DOCTOR OF PHILOSOPHY

Towards population-scale proteomics to study molecular phenotypes in health and disease

Brenes Murillo, Alejandro

Award date:
2023

Licence:
CC BY-NC-ND

Link to publication

Towards population-scale proteomics to study molecular phenotypes in health and disease

Alejandro J. Brenes

Thesis submitted for the degree of
Doctor of Philosophy by publications
School of Life Sciences
University of Dundee
February 2023

Table of Contents

Table of Figures

Abbreviations

40S: Small ribosomal subunit complex

60S: Large ribosomal subunit complex

80S: Full eukaryotic ribosome composed of the 40S and 60S

AG: Azurophilic granule

AGC: Automatic gain control

ARDS: Acute respiratory distress syndrome

ASE: Allele specific expression

AUC: Area under the curve

CII: Co-isolation interference

COPD: Chronic obstructive pulmonary disease

COVID19: Coronavirus disease 19

CSV: Comma separated values

CTL: Cytotoxic lymphocyte

CV: Coefficient of variation

DDA: Data dependent acquisition

DIA: Data independent acquisition

DNA: Deoxyribonucleic acid

EBI: European Bioinformatics Institute

eIF2: Eukaryotic initiation factor 2

eIF4F: Eukaryotic initiation factor 4f

IEL: Intraepithelial lymphocytes

ELISA: Enzyme-linked immunosorbent assay

ENA: European nucleotide archive

Ensembl: Ensembl genome database project

EPD: The Encyclopedia of Proteome Dynamics

ESC: embryonic stem cell

ESI: Electrospray ionization

ETC: Electron transport chain

FACS: Fluorescence-assisted cell sorting

FDR: False discovery rate

GMP: Granulocyte macrophage progenitor

GO: Gene ontology

hESC: human embryonic stem cell

hiPSC: Human induced pluripotent stem cell

HipSci: Human induced pluripotent stem cell initiative

HPLC: High performance liquid chromatography

iBAQ: Intensity based absolute quantification

IFN: Interferon

IFN-I: Type I IFN

ImmPRes: The Immunological Proteome Resource

iPSC: Induced pluripotent stem cell

ITAM: Immunoreceptor tyrosine-based activating motif

ITIM: Immunoreceptor tyrosine-based inhibitory motif

iTRAQ: Isobaric tag for relative and absolute quantification

KEGG: Kyoto encyclopedia of genes and genomes

FAO: Fatty acid oxidation

FG: Ficolin granule

LAD: Leukocyte adhesion deficiency

LC: Liquid chromatography

LC-MS: Liquid chromatography mass spectrometry

LDN: Low density neutrophils

lncRNA: Long non-coding RNA

LRTI: Lower respiratory tract infection

MHC: Major histocompatibility complex

mRNA: Messenger RNA

MS: mass spectrometry

MS2: Tandem mass spectra

MS3: MS2 with an additional fragmentation and isolation event

MS/MS: Tandem mass spectra

M/Z: Mass over charge

mTORC1: Mammalian target of rapamycin complex 1

NADPH: Nicotinamide dinucleotide phosphate

NET: Neutrophil extracellular trap

ORA: Overrepresentation analysis

OT1: Strain of transgenic mice where T cells express a TCR complex specific for ovalbumin peptide SIINFEKL in the context of H2K^b

P14: Strain of transgenic mice where T cells express TCR complex specific for the gp33 peptide from lymphocytic choriomeningitis virus

PBMC: Peripheral blood mononuclear cell

PCR: Polymerase chain reaction

PPI: Protein-protein interaction

PRIDE: Proteomics identifications database

PSC: Pluripotent stem cell

PSM: Peptide spectrum match

PTM: post-translational modification

QC: Quality control

RII: Reporter ion interference

RNA: Ribonucleic acid

RNAseq: RNA sequencing

ROC: Receiver operating characteristic

RT-PCR: Reverse transcription PCR

R-WHO-1: Patients fully recovered from COVID19 after 29 days

R-WHO2-3: Patients not fully recovered from COVID19 after 29 days

SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2

SEC: Size exclusion chromatography

SG: Specific granule

SILAC: Stable isotope labelling by amino acids in cell culture

SV: Secretory vesicle

TCA: tricarboxylic acid cycle, critric acid cycle, Krebs cycle

TCR: T cell receptor

TMT: Tandem Mass Tags

TPM: Transcripts per million

Treg: Regulatory T cell

Th: T helper cell

UniProt: Universal Protein Resource

URP: Unique plus razor peptides

WHO: World health organization

WHO3: Hospitalised COVID19 patients that did not require supplemental oxygen at admission

WHO4: Hospitalised COVID19 patients that required supplemental oxygen

WHO5-6: Hospitalised COVID19 patients that required high-flow oxygen or mechanical ventilation

XCI: X-chromosome inactivation

Xic: X-inactivation centre

Acknowledgements

The start of my scientific career happened by accident, I stumbled on to the field by chance, but once I did, I knew this was what I wanted to do. This would never have been possible if not for the support of my family, to them I owe it all. It also wouldn't have happened had I not found a job in the School of Life Sciences in Angus's lab. Angus provided that initial guidance and mentoring that got me into research, and for that I am extremely grateful. Doreen and the School of Life Sciences management also played a vital role in the shaping of this work by allowing me to undertake a degree format that was catered towards older researchers, by making use of my previous experience and outputs. Doreen in particular has been an incredibly supportive supervisor, mentor and friend, I would not have succeeded without her guidance.

Along my short scientific career, I had great experiences interacting with my peers and lab colleagues and their help/advice with my projects has been sincerely appreciated. In particular I have to thank Christina, Laura, Linda, Julia, Andy, Kate, Haru, Hao and Jens for putting up with me and helping me over the years. I've also been lucky to have had many great collaborators that assisted me with many of my projects, so thanks to you as well.

A project this long also takes a toll on one's personal life. So, it is important that I recognise those who supported me all throughout; my family, my partner, my friends, my bikes, cycling/running club mates and racing pals. Thanks for the encouragement, the diners, the coffees and the beers. From the lazy days to the wee runs/hikes, from the chilled social rides to time trials and doing a half iron man together. You have all played a vital part of my work/life balance over the last few years, specially over the pandemic and post-lockdown periods. Thanks for being there for me.

Finally, I would also like to extend my gratitude to the Wellcome Trust whose funding made this all possible. In particular for the ISFF which funded me, the Strategic Award and the principal fellowships to Angus and Doreen that enabled the vast majority of my scientific outputs over the last few years.

