffmann. **Comprehensive comparison of transcriptomes in SARS-CoV-2 infection: alternative entry routes and innate immune responses.** bioRxiv (2021): 2021-01
https://doi.org/10.1101/2021.01.07.425716

Comprehensive comparison of transcriptomes in SARS-CoV-2 infection: alternative entry routes and innate immune responses

Yingying Cao[1*], Xintian Xu[2], Simo Kitanovski[1], Lina Song[3], Jun Wang[1], Pei Hao[2,4*], Daniel Hoffmann[1*]

[1]Bioinformatics and Computational Biophysics,
Faculty of Biology and Center for Medical Biotechnology,
University of Duisburg-Essen, Essen 45141, Germany
[2]Key Laboratory of Molecular Virology and Immunology,
Institut Pasteur of Shanghai, Center for Biosafety Mega-Science,
Chinese Academy of Sciences, Shanghai 200031, China
[3]Translational Skin Cancer Research,
German Consortium for Translational Cancer Research, Essen, Germany
[4]The Joint Program in Infection and Immunity:
a. Guangzhou Women and Children's Medical Center,
Guangzhou Medical University, Guangzhou 510623, China;
b. Institut Pasteur of Shanghai,
Chinese Academy of Sciences, Shanghai 200031, China

*To whom correspondence should be addressed;
E-mail: daniel.hoffmann@uni-due.de, phao@ips.ac.cn, yingying.cao@uni-due.de.

The pathogenesis of COVID-19 emerges as complex, with multiple factors leading to injury of different organs. Several studies on underlying cellular processes have produced contradictory claims, e.g. on SARS-CoV-2 cell entry or innate immune responses. However, clarity in these matters is imper-

ative for therapy development. We therefore performed a meta-study with a diverse set of transcriptomes under infections with SARS-CoV-2, SARS-CoV and MERS-CoV, including data from different cells and COVID-19 patients. Using these data, we investigated viral entry routes and innate immune responses. First, our analyses support the existence of cell entry mechanisms for SARS and SARS-CoV-2 other than the ACE2 route with evidence of inefficient infection of cells without expression of ACE2; expression of TMPRSS2/TPMRSS4 is unnecessary for efficient SARS-CoV-2 infection with evidence of efficient infection of A549 cells transduced with a vector expressing human ACE2. Second, we find that innate immune responses in terms of interferons and interferon simulated genes are strong in relevant cells, for example Calu3 cells, but vary markedly with cell type, virus dose, and virus type.

Introduction

Coronaviruses are non-segmented positive-sense RNA viruses with a genome of around 30 kilobases. The genome has a 5' cap structure along with a 3' poly (A) tail, which acts as mRNA for translation of the replicase polyproteins. The replicase gene occupies approximately two thirds of the entire genome and encodes 16 non-structural proteins (nsps). The remaining third of the genome contains open reading frames (orfs) that encode accessory proteins and four structural proteins, including spike (S), envelope (E), membrane (M), and nucleocapsid (N) (*1*).

Over the past 20 years, three epidemics or pandemics of life-threatening diseases have been caused by three closely related coronaviruses – severe acute respiratory syndrome coronavirus (SARS-CoV), which emerged with nearly 10 % mortality (*2, 3*) in 2002-2003 and spread to 26 countries before being contained; Middle East respiratory syndrome coronavirus (MERS-CoV), with mortality around 34 % (*4, 5*) starting in 2012 and since then spreading to 27 countries;

SARS-CoV-2, emerging in late 2019 (*6*), which has caused many millions of confirmed cases and > 1 million deaths worldwide (*7*). Infection with SARS-CoV, MERS-CoV or SARS-CoV-2 can cause a severe acute respiratory illness with similar symptoms, including fever, cough, and shortness of breath.

SARS-CoV-2 is a new coronavirus, but its similarity to SARS-CoV (amino acid sequences about 76% identical (*8*)) and MERS-CoV suggests comparisons to these earlier epidemics. Despite the difference in the total number of cases caused by SARS-CoV and SARS-CoV-2 (*3, 7*) due to different transmission rates, the outbreak caused by SARS-CoV-2 resembles the outbreak of SARS: both emerged in winter and were linked to exposure to wild animals sold at markets. Although MERS-CoV has high morbidity and mortality rates, lack of autopsies from MERS-CoV cases has hindered our understanding of MERS-CoV pathogenesis in humans.

Until now there are no specific anti-SARS-CoV-2, anti-SARS-CoV or anti-MERS-CoV therapeutics approved for human use. There are several points of attack for potential anti-SARS-CoV-2/SARS-CoV/MERS-CoV therapies, e.g. intervention on cell entry mechanisms to prevent virus invasion, or acting on the host immune system to kill the infected cells and thus prevent replication of the invading viruses. A better understanding of virus entry mechanisms and the immune responses can therefore guide the development of novel therapeutics.

Virus entry into host cells is the first step of the viral life cycle. It is an essential component of cross-species transmission and an important determinant of virus pathogenesis and infectivity (*9, 10*), and also constitutes an antiviral target for treatment and prevention (*11*). It seems that SARS-CoV and SARS-CoV-2 use similar virus entry mechanisms (*12*). The infection of SARS-CoV or SARS-CoV-2 in target cells was initially identified to occur by cell-surface membrane fusion (*13, 14*). Some later studies have shown that SARS-CoV can infect cells through receptor mediated endocytosis (*15, 16*) as well. Both mechanisms require the S protein of SARS-CoV or SARS-CoV-2 to bind to angiotensin converting enzyme 2 (ACE2), and S protein of MERS-

CoV to dipeptidyl peptidase 4 (DPP4) (*17*), respectively, through their receptor-binding domain (RBD) (*18*). In addition to ACE2 and DPP4, some recent studies suggest that there are possible other coronavirus-associated receptors and factors that facilitate the infection of SARS-CoV-2 (*19*), including the cell surface proteins Basignin (BSG or CD147) (*20*), and CD209 (*21*). Recently, clinical data have revealed that SARS-CoV-2 can infect several organs where ACE2 expression could not be detected in healthy individuals (*22, 23*), which highlights the need of closer inspection of virus entry mechanisms.

The binding of S protein to a cell-surface receptor is not sufficient for infection of host cell (*24*). In the cell-surface membrane fusion mechanism, after binding to the receptor, the S protein requires proteolytic activation by cell surface proteases like TMPRSS2, TPMRSS4, or other members of the TMPRSS family (*14, 25, 26*), followed by the fusion of virus and target cell membranes. In the alternative receptor mediated endocytosis mechanism, the endocytosed virion is subjected to an activation step in the endosome, resulting in the fusion of virus and endosome membranes and the release of the viral genome into the cytoplasm. The endosomal cysteine proteases cathepsin B (CTSB) and cathepsin L (CTSL) (*27*) might be involved in the fusion of virus and endosome membranes. Availability of these proteases in target cells largely determines whether viruses infect the cells through cell-surface membrane fusion or receptor mediated endocytosis. How the presence of these proteases impacts efficiency of infection with SARS-CoV-2, SARS-CoV and MERS-CoV, still remains elusive.

When the virus enters a cell, it may trigger an innate immune response, a crucial component of the defense against viral invasion. Compounds that regulate innate immune responses can be introduced as antiviral agents (*10*). The innate immune system is initialized as pattern recognition receptors (PRRs) such as Toll-like receptors (TLRs) and cytoplasmic retinoic acid-inducible gene I (RIG-I) like receptors (RLRs) recognize molecular structures of the invading virus (*28, 29*). This pattern recognition activates several signaling pathways and then

downstream transcription factors such as interferon regulator factors (IRFs) and nuclear factor κB (NF-κB). Transcriptional activation of IRFs and NF-κB stimulates the expression of type I (α or β) and type III (λ) interferons (IFNs). IFN-α (IFNA1, IFNA2, etc), IFN-β (IFNB1) and IFN-λ (IFNL1-4) are important cytokines of the innate immune responses. IFNs bind and induce signaling through their corresponding receptors (IFNAR for IFN-α/β and IFNLR for IFN-λ), and subsequently induce expression of IFN-simulated genes (ISGs) (e.g. MX1, ISG15 and OASL) and pro-inflammatory chemokines (e.g. CXCL8 and CCL2) to suppress viral replication and dissemination (*30, 31*). Dysregulated inflammatory host response results in acute respiratory distress syndrome (ARDS), a leading cause of COVID-19 mortality (*32*).

One attractive therapy option to combat COVID-19 is to harness the IFN-mediated innate immune responses. Clinical trials with type I and type III IFNs for treatment of COVID-19 have been conducted and many more are still ongoing (*33, 34*). In this regard, the kinetics of the secretion of IFNs in the course of SARS-CoV-2 infection needs to be defined. Unfortunately, some results on the host innate immune responses to SARS-CoV-2 are apparently at odds with each other (*35–39*), e.g. it is unclear whether SARS-CoV-2 infection induces low IFNs and moderate ISGs (*35*), or robust IFN responses and markedly elevated expression of ISGs (*36–39*). This has to be clarified. The use of IFNs as a treatment in COVID-19 is now a subject of debate as well (*40*). Thus, the kinetics of IFN secretion relative to the kinetics of virus replication need to be thoroughly examined to better understand the biology of IFNs in the course of SARS-CoV-2 infection and thus provide guidance to identify the temporal window of therapeutic opportunity.

We have collected and analyzed a diverse set of publicly available transcriptome data (*35, 41–45*): (1) bulk RNA-Seq data with different types of cells, including human non-small cell lung carcinoma cell line (H1299), human lung fibroblast-derived cells (MRC5), human alveolar basal epithelial carcinoma cell line (A549), A549 cells transduced with a vector expressing

human ACE2 (A549-ACE2), primary normal human bronchial epithelial cells (NHBE), heterogeneous human epithelial colorectal adenocarcinoma cells (Caco2), and African green monkey (*Chlorocebus sabaeus*) kidney epithelial cells (Vero E6) infected with SARS-CoV-2, SARS-CoV and MERS-CoV (Table 1); (2) RNA-Seq data of lung samples, peripheral blood mononuclear cell (PBMC) samples, and bronchoalveolar lavage fluid (BALF) samples of COVID-19 patients and their corresponding healthy controls (Table 1 and Table 2). Using this collection, we systemically evaluated the replication and transcription status of virus in these cells, expression levels of coronavirus-associated receptors and factors, as well as the innate immune responses of these cells during virus infection.

Results

Different infection efficiency of SARS-CoV-2, SARS-CoV and MERS-CoV in different cell types

The RNA-Seq data for all samples can be aligned to the genome of the corresponding virus to evaluate the infection efficiency in cells, estimated by the mapping rate to the virus genome, i.e. the percentages of viral RNAs in intracellular RNAs. To assess the infection efficiency of SARS-CoV-2, SARS-CoV, and MERS-CoV in different types of cells, we collected and analyzed a comprehensive public datasets of RNA-Seq data of cells infected with these viruses at 24 hours post infection (hpi) with comparable multiplicity of cellular infection (MOI) (Table 1). MOI refers to the number of viruses that are added per cell in infection experiments. For example, if 2000 viruses are added to 1000 cells, the MOI is 2.

Our analysis shows that the infection efficiency of viruses can be both cell type dependent and virus dose dependent (Fig. 1). MERS-CoV can efficiently infect MRC5 and Vero E6 cells. However, the infection efficiency is influenced strongly by MOI in the same type of cells. Cells infected with low MOI, say 0.1, have significantly lower mapping rates than those with high

MOI, say 3 (Fig. 1). For SARS-CoV and SARS-CoV-2, the infection efficiency is influenced strongly by cell type. For SARS-CoV-2, there is efficient virus infection in A549-ACE2, Calu3, Caco2, and Vero E6 cells, but not in A549, H1299, or NHBE cells (Fig. 1 and Table S1). The mapping rates in A549, H1299, and NHBE cells are low even at high MOIs (Fig. 1 and Table S1). Similar to SARS-CoV-2, the infection by SARS-CoV is also cell type dependent, Vero E6 cells and Calu3 cells show high mapping rates to SARS-CoV genome, but the mapping rates of SARS-CoV in MRC5 and H1299 cells are close to zero even at the high MOI of 3 (Fig. 1 and Table S1). Since "total RNA" (see Methods/Data collection) includes additional negative-strand templates of virus, the mapping rates are usually much higher than those that used the PolyA+ selection method in the same condition (Fig. 1 and Table S1).

Evidence for multiple entry mechanisms for SARS-CoV-2 and SARS-CoV

To examine the detailed replication and transcription status of these viruses in the cells, we calculated the number of reads (depth) mapped to each site of the corresponding virus genome (Fig. 2). For better comparison, these read numbers were $\log_{10}$ transformed. The replication and transcription of MERS-CoV, SARS-CoV-2 and SARS-CoV share an uneven pattern of expression along the genome, typically with a minimum depth in the first half of the viral genome, and the maximum towards the end. Among the parts with very high levels, there are especially coding regions for structural proteins, including S, E, M, and N proteins, as well as the first coding regions with nsp1 and nsp2. Interestingly, there is an exception for BALF samples in COVID-19 patients, which show a more irregular, fluctuating behavior along the genome (Fig. 2B). The deviation from the cellular expression pattern is not surprising because BALF is not a well-organized tissue but a mixture of many components, some of which will probably digest viral RNA.

Interestingly, the mentioned uneven transcription pattern of efficient infections with SARS-

CoV-2, SARS-CoV, and MERS-CoV, is also visible for inefficient infection with SARS-CoV-2 in A549, NHBE, and H1299 cells, and SARS-CoV in H1299 and MRC5 cells (Fig. 2C, D), although there the total mapping rates to their corresponding virus genomes are much lower (Fig. 1).

To further elucidate the corresponding entry mechanisms for different types of cells, we examined the expression levels of those receptors and proteases that have already been described as facilitating target cell infection (Fig. 3).

Our analysis shows that MERS-CoV can efficiently infect MRC5 and Vero E6 cells (Fig. 1 and Fig. 2E) that both express DPP4 (Fig. 3A), though compared to Vero E6 cells, MRC5 cells infected with MERS-CoV have higher expression levels of DPP4 (Fig. 3A), but lower mapping rates to the virus genome (Fig. 1). These observations show that higher expression levels of the receptor (DPP4) do not guarantee higher MERS-CoV infection efficiency in cells. This is also true for SARS-CoV-2 receptor ACE2, which is expressed three orders of magnitudes higher in A549-ACE2 cells than in Vero E6 cells (Fig. 3B), while both cells produce about the same amount of virus (Fig. 1).

Although SARS-CoV-2 can efficiently infect A549-ACE2 cells (Fig. 1 and Fig. 2), there is no expression of TMPRSS2 or TMPRSS4 (Fig. 3C, D), needed for the canonical cell-surface membrane fusion mechanism (Fig. 3J). However, there are considerable expression levels of CTSB and CTSL (Fig. 3E, F), which are involved in endocytosis (Fig. 3J).

In A549, H1299, and MRC5 cells, which do express small amounts of SARS-CoV-2 and SARS-CoV virus (Fig. 1, Fig. 2C, D), there is no ACE2 expression at all (Fig. 3B). This could point to an alternative ACE2-independent entry mechanism for SARS-CoV-2 and SARS-CoV (Fig. 3J). Since there were already reports about alternative SARS-CoV-2 receptors such as BSG/CD147 and CD209 (*20, 21*), we examined their expressions in these cells as well (Fig. 3G, H). For all cells, the expression of BSG is at the same level of 2-3 (Fig. 3G), and the expression

of CD209 is very low. Certainly, CD209 and BSG alone cannot explain the differences in virus expression (Fig. 1), nor can we exclude other low efficiency entry mechanisms. It could e.g. be that relatively inefficient alternative entry paths are often present but in some cells masked by more efficient entry via ACE2/TPMRSS.

To gain a comprehensive overview we clustered cells with respect to gene expression levels of coronavirus-associated receptors and factors (Fig. 3I), and summarized conceivable mechanisms accordingly (Fig. 3J). Since all cells show high expression levels of CTSB and CTSL, the major differences between these cells lie in the expression levels of ACE2, TMPRSS2 and TPMRSS4.