Declaration

I declare that, unless otherwise stated, the work described in this thesis was carried out by me in the laboratories of Professor Doreen A. Cantrell and Professor Angus I. Lamond. This thesis was written by me under the supervision of Professor Doreen A. Cantrell and Professor Angus I. Lamond and all references have been consulted by me. To my knowledge, this thesis is a unique record of publications to which I am first or last author and have not been submitted for any other degree at this or any other university.

Alejandro J. Brenes
September 2022

Summary

Proteomics has reached a turning point where the datasets are now approaching population-scale, this was achieved via faster and more sensitive instruments, improved sample handling and robust software tools. The population-scale era creates the opportunity to further study the molecular differences across cell lines, cell types or tissues as well as phenotypes of health and disease in human patients. In this thesis we utilize different methodologies to perform large-scale proteomic analyses, including a data dependent acquisition (DDA) based dataset using tandem mass tags (TMT) and a label-free data independent acquisition (DIA) one. The two studies looked at different systems, including induced pluripotent stem cells (hiPSCs) as well as peripheral blood neutrophils derived from both healthy and diseased patients and provide new biological insights including new disease specific biomarkers as well as novel drug targets.

On a technical aspect, this work revealed novel considerations and limitations of using TMT for large-scale proteomic experiments, providing a framework to improve future studies and minimise the potential issues involved. Furthermore, the TMT-based hiPSC dataset revealed novel protein level effects caused by the erosion of X chromosome inactivation (XCI) in healthy female hiPSCs. The data show the erosion of XCI increases abundance of not just X chromosome proteins, but over 2,000 proteins derived from all other chromosomes, thereby significantly increasing the protein content in these eroded female lines compared to male and non-eroded female lines.

The DIA-based large-scale proteomic characterisation of neutrophils derived from control and COVID19 patients revealed the effects of SARS-CoV2 infection in human neutrophils at the early infection and recovery phase. It highlighted a core signature present in the early infection timepoints in COVID19 patients, as well as some transient and some persistent changes on the neutrophil proteomes caused by COVID19. The study also highlighted the potential for patient stratification and precision medicine, as it detected important proteomic changes that were only present in patients with critically severe COVID19, and empowered treatment options that could help improve clinical trajectories for these patients. Furthermore, the data provided molecular insights into the delayed recovery state seen in a subset of COVID19 patients, as this patient group displayed a dysfunctional neutrophil phenotype reminiscent of what is seen in chronic diseases like COPD.

Finally, this thesis explores the value of web applications (web app) to visualise and explore the proteomic data. Proteomic datasets are growing in size and complexity, thus finding ways to share and visualise this complex data in an intuitive format for non-specialist users becomes an important goal. As a solution we built two web apps, the Encyclopedia of Proteome Dynamics (EPD) and the Immunological Proteome Resource (ImmPRes), to ensure open access to the to the raw and processed proteomic data, as well as easy visualisation/exploration via graphical interfaces. ImmPRes provides access to the biggest collection of leukocyte proteomic data, integrating data from innate and adaptive cells as well as multiple datasets characterising different T cell populations along with multiple signalling pathways that modulate their functions, all contained within a simple web application that is easily accessible for all immunologists.

Aims of the narrative thesis

- Highlighting how large-scale proteomics was used to generate new insights into phenotypes of health and disease in stem cells and neutrophils.
- Providing novel technical insights into the issues and limitations of using TMT for large-scale proteomics.
- Providing guidelines to design and execute large-scale TMT experiments
- Characterising the molecular phenotype of human induced pluripotent stem cells derived from healthy patients.
- Highlighting novel insights into neutrophil phenotypes in COVID19 patients discovered by large-scale proteomics.
- Describing two web applications created specifically for sharing and re-using large-scale proteomics derived data.
- Highlighting the importance of making proteomics data easily accessible and interpretable for the wider research community

Chapter 1 – Introduction

Proteomics has been widely used in clinical as well as immunological research over the last few years. Providing new insight to further stratify heterogenous diseases like cancer(Krug et al., 2020), as well as complementing our understanding of how the immune system works(Howden et al., 2019). As we enter the age of population-scale datasets, where thousands of samples are analysed in a single study, the stratification of patients and stages of disease becomes feasible, as seen recently by a study characterising over 900 human cancer cell lines(Goncalves et al., 2022).

In this thesis we explore using two different acquisition methods, data dependent (DDA) with a tandem mass tag (TMT) based system and label-free data independent acquisition to perform large-scale proteomic characterisations of human induced pluripotent stem cells (hiPSCs) and peripheral blood neutrophils. These large-scale datasets allowed us to explore some of the technological limitations relating to large-scale proteomics and enabled us to probe the variation and phenotypes in healthy hiPSCs, as well as characterising the effects of SARS-CoV2 infection in human neutrophils.

Furthermore, the thesis focusses on the important element of sharing and visualising the data derived from proteomics experiments. To achieve this goal, two web applications were built, specifically catered to sharing and browsing large-scale proteomics data for the wider research community via easily accessible graphical interfaces.