Cell-surface membrane fusion (Fig. 3J, 1a) might be mainly used in SARS-CoV-2 infection of Calu3, Caco2, and NHBE cells where there are low to moderate expression of ACE2 and moderate expression of TMPRSS2 and TMPRSS4. Endocytosis (Fig. 3J, 1b) might be mainly used in SARS-CoV-2 infection of A549-ACE2 cells where ACE2 is expressed at high levels but there is no expression of TMPRSS2 or TMPRSS4. An alternative ACE2-independent way (Fig. 3J, 1c) in absence of ACE2, TMPRSS2, or TMPRSS4 could be mainly employed in SARS-CoV-2 infection of MRC5, A549, and H1299 cells. Note that although the expression pattern of coronavirus-associated receptors and factors of NHBE cells is similar to that in Caco2 cells, NHBE cells are not infected efficiently by SARS-CoV-2. Vero E6 cells have moderate expression of ACE2, and low expression of TMPRSS2 and TMPRSS4, so all these entry mechanisms mentioned above could contribute to SARS-CoV-2 infection of Vero E6 cells.

Strength of IFN/ISG response varies between cell lines and viruses, with strong response to SARS-CoV-2 in relevant cells

As a virus enters a cell, it may trigger an innate immune response, i.e. the cell may start expression of various types of innate immunity molecules at different strengths. There is currently

an intense debate about which of these molecules, especially IFNs and ISGs, are expressed how strongly (*35–39*). We therefore focused in our analysis on innate immunity molecules such as IFNs, ISGs, and pro-inflammatory cytokines. To broaden the basis for conclusions, we analyzed, apart from cell lines, bulk RNA-Seq data of lung, PBMC, and BALF samples of COVID-19 patients, and single-cell RNA-Seq data of BALF samples from moderate and severe COVID-19 patients; for each type of patient data, we also included healthy controls. Gene expressions were compared quantitatively in terms of TPM (transcripts per million), as well as log fold changes (logFC) with respect to healthy controls (human samples) or mock-infected cultures (cell lines) (Fig. S1, Fig. S2).

The heatmap and clustering dendrogram of the logFC of IFNs, ISGs and pro-inflammatory cytokines in Fig. 4A reveal broadly two groups of samples with fundamentally different expression of ISGs, IFNs, and pro-inflammatory cytokines.

The top cluster in Fig. 4A are samples that show weaker innate immune response, including the two PBMC samples of COVID-19 patients, A549, NHBE, Caco2, and H1299 cells infected with SARS-CoV-2 and A549-ACE2 cells infected with SARS-CoV-2 at lower MOI (0.2), MRC5 cells infected with SARS, MRC5 and Vero E6 cells infected with MERS. The bottom cluster in Fig. 4A are samples that show stronger innate immune response, including BALF and lung samples of COVID-19 patients, Calu3 cells infected with SARS-CoV-2, A549-ACE2 cells infected with SARS-CoV-2 at higher MOI (2), as well as Vero E6 cells infected with SARS-CoV-2 and SARS. Most of the samples in the bottom part show markedly elevated levels of ISGs and elevated pro-inflammatory cytokines. An exception in the bottom cluster are four samples, namely Lung.1/2 and BALF.1/2, with a mixture of up- and down-regulation of ISGs and pro-inflammatory cytokines. In this respect, these four samples from patients with unknown COVID-19 severity differ from the BALF samples from moderate and severe COVID-19 patients.

The expression levels of IFNs are not upregulated either in most of these lung, PBMC and BALF samples of COVID-19 patients where no information about the severity of infection of these COVID-19 patients are available. However, we estimated the severity of their infection by aligning all the samples to SARS-COV-2 virus genome. There are no (0.00%) reads mapping to the SARS-CoV-2 genome in the PBMC samples. For the two BALF samples, there are low mapping rates (1.56% and 0.65%) to SARS-CoV-2 genome. The expression levels of ACE2 in these tissues (PBMC, lung and BALF samples) of healthy individuals are around zero (Fig. S8), which explains why there are almost no virus reads in these tissues.

One of the two lung samples (accession number: SAMN14563387) has slightly upregulated IFNL1 (Fig. S6), which had been ignored in the original publication (*35*), although the total mapping rates to virus genome are both 0.00% for these two lung samples. We then checked the detailed coverage along the virus genome. There were a small number of virus reads aligned to SARS-CoV-2 genome in this sample (Fig. S7). Different from other lung samples that did not express ACE2, this lung sample expressed ACE2 at a considerable level (5.45 TPM, Table S2). This result implies that when SARS-CoV-2 enters into lung successfully, or when the lung tissue chosen for sequencing are successfully infected by SARS-CoV-2, IFNs (at least IFNL1) can be upregulated.

Calu3 cells infected with SARS-CoV and SARS-CoV-2, and A549-ACE2 cells infected with SARS-CoV-2 at a high MOI of 2 have upregulated IFNB1, IFNL1, IFNL2 and IFNL3 (Fig. 4B-E). A549, H1299, NHBE (Fig. 4B-E), and MRC5 cells (Fig. S3), which do not support efficient virus infection, show no upregulation of IFNs. Low levels of IFN expression are also observed in Caco2 cells, which are efficiently infected with SARS-CoV and SARS-CoV-2. The same is true for A549-ACE2 cells infected with SARS-CoV-2 at low MOI of 0.2. In Vero E6 cells IFNL1 is upregulated as well in infected with SARS-CoV and SARS-CoV-2, but not with MERS-CoV (Fig. 4F). In BALF samples of moderate and severe COVID-19 patients,

upregulation of IFNs was not as obvious as in Calu3 cells, but is still present in some patients. These observations demonstrate that the innate immune response depends in complex ways on cell line, viral dose, and virus.

Several studies (*36–39*) reported robust IFN responses and markedly elevated expression of ISGs in SARS-CoV-2 infection of different cells and patient samples. Conversely, the study by (*35*) concluded that weak IFN response and moderate ISG expression are characteristic for SARS-CoV-2 infection. This apparent contradiction can be resolved if we consider that Ref. (*35*) generalized from patient samples and cells that were only weakly infected, and that in such cases the host, in fact, responds with low levels of IFNs and ISGs. On the other hand, Ref. (*35*) treated efficiently infected cells, such as Calu3 and A549-ACE2 (at MOI 0f 2) as exceptions. However, our meta-analysis shows that these are not exceptions but typical for severely infected target cells that have robust IFN responses and ISG expressions (cluster 2 in Fig. 4A).

Discussion

One attractive potential anti-SARS-CoV-2 therapy is intervention in the cell entry mechanisms (*12*). However, the entry mechanisms of SARS-CoV-2 into human cells are partly unknown. During the last few months scientists have confirmed that SARS-CoV-2 and SARS-CoV both use human ACE2 as entry receptor, and human proteases like TMPRSS2 and TMPRSS4 (*8, 14, 25*), and lysosomal proteases like CTSB and CTSL (*27*) as entry activators. Since ACE2 is beneficial in cardiovascular diseases such as hypertension or heart failure (*46*), treatments targeting ACE2 could have a negative effect. Inhibitors of CTSL (*47*) or TMPRSS2 (*14*) are seen as potential treatment options for SARS-CoV and SARS-CoV-2. However, recently alternate coronavirus-associated receptors and factors including BSG/CD147 (*20*) and CD209 (*21*) have been proposed to facilitate virus invasion. Additionally, clinical data of SARS-CoV-2 infection

have shown that SARS-CoV-2 can infect several organs where ACE2 expression could not be detected (*22, 23*), urging us to explore other potential entry routes.

First, our analyses here have shown that even without expression of TMPRRS2 or TMPRSS4, high SARS-CoV-2 infection efficiency in cells is possible (Fig. 1A, C) with considerable expression levels of CTSB and CTSL (Fig. 2E, F). This suggests receptor mediated endocytosis (*15, 16, 27*) as an alternative major entry mechanism. Given this TMPRSS-independent route, TMPRSS inhibitors will likely not provide complete protection. The studies designed to predict the tropism of SARS-CoV-2 by profiling the expression levels of ACE2 and TMPRSS2 across healthy tissues (*48, 49*) may need to be reconsidered as well.

Second, the evidence presented in our study suggests further, possibly undiscovered entry mechanism for SARS-CoV-2 and SARS-CoV (Fig. 2). Although BSG/CD147 has been recently proposed as an alternate receptor (*20*), later experiments reported there was no evidence supporting the role of BSG/CD147 as a putative spike-binding receptor (*50*). The expression patterns of BSG/CD147 in different types of cells observed in our study could not explain the difference in virus loads observed in these cells either. CD209 and CD209L were recently reported as attachment factors to contribute to SARS-CoV-2 infection in human cells as well (*21*). However, CD209 expression in the cell lines included here is low. Another reasonable hypothesis could be that the inefficient ACE2-independent entry mechanism we observed could be macropinocytosis, one endocytic pathway that does not require receptors (*51*). Until now there is still no direct evidence for macropinocytosis involvement in SARS-CoV-2 and SARS-CoV entry mechanism. To confirm such an involvment, specific experiments are needed. Moreover, this ACE2-independent entry mechanism, only enables inefficient infection by SARS-CoV and SARS-CoV-2 (Fig. 2) and therefore cannot be a major entry mechanism.

Fig. 3J summarizes the outcomes of our study with respect to entry mechanisms. The observations with the broad range of transcriptome data can only be explained if there are several

entry routes. This is certainly a challenge to be reckoned with in the development of antiviral therapeutics (*52*).

Another attractive potential anti-SARS-CoV-2 point of attack is supporting the human innate immune system to kill the infected cells and, thus disrupt viral replication. Not surprisingly, research in this area is flourishing but sometimes generates conflicting results, especially on the involvement of type I and III IFNs and ISGs (*35–39*). The results of our analyses could help to dissolve the confusion on the involvement of IFNs and ISGs.

We found that immune responses in Calu3 cells infected with SARS-CoV and SARS-CoV-2 resemble those of BALF samples of moderate and severe COVID-19 patients, with elevated levels of type I and III IFNs, robust ISG induction as well as markedly elevated pro-inflammatory cytokines, in agreement with recent studies (*36–39*). However this picture differs from the one reported by (*35*) with low levels of IFNs and moderate ISGs. This latter study was partially based on A549 cells and NHBE cells with nearly no ACE2 expression and very low mapping rate to the viral genome, and lung samples of two patients (both show 0.00% mapping rate to virus genome). Hence, given that there was no efficient virus infection in theses cells, the low levels of IFNs and ISGs were to be expected. However, in one of the lung samples sequenced by (*35*) (accession number: SAMN14563387), we observed a slight upregulation of IFNL1 (Fig. S6), which was ignored in the original publication, together with considerable ACE2 expression (Table S2) (5.45 TPM), and a few virus reads aligned to SARS-CoV-2 genome (Fig. S7). This results suggests that levels of IFNs are ISGs are associated with viral load and severity of virus infection.

We found low induction of IFNs and moderate expression of ISGs in PBMC samples and BALF samples of COVID-19 patients (Fig. 4, Fig. S5). In these PBMC samples, there are no (0.00%) virus reads mapping to the SARS-CoV-2 genome. The failure to detect virus reads in these three PBMC samples can be explained by the absence of efficient entry routes (e.g.

no expression of ACE2 in PBMC samples of healthy individuals, Fig. S8), or with the cell types being otherwise incompatible with viral replication. This observation is consistent with the studies on SARS-CoV (*53–55*) with abortive infections of macrophages, monocytes, and dendritic cells; moreover, replication of SARS-CoV in PBMC samples is also self-limiting. However, due to the limited number of PBMC, BALF and lung samples included in this study, and the lack of the information of infection stage and infection severity of these COVID-19 patients, the assessment of IFNs and ISGs as well as the infection of SARS-CoV-2 in these samples may not be representative of host response against SASR-CoV-2. Future studies that include also other affected organs of more patients with different infection stages and severity are necessary for a better understanding of the immune responses.

Several unexpected observations need further investigations. First, A549-ACE2 and Caco2 cells are efficiently infected with low MOI of 0.2 and 0.3, respectively, (Fig. 1), but fail to upregulate INF expression (Fig. 4B-E). Their cellular immune responses are more similar to those of cells that cannot support efficient virus infection (Fig. 4A). These results suggest that in Caco2 and A549-ACE2 cells the invasion of SARS-CoV-2 or SARS-CoV at low MOI shuts down or fails to activate the innate immune system.

Based on the results observed above, multiple factors including disease severity, different organs, cell types and virus dose contribute to the variability in the innate immune responses. For a better characterization of the innate immune responses, a more comprehensive profiling is necessary, including of patients with infections in different stages, different levels of severity, and different clinical outcomes of the infection. Further, a larger array of cell types should be profiled over time after infection with different virus doses. In this way we would be better able to understand the kinetics of IFNs and ISGs in response to SARS-CoV-2 infection.

In summary, our study has comparatively analyzed an extensive data collection from different cell types infected with SARS-CoV-2, SARS-CoV and MERS-CoV, and from COVID-19

patients. We have presented evidence for multiple SARS-CoV-2 entry mechanisms. We could also dissolve apparent conflicts on innate immune responses in SARS-CoV-2 infection (*35–39*), by drawing upon a larger set of cell types and infection severity. The results emphasize the complexity of interactions between host and SARS-CoV-2, offer new insights into pathogenesis of SARS-CoV-2, and can inform development of antiviral drugs.

Materials and Methods

Data collection

After the successful release of the virus genome into the cytoplasm, a negative-strand genomic-length RNA is synthesized as the template for replication. Negative-strand subgenome-length mRNAs are formed as well from the virus genome as discontinuous RNAs, and used as the templates for transcription. In the public data we collected for the analysis, there are two main library preparation methods to remove the highly abundant ribosomal RNAs (rRNA) from total RNA before sequencing. One is polyA+ selection, the other is rRNA-depletion (*56*). It is known that coronavirus genomic and subgenomic mRNAs carry a polyA tail at their 3' ends, so in the polyA+ RNA-Seq, we have (1) virus genomic sequence from virus replication, i.e. replicated genomic RNAs from negative-strand as template, and (2) subgenomic mRNAs from virus transcription; in the rRNA-depletion RNA-Seq we have (1) virus genomic sequence from virus replication: both replicated genomic RNAs from negative-strand as template and the negative-strand templates themselves, and (2) subgenomic mRNAs from virus transcription. PolyA+ selection was used if not specifically stated in this study, "total RNA" is used to specify that the rRNA-depletion method was used to prepare the sequencing libraries.

The raw FASTQ data of different cell types infected with SARS-CoV-2, SARS-CoV and MERS-CoV, and lung samples of COVID-19 patients and healthy controls were retrieved from NCBI (*57*) (https://www.ncbi.nlm.nih.gov/) and ENA (*58*) (https://www.ebi.ac.uk/ena) (acces-

sion numbers GSE147507 (*35*), GSE56189, GSE148729 (*41*) and GSE153940 (*59*)). The raw FASTQ data of PBMC and BALF samples of COVID-19 patients and corresponding controls were downloaded from BIG Data Center (*60*) (https://bigd.big.ac.cn/) (accession number CRA002390) (*42*), and the raw FASTQ data for BALF healthy control samples were downloaded from NCBI (accession numbers SRR10571724, SRR10571730, and SRR10571732 under project PRJNA434133 (*43*)). The preprocessed single cell RNA-Seq data of BALF samples from 6 severe COVID-19 patients and 3 moderate COVID-19 patients were downloaded from NCBI with accession number GSE145926 (*44*). The preprocessed single cell RNA-Seq data of BALF sample from a healthy control was retrieved from NCBI (accession number GSM3660650 under project PRJNA526088 (*45*)). Detailed information about these public datasets are available in the supplementary file: Supplementary.pdf

For analysis, the human GRCh38 release 99 transcriptome and the green monkey (*Chlorocebus sabaeus*) ChlSab1.1 release 99 transcriptome and their corresponding annotation GTF files were downloaded from ENSEMBL (*61*) (https://www.ensembl.org). The reference virus genomes were downloaded from NCBI: SARS-CoV-2 (GenBank: MN985325.1), SARS-CoV (GenBank: AY278741.1), MERS-CoV (GenBank: JX869059.2).

Data analysis workflow

The workflow of this study is summarized in Fig. S1 and Fig. S2 in the supplementary file: Supplementary.pdf. The quality of the raw FASTQ data was examined with FastQC (*62*). Trimmomatic-0.36 (*63*) was used to remove adapters and filter out low quality reads with parameters "-threads 4 -phred33 ILLUMINACLIP:adapters.fasta:2:30:10 HEADCROP:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:36". The clean RNA sequencing reads were then pseudo-aligned to reference transcriptome and quantified using Kallisto (version 0.43.1) (*64*) with parameters "-b 30 –single -l 180 -s 20" for single-end sequencing data

and with parameter "-b 30" for paired-end sequencing data. Expression levels were calculated and summarized as transcripts per million (TPM) on gene levels with Sleuth (*65*), and logFC was then calculated for each condition. The single cell RNA-Seq data were summarized across all cells to obtain "pseudo-bulk" samples. R packages EDASeq (*66*) and org.Hs.eg.db (*67*) were used to obtain gene length, and TPM was calculated with the "calculateTPM" function of R package scater (*68*). logFC was then calculated for each patient.

The clean RNA-Seq data were also aligned to the virus genome with Bowtie 2 (*69*) (version 2.2.6) and the aligned BAM files were created, and the mapping rates to the virus genomes were obtained as well. SAMtools (*70*) (version 1.5) was then used for sorting and indexing the aligned BAM files. The "SAMtools depth" command was used to produce the number of aligned reads per site along the virus genome.

The heatmap in Fig. 3I was made by pheatmap R package (*71*), "complete" clustering method was used for clustering the rows and "euclidean" distance was used to measure the cluster distance. The heatmap in Fig. 4A was made by ComplexHeatmap R package (*72*). "complete" clustering method was used for clustering the rows and columns and "euclidean" distance was used to measure the cluster distance.

References

1. A. R. Fehr, S. Perlman, *Coronaviruses* (Springer, 2015), pp. 1–23.

2. T. Kuiken, R. A. Fouchier, M. Schutten, G. F. Rimmelzwaan, G. Van Amerongen, D. Van Riel, J. D. Laman, T. De Jong, G. Van Doornum, W. Lim, A. E. Ling, P. K. Chan, J. S. Tam, M. C. Zambon, R. Gopal, C. Drosten, S. Van Der Werf, N. Escriou, J. C. Manuguerra, K. Stöhr, J. S. Peiris, A. D. Osterhaus, Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome. *The Lancet* **362**, 263–270 (2003).

3. WHO, Summary of probable sars cases with onset of illness from 1 november 2002 to 31 july 2003.

4. A. M. Zaki, S. Van Boheemen, T. M. Bestebroer, A. D. Osterhaus, R. A. Fouchier, Isolation of a novel coronavirus from a man with pneumonia in saudi arabia. *New England Journal of Medicine* **367**, 1814–1820 (2012).

5. WHO, Middle east respiratory syndrome coronavirus (mers-cov) âĂŞ saudi arabia.

6. F. Wu, S. Zhao, B. Yu, Y. M. Chen, W. Wang, Z. G. Song, Y. Hu, Z. W. Tao, J. H. Tian, Y. Y. Pei, M. L. Yuan, Y. L. Zhang, F. H. Dai, Y. Liu, Q. M. Wang, J. J. Zheng, L. Xu, E. C. Holmes, Y. Z. Zhang, A new coronavirus associated with human respiratory disease in china. *Nature* **579**, 265–269 (2020).

7. WHO, Who coronavirus disease (covid-19) dashboard.

8. X. Xu, P. Chen, J. Wang, J. Feng, H. Zhou, X. Li, W. Zhong, P. Hao, Evolution of the novel coronavirus from the ongoing wuhan outbreak and modeling of its spike protein for risk of human transmission. *Science China Life Sciences* **63**, 457–460 (2020).

9. S. Belouzard, J. K. Millet, B. N. Licitra, G. R. Whittaker, Mechanisms of coronavirus cell entry mediated by the viral spike protein. *Viruses* **4**, 1011–1033 (2012).

10. Z. Lou, Y. Sun, Z. Rao, Current progress in antiviral strategies. *Trends in pharmacological sciences* **35**, 86–102 (2014).

11. E. Teissier, F. Penin, E.-I. Pécheur, Targeting cell entry of enveloped viruses as an antiviral strategy. *Molecules* **16**, 221–250 (2011).

12. I. S. Mahmoud, Y. B. Jarrar, W. Alshaer, S. Ismail, Sars-cov-2 entry in host cells-multiple targets for treatment and prevention. *Biochimie* (2020).

13. Z. Qinfen, C. Jinming, H. Xiaojun, Z. Huanying, H. Jicheng, F. Ling, L. Kunpeng, Z. Jingqiang, The life cycle of sars coronavirus in vero e6 cells. *Journal of medical virology* **73**, 332–337 (2004).

14. M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N. H. Wu, A. Nitsche, M. A. Müller, C. Drosten, S. Pöhlmann, Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *Cell* (2020).

15. Z.-Y. Yang, Y. Huang, L. Ganesh, K. Leung, W.-P. Kong, O. Schwartz, K. Subbarao, G. J. Nabel, ph-dependent entry of severe acute respiratory syndrome coronavirus is mediated by the spike glycoprotein and enhanced by dendritic cell transfer through dc-sign. *Journal of virology* **78**, 5642–5650 (2004).

16. H. Wang, P. Yang, K. Liu, F. Guo, Y. Zhang, G. Zhang, C. Jiang, Sars coronavirus entry into host cells through a novel clathrin-and caveolae-independent endocytic pathway. *Cell research* **18**, 290–301 (2008).

17. W. Widagdo, S. Sooksawasdi Na Ayudhya, G. B. Hundie, B. L. Haagmans, Host determinants of mers-cov transmission and pathogenesis. *Viruses* **11**, 280 (2019).

18. F. Li, Structure, function, and evolution of coronavirus spike proteins. *Annual review of virology* **3**, 237–261 (2016).

19. M. Singh, V. Bansal, C. Feschotte, A single-cell rna expression map of human coronavirus entry factors. *bioRxiv* (2020).

20. K. Wang, W. Chen, Y.-S. Zhou, J.-Q. Lian, Z. Zhang, P. Du, L. Gong, Y. Zhang, H.-Y. Cui, J.-J. Geng, B. Wang, X.-X. Sun, C.-F. Wang, X. Yang, P. Lin, Y.-Q. Deng, D. Wei, X.-M.

Yang, Y.-M. Zhu, K. Zhang, Z.-H. Zheng, J.-L. Miao, T. Guo, Y. Shi, J. Zhang, L. Fu, Q.-Y. Wang, H. Bian, P. Zhu, Z.-N. Chen, Sars-cov-2 invades host cells via a novel route: Cd147-spike protein. *BioRxiv* (2020).

21. R. Amraie, M. A. Napoleon, W. Yin, J. Berrigan, E. Suder, G. Zhao, J. Olejnik, S. Gummuluru, E. Muhlberger, V. Chitalia, N. Rahimi, Cd209l/l-sign and cd209/dc-sign act as receptors for sars-cov-2 and are differentially expressed in lung and kidney epithelial and endothelial cells. *BioRxiv* (2020).

22. F. Hikmet, L. Méar, Å. Edvinsson, P. Micke, M. Uhlén, C. Lindskog, The protein expression profile of ace2 in human tissues. *Molecular Systems Biology* **16**, e9610 (2020).

23. L. Zou, F. Ruan, M. Huang, L. Liang, H. Huang, Z. Hong, J. Yu, M. Kang, Y. Song, J. Xia, Q. Guo, T. Song, J. He, H. L. Yen, M. Peiris, J. Wu, Sars-cov-2 viral load in upper respiratory specimens of infected patients. *New England Journal of Medicine* **382**, 1177–1179 (2020).

24. G. Simmons, J. D. Reeves, A. J. Rennekamp, S. M. Amberg, A. J. Piefer, P. Bates, Characterization of severe acute respiratory syndrome-associated coronavirus (sars-cov) spike glycoprotein-mediated viral entry. *Proceedings of the National Academy of Sciences* **101**, 4240–4245 (2004).

25. R. Zang, M. F. G. Castro, B. T. McCune, Q. Zeng, P. W. Rothlauf, N. M. Sonnek, Z. Liu, K. F. Brulois, X. Wang, H. B. Greenberg, M. S. Diamond, M. A. Ciorba, S. P. Whelan, S. Ding, Tmprss2 and tmprss4 promote sars-cov-2 infection of human small intestinal enterocytes. *Science immunology* **5** (2020).

26. P. Zmora, M. Hoffmann, H. Kollmus, A.-S. Moldenhauer, O. Danov, A. Braun, M. Winkler, K. Schughart, S. Pöhlmann, Tmprss11a activates the influenza a virus hemagglutinin and

the mers coronavirus spike protein and is insensitive against blockade by hai-1. *Journal of Biological Chemistry* **293**, 13863–13873 (2018).

27. X. Ou, Y. Liu, X. Lei, P. Li, D. Mi, L. Ren, L. Guo, R. Guo, T. Chen, J. Hu, Z. Xiang, Z. Mu, X. Chen, J. Chen, K. Hu, Q. Jin, J. Wang, Z. Qian, Characterization of spike glycoprotein of sars-cov-2 on virus entry and its immune cross-reactivity with sars-cov. *Nature communications* **11**, 1–12 (2020).

28. Y.-M. Loo, M. Gale Jr, Immune signaling by rig-i-like receptors. *Immunity* **34**, 680–692 (2011).

29. A. G. Bowie, I. R. Haga, The role of toll-like receptors in the host response to viruses. *Molecular immunology* **42**, 859–867 (2005).

30. C. Chiang, M. U. Gack, Post-translational control of intracellular pathogen sensing pathways. *Trends in immunology* **38**, 39–52 (2017).

31. A. Park, A. Iwasaki, Type i and type iii interferons–induction, signaling, evasion, and application to combat covid-19. *Cell Host & Microbe* (2020).

32. Q. Ruan, K. Yang, W. Wang, L. Jiang, J. Song, Clinical predictors of mortality due to covid-19 based on an analysis of data of 150 patients from wuhan, china. *Intensive care medicine* **46**, 846–848 (2020).

33. I. F. N. Hung, K. C. Lung, E. Y. K. Tso, R. Liu, T. W. H. Chung, M. Y. Chu, Y. Y. Ng, J. Lo, J. Chan, A. R. Tam, H. P. Shum, V. Chan, A. K. L. Wu, K. M. Sin, W. S. Leung, W. L. Law, D. C. Lung, S. Sin, P. Yeung, C. C. Y. Yip, R. R. Zhang, A. Y. F. Fung, E. Y. W. Yan, K. H. Leung, J. D. Ip, A. W. H. Chu, W. M. Chan, A. C. K. Ng, R. Lee, K. Fung, A. Yeung, T. C. Wu, J. W. M. Chan, W. W. Yan, W. M. Chan, J. F. W. Chan, A. K. W. Lie,

O. T. Y. Tsang, V. C. C. Cheng, T. L. Que, C. S. Lau, K. H. Chan, K. K. W. To, K. Y. Yuen, Triple combination of interferon beta-1b, lopinavir–ritonavir, and ribavirin in the treatment of patients admitted to hospital with covid-19: an open-label, randomised, phase 2 trial. *The Lancet* **395**, 1695–1704 (2020).

34. E. Andreakos, S. Tsiodras, Covid-19: lambda interferon against viral load and hyperinflammation. *EMBO Molecular Medicine* p. e12465 (2020).

35. D. Blanco-Melo, B. E. Nilsson-Payant, W. C. Liu, S. Uhl, D. Hoagland, R. Møller, T. X. Jordan, K. Oishi, M. Panis, D. Sachs, T. T. Wang, R. E. Schwartz, J. K. Lim, R. A. Albrecht, B. R. TenOever, Imbalanced host response to sars-cov-2 drives development of covid-19. *Cell* (2020).

36. Z. Zhou, L. Ren, L. Zhang, J. Zhong, Y. Xiao, Z. Jia, L. Guo, J. Yang, C. Wang, S. Jiang, D. Yang, G. Zhang, H. Li, F. Chen, Y. Xu, M. Chen, Z. Gao, J. Yang, J. Dong, B. Liu, X. Zhang, W. Wang, K. He, Q. Jin, M. Li, J. Wang, Heightened innate immune responses in the respiratory tract of covid-19 patients. *Cell Host & Microbe* (2020).

37. A. Broggi, S. Ghosh, B. Sposito, R. Spreafico, F. Balzarini, A. Lo Cascio, N. Clementi, M. de Santis, N. Mancini, F. Granucci, I. Zanoni, Type iii interferons disrupt the lung epithelial barrier upon viral recognition. *Science* (2020).

38. L. Wei, S. Ming, B. Zou, Y. Wu, Z. Hong, Z. Li, X. Zheng, M. Huang, L. Luo, J. Liang, X. Wen, T. Chen, Q. Liang, L. Kuang, H. Shan, X. Huang, Viral invasion and type i interferon response characterize the immunophenotypes during covid-19 infection. *Available at SSRN 3555695* (2020).

39. J. Y. Zhang, X. M. Wang, X. Xing, Z. Xu, C. Zhang, J. W. Song, X. Fan, P. Xia, J. L. Fu, S. Y. Wang, R. N. Xu, X. P. Dai, L. Shi, L. Huang, T. J. Jiang, M. Shi, Y. Zhang, A. Zumla,

M. Maeurer, F. Bai, F. S. Wang, Single-cell landscape of immunological responses in patients with covid-19. *Nature Immunology* pp. 1–12 (2020).

40. E. Sallard, F. X. Lescure, Y. Yazdanpanah, F. Mentre, N. Peiffer-Smadja, Type 1 interferons as a potential treatment against covid-19. *Antiviral Research* p. 104791 (2020).

41. E. Wyler, K. Mösbauer, V. Franke, A. Diag, T. G. Lina, R. Arsie, F. Klironomos, D. Koppstein, S. Ayoub, C. Buccitelli, A. Richter, I. Legnini, A. Ivanov, T. Mari, S. D. Giudice, P. P. Jan, A. M. Marcel, D. Niemeyer, M. Selbach, A. Akalin, N. Rajewsky, C. Drosten, M. Landthaler, Bulk and single-cell gene expression profiling of sars-cov-2 infected human cell lines identifies molecular targets for therapeutic intervention. *bioRxiv* (2020).

42. Y. Xiong, Y. Liu, L. Cao, D. Wang, M. Guo, A. Jiang, D. Guo, W. Hu, J. Yang, Z. Tang, H. Wu, Y. Lin, M. Zhang, Q. Zhang, M. Shi, Y. Liu, Y. Zhou, K. Lan, Y. Chen, Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in covid-19 patients. *Emerging microbes & infections* **9**, 761–770 (2020).

43. D. Michalovich, N. Rodriguez-Perez, S. Smolinska, M. Pirozynski, D. Mayhew, S. Uddin, S. Van Horn, M. Sokolowska, C. Altunbulakli, A. Eljaszewicz, B. Pugin, W. Barcik, M. Kurnik-Lucka, K. A. Saunders, K. D. Simpson, P. Schmid-Grendelmeier, R. Ferstl, R. Frei, N. Sievi, M. Kohler, P. Gajdanowicz, K. B. Graversen, K. Lindholm Bøgh, M. Jutel, J. R. Brown, C. A. Akdis, E. M. Hessel, L. O'Mahony, Obesity and disease severity magnify disturbed microbiome-immune interactions in asthma patients. *Nature communications* **10**, 1–14 (2019).

44. M. Liao, Y. Liu, J. Yuan, Y. Wen, G. Xu, J. Zhao, L. Cheng, J. Li, X. Wang, F. Wang, L. Liu, I. Amit, S. Zhang, Z. Zhang, Single-cell landscape of bronchoalveolar immune cells in patients with covid-19. *Nature medicine* pp. 1–3 (2020).

45. C. Morse, T. Tabib, J. Sembrat, K. L. Buschur, H. T. Bittar, E. Valenzi, Y. Jiang, D. J. Kass, K. Gibson, W. Chen, A. Mora, P. V. Benos, M. Rojas, R. Lafyatis, Proliferating spp1/mertk-expressing macrophages in idiopathic pulmonary fibrosis. *European Respiratory Journal* **54** (2019).

46. C. Tikellis, M. Thomas, Angiotensin-converting enzyme 2 (ace2) is a key modulator of the renin angiotensin system in health and disease. *International journal of peptides* **2012** (2012).