1.1 Mass spectrometry-based proteomics

The idea of doing global protein analysis is not recent, however it was not until the 90s that it became a reality. The formal definition of proteomics was proposed in 1995 and it described it as the large-scale study and characterization of the entire set of proteins, or proteome, expressed in a cell line, tissue or even a whole organism(Wilkins et al., 1996, Graves and Haystead, 2002). With the years it has become a vital dimension to study molecular and cellular biology(Aebersold and Mann, 2003), by shifting the focus from genes to proteins it became possible to expand the dimensions of analysis(Larance and Lamond, 2015). Proteomics enabled experiments doing a systematic analysis of post-translational modifications (PTMS) like phosphorylation (Zhou et al., 2001, Beausoleil et al., 2006, Villen et al., 2007, Dephoure et al., 2008, Bekker-Jensen et al., 2020, Collins et al., 2007, Martinez-Val et al., 2021, Huang et al., 2019, Searle et al., 2019, Ochoa et al., 2020, Urbaniak et al., 2013, Nett et al., 2009), as well as studies exploring of protein-protein interactions (PPI) using affinity purification (Fleming et al., 2016, Sullivan et al., 2013, Trinkle-Mulcahy et al., 2008, Mellacheruvu et al., 2013, Ji et al., 2021) , size exclusion chromatography (SEC) based proteomic analysis (Leitner et al., 2012, Kirkwood et al., 2013, Crozier et al., 2017, Larance et al., 2016) or proximity labelling (Gingras et al., 2007, Go et al., 2021) . It also enabled breakthroughs in protein subcellular localisation via methods like LOPIT(Thul et al., 2017, Dunkley et al., 2004, Christoforou et al., 2016, Dunkley et al., 2006, Sadowski et al., 2006, Mulvey et al., 2017, Christopher et al., 2021, Itzhak et al., 2016, Guther et al., 2014, Mardakheh et al., 2016, Geladaki et al., 2019), protein synthesis and turnover with metabolic labelling (Boisvert et al., 2012, Ly et al., 2018, Tinti et al., 2019, Doherty et al., 2009, Schwanhausser et al., 2011, Welle et al., 2016, Martin-Perez and Villen, 2017, Zecha et al., 2018, Rolfs et al., 2021), cell cycle studies(Crozier et al., 2018, Ly et al., 2015, Ly et al., 2017, Ginno et al., 2018, Herr et al., 2020) and even isoform/proteoform specific analyses(Aebersold et al., 2018, Zecha et al., 2018, Ahmad et al., 2012, Bludau et al., 2021). The most prevalent and comprehensive methodology used for proteomics is mass spectrometry(Aebersold and Mann, 2003, Aebersold and Mann, 2016), from which other recent advances such as mass cytometry were derived(Bandura et al., 2009). Mass spectrometers are used to detect the presence and quantity of multiple biomolecules, such as peptides, proteins, or metabolites, by using multiple dimensions such as mass, charge and more recently collisional cross section(Meier et al., 2021).

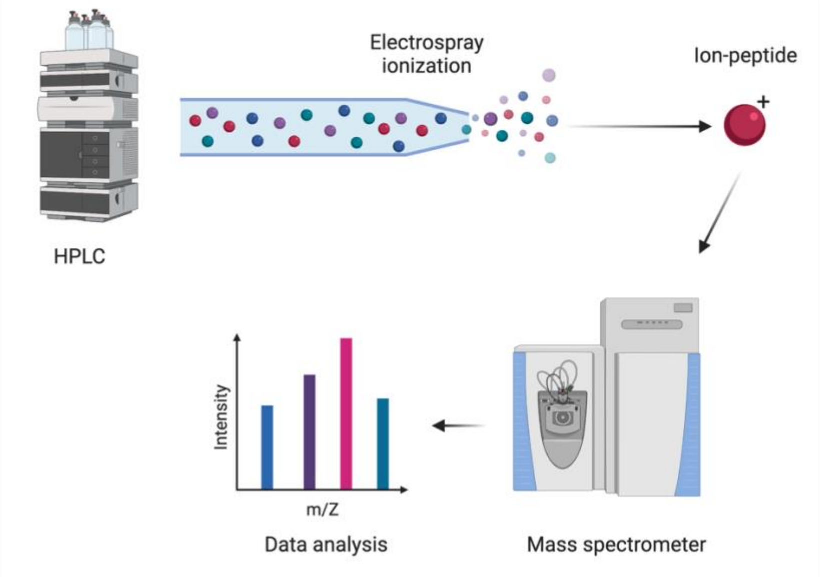

Figure 1. 1- Schematic showing an electrospray ionization based high performance liquid chromatography-based mass spectrometry where the peptide ions are injected into the mass spectrometer to be identified and quantified.

One most prevalent ionization methods used for mass spectrometry is electrospray ionization (ESI). ESI is based on supplying a high voltage to a needle placed in front of a mass spectrometer and creating an aerosol from a liquid(Mann, 2019) (Fig 1.1). With ESI producing the first mass spectra of intact proteins(Fenn et al., 1989), the field of mass spectrometry-based proteomics grew exponentially. ESI also provided the ideal mechanism to couple mass spectrometry to liquid chromatography (Fig 1.1), with high performance liquid chromatography (HPLC) now being a vital part of many mass spectrometry-based proteomics workflows.

However, it was also the computational solutions that enabled the field to grow and flourish. The first algorithm that calculated the theoretical set of fragment ions that could be produced by predetermined cleavage rules and compare the MS/MS spectrum to the theoretical spectra to identify the amino acid sequences present in complex lysates(Eng et al., 1994) was also enormously important for the field. Furthermore, multiple other important algorithms were also developed and were integrated into vital software tools(Perkins et al., 1999, Craig and Beavis, 2004, Kim et al., 2010) which underpin current bottom-up proteomics. This was complemented by important statistical methods to be applied to protein identification(Nesvizhskii et al., 2003) and the development of a model to calculate false discovery rates(Elias and Gygi, 2007) , which still underpins the working of many bottom-up proteomic search engines to this day.

1.1.1 Bottom-up proteomics

The proteomics field has two main methodologies that are used, they are termed top-down and bottom-up proteomics. Top-down proteomics derived its name from the 2-D gels that were used to fractionate a complex lysate into protein spots that were evaluated via staining, this approach focussed large on analysing intact proteins(Kelleher, 2004, Reid and McLuckey, 2002) without requiring digestion (Fig. 1.2). To this day the term top-down proteomics still refers to the same concept, however it is possible to direct the analysis of hundreds of intact proteins in a sample using mass spectrometry(Toby et al., 2019).