47. G. Simmons, D. N. Gosalia, A. J. Rennekamp, J. D. Reeves, S. L. Diamond, P. Bates, Inhibitors of cathepsin l prevent severe acute respiratory syndrome coronavirus entry. *Proceedings of the National Academy of Sciences* **102**, 11876–11881 (2005).

48. S. Lukassen, R. L. Chua, T. Trefzer, N. C. Kahn, M. A. Schneider, T. Muley, H. Winter, M. Meister, C. Veith, A. W. Boots, B. P. Hennig, M. Kreuter, C. Conrad, R. Eils, Sars-cov-2 receptor ace 2 and tmprss 2 are primarily expressed in bronchial transient secretory cells. *The EMBO journal* **39**, e105114 (2020).

49. R. Ueha, T. Sato, T. Goto, A. Yamauchi, K. Kondo, T. Yamasoba, Expression of ace2 and tmprss2 proteins in the upper and lower aerodigestive tracts of rats. *bioRxiv* (2020).

50. J. Shilts, G. J. Wright, No evidence for basigin/cd147 as a direct sars-cov-2 spike binding receptor. *bioRxiv* (2020).

51. J. Mercer, A. Helenius, Virus entry by macropinocytosis. *Nature cell biology* **11**, 510–520 (2009).

52. D. L. McKee, A. Sternberg, U. Stange, S. Laufer, C. Naujokat, Candidate drugs against sars-cov-2 and covid-19. *Pharmacological Research* p. 104859 (2020).

53. H. K. Law, C. Y. Cheung, H. Y. Ng, S. F. Sia, Y. O. Chan, W. Luk, J. M. Nicholls, J. Peiris, Y. L. Lau, Chemokine up-regulation in sars-coronavirus–infected, monocyte-derived human dendritic cells. *Blood* **106**, 2366–2374 (2005).

54. C. Y. Cheung, L. L. M. Poon, I. H. Y. Ng, W. Luk, S.-F. Sia, M. H. S. Wu, K.-H. Chan, K.-Y. Yuen, S. Gordon, Y. Guan, J. S. M. Peiris, Cytokine responses in severe acute respiratory syndrome coronavirus-infected macrophages in vitro: possible relevance to pathogenesis. *Journal of virology* **79**, 7819–7826 (2005).

55. L. Li, J. Wo, J. Shao, H. Zhu, N. Wu, M. Li, H. Yao, M. Hu, R. H. Dennin, Sars-coronavirus replicates in mononuclear cells of peripheral blood (pbmcs) from sars patients. *Journal of Clinical Virology* **28**, 239–244 (2003).

56. W. Zhao, X. He, K. A. Hoadley, J. S. Parker, D. N. Hayes, C. M. Perou, Comparison of rna-seq by poly (a) capture, ribosomal rna depletion, and dna microarray for expression profiling. *BMC genomics* **15**, 1–11 (2014).

57. E. W. Sayers, R. Agarwala, E. E. Bolton, J. R. Brister, K. Canese, K. Clark, R. Connor, N. Fiorini, K. Funk, T. Hefferon, J. B. Holmes, S. Kim, A. Kimchi, P. A. Kitts, S. Lathrop, Z. Lu, T. L. Madden, A. Marchler-Bauer, L. Phan, V. A. Schneider, C. L. Schoch, K. D. Pruitt, J. Ostell, Database resources of the national center for biotechnology information. *Nucleic acids research* **36**, D13–D21 (2007).

58. R. Leinonen, R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tárraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, G. Hoad, M. Jang, N. Pakseresht, S. Plaister, R. Radhakrishnan, K. Reddy, S. Sobhany, P. T. Hoopen, R. Vaughan, V. Zalunin, G. Cochrane, The european nucleotide archive. *Nucleic acids research* **39**, D28–D31 (2010).

59. L. Riva, S. Yuan, X. Yin, L. Martin-Sancho, N. Matsunaga, L. Pache, S. Burgstaller-Muehlbacher, P. D. De Jesus, P. Teriete, M. V. Hull, M. W. Chang, J. F. W. Chan, J. Cao, V. K. M. Poon, K. M. Herbert, K. Cheng, T. T. H. Nguyen, A. Rubanov, Y. Pu, C. Nguyen, A. Choi, R. Rathnasinghe, M. Schotsaert, L. Miorin, M. Dejosez, T. P. Zwaka, K. Y. Sit, L. Martinez-Sobrido, W. C. Liu, K. M. White, M. E. Chapman, E. K. Lendy, R. J. Glynne, R. Albrecht, E. Ruppin, A. D. Mesecar, J. R. Johnson, C. Benner, R. Sun, P. G. Schultz, A. I. Su, A. García-Sastre, A. K. Chatterjee, K. Y. Yuen, S. K. Chanda, Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing. *Nature* **586**, 113–119 (2020).

60. Z. Zhang, *et al.*, Database resources of the national genomics data center in 2020. *Nucleic Acids Research* **48**, D24 (2020).

61. A. D. Yates, *et al.*, Ensembl 2020. *Nucleic acids research* **48**, D682–D688 (2020).

62. S. Andrews, Fastqc: a quality control tool for high throughput sequence data (2010).

63. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

64. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic rna-seq quantification. *Nature biotechnology* **34**, 525–527 (2016).

65. H. Pimentel, N. L. Bray, S. Puente, P. Melsted, L. Pachter, Differential analysis of rna-seq incorporating quantification uncertainty. *Nature methods* **14**, 687 (2017).

66. D. Risso, K. Schwartz, G. Sherlock, S. Dudoit, Gc-content normalization for rna-seq data. *BMC bioinformatics* **12**, 480 (2011).

67. M. Carlson, S. Falcon, H. Pages, N. Li, org. hs. eg. db: Genome wide annotation for human. *R package version* **3** (2017).

68. D. J. McCarthy, K. R. Campbell, A. T. Lun, Q. F. Wills, Scater: pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. *Bioinformatics* **33**, 1179–1186 (2017).

69. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with bowtie 2. *Nature methods* **9**, 357 (2012).

70. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079 (2009).

71. R. Kolde, *pheatmap: Pretty Heatmaps* (2019). R package version 1.0.12.

72. Z. Gu, R. Eils, M. Schlesner, Complex heatmaps reveal patterns and correlations in multi-dimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

Acknowledgements: The authors thank professor Ke Xu from Wuhan University and professor Dimitri Lavillette from Institut Pasteur of Shanghai for helpful conversations.

Funding: This work was partially funded by grant 01Kl20185B (SECOVIT) of the German Federal Ministry of Education and Research.

Author Contributions: Pei Hao and Yingying Cao conceived the research. Daniel Hoffmann, Pei Hao, and Yingying Cao designed the analyses. Yingying Cao, Xintian Xu conducted the analyses. All authors wrote the manuscript.

Competing Interests: The authors declare that they have no competing financial interests.

Data and materials availability: Additional data and materials are available online.

Figures and Tables:

Table 1. Data of cell lines (cells) included in this study

Virus	Virus strain	Virus dose (MOI)	Time	Replicates	Species of origin	Cell type	Library preparation	Accession number
SARS-CoV-2	USA-WA1/2020	2	24h	3	Homo sapiens	NHBE	polyA+ selection	GSE147507
Mock	Mock	Mock	24h	3	Homo sapiens	NHBE	polyA+ selection	GSE147507
SARS-CoV-2	USA-WA1/2020	0.2	24h	3	Homo sapiens	A549	polyA+ selection	GSE147507
Mock	Mock	Mock	24h	3	Homo sapiens	A549	polyA+ selection	GSE147507
SARS-CoV-2	USA-WA1/2020	2	24h	3	Homo sapiens	A549	polyA+ selection	GSE147507
Mock	Mock	Mock	24h	3	Homo sapiens	A549	polyA+ selection	GSE147507
SARS-CoV-2	USA-WA1/2020	0.2	24h	3	Homo sapiens	A549-ACE2	polyA+ selection	GSE147507
Mock	Mock	Mock	24h	3	Homo sapiens	A549-ACE2	polyA+ selection	GSE147507
SARS-CoV-2	USA-WA1/2020	2	24h	3	Homo sapiens	A549-ACE2	polyA+ selection	GSE147507
Mock	Mock	Mock	24h	3	Homo sapiens	A549-ACE2	polyA+ selection	GSE147507
SARS-CoV-2	USA-WA1/2020	2	24h	3	Homo sapiens	Calu3	polyA+ selection	GSE147507
Mock	Mock	Mock	24h	3	Homo sapiens	Calu3	polyA+ selection	GSE147507
SARS-CoV-2	Munich/BavPat1/2020	0.3	24h	2	Homo sapiens	Calu3	rRNA-depletion	GSE148729
Mock	Mock	Mock	24h	2	Homo sapiens	Calu3	rRNA-depletion	GSE148729
SARS-CoV-2	Munich/BavPat1/2020	0.3	24h	2	Homo sapiens	Calu3	polyA+ selection	GSE148729
Mock	Mock	Mock	24h	2	Homo sapiens	Calu3	polyA+ selection	GSE148729
SARS-CoV-2	Munich/BavPat1/2020	0.3	24h	2	Homo sapiens	Caco2	polyA+ selection	GSE148729
Mock	Mock	Mock	24h	2	Homo sapiens	Caco2	polyA+ selection	GSE148729
SARS-CoV-2	Munich/BavPat1/2020	0.3	24h	2	Homo sapiens	H1299	polyA+ selection	GSE148729
Mock	Mock	Mock	36h^	2	Homo sapiens	H1299	polyA+ selection	GSE148729
SARS-CoV-2	USA-WA1/2020	0.3	24h	2*	Chlorocebus sabaeus	Vero E6	rRNA-depletion	GSE153940
Mock	Mock	Mock	24h	3	Chlorocebus sabaeus	Vero E6	rRNA-depletion	GSE153940
SARS-CoV	Frankfurt strain	0.3	24h	2	Homo sapiens	Calu3	polyA+ selection	GSE148729
SARS-CoV	Frankfurt strain	0.3	24h	2	Homo sapiens	Calu3	rRNA-depletion	GSE148729
SARS-CoV	Frankfurt strain	0.3	24h	2	Homo sapiens	Caco2	polyA+ selection	GSE148729
SARS-CoV	Frankfurt strain	0.3	24h	2	Homo sapiens	H1299	polyA+ selection	GSE148729
SARS-CoV	Urbani strain	0.1	24h	3	Homo sapiens	MRC5	polyA+ selection	GSE56189
SARS-CoV	Urbani strain	3	24h	3	Homo sapiens	MRC5	polyA+ selection	GSE56189
SARS-CoV	Urbani strain	0.1	24h	3	Chlorocebus sabaeus	Vero E6	polyA+ selection	GSE56189
SARS-CoV	Urbani strain	3	24h	3	Chlorocebus sabaeus	Vero E6	polyA+ selection	GSE56189
MERS-CoV	EMC/2012	0.1	24h	3	Homo sapiens	MRC5	polyA+ selection	GSE56189
MERS-CoV	EMC/2012	3	24h	3	Homo sapiens	MRC5	polyA+ selection	GSE56189
MERS-CoV	EMC/2012	0.1	24h	3	Chlorocebus sabaeus	Vero E6	polyA+ selection	GSE56189
MERS-CoV	EMC/2012	3	24h	3	Chlorocebus sabaeus	Vero E6	polyA+ selection	GSE56189
Mock	Mock	Mock	24h	3	Homo sapiens	MRC5	polyA+ selection	GSE56189
Mock	Mock	Mock	24h	3	Homo sapiens	Vero E6	polyA+ selection	GSE56189

^No corresponding 24h mock control samples for H1299 cells, 36h mock control samples were used instead.
* There are three replicates, but when the manuscript was in preparation only two of them are available for downloading.

Table 2. Data of COVID-19 patients included in this study

Individuals	Tissue	Data Type	Accession number
2	bronchoalveolar lavage fluid from COVID-19 patients	bulk RNA-Seq	CRA002390
3	bronchoalveolar lavage fluid from healthy negative control	bulk RNA-Seq	PRJNA434133^
3	peripheral blood mononuclear cells from COVID-19 patients	bulk RNA-Seq	CRA002390
3	peripheral blood mononuclear cells from healthy negative control	bulk RNA-Seq	CRA002390
2	lung biopsy from postmortem COVID-19 patients	bulk RNA-Seq	GSE147507
2	lung biopsy from healthy negative control	bulk RNA-Seq	GSE147507
6	bronchoalveolar lavage fluid from COVID-19 patients (severe)	single cell RNA-Seq	GSE145926
3	bronchoalveolar lavage fluid from COVID-19 patients (moderate)	single cell RNA-Seq	GSE145926
1	bronchoalveolar lavage fluid from healthy negative control	single cell RNA-Seq	PRJNA526088*

^Three samples under project PRJNA434133: SRR10571724, SRR10571730, and SRR10571732 were used.
* One sample with accession number GSM3660650 under project PRJNA526088 was used.

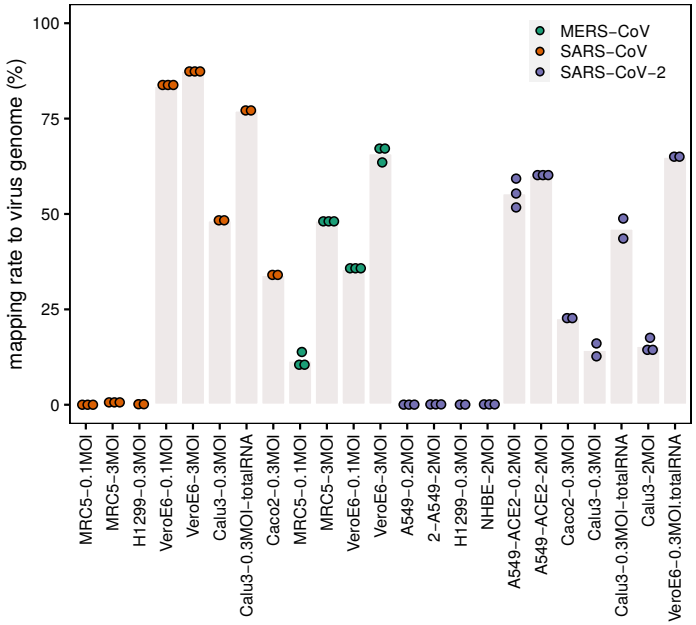


Fig. 1. Mapping rate to virus genome. The dots represent the mapping rates to the virus genome for each individual replicate under the given conditions (cell line, MOI, and virus). Bar heights are mean mapping rates to the virus genome for each condition.