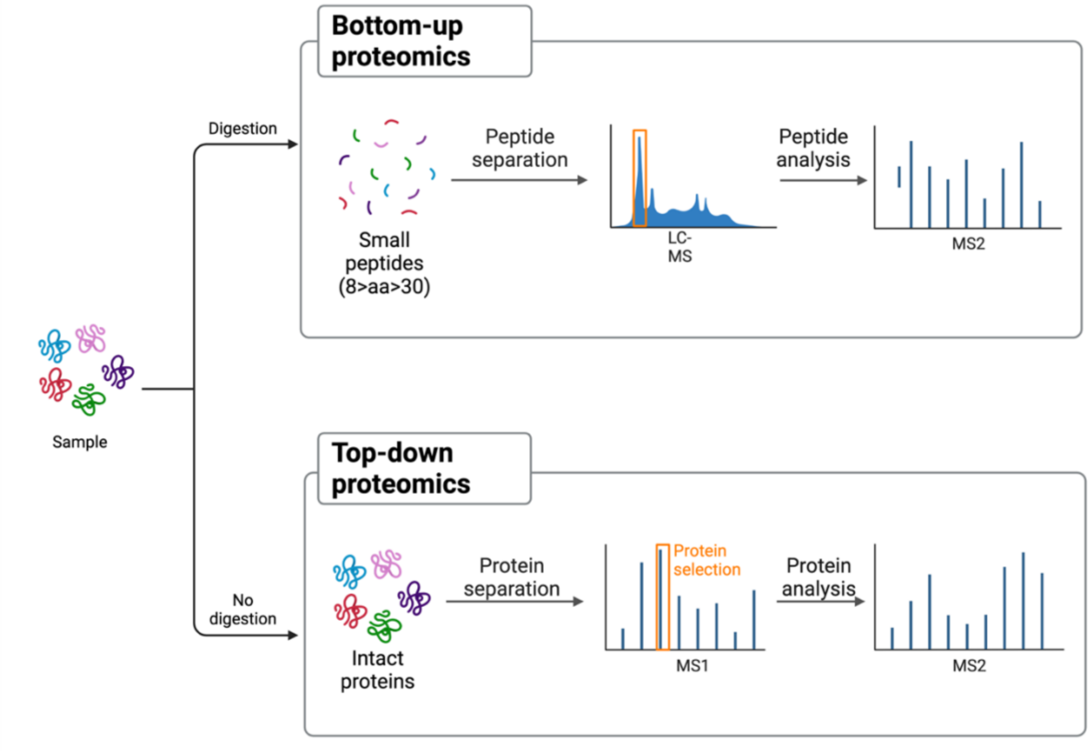

Figure 1.2 – Schematic showing the mode of action of the two main non-targeted proteomics methods. Top-down methods analyse intact proteins and proteoforms, bottom-up methods analyse peptides after they have been digested from a sample and injected into the mass spectrometer.

In bottom-up proteomics, also frequently referred to as Shotgun proteomics(Wolters et al., 2001), proteins are not directly characterised, as they are cleaved enzymatically before they are analysed on the mass spectrometer(Zhang et al., 2013b) (Fig. 1.2). This means that proteins are never directly measured, instead the different peptides that are produced by proteolysis are what is analysed in the mass spectrometer. Bottom-up proteomics peptide identification is performed by comparing the MS/MS derived from the peptide analysis on the mass spectrometer with theoretical MS/MS spectra generated *in silico*(Zhang et al., 2013b). As proteins themselves are never measured, their identification and quantification is inferred based on the pool of peptides that were identified across all the samples, something that is referred to as protein inference, which is a known issue and limitation within bottom-up proteomics(Baldwin, 2004). Assembling peptide identifications back into proteins is not a trivial task, as it has been shown that multiple proteins can contain the same peptide sequences, which leads to ambiguities in the identities of the proteins present as well as ambiguities in the quantification of the proteins(Nesvizhskii and Aebersold, 2005).

1.1.2 Protein inference and quantification in bottom-up proteomics

Within Bottom-up proteomics the peptides that are identified fall into two main categories; (i) those that uniquely map to a single protein within the search database that is provided, which are referred to as 'unique peptides' and (ii) those whose sequences match multiple proteins within the search database, and these are referred to as 'shared peptides'(Nesvizhskii and Aebersold, 2005). There are several approaches that have been developed through-out the years to assign peptides to their corresponding proteins(Zhang et al., 2007, Li et al., 2009, Tabb et al., 2002, Nesvizhskii et al., 2003), these approaches affect both protein identification as well as quantification.

One important concept which is frequently used by proteomic tools is protein grouping(Serang and Noble, 2012). Protein grouping implies the merging of multiple proteins, which, when represented as a graph model, are the nodes that are adjacent to an identical set of peptides(Serang and Noble, 2012). These proteins are aggregated and posteriorly treated as a single element. The selection of the protein that is considered to be the "master" or "lead" protein within the group can vary from tool to tool, for example MaxQuant will select the protein with the shortest sequence that contains the peptides within the protein group. However, even when using protein groups, it is still common to have 'shared peptides' across different protein groups within a dataset. The assignment of these peptides has a significant effect on the protein level quantification. Many tools assign the whole quantified peptide peak of a shared peptide to a single protein group, in effect using a "winner takes all" approach(Saltzman et al., 2018). This can cause artificial quantitative differences in protein groups when they share a large proportion of peptides, e.g. histones.

1.1.3 Data Dependent Acquisition (DDA)

If we drill into non-targeted bottom-up proteomics, we find there are two main acquisition strategies that are currently being utilised: data dependent acquisition (DDA) and data independent acquisition (DIA). Data dependent acquisition was the original method that was utilized for bottom-up proteomics(Washburn et al., 2001, Pandey and Mann, 2000). The name of this acquisition method is derived from its selective fragmentation nature, where a number of peptide ions (typically between 10-30), typically the ones with highest intensity(Michalski et al., 2011), are selected from an MS1 survey scan and get posteriorly selected for MS2 fragmentation. The method can be further refined by selecting a dynamic exclusion window which prevent fragmenting the same ions concurrently. The diagram displays this process by exemplifying a Top 2 scenario (Fig. 1.3). The advantage of DDA is that the MS2 tends to contain ions from the same peptide, which makes it simpler to deconvolute$^{(\text{Davies et al., 2021})}$.

The main disadvantage of DDA is that the stochastic nature of the ion selection means that the data regularly suffers from a higher incidence of missing values and reduced precision of the quantitation. It is common that the same ions are not fragment in subsequent runs even in the case of identical technical replicates. To minimise these effects, the complexity of the samples that are analysed via DDA is reduced by employing off-line fractionation of the samples using an HPLC. The fractionation strategy increases the depth of proteome coverage, thereby reducing missing values and improving the quantification, but this is achieved at the cost of reduced throughput as well as increased costs.

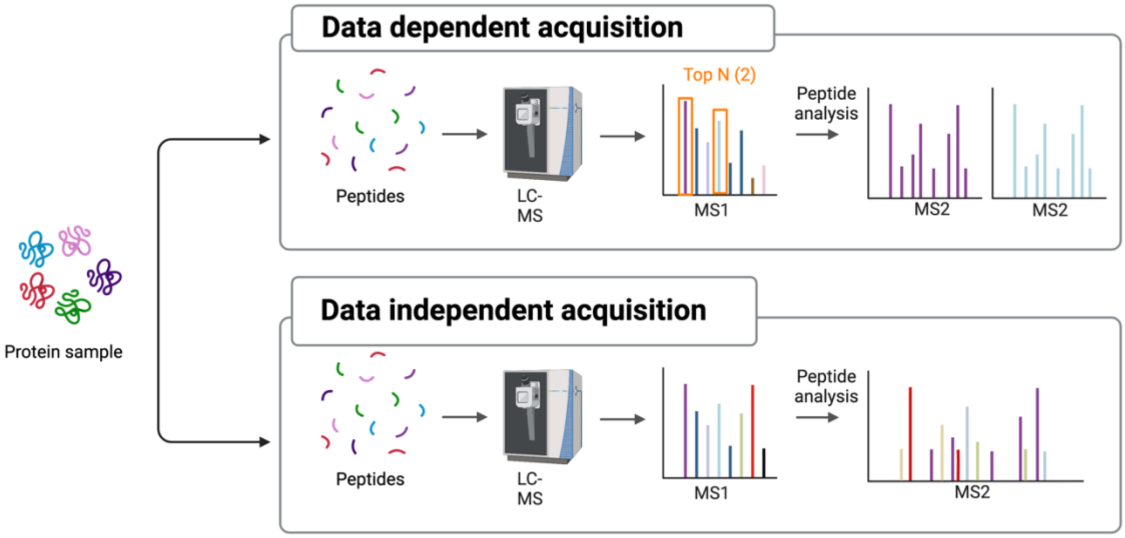

Figure 1.3 – Schematic showing the main two bottom-up proteomics workflows, data dependent acquisition (DDA) and data independent acquisition (DIA). The workflow leading up to the injection of the samples in the MS, protein extraction and digestion is the same in both methodologies. The differences between DDA and DIA occur in the acquisition mode within the MS, as highlighted by the figure.

1.1.4 Data independent acquisition (DIA)

To address some of the shortcomings present in DDA, different acquisition methods where the selection of ions for fragmentation did not exclusively depend on the selection a subset of ions based on intensity were developed. These methods are now referred to as data independent acquisition as originally coined by yates et al. These first methods were developed more than a decade ago and followed two main approaches.

Some of them worked with 1 MS1 scan combined with 1 MS2 scan and a 800Da window to fragment all ions (Geiger et al., 2010, Purvine et al., 2003, Plumb et al., 2006) while others used multiple, up to 32, narrower sequential acquisition windows (2.5 – 25Da) where they would fragment all ions (Gillet et al., 2012, Venable et al., 2004, Panchaud et al., 2009).

Modern DIA methods tend to follow the sequential window approaches and unlike DDA, they do not select a subset of ions to fragment in an MS2 scan, instead they fragment all the precursors ions present with a m/z window (Fig. 1.3). The instrument is then used to cycle over multiple windows to cover all the m/z range of interest. Modern DIA has been reported to provide more consistent proteomic coverage, with reduced missing values and higher precision in the quantification compared to DDA(Dowell et al., 2021).

The reduced missing values attributed to the fragment all strategy used in modern DIA comes at a cost, as DIA has been shown to produce more convoluted MS2 spectra compared to DDA. The convoluted spectra make the computational challenge of identifying the peptides present within the sample more difficult than with DDA. This meant that in the earlier implementations of DIA software tools the use of spectral libraries was required(Weisbrod et al., 2012, Rost et al., 2014, Bruderer et al., 2015, Wang et al., 2015).

1.1.5 Spectral libraries and library free DIA

Spectral libraries in the ‘peptide centric’ approach frequently originate from a collection of data that were generated in DDA mode and with extensive fractionation and multiple replicates. These data would then be processed using the usual DDA pipelines, based on sequence database searching, by software tools like MaxQuant(Cox and Mann, 2008), MSFragger(Kong et al., 2017) or Proteome Discoverer. The DDA software tools would produce a list of peptide spectrum matches, which would then be filtered and stored as spectral libraries for use in downstream DIA analysis(Deutsch et al., 2018). DIA software tools(Bruderer et al., 2015, Demichev et al., 2020, MacLean et al., 2010, Rost et al., 2014) would then use these libraries to score entries based on the retention time, keeping the highest scoring result for each peptide, which would then be used to search the DIA data. The main limitations of spectral libraries are that they can only be used to search for peptides detected in the library, whatever is not found in the library will not be detected in the posterior DIA analysis(Searle et al., 2020). Furthermore, the libraries work best when they are specific to the system of study, with mid-sized project specific libraries providing better results and FDR control compared to large pan-tissue libraries(Bruderer et al., 2017), and require considerable time and effort to generate a new library for each new cell type/tissue that is being studied. Finally, spectral libraries generated in a specific instrument platform, such as Orbitraps or time of flight instruments, are not easily reusable in a different instrument platform(Bruderer et al., 2017).

To overcome these limitations more emphasis has been focused on methods which do not require libraries to analyse DIA data(Ting et al., 2017, Tsou et al., 2015, Tiwary et al., 2019). These methods are referred to as "library free" methods. And while the library free methods do not always obtain the same depth as the project specific library, they have more recently been shown to produce comparable results, especially when large numbers of samples are used. It is now also possible to use hybrid libraries as well. The hybrid libraries are a mixture of DDA based spectral libraries along with a library-free search of DIA data. The combination of DDA and DIA based libraries has shown to produce the best results in the identification level(Muntel et al., 2019a).

1.1.6 Label free and labelled strategies in proteomics

The previously described workflows, both for DDA and DIA displayed the basic label free strategies. As previously mentioned, label free DDA requires extensive fractionation which significantly reduced throughput. As such multiple approaches have been developed over the years using specific metabolic or chemical labels which allow multiple samples to be analysed in the same run in the mass spectrometer(Gygi et al., 1999, Munchbach et al., 2000, Ong et al., 2002, Thompson et al., 2003, Ross et al., 2004, Beynon and Pratt, 2005). Within metabolic labelling the most popular approach is stable isotope labelling of amino acids in cell culture (SILAC)(Ong et al., 2002). SILAC involves growing cell lines in a media which contain amino acids labelled with the different isotopes, which increases the cost, time and difficulty of the labelling. However, as a trade-off it allows up to three-fold increased throughput by labelling "light", "medium" and "heavy" isotopes in the cell cultures.

For chemical labelling, the most popular methods involve using isobaric tags. Isobaric were originally developed almost two decades ago with the most commonly used versions being Tandem Mass Tags(Thompson et al., 2003) (TMT) and isobaric tag for relative and absolute quantitation(Ross et al., 2004) (iTRAQ). As was the case with SILAC, isobaric labels aimed to increase throughput. However, unlike SILAC, the tags here do not require metabolic labelling, instead they use a set of chemicals that label peptides after they have been enzymatically cleaved.