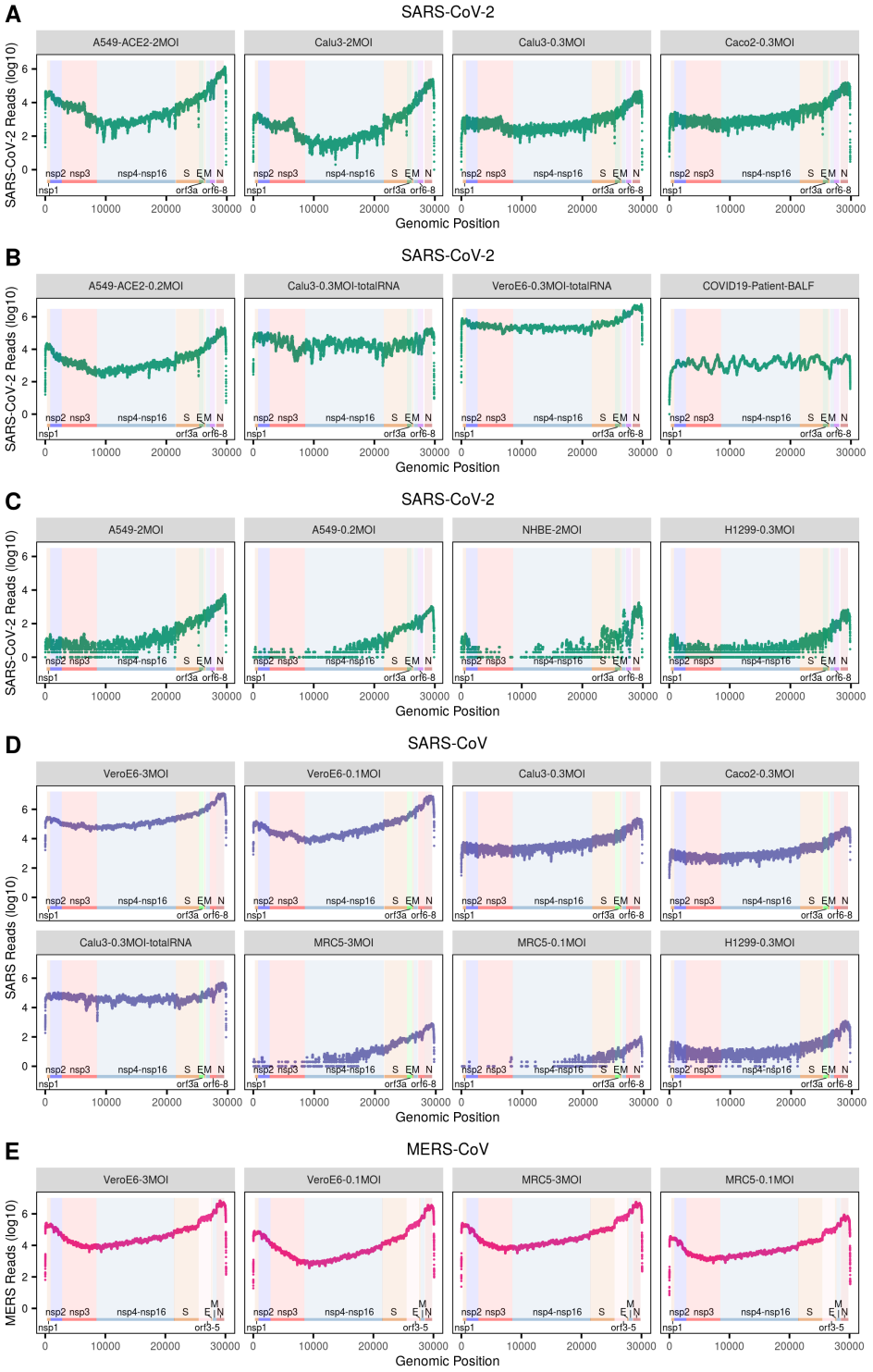


Fig. 2. The number of reads mapped to the corresponding virus genome. (**A-E**) The dot plots show the number of reads mapped to each site of the corresponding virus genome. The annotation of the genome of each virus is from NCBI (SARS: GCF_000864885.1, SARS-CoV-2: GCF_009858895.2, MERS: GCF_000901155.1). Labels in grey title bars correspond to conditions as in Fig. 1.

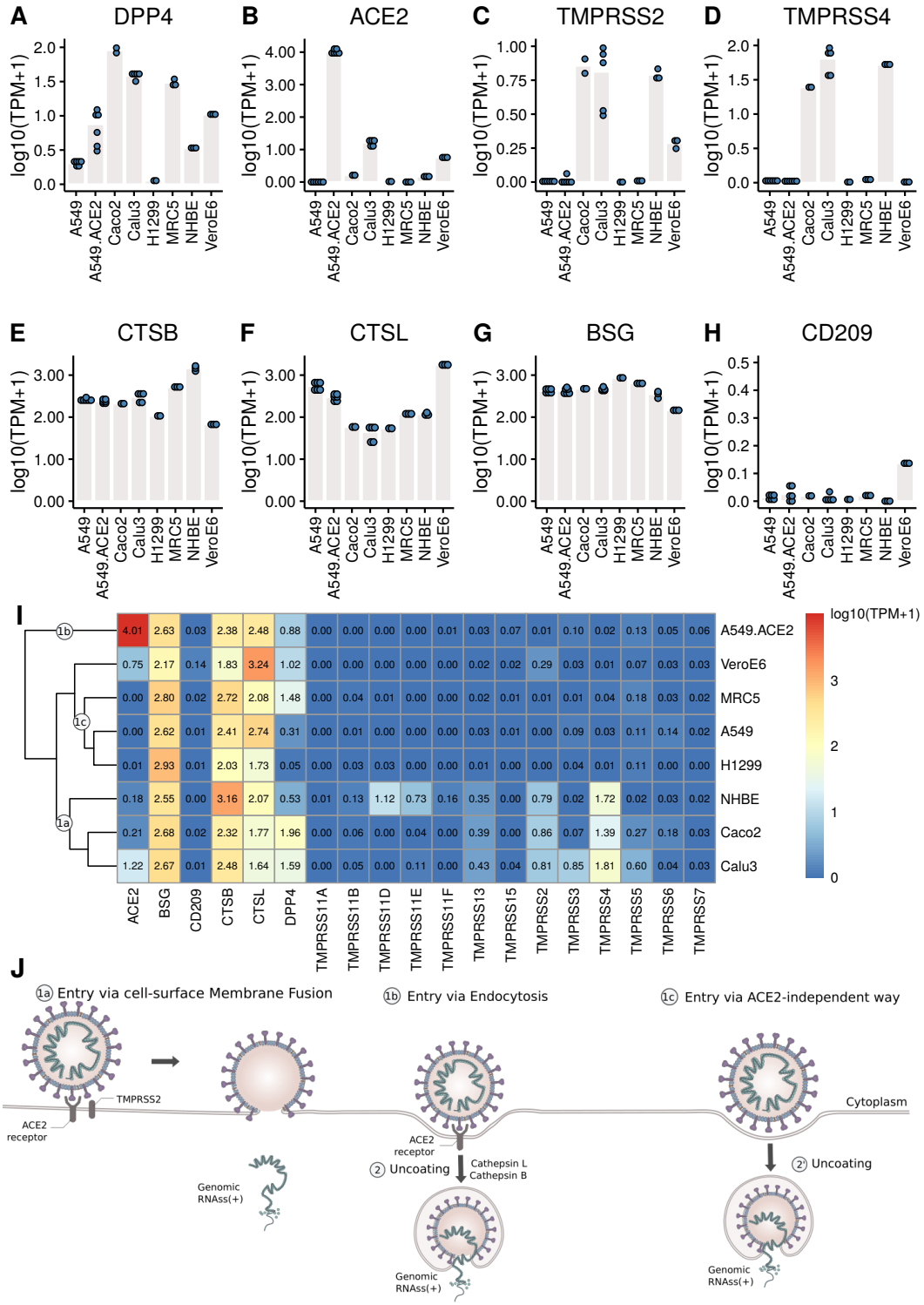


Fig. 3. The expression levels of the receptors and proteases. (**A-H**) Each dot represents the expression value in each sample. (**I**) Heatmap of the expression levels of coronavirus associated **receptors** and factors of different cell types. Labels 1a, 1b, 1c mark cell clusters that likely share entry routes sketched in panel **J**. (**J**) Entry mechanisms involved in SARS-CoV-2 entry into cells. Schematic is based on a figure by Vega Asensio - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=88682468.

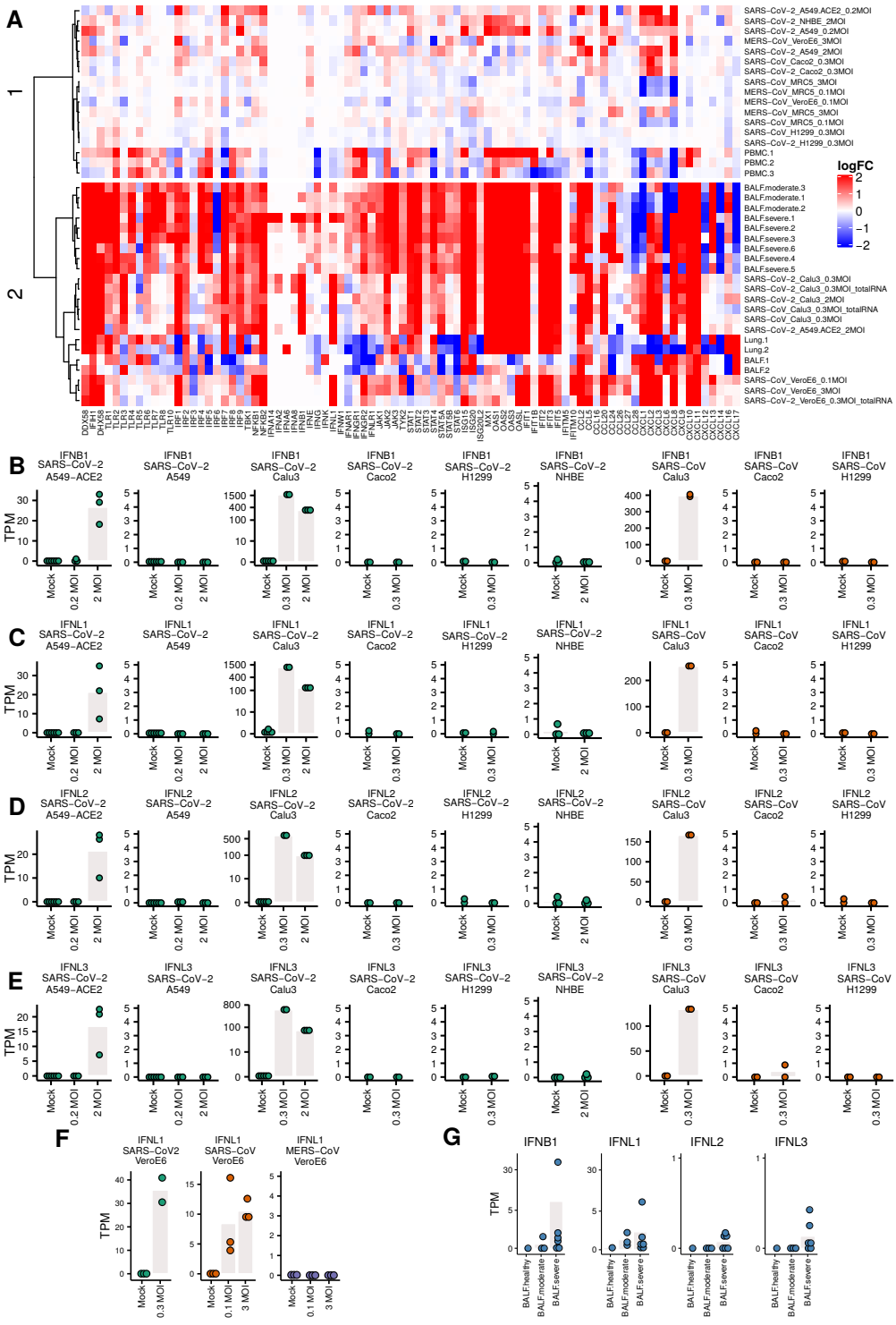


Fig. 4. Expression levels of genes related to immune responses (**A**) Heatmap of the logFC of IFNs, ISGs and pro-inflammatory cytokines. The clustering of samples produces a cluster 1 (top) with little IFN/ISG expression comprising MERS infections and non-infectable cells/SARS-CoV-1/2 (except for Caco2 cells), and a cluster 2 (bottom) strong IFN/ISG expression with SARS-CoV-1/2 infectable cells and patient samples. (**B-G**) Expression levels of IFNs. Each dot represents the expression value of a sample. Bars indicate mean expression levels (in TPM) of respective IFN at different MOI values.

Supplementary

Yingying Cao[1*], Xintian Xu[2], Simo Kitanovski[1], Lina Song[3], Jun Wang[1], Pei Hao[2,4*], Daniel Hoffmann[1*]

[1]Bioinformatics and Computational Biophysics,
Faculty of Biology and Center for Medical Biotechnology,
University of Duisburg-Essen, Essen 45141, Germany
[2]Key Laboratory of Molecular Virology and Immunology,
Institut Pasteur of Shanghai, Center for Biosafety Mega-Science,
Chinese Academy of Sciences, Shanghai 200031, China
[3]Translational Skin Cancer Research,
German Consortium for Translational Cancer Research, Essen, Germany
[4]The Joint Program in Infection and Immunity:
a. Guangzhou Women and Children's Medical Center,
Guangzhou Medical University, Guangzhou 510623, China;
b. Institut Pasteur of Shanghai,
Chinese Academy of Sciences, Shanghai 200031, China

[*]To whom correspondence should be addressed;
E-mail: daniel.hoffmann@uni-due.de, phao@ips.ac.cn, yingying.cao@uni-due.de.

1 Additional information about public data

All data can be downloaded from public repositories, the three main sources are NCBI (*1*) (https://www.ncbi.nlm.nih.gov/) and ENA (*2*) (https://www.ebi.ac.uk/ena) and BIG Data Center (*3*) (https://bigd.big.ac.cn/).

1.1 GSE147507 dataset (*4*)

From this dataset we downloaded: Biological triplicates of primary human lung epithelium (NHBE) which were mock treated or infected with SARS-CoV-2 (USA-WA1/2020) at an MOI

of 2; Biological triplicates of transformed lung alveolar (A549) cells which were mock treated or infected with SARS-CoV-2 (USA-WA1/2020) at an MOI of 0.2 or 2; Biological triplicates of transformed lung alveolar (A549) transduced with a vector expressing human ACE2, which were also mock treated or infected with SARS-CoV-2 (USA-WA1/2020) at an MOI of 0.2 or 2; Biological triplicates of transformed lung-derived Calu-3 cells which were mock treated or infected with SARS-CoV-2 (USA-WA1/2020) at an MOI of 2; COVID-19 patient samples: Uninfected human lung biopsies derived from one male (age 72) and one female (age 60) and used as control biological replicates, and lung samples derived from a single male COVID-19 deceased patient (age 74) which were processed in technical replicates. Library preparation method polyA+ selection was used to remove rRNAs before sequencing.

1.2 GSE148729 dataset (*5*)

From this dataset we downloaded biological replicates of Calu-3, Caco-2 and H1299 cells which were mock treated or infected with SARS-CoV-2 (patient isolate BetaCoV/Munich/BavPat1/2020/EPI_ISL_406862) or SARS-CoV (Frankfurt strain) at an MOI of 0.3. Library preparation method polyA+ selection was used to remove rRNAs before sequencing Caco-2 and H1299 cells. For Calu-3 cells, two library preparation method polyA+ selection and rRNA-depletion were used respectively to remove rRNAs before sequencing.

1.3 GSE153940 dataset

From this dataset we downloaded RNA sequencing data of Vero E6 cells which were either mock-infected or infected with SARS-CoV-2 USA-WA1/2020 (MOI = 0.3) with three replicates. However, when we downloaded the data one sample with accession number GSM4658806 was not available for downloading. Cells were harvested at 24 hours after infection, and rRNA-depletion method was used to extract RNA for sequencing.

1.4 GSE56189 dataset

From this dataset we downloaded: Biological triplicates of MRC5 and Vero E6 cells which were mock treated or infected with SARS-CoV (Urbani strain) or MERS-CoV (EMC/2012) at an MOI of 0.1 or 3. Library preparation method polyA+ selection was used to remove rRNAs before sequencing.

1.5 CRA002390 dataset (*6*)

This dataset is public available in https://bigd.big.ac.cn/gsa/browse/CRA002390. From this dataset we downloaded: The raw FASTQ data of PBMC and BALF samples of COVID-19 patients and corresponding PBMC controls.

1.6 PRJNA434133 dataset (*7*)

From this dataset we downloaded the raw FASTQ data for BALF healthy control samples with accession numbers SRR10571724, SRR10571730, and SRR10571732.

1.7 GSE145926 dataset (*8*)

From this dataset we downloaded the preprocessed single cell RNA-Seq data of BALF samples from 6 severe COVID-19 patients and 3 mild COVID-19 patients.

1.8 PRJNA526088 dataset (*9*)

From this dataset we downloaded the preprocessed single cell RNA-Seq data of BALF sample from a healthy control with accession number GSM3660650.

2 Supplementary figures

Fig. S1:

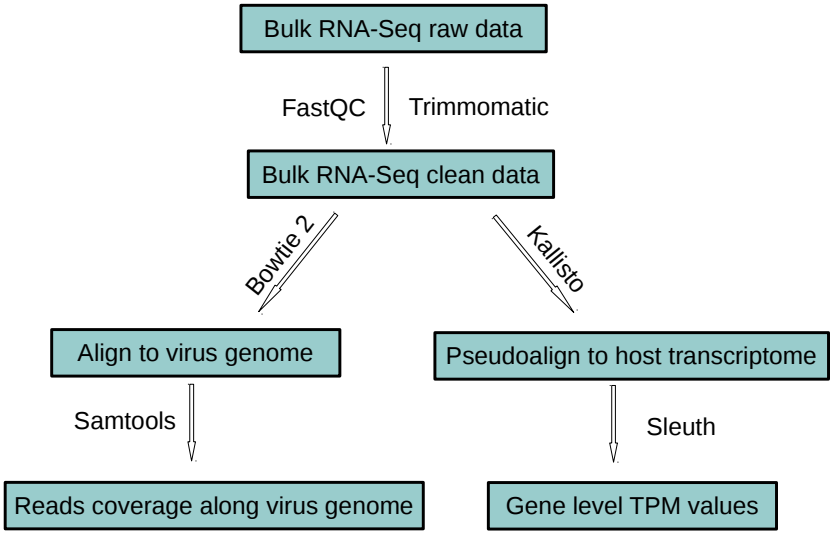


Fig. S2:

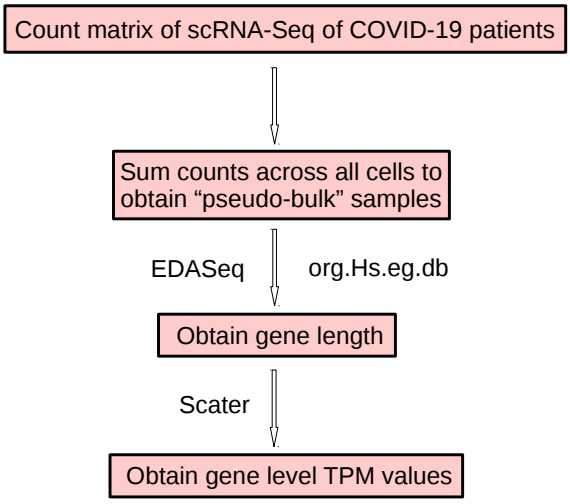


Fig. S3:

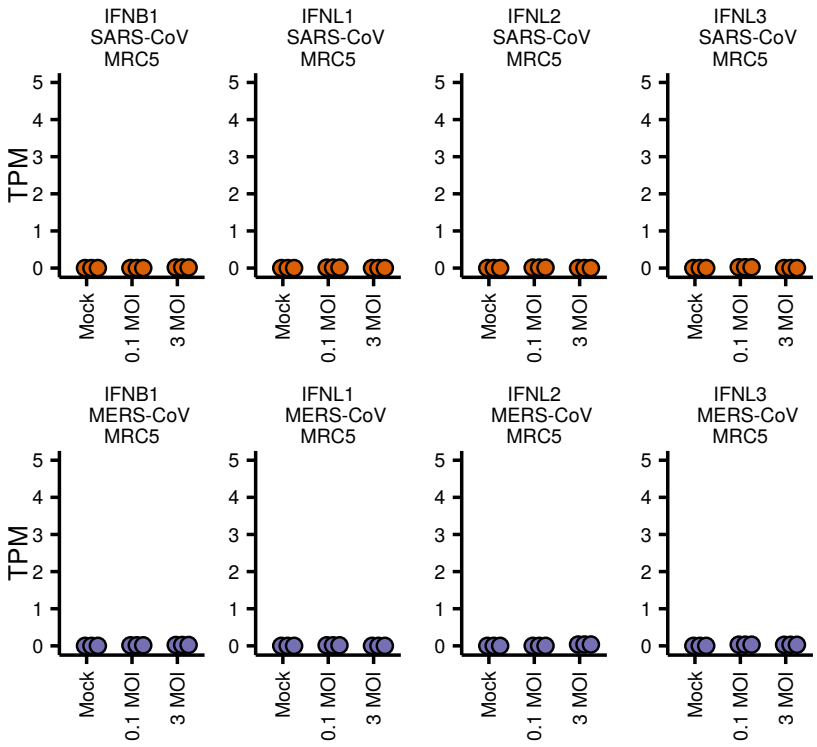

IFNB1
SARS-CoV
MRC5
IFNL1
SARS-CoV
MRC5
IFNL2
SARS-CoV
MRC5
IFNL3
SARS-CoV
MRC5
IFNB1
MERS-CoV
MRC5
IFNL1
MERS-CoV
MRC5
IFNL2
MERS-CoV
MRC5
IFNL3
MERS-CoV
MRC5
TPM
0
1
2
3
4
5
Mock
0.1 MOI
3 MOI

Fig. S4:

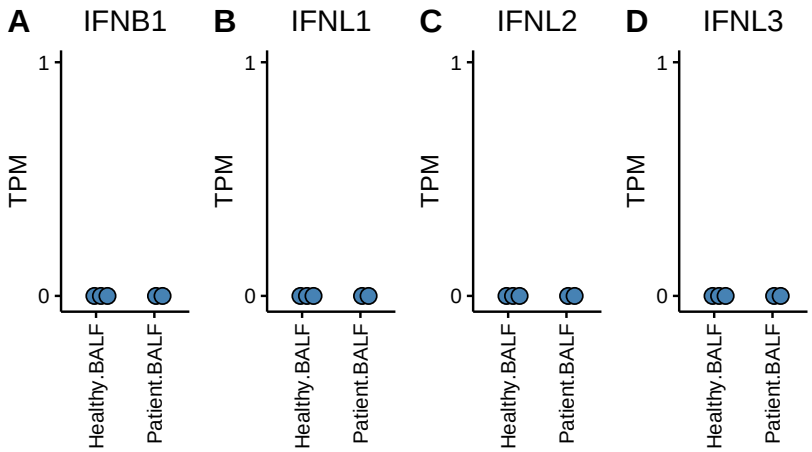

A
IFNB1
B
IFNL1
C
IFNL2
D
IFNL3
TPM
1
0
Healthy.BALF
Patient.BALF

Fig. S5:

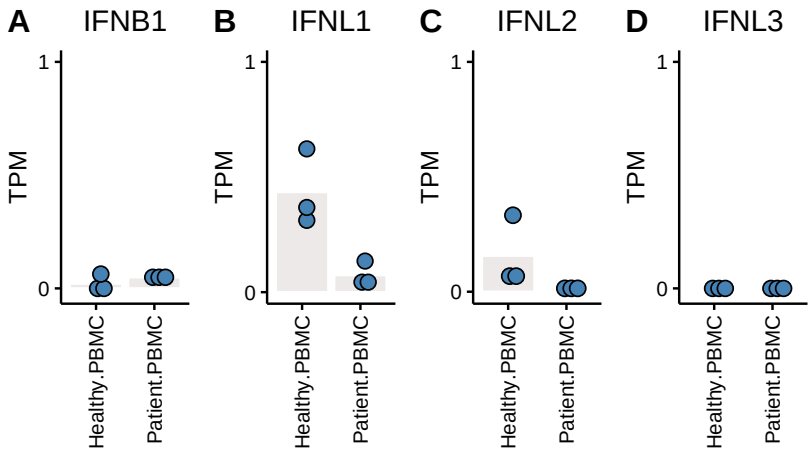

A
IFNB1
B
IFNL1
C
IFNL2
D
IFNL3
TPM
1
0
Healthy.PBMC
Patient.PBMC

Fig. S6:

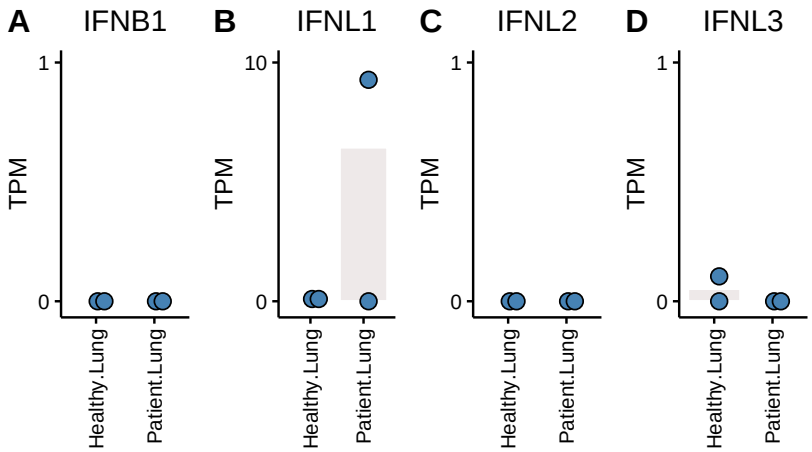

A
IFNB1
B
IFNL1
C
IFNL2
D
IFNL3
TPM
1
0
10
Healthy.Lung
Patient.Lung

Fig. S7:

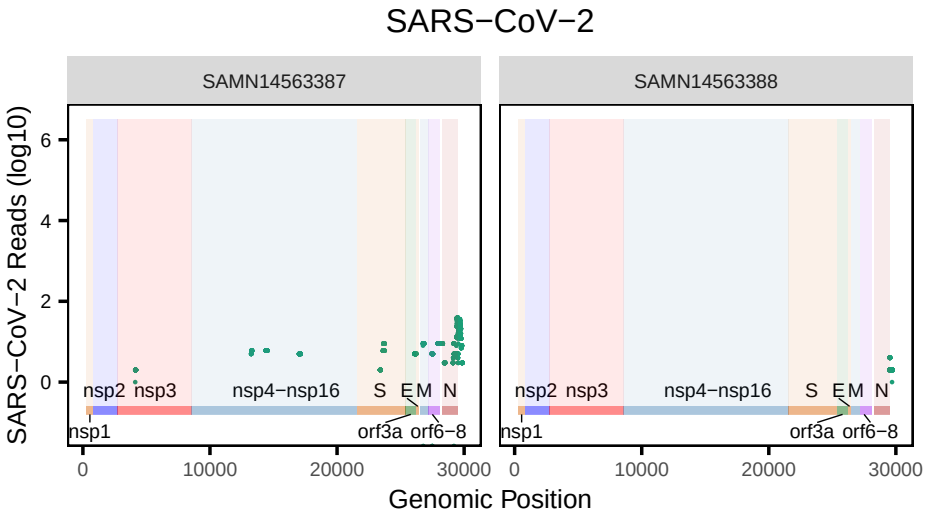
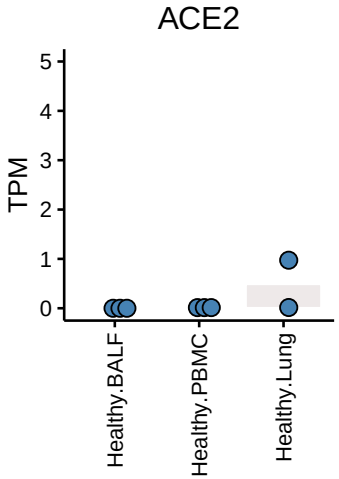

SARS−CoV−2
SAMN14563387
SAMN14563388
SARS−CoV−2 Reads (log10)
6
4
2
0
nsp2 nsp3
nsp4−nsp16
S
E
M
N
nsp1
orf3a
orf6−8
0
10000
20000
30000
Genomic Position

Fig. S8:



References

1. D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, *et al.*, Database resources of the national center for biotechnology information. *Nucleic acids research* **36**, D13–D21 (2007).

2. R. Leinonen, R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tárraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, *et al.*, The european nucleotide archive. *Nucleic acids research* **39**, D28–D31 (2010).

3. N. G. D. C. Members, *et al.*, Database resources of the national genomics data center in 2020. *Nucleic Acids Research* **48**, D24 (2020).

4. D. Blanco-Melo, B. E. Nilsson-Payant, W.-C. Liu, S. Uhl, D. Hoagland, R. Møller, T. X. Jordan, K. Oishi, M. Panis, D. Sachs, *et al.*, Imbalanced host response to sars-cov-2 drives development of covid-19. *Cell* (2020).

5. E. Wyler, K. Mösbauer, V. Franke, A. Diag, L. T. Gottula, R. Arsie, F. Klironomos, D. Koppstein, S. Ayoub, C. Buccitelli, *et al.*, Bulk and single-cell gene expression profiling of sars-cov-2 infected human cell lines identifies molecular targets for therapeutic intervention. *bioRxiv* (2020).

6. Y. Xiong, Y. Liu, L. Cao, D. Wang, M. Guo, A. Jiang, D. Guo, W. Hu, J. Yang, Z. Tang, *et al.*, Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in covid-19 patients. *Emerging microbes & infections* **9**, 761–770 (2020).

7. D. Michalovich, N. Rodriguez-Perez, S. Smolinska, M. Pirozynski, D. Mayhew, S. Uddin, S. Van Horn, M. Sokolowska, C. Altunbulakli, A. Eljaszewicz, *et al.*, Obesity and disease severity magnify disturbed microbiome-immune interactions in asthma patients. *Nature communications* **10**, 1–14 (2019).

8. M. Liao, Y. Liu, J. Yuan, Y. Wen, G. Xu, J. Zhao, L. Cheng, J. Li, X. Wang, F. Wang, *et al.*, Single-cell landscape of bronchoalveolar immune cells in patients with covid-19. *Nature medicine* pp. 1–3 (2020).

9. C. Morse, T. Tabib, J. Sembrat, K. L. Buschur, H. T. Bittar, E. Valenzi, Y. Jiang, D. J. Kass, K. Gibson, W. Chen, *et al.*, Proliferating spp1/mertk-expressing macrophages in idiopathic pulmonary fibrosis. *European Respiratory Journal* **54** (2019).

Chapter 4

Discussion & outlook

It is an exciting time for modern biology with powerful performance of sequencing technologies and their rapid, active development. However, computational challenges come along with the advantages. With the decreasing cost of sequencing and the advances of high-throughput sequencing technologies in multi-omics such as genomics, epigenomics, transcriptomics, and proteomics (Sun and Hu, 2016), it is imperative to take an integrative approach to gain insights into the complex biological processes at multiple levels simultaneously to overcome the limitation of single layer biological information provided by single-omics. For example, not only can we evaluate the differential gene expression levels in different conditions with RNA-seq data, we can also examine the changes in signals of ChIP-seq of different histone modification marks to better understand the gene regulation during development or differentiation. However, such integrative tools are still very limited. In (Cao et al., 2020), we developed an R package named intePareto to conduct such a task to prioritize genes with consistent changes in both RNA-seq and ChIP-seq data of different histone modifications across different conditions using Pareto optimization. Our easy-to-use approach to integrate such data and can be extended and generalized to other data types as well in the future as long as the data can be quantified and logFC can be calculated. An integration method as we proposed here will yield better understanding and clearer picture of the biological processions for example during development or disease progression.

One of the main challenges in the computational analysis of scRNA-seq data is the abundance of observed zeros, or dropouts. Over the past few years, lots of imputation tools (Gong et al., 2018; Huang et al., 2018; Li and Li, 2018; Van Dijk et al., 2018) have been developed for the computational correction of zeros, and zero-inflation model is also universally widely used for modeling scRNA-seq data (Kharchenko et al., 2014; Lopez et al., 2018; Pierson and Yau, 2015) without further detailed exploration of source of zero-inflation. Imputation has the danger of introducing flase signals, oversmoothing of the data and removing biologically meaningfully variation in gene expression which could reflect the natural cellular heterogeneity (Andrews and Hemberg, 2018; Hou et al., 2020). Svensson (Svensson, 2020) also points out previous misunderstanding of "zero-

inflation" scRNA-seq data from droplet-based protocols. In (Cao et al., 2021a), we further demonstrate that the observed zero-inflation in read counts are mainly attributed to amplification bias not due to sequencing platforms by taking advantage of the fact that one can count reads as well as UMIs at the same time and showing that in the same data set, the read counts are zero-inflated, while the UMI counts are barely zero-inflated. Recently several studies show that the variation in the observed zeros actually reflects biological variation (Bouland et al., 2021; Choi et al., 2020; Qiu, 2020), of which by binarizing single-cell RNA-seq count data as zeros and non-zeros to do further downstream analysis (Qiu, 2020), or conducting differential dropout analysis to identify the biological differences (Bouland et al., 2021). These most recent studies, including ours, all argue against imputation and suggest zeros are informative should remain as zeros without modification for further examination of transcriptional and cellular diversity inherent in the original data.