1.1.7 Tandem mass tags (TMT)

In the case of TMT, the tags contain a reporter, cleavable linker, a mass normalisation and an amine-reactive group region, with the newer TMTpro 18plex shown above (Fig. 1.4), with the amine-reactive group binding the N-terminus of an available lysine within each peptide. Each version of the isobaric tags has an identical mass when all 4 regions are combined. The different tags have a distinct heavy isotope distribution and also a different mass normalisation region which is used to compensate the differences in mass. This means that the different tagged samples have peptides with identical peaks at the MS1 level (Fig. 1.5), producing a simpler MS1 spectra compared to SILAC which has new MS1 peaks for each additional labelled sample (Pappireddi et al., 2019). The quantification of each of the samples occurs only after the ions are fragmented at the MS2 or MS3 level, which is when a specific reporter ion for each of the individual tags can be identified and then used for relative quantitation.

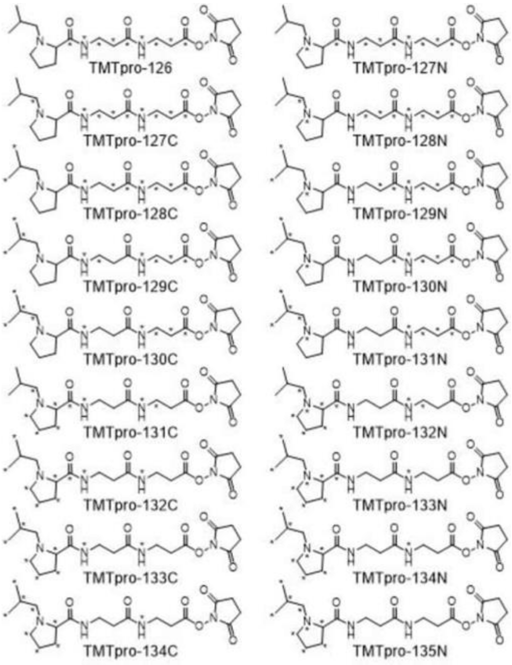

Figure 1.4 – Chemical structure of all TMTpro^{TM} 18-plex reagents. Reproduced from ThermoFisher Scientific

TMT workflows have similar starting steps to the label free workflows (Fig. 1.5). The different samples are individually processed, and the proteins extracted and digested. It is after this point that the different samples are be labelled with their corresponding TMT tag (Fig. 1.5). After labelling, the samples are then combined and analysed together in the same run within the mass spectrometer. This simplified labelling process, combined with increased throughput has made TMT one of the most popular techniques to use within DDA-based bottom-up proteomics experiments. Currently TMT has increased the multiplexing capacity considerably and offers a 6-plex, 10-plex, 11-plex, 16-plex(Li et al., 2020) and 18-plex capacity(Li et al., 2021).

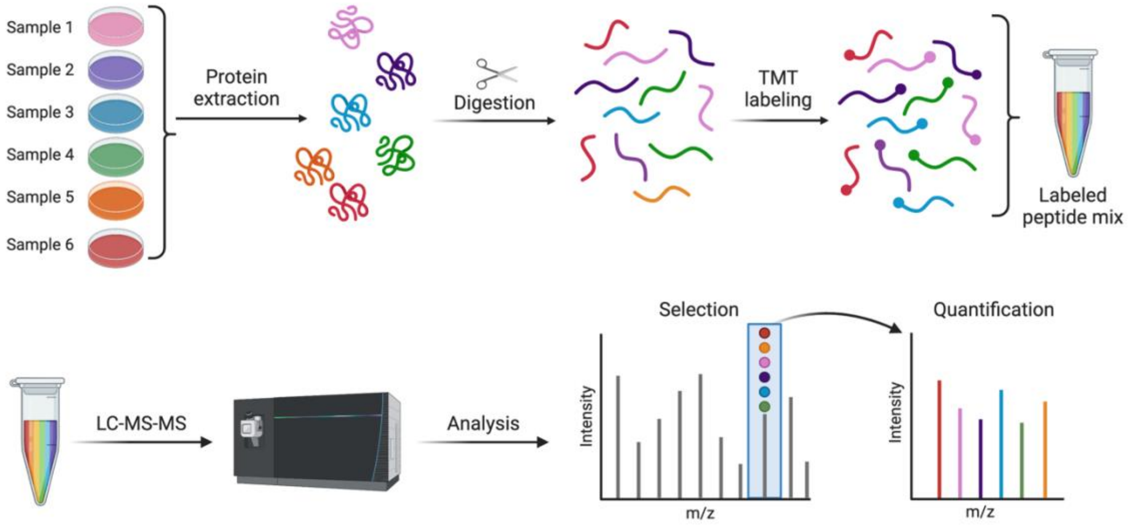

Figure 1.5 – Schematic showing the typical TMT workflow starting from protein extraction and digestion and leading up to the labelling of all individual samples which are then combined and injected into the mass spectrometer. It highlights that the isobaric tags have identical mass at the MS1 level and it's only after fragmentation that the relative intensities of the reporters (which represent the different samples) can be obtained.

The increased multiplexing capacity means that many experiments will have all conditions and all replicates contained within the same multiplexed batch/experiment. As such all conditions and replicates are analysed together in the mass-spectrometers, which means the same peptides are detected at the exact same area peak, leading to very low number of missing values and a high level of precision in their quantification(Isasa et al., 2015). How TMT behaves in large-scale proteomics experiments where multiple TMT experiments are required is explored in chapter 2.

1.2 Biological systems: Embryonic stem cells and leukocytes

Within this thesis we used large-scale proteomics to analyse different biological systems. Therefore, to understand and interpret the results obtained from the proteomic data, it is important to understand the basics of the biological systems. This section provides a basic introduction into of human pluripotent stem cells, both embryonic and induced, as well as the different innate and adaptive leukocyte populations which are studied within this thesis.

1.2.1 Human embryonic and induced pluripotent stem cells

Human embryonic stem cells (hESC) are derived from the inner cell mass of a blastocyte, also known as a pre-implantation embryo(Smith, 2001). Embryos reach the blastocyte stage within 4-5 days and at that specific timepoint consist of a range of 50 to 150 cells. hESCs derived from the blastocyte show prolonged undifferentiated potential, as well as the ability to differentiate into the three main embryonic germ layers(Thomson et al., 1998). Having the ability to generate specialized differentiated cells makes hESCs excellent models to study disease mechanisms, development and differentiation. However, their use remains restricted by regulations, as acquiring the inner cell mass of the blastocyte leads to its destruction, highlighting ethical considerations of hESC use(Volarevic et al., 2018).