We believe differential dropout analysis can be a quite interesting direction in addition to differential gene expression analysis, for example more further analyses can be conducted by taking advantage of the full, quantitative data in addition to the binarization of zeros and non-zeros. Another interesting direction can be denoising of read counts data. As we have already showed in (Cao et al., 2021a), we can take advantages of read counts as well as UMIs in the same data set to examine the behavior of amplification bias and then apply this relationship to denoise read counts from protocols that can not incorporate UMIs. One similar method (Townes and Irizarry, 2020) has been published recently but the influence of this practice on the further downstream analysis still needs further careful examination and benchmark.

In (Cao et al., 2021b) and (Wang et al., 2020), we have conducted comprehensive large scale data analyses to examine the biological processes during SARS-CoV-2 infection. We include RNA-seq data of infections of different similar coronaviruses including SARS-CoV, MERS-CoV and SARS-CoV-2; different cell types; and different viral doses to examine the innate immune responses as well as the entry of virus in different cell types. Our analysis supports the existence of alternative entry routes of SARS-CoV-2 and SARS. We also find that innate immune responses in terms of IFNs and ISGs vary strongly as function of cell types, virus doses, and virus types. We believe such a comprehensive analysis is needed to better understand not only virus infection but also other diseases as well. Conclusion made from single study can be limited and a comprehensive comparison study like this can provide us unexpected novel information unobserved or hidden in a single study.

The public sequencing data grow exponentially (Cook et al., 2020; Svensson et al., 2018). They are unprecedented, vast and valuable resources which are free publicly available for example at EMBL-EBI (European Molecular Biology Laboratory-European Bioinformatics Institute) archive (`https://www.ebi.ac.uk/`), GEO (`https://www.ncbi.nlm.nih.gov/geo/`) public repository and NGDC(National Genomics Data Center, `https:`

//bigd.big.ac.cn/) for computational method development as well as for biological discovery as we have shown in the examples above. We believe in the future more and more integrative sequencing data analyses, comprehensive large scale meta-studies and method development studies based on the public available sequencing data sets will further promote advances in biological and medical research.

Bibliography

Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. The structure and function of dna. In *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

Robert A Amezquita, Aaron TL Lun, Etienne Becht, Vince J Carey, Lindsay N Carpp, Ludwig Geistlinger, Federico Martini, Kevin Rue-Albrecht, Davide Risso, Charlotte Soneson, et al. Orchestrating single-cell analysis with bioconductor. *Nature methods*, pages 1–9, 2019.

Benedict Anchang, Tom DP Hart, Sean C Bendall, Peng Qiu, Zach Bjornson, Michael Linderman, Garry P Nolan, and Sylvia K Plevritis. Visualization and cellular hierarchy inference of single-cell data using spade. *Nature protocols*, 11(7):1264–1279, 2016.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.

Tallulah S Andrews and Martin Hemberg. False signals induced by single-cell imputation. *F1000Research*, 7, 2018.

Andrew J Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell research*, 21(3):381–395, 2011.

Andrew J Bannister, Robert Schneider, Fiona A Myers, Alan W Thorne, Colyn Crane-Robinson, and Tony Kouzarides. Spatial distribution of di-and tri-methyl lysine 36 of histone h3 at active genes. *Journal of Biological Chemistry*, 280(18):17732–17736, 2005.

Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007.

Christian Beisel and Renato Paro. Silencing chromatin: comparing modes and mechanisms. *Nature Reviews Genetics*, 12(2):123–135, 2011.

Sandrine Belouzard, Jean K Millet, Beth N Licitra, and Gary R Whittaker. Mechanisms of coronavirus cell entry mediated by the viral spike protein. *Viruses*, 4(6):1011–1033, 2012.

Elizaveta V Benevolenskaya. Histone h3k4 demethylases are essential in development and differentiation. *Biochemistry and cell biology*, 85(4):435–443, 2007.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Daniel Blanco-Melo, Benjamin E. Nilsson-Payant, Wen Chun Liu, Skyler Uhl, Daisy Hoagland, Rasmus Møller, Tristan X. Jordan, Kohei Oishi, Maryline Panis, David Sachs, Taia T. Wang, Robert E. Schwartz, Jean K. Lim, Randy A. Albrecht, and Benjamin R. TenOever. Imbalanced host response to sars-cov-2 drives development of covid-19. *Cell*, 2020.

Gerard A Bouland, Ahmed Mahfouz, and Marcel JT Reinders. Differential dropout analysis captures biological variation in single-cell rna sequencing data. *bioRxiv*, 2021.

Andrew G Bowie and Ismar R Haga. The role of toll-like receptors in the host response to viruses. *Molecular immunology*, 42(8):859–867, 2005.

Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5):525–527, 2016.

Achille Broggi, Sreya Ghosh, Benedetta Sposito, Roberto Spreafico, Fabio Balzarini, Antonino Lo Cascio, Nicola Clementi, Maria de Santis, Nicasio Mancini, Francesca Granucci, and Ivan Zanoni. Type iii interferons disrupt the lung epithelial barrier upon viral recognition. *Science*, 2020.

Yingying Cao, Simo Kitanovski, and Daniel Hoffmann. intepareto: An r package for integrative analyses of rna-seq and chip-seq data. *BMC genomics*, 21(11):1–9, 2020.

Yingying Cao, Simo Kitanovski, Ralf Küppers, and Daniel Hoffmann. Umi or not umi, that is the question for scrna-seq zero-inflation. *Nature Biotechnology*, pages 1–2, 2021a.

Yingying Cao, Xintian Xu, Simo Kitanovski, Lina Song, Jun Wang, Pei Hao, and Daniel Hoffmann. Comprehensive comparison of transcriptomes in sars-cov-2 infection: alternative entry routes and innate immune responses. *bioRxiv*, pages 2021–01, 2021b.

James J Chen, Paula K Robeson, and Michael J Schell. The false discovery rate: a key concept in large-scale genetic studies. *Cancer Control*, 17(1):58–62, 2010.

Wenan Chen, Yan Li, John Easton, David Finkelstein, Gang Wu, and Xiang Chen. Umi-count modeling and differential expression analysis for single-cell rna sequencing. *Genome biology*, 19(1):70, 2018.

Yiwen Chen, Nicolas Negre, Qunhua Li, Joanna O Mieczkowska, Matthew Slattery, Tao Liu, Yong Zhang, Tae-Kyung Kim, Housheng Hansen He, Jennifer Zieba, et al. Systematic evaluation of factors influencing chip-seq fidelity. *Nature methods*, 9(6):609, 2012.

Cindy Chiang and Michaela U Gack. Post-translational control of intracellular pathogen sensing pathways. *Trends in immunology*, 38(1):39–52, 2017.

Kwangbom Choi, Yang Chen, Daniel A Skelly, and Gary A Churchill. Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. *Genome biology*, 21(1):1–16, 2020.

Charles E Cook, Oana Stroe, Guy Cochrane, Ewan Birney, and Rolf Apweiler. The european bioinformatics institute in 2020: building a global infrastructure of interconnected data resources for the life sciences. *Nucleic acids research*, 48(D1):D17–D23, 2020.

Menno P Creyghton, Albert W Cheng, G Grant Welstead, Tristan Kooistra, Bryce W Carey, Eveline J Steine, Jacob Hanna, Michael A Lodato, Garrett M Frampton, Phillip A Sharp, et al. Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107 (50):21931–21936, 2010.

Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

Xianjun Dong, Melissa C Greven, Anshul Kundaje, Sarah Djebali, James B Brown, Chao Cheng, Thomas R Gingeras, Mark Gerstein, Roderic Guigó, Ewan Birney, et al. Modeling gene expression using chromatin features in various cellular contexts. *Genome biology*, 13(9):R53, 2012.

Olive Jean Dunn. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961.

Ciaran Evans, Johanna Hardin, and Daniel M Stoebel. Selecting between-sample rna-seq normalization methods from the perspective of their assumptions. *Briefings in bioinformatics*, 19(5):776–792, 2018.

Anthony R Fehr and Stanley Perlman. Coronaviruses: an overview of their replication and pathogenesis. In *Coronaviruses*, pages 1–23. Springer, 2015.

Silvia Giulia Galfre and Francesco Morandin. A mathematical framework for raw counts of single-cell rna-seq data analysis. *arXiv preprint arXiv:2002.02933*, 2020.

ER Gibney and CM Nolan. Epigenetics and gene expression. *Heredity*, 105(1):4–13, 2010.

Tomás Gomes, Sarah A Teichmann, and Carlos Talavera-López. Immunology driven by large-scale single-cell sequencing. *Trends in immunology*, 2019.

Wuming Gong, Il-Youp Kwak, Pruthvi Pota, Naoko Koyano-Nakagawa, and Daniel J Garry. Drimpute: imputing dropout events in single cell rna sequencing data. *BMC bioinformatics*, 19(1):1–10, 2018.

Laura González-Silva, Laura Quevedo, and Ignacio Varela. Tumor functional heterogeneity unraveled by scrna-seq technologies. *Trends in cancer*, 6(1):13–19, 2020.

Michael Hagemann-Jensen, Christoph Ziegenhain, Ping Chen, Daniel Ramsköld, Gert-Jan Hendriks, Anton JM Larsson, Omid R Faridani, and Rickard Sandberg. Single-cell rna counting at allele and isoform resolution using smart-seq3. *Nature Biotechnology*, 38(6):708–714, 2020.

Arif Harmanci, Joel Rozowsky, and Mark Gerstein. Music: identification of enriched regions in chip-seq experiments using a mappability-corrected multiscale signal processing framework. *Genome biology*, 15(10):1–15, 2014.

Tamar Hashimshony, Naftalie Senderovich, Gal Avital, Agnes Klochendler, Yaron de Leeuw, Leon Anavy, Dave Gennert, Shuqiang Li, Kenneth J Livak, Orit Rozenblatt-Rosen, et al. Cel-seq2: sensitive highly-multiplexed single-cell rna-seq. *Genome biology*, 17(1):77, 2016.

Feria Hikmet, Loren Méar, Åsa Edvinsson, Patrick Micke, Mathias Uhlén, and Cecilia Lindskog. The protein expression profile of ace2 in human tissues. *Molecular Systems Biology*, 16(7):e9610, 2020.

Stephen A Hoang, Xiaojiang Xu, and Stefan Bekiranov. Quantification of histone modification chip-seq enrichment for data mining and machine learning applications. *BMC research notes*, 4(1):288, 2011.

Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S. Schiergens, Georg Herrler, Nai Huei Wu, Andreas Nitsche, Marcel A. Müller, Christian Drosten, and Stefan Pöhlmann. Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *Cell*, 2020.

Wenpin Hou, Zhicheng Ji, Hongkai Ji, and Stephanie C Hicks. A systematic evaluation of single-cell rna-sequencing imputation methods. *Genome biology*, 21(1):1–30, 2020.

Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. Saver: gene expression recovery for single-cell rna sequencing. *Nature methods*, 15(7):539–542, 2018.

Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell rna-seq with unique molecular identifiers. *Nature methods*, 11(2):163, 2014.

Thomas Jenuwein and C David Allis. Translating the histone code. *Science*, 293(5532): 1074–1080, 2001.

Rosa Karlić, Ho-Ryun Chung, Julia Lasserre, Kristian Vlahovičcek, and Martin Vingron. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*, 107(7):2926–2931, 2010.

Takumi Kawasaki and Taro Kawai. Toll-like receptor signaling pathways. *Frontiers in immunology*, 5:461, 2014.

Raghuvir Keni, Anila Alexander, Pawan Ganesh Nayak, Jayesh Mudgal, and Krishnadas Nandakumar. Covid-19: emergence, spread, possible treatments, and global burden. *Frontiers in public health*, 8:216, 2020.

Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–742, 2014.

Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):1–13, 2013.

Daehwan Kim, Ben Langmead, and Steven L Salzberg. Hisat: a fast spliced aligner with low memory requirements. *Nature methods*, 12(4):357–360, 2015.

Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, 9(1):72–74, 2012.

Christoph M Koch, Robert M Andrews, Paul Flicek, Shane C Dillon, Ulaş Karaöz, Gayle K Clelland, Sarah Wilcox, David M Beare, Joanna C Fowler, Phillippe Couttet, et al. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome research*, 17(6):691–707, 2007.

Tony Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007a.

Tony Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007b.

Thijs Kuiken, Ron A.M. Fouchier, Martin Schutten, Guus F. Rimmelzwaan, Geert Van Amerongen, Debby Van Riel, Jon D. Laman, Ton De Jong, Gerard Van Doornum, Wilina Lim, Ai Ee Ling, Paul K.S. Chan, John S. Tam, Maria C. Zambon, Robin Gopal,

Christian Drosten, Sylvie Van Der Werf, Nicolas Escriou, Jean Claude Manuguerra, Klaus Stöhr, J. S.Malik Peiris, and Albert D.M.E. Osterhaus. Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome. *The Lancet*, 362 (9380):263–270, 2003.

Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357, 2012.

Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10 (3):1–10, 2009.

Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.

Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications*, 9(1):1–9, 2018.

Yang Liao, Gordon K Smyth, and Wei Shi. The r package rsubread is easier, faster, cheaper and better for alignment and quantification of rna sequencing reads. *Nucleic acids research*, 47(8):e47–e47, 2019.

Yueh-Ming Loo and Michael Gale Jr. Immune signaling by rig-i-like receptors. *Immunity*, 34(5):680–692, 2011.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Zhiyong Lou, Yuna Sun, and Zihe Rao. Current progress in antiviral strategies. *Trends in pharmacological sciences*, 35(2):86–102, 2014.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

Karolin Luger, Armin W Mäder, Robin K Richmond, David F Sargent, and Timothy J Richmond. Crystal structure of the nucleosome core particle at 2.8 å resolution. *Nature*, 389(6648):251–260, 1997.

Aaron TL Lun and Gordon K Smyth. csaw: a bioconductor package for differential binding analysis of chip-seq data using sliding windows. *Nucleic acids research*, 44(5): e45–e45, 2016.

Ismail Sami Mahmoud, Yazun Bashir Jarrar, Walhan Alshaer, and Said Ismail. Sars-cov-2 entry in host cells-multiple targets for treatment and prevention. *Biochimie*, 2020.

J.L. Manley and D.C. Di Giammartino. mrna polyadenylation in eukaryotes. In William J. Lennarz and M. Daniel Lane, editors, *Encyclopedia of Biological Chemistry (Second Edition)*, pages 188–193. Academic Press, Waltham, second edition edition, 2013. ISBN 978-0-12-378631-9. doi: https://doi.org/10.1016/B978-0-12-378630-2.00624-1. URL `https://www.sciencedirect.com/science/article/pii/B9780123786302006241`.

Anne C Mirabella, Benjamin M Foster, and Till Bartke. Chromatin deregulation in disease. *Chromosoma*, 125(1):75–93, 2016.

Kiyoshisa Mizumoto and Yoshito Kaziro. Messenger rna capping enzymes from eukaryotic cells. *Progress in nucleic acid research and molecular biology*, 34:1–28, 1987.

Ryuichiro Nakato and Katsuhiko Shirahige. Recent advances in chip-seq analysis: from quality management to whole-genome annotation. *Briefings in bioinformatics*, 18(2): 279–290, 2017.

Victor Naumenko, Madison Turk, Craig N Jenne, and Seok-Joo Kim. Neutrophils in viral infection. *Cell and tissue research*, 371(3):505–516, 2018.

Alicia Oshlack and Matthew J Wakefield. Transcript length bias in rna-seq data confounds systems biology. *Biology direct*, 4(1):1–10, 2009.

Efthymia Papalexi and Rahul Satija. Single-cell rna sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, 18(1):35, 2018.

Annsea Park and Akiko Iwasaki. Type i and type iii interferons–induction, signaling, evasion, and application to combat covid-19. *Cell Host & Microbe*, 2020.

Peter J Park. Chip–seq: advantages and challenges of a maturing technology. *Nature reviews genetics*, 10(10):669–680, 2009.

Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4):417, 2017.

Craig L Peterson and Marc-André Laniel. Histones and histone modifications. *Current Biology*, 14(14):R546–R551, 2004.

Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, 9(1):171–181, 2014.

Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16(1):1–10, 2015.

Harold Pimentel, Nicolas L Bray, Suzette Puente, Páll Melsted, and Lior Pachter. Differential analysis of rna-seq incorporating quantification uncertainty. *Nature methods*, 14 (7):687, 2017.

Peng Qiu. Embracing the dropouts in single-cell rna-seq analysis. *Nature communications*, 11(1):1–9, 2020.