Over a decade ago, methods allowing the reprogramming of somatic cells by transferring their nuclear content into oocytes(Wilmut et al., 1997) or by fusion of somatic cells with ESCs(Tada et al., 2001) were discovered. Both methods became steppingstones to the pivotal discovery of induced pluripotent stem cells that could be derived from fibroblast cultures, both in human and in mice(Takahashi et al., 2007, Takahashi and Yamanaka, 2006, Yu et al., 2007). Researchers showed that by exogenously expressing a small set of key transcription factors (Oct4, Sox2, c-Myc and Klf4), without the need for a transfer, they could reprogram fibroblasts back into a pluripotent state(Takahashi et al., 2007, Takahashi and Yamanaka, 2006, Yu et al., 2007). These human induced pluripotent stem cells (hiPSCs) were characterised by their capacity for self-renewal and ability to differentiate into the three main germ layers and show many key features of their physiological embryonic stem cell counterparts, while avoiding many of the ethical issues regarding the use of stem cells derived from embryos.

Since the original protocol, new methods have been devised to reprogram multiple type of somatic cells beyond fibroblasts, such as keratinocytes(Maherali et al., 2008), peripheral blood cells(Staerk et al., 2010) and even exfoliated epithelial cells collected from urine(Zhou et al., 2012). Furthermore the methods used for reprogramming have expanded considerably, including alternative approaches that avoid using the transcription factor c-Myc, due to its cancerogenic potential(Nakagawa et al., 2008), and using alternative vectors like adenoviruses(Zhou and Freed, 2009) or plasmids(Okita et al., 2008). hiPSCs are now being used as alternatives to hESCs in regenerative medicine(Kimbrel and Lanza, 2015) and disease modelling, including studies on monogenic disorders(Ebert et al., 2009, Lee et al., 2009), as well as models to study some late onset diseases(Liu et al., 2012).

1.2.2 Primed vs naïve pluripotency

Both human and mouse embryonic stem cells are described as pluripotent, which is defined as the ability to give rise to cells from the three embryonic germ layers(Hanna et al., 2010). Pluripotency has been described as a transient state *in vivo* which can be maintained in vitro by exogenously supplying the right signalling cues(Weinberger et al., 2016). It is also a dynamic state that exists in a spectrum ranging from naïve pluripotency, defined to resemble the pre-implantation embryonic state, to primed pluripotency, defined to resemble the post-implantation embryonic state(Weinberger et al., 2016). For ESCs the types of pluripotency represent the tissue from which they were derived as well as the growth conditions in which they are maintained. Whereas in iPSCs they mostly represent the reprogramming and culture conditions.

The differences between naïve and primed pluripotent states are reflected in the epigenetic state and the differentiation capacity of the cells(Dong et al., 2019). Naïve cells have distinct methylation patterns on their histones, display two active X chromosomes and can differentiate into extraembryonic tissues(Osnato et al., 2021). Primed cells on the other hand already display functional X chromosome inactivation (XCI), where one X chromosome is transcriptionally repressed. Mouse iPSCs and mESCs represent a naïve state of pluripotency(Hanna et al., 2009), whereas hESCs and hiPSCs represent a more primed state of pluripotency. It is however possible to generate human naïve PSCs by inducing primed PSCs or by directing reprogramming somatic cells under specific culture conditions.

1.2.3 HipSci consortium

The Human Induced Pluripotent Stem Cell Initiative (HipSci; https://www.hipsci.org/) was a collaborative effort between searchers in Sanger Centre, European Bioinformatics Institute (EBI), Kings College and the University of Dundee. It combined genomics, proteomics, cell biology and clinical genetics to create a UK national hiPSC resource. The HipSci consortium was a pioneer in the induced pluripotent stem cell field in the United Kingdom. It established a robust pipeline to reprogram somatic cells back into induced pluripotent stem cells, testing reprogramming protocols from multiple different cell types as well as different reprogramming methods and growth media. The HipSci consortium produced over 750 different hiPSC lines, all in a state of primed pluripotency, derived from over 300 donors, with over 50% of the donors having at least two lines generated from different biopsies. Each hiPSC line was subjected to stringent quality control measures including genotyping using 'gtarray' and a cellular differentiation assay as well as a pluripotency assay. (Kilpinen et al., 2017). The hiPSC lines that passed all QC thresholds were then banked and are currently accessible to be used research purposes (https://www.hipsci.org/). These same lines were then subjected to large-scale multi-omic analysis, including imaging, methylation, RNAseq and proteomics. All the hiPSC lines used within this thesis were primed PSCs generated by the HipSci consortium.

1.2.4 Innate immune system

Our body possess physical barriers to stop the entry of pathogens, and these include the skin, epithelial cells, and mucus layers(Ganz, 2003, Janeway et al., 2001). When these barriers are breached, the immune system is directed to its role of controlling and destroying the invading pathogens. Historically the immune response has been stratified in two main categories: the innate and the adaptive immune response. In humans and other vertebrates, the first response to invading pathogens is the innate immune system. The innate immune system acts to eliminate or contain the threat while avoiding excessive damage to healthy tissue and cells(Paludan et al., 2021). Innate immunity largely depends on macrophages and granulocytes(Janeway et al., 2001), these effectors can operate to engulf or kill bacteria, limit viral entry, replication and assembly, remove infected cells as well as priming the adaptive immune response through the action of dendritic cells and other innate antigen presenting cells.

Their function is regulated via a set of pattern recognition receptors (PRR) that are encoded within the germ line which allow these cells to detect and respond to a set of common structures present in pathogens, like the pathogen-associated molecular patterns (PAMPS) that are not present in the host(Janeway and Medzhitov, 2002) as well as damage-associated molecular patterns (DAMPS). Innate effectors derive from hematopoietic stem cells (HSCs) that are present in the medullary portion of the bone marrow(Kondo, 2010). HSCs give rise to two types of progenitor cells, the common myeloid progenitors and the common lymphoid progenitors. It is from these myeloid progenitors that the effector cells of the innate immune system are derived. The list of effector cells include dendritic cells, mast cells, macrophages and granulocytes like eosinophils, basophils and neutrophils(De Kleer et al., 2014).