Xianwen Ren, Wen Wen, Xiaoying Fan, Wenhong Hou, Bin Su, Pengfei Cai, Jiesheng Li, Yang Liu, Fei Tang, Fan Zhang, et al. Covid-19 immune features revealed by a large-scale single cell transcriptome atlas. *Cell*, 2021.

Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. Gc-content normalization for rna-seq data. *BMC bioinformatics*, 12(1):480, 2011.

Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):1–9, 2010.

Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

Sayed Mohammad Ebrahim Sahraeian, Marghoob Mohiyuddin, Robert Sebra, Hagen Tilgner, Pegah T Afshar, Kin Fai Au, Narges Bani Asadi, Mark B Gerstein, Wing Hung Wong, Michael P Snyder, et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum rna-seq analysis. *Nature communications*, 8(1):1–15, 2017.

Valeria Saladino, Davide Algeri, and Vincenzo Auriemma. The psychological and social impact of covid-19: new perspectives of well-being. *Frontiers in psychology*, 11:2550, 2020.

Rickard Sandberg. Entering the era of single-cell transcriptomics in biology and medicine. *Nature methods*, 11(1):22–24, 2014.

Sanket Shah, Mudasir Rashid, Tripti Verma, and Sanjay Gupta. Chapter 8 - chromatin, histones, and histone modifications in health and disease. In Diego A. Forero and George P. Patrinos, editors, *Genome Plasticity in Health and Disease*, Translational and Applied Genomics, pages 109–135. Academic Press, 2020. ISBN 978-0-12-817819-5. doi: https://doi.org/10.1016/B978-0-12-817819-5.00008-5. URL `https://www.sciencedirect.com/science/article/pii/B9780128178195000085`.

Robert J Sims III, Subhrangsu S Mandal, and Danny Reinberg. Recent highlights of rna-polymerase-ii-mediated transcription. *Current opinion in cell biology*, 16(3):263–271, 2004.

Ritambhara Singh, Jack Lanchantin, Gabriel Robins, and Yanjun Qi. Deepchrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, 32 (17):i639–i648, 2016.

Avi Srivastava, Laraib Malik, Hirak Sarkar, Mohsen Zakeri, Fatemeh Almodaresi, Charlotte Soneson, Michael I Love, Carl Kingsford, and Rob Patro. Alignment and mapping methodology influence transcript abundance estimation. *Genome biology*, 21(1):1–29, 2020.

Rory Stark, Gordon Brown, et al. Diffbind: differential binding analysis of chip-seq peak data. *R package version*, 100(4.3), 2011.

Sebastian Steinhauser, Nils Kurzawa, Roland Eils, and Carl Herrmann. A comprehensive comparison of tools for differential chip-seq analysis. *Briefings in bioinformatics*, 17(6): 953–966, 2016.

Brian D Strahl and C David Allis. The language of covalent histone modifications. *Nature*, 403(6765):41–45, 2000.

Yan V Sun and Yi-Juan Hu. Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Advances in genetics*, 93:147–190, 2016.

Valentine Svensson. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, pages 1–4, 2020.

Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell rna-seq in the past decade. *Nature protocols*, 13(4):599–604, 2018.

Bosiljka Tasic. Single cell transcriptomics in neuroscience: cell classification and beyond. *Current opinion in neurobiology*, 50:242–249, 2018.

Elodie Teissier, François Penin, and Eve-Isabelle Pécheur. Targeting cell entry of enveloped viruses as an antiviral strategy. *Molecules*, 16(1):221–250, 2011.

Reuben Thomas, Sean Thomas, Alisha K Holloway, and Katherine S Pollard. Features that define the best chip-seq peak calling algorithms. *Briefings in bioinformatics*, 18 (3):441–450, 2017.

F William Townes and Rafael A Irizarry. Quantile normalization of single-cell rna-seq read counts without unique molecular identifiers. *Genome biology*, 21(1):1–17, 2020.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nature biotechnology*, 32(4):381, 2014.

David Van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174 (3):716–729, 2018.

Jun Wang, Qian Li, Yongmei Yin, Yingying Zhang, Yingying Cao, Xiaoming Lin, Lihua Huang, Daniel Hoffmann, Mengji Lu, and Yuanwang Qiu. Excessive neutrophils and neutrophil extracellular traps in covid-19. *Frontiers in immunology*, 11:2063, 2020.

Zhibin Wang, Chongzhi Zang, Jeffrey A Rosenfeld, Dustin E Schones, Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Weiqun Peng, Michael Q Zhang, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature genetics*, 40(7):897, 2008.

Lai Wei, Siqi Ming, Bin Zou, Yongjian Wu, Zhongsi Hong, Zhonghe Li, Xiaobin Zheng, Mingxing Huang, Liyun Luo, Juanran Liang, Xiaofeng Wen, Tingting Chen, Qiaoxing Liang, Liangjian Kuang, Hong Shan, and Xi Huang. Viral invasion and type i interferon response characterize the immunophenotypes during covid-19 infection. *Available at SSRN 3555695*, 2020.

Cindy L Will and Reinhard Lührmann. Protein functions in pre-mrna splicing. *Current opinion in cell biology*, 9(3):320–328, 1997.

Douglas C Wu, Jun Yao, Kevin S Ho, Alan M Lambowitz, and Claus O Wilke. Limitations of alignment-free tools in total rna-seq quantification. *BMC genomics*, 19(1):1–14, 2018.

Fan Wu, Su Zhao, Bin Yu, Yan Mei Chen, Wen Wang, Zhi Gang Song, Yi Hu, Zhao Wu Tao, Jun Hua Tian, Yuan Yuan Pei, Ming Li Yuan, Yu Ling Zhang, Fa Hui Dai, Yi Liu, Qi Min Wang, Jiao Jiao Zheng, Lin Xu, Edward C. Holmes, and Yong Zhen Zhang. A new coronavirus associated with human respiratory disease in china. *Nature*, 579 (7798):265–269, 2020.

Manuela Wuelling, Christoph Neu, Andrea M Thiesen, Simo Kitanovski, Yingying Cao, Anja Lange, Astrid M Westendorf, Daniel Hoffmann, and Andrea Vortkamp. Epigenetic mechanisms mediating cell state transitions in chondrocytes. *bioRxiv*, 2020.

Haipeng Xing, Yifan Mo, Will Liao, and Michael Q Zhang. Genome-wide localization of protein-dna binding and histone modification by a bayesian change-point method with chip-seq data. *PLoS Comput Biol*, 8(7):e1002613, 2012.

Xintian Xu, Ping Chen, Jingfang Wang, Jiannan Feng, Hui Zhou, Xuan Li, Wu Zhong, and Pei Hao. Evolution of the novel coronavirus from the ongoing wuhan outbreak and modeling of its spike protein for risk of human transmission. *Science China Life Sciences*, 63(3):457–460, 2020.

Ali M Zaki, Sander Van Boheemen, Theo M Bestebroer, Albert DME Osterhaus, and Ron AM Fouchier. Isolation of a novel coronavirus from a man with pneumonia in saudi arabia. *New England Journal of Medicine*, 367(19):1814–1820, 2012.

Wanwen Zeng, Yong Wang, and Rui Jiang. Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network. *Bioinformatics*, 36(2):496–503, 2020.

Ji Yuan Zhang, Xiang Ming Wang, Xudong Xing, Zhe Xu, Chao Zhang, Jin Wen Song, Xing Fan, Peng Xia, Jun Liang Fu, Si Yu Wang, Ruo Nan Xu, Xiao Peng Dai, Lei Shi, Lei Huang, Tian Jun Jiang, Ming Shi, Yuxia Zhang, Alimuddin Zumla, Markus Maeurer, Fan Bai, and Fu Sheng Wang. Single-cell landscape of immunological responses in patients with covid-19. *Nature Immunology*, pages 1–12, 2020.

Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of chip-seq (macs). *Genome biology*, 9(9):1–9, 2008.

Shanrong Zhao, Ying Zhang, Ramya Gamini, Baohong Zhang, and David von Schack. Evaluation of two main rna-seq approaches for gene quantification in clinical rna sequencing: polya+ selection versus rrna depletion. *Scientific reports*, 8(1):1–12, 2018.

Shanrong Zhao, Zhan Ye, and Robert Stanton. Misuse of rpkm or tpm normalization when comparing across samples and sequencing protocols. *Rna*, 26(8):903–909, 2020.

Chunhong Zheng, Liangtao Zheng, Jae-Kwang Yoo, Huahu Guo, Yuanyuan Zhang, Xinyi Guo, Boxi Kang, Ruozhen Hu, Julie Y Huang, Qiming Zhang, et al. Landscape of infiltrating t cells in liver cancer revealed by single-cell sequencing. *Cell*, 169(7):1342–1356, 2017a.

Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):1–12, 2017b.

Zhuo Zhou, Lili Ren, Li Zhang, Jiaxin Zhong, Yan Xiao, Zhilong Jia, Li Guo, Jing Yang, Chun Wang, Shuai Jiang, Donghong Yang, Guoliang Zhang, Hongru Li, Fuhui Chen, Yu Xu, Mingwei Chen, Zhancheng Gao, Jian Yang, Jie Dong, Bo Liu, Xiannian Zhang, Weidong Wang, Kunlun He, Qi Jin, Mingkun Li, and Jianwei Wang. Heightened innate immune responses in the respiratory tract of covid-19 patients. *Cell Host & Microbe*, 2020.

Lirong Zou, Feng Ruan, Mingxing Huang, Lijun Liang, Huitao Huang, Zhongsi Hong, Jianxiang Yu, Min Kang, Yingchao Song, Jinyu Xia, Qianfang Guo, Tie Song, Jianfeng He, Hui Ling Yen, Malik Peiris, and Jie Wu. Sars-cov-2 viral load in upper respiratory

specimens of infected patients. *New England Journal of Medicine*, 382(12):1177–1179, 2020.

Appendix A

List of Figures, Tables and Abbreviations

List of Figures

List of Tables

List of Abbreviations

A adenine

ACE2 angiotensin converting enzyme 2

C cytosine

ChIP-seq chromatin immunoprecipitation followed by next generation sequencing

CPM counts per million

DE differential expression

E envelope

EMBL-EBI European Molecular Biology Laboratory-European Bioinformatics Institute

G guanine G

GEO Gene Expression Omnibus

HT-seq high throughput sequencing

IFNs interferons

IRFs interferon regulator factors

ISGs IFN-simulated genes

IVT in vitro transcription

LGP2 Laboratory of Genetics and Physiology 2

logFC *log* fold change

LOWESS locally weighted scatterplot smoothing

LRT likelihood ration test

M membrane

MDA5 melanoma differentiation-associated gene 5

MERS-CoV Middle East respiratory syndrome coronavirus

mESCs mouse embryonic stem cells

mNRA messenger RNA

N nucleocapsid

NF-κB nuclear factor κB

nsps non-structural proteins

PCR polymerase chain reaction

pre-mRNA precursor mRNA

PRRs pattern recognition receptors

PTMs post-translational modifications

RBD receptor-binding domain

RIG-I retinoic acid-inducible gene I

RLRs RIG-I like receptors

RNAPII RNA polymerase II

RNA-seq RNA sequencing

RPKM/FPKM reads/fragments per kilobase per million reads mapped

RPM reads per million

S spike

SAM Sequence Alignment Map

SARS-CoV severe acute respiratory syndrome coronavirus

scRNA-seq single cell RNA sequencing

T thymine

TLRs Toll-like receptors

TMM Trimmed Mean of M-values

TPM transcripts per million

TSS transcription start site

U uracil

UMI Unique Molecular Identifier

10X 10X Genomics Chromium

Declaration of Contribution

Kumulative Dissertation/Beteiligung an Veröffentlichungen

Kumulative Dissertation von Frau Yingying Cao

Autorenbeiträge

Titel der Publikation: intePareto: an R package for integrative analyses of RNA-seq and ChIP-seq data

Autoren: Yingying Cao, Simo Kitanovski, and Daniel Hoffmann.

Anteile:

- Konzept - %: 70
- Durchführung der Experimente - % NA
- Datenanalyse - % 100
- Artenanalyse -% NA
- Statistische Analyse -% 100
- Manuskripterstellung -% 50
- Überarbeitung des Manuskripts -% 50

Unterschrift Doktorand/in	Unterschrift Betreuer/in

Kumulative Dissertation/Beteiligung an Veröffentlichungen

Kumulative Dissertation von Frau Yingying Cao

Autorenbeiträge

Titel der Publikation: UMI or not UMI, that is the question for scRNA-seq zero-inflation

Autoren: Yingying Cao, Simo Kitanovski, Ralf Küppers, and Daniel Hoffmann.

Anteile:

- Konzept - %: 65
- Durchführung der Experimente - % 0
- Datenanalyse - % 90
- Artenanalyse -% NA
- Statistische Analyse -% 80
- Manuskripterstellung -% 50
- Überarbeitung des Manuskripts -% 50

Unterschrift Doktorand/in	Unterschrift Betreuer/in

Kumulative Dissertation/Beteiligung an Veröffentlichungen

Kumulative Dissertation von Frau Yingying Cao

Autorenbeiträge

Titel der Publikation: Excessive Neutrophils and Neutrophil Extracellular Traps in COVID-19

Autoren: Jun Wang, Qian Li, Yongmei Yin, Yingying Zhang, Yingying Cao, Xiaoming Lin, Lihua Huang, Daniel Hoffmann, Mengji Lu, and Yuanwang Qiu.

Anteile:

- Konzept - %: 5
- Durchführung der Experimente - % 0
- Datenanalyse - % 15
- Artenanalyse -% NA
- Statistische Analyse -% 15
- Manuskripterstellung -% 10
- Überarbeitung des Manuskripts -% 10

Unterschrift Doktorand/in	Unterschrift Betreuer/in

Kumulative Dissertation/Beteiligung an Veröffentlichungen

Kumulative Dissertation von Frau Yingying Cao

Autorenbeiträge

Titel der Publikation: Comprehensive comparison of transcriptomes in SARS-CoV-2 infection: alternative entry routes and innate immune responses

Autoren: Yingying Cao, Xintian Xu, Simo Kitanovski, Lina Song, Jun Wang, Pei Hao, and Daniel Hoffmann.

Anteile:

- Konzept - %: 70
- Durchführung der Experimente - % NA
- Datenanalyse - % 95
- Artenanalyse -% NA
- Statistische Analyse -% 100
- Manuskripterstellung -% 50
- Überarbeitung des Manuskripts -% 50

Unterschrift Doktorand/in	Unterschrift Betreuer/in

Declarations

Erklärung:
Hiermit erkläre ich, gem. § 7 Abs. (2) d) + f) der Promotionsordnung der Fakultät für Biologie zur Erlangung des Dr. rer. nat., dass ich die vorliegende Dissertation selbständig verfasst und mich keiner anderen als der angegebenen Hilfsmittel bedient, bei der Abfassung der Dissertation nur die angegeben Hilfsmittel benutzt und alle wörtlich oder inhaltlich übernommenen Stellen als solche gekennzeichnet habe.

Essen, den _____ _____
Unterschrift des/r Doktoranden/in

Erklärung:
Hiermit erkläre ich, gem. § 7 Abs. (2) e) + g) der Promotionsordnung der Fakultät für Biologie zur Erlangung des Dr. rer. nat., dass ich keine anderen Promotionen bzw. Promotionsversuche in der Vergangenheit durchgeführt habe und dass diese Arbeit von keiner anderen Fakultät/Fachbereich abgelehnt worden ist.

Essen, den _____ _____
Unterschrift des/r Doktoranden/in

Erklärung:
Hiermit erkläre ich, gem. § 6 Abs. (2) g) der Promotionsordnung der Fakultät für Biologie zur Erlangung der Dr. rer. nat., dass ich das Arbeitsgebiet, dem das Thema "Computational analysis and interpretation of multi-omics data" zuzuordnen ist, in Forschung und Lehre vertrete und den Antrag von Yingying Cao befürworte und die Betreuung auch im Falle eines Weggangs, wenn nicht wichtige Gründe dem entgegenstehen, weiterführen werde.

_____
Name des Mitglieds der Universität Duisburg-Essen in Druckbuchstaben

Essen, den _____
Unterschrift eines Mitglieds der Universität Duisburg-Essen