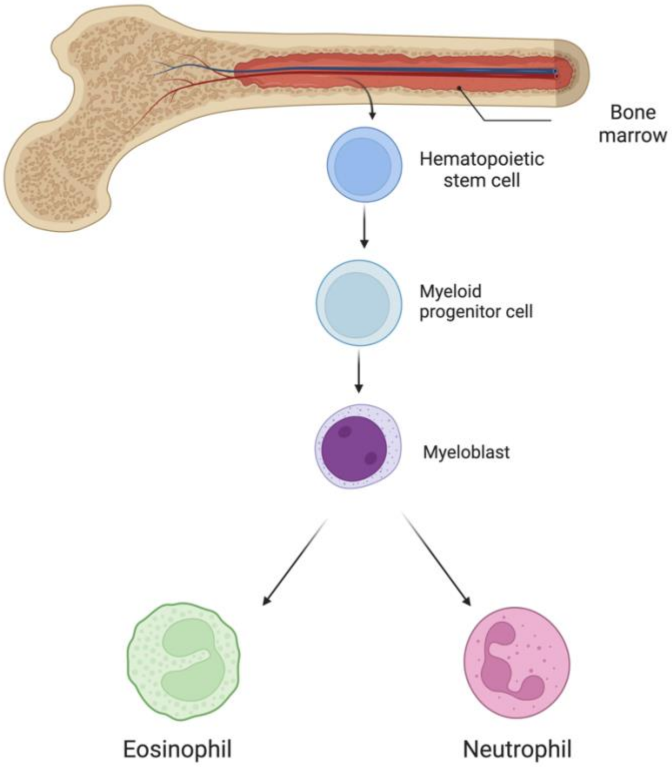

Figure 1.6: Schematic showing the neutrophil and eosinophil developmental process within the bone marrow, starting from the hematopoietic stem cell and lead to their eventual differentiation into the myeloblast, from which both cell types are derived.

1.2.5 Neutrophils

Neutrophils are believed to be short lived effector cells that are part of the innate immune system. Mature neutrophils are replenished within the bone marrow, have lobulated nuclei and are not undergoing cell division(Lawrence et al., 2018). They are the most abundant leukocyte that is found within the human blood, where under certain conditions they can make up over 70% of all white blood cells. They are granulocytes that are capable of effective antimicrobial activity via 3 main mechanisms. The first release of their granule content, referred to as degranulation, the second is phagocytosis, then engulfment of a pathogen, and the last is formation of neutrophil extracellular traps (NETs). These effector functions need to be tightly regulated as their activity can cause tissue damage and chronic inflammation(Soehnlein et al., 2017).

Neutrophils were thought to originate exclusively from myeloid progenitors (Fig. 1.6), however it is now believed they can also be derived from lymphoid-primed multipotent progenitors(Hong, 2017). Neutrophil differentiation is thought to begin between the myeloblast to promyelocyte stage(Bartels et al., 2015). Cells become committed during the transition from myelocytes to metamyelocytes, and it is also at this point that cell division stops(Cowland and Borregaard, 2016). Granulopoiesis, the formation of granules, is a stepwise process which goes hand in hand with neutrophil differentiation. The first granules to be synthesised are the azurophilic or primary granules, which arise in the myeloblast and promyelocyte stage. At this point in the development the cells still contain a mostly round nucleus(Lawrence et al., 2018). They are followed the specific or secondary granules which occurs in myelocytes and metamyelocytes, in this stage the nucleus starts to lose its rounded shape. Gelatinase or tertiary granules are the next to be produced and this occurs as neutrophils start to display band-like nuclear shape. The last granules to be produced are ficolin-1 granules and secretory vesicles which are present in mature neutrophils that have the characteristic segmented nucleus(Faurschou and Borregaard, 2003).

1.2.6 Eosinophils

Eosinophils are set of leukocytes and part of the innate immune system. They are similar to neutrophils as they are both bone marrow-derived and are both granulocytes derived from myeloid progenitors (Fig. 1.6). However unlike neutrophils, eosinophils are derived from GATA1$^+$ precursors(Drissen et al., 2016) with a granule content also that differs considerably from neutrophils, as they have been reported to contain a single population of granules(Weller and Spencer, 2017). Some of these granule proteins, mainly PRG2 (major basic protein), RNASE3 (ECP) and RNASE2 (EDN) have been shown to have antibacterial properties(Lehrer et al., 1989). However their granule functions have also been linked to organ development(Gouon-Evans et al., 2000), lymphocyte recruitment(Jacobsen et al., 2008) and tissue repair(Todd et al., 1991). Eosinophils have also been shown to store a wide array of cytokines like IFNy, IL4 and IL10 among others, pre-packaged within their granules(Beil et al., 1993).

Eosinophils are also considerably less abundant than neutrophils, making up less than 5% of all circulating leukocytes in human blood(Weller and Spencer, 2017). They have been classically shown to be involved in immune responses to helminth infection and allergens(Voehringer et al., 2007), with increased number of blood eosinophils reported under both conditions(Huang et al., 2014). However, their functions and residency properties reach far beyond controlling parasites in the bloodstream. Eosinophils are tissue dwelling cells with reports of them establishing residency within intestinal tract(Mishra et al., 1999) where their absence is linked with issues in the mucosal barrier integrity(Chu et al., 2014), in the muscle where they are associated with muscle regeneration(Heredia et al., 2013), in the liver where they are also linked to regenerative functions(Goh et al., 2013) and in adipose tissue where they are linked to tissue homeostasis(Wu et al., 2011).

1.2.7 Emergency granulopoiesis

As neutrophils are believed to be short-lived cells which need to be continuously replenished in steady-state conditions, it is estimated that 10^9 cells/kg of body weight are produced every single day(Lawrence et al., 2018). However, the hematopoietic system is capable of performing adaptions in response to stress, such as during a severe infection, by increasing the overall production of specific cell types. Emergency granulopoiesis is an example of such an adaptation, where there is increased generation of neutrophils from myeloid progenitors in the bone marrow(Manz and Boettcher, 2014). Emergency granulopoiesis has been shown to lead to neutrophilia, an increase in the number of neutrophils in the blood, and to the emergence of a population of immature neutrophils in circulation. These immature neutrophils would be found in the bone marrow under normal conditions, however in emergency granulopoiesis they are reported to be present the blood stream. This occurrence has been associated to severe disease in areas as diverse as cancer(Sagiv et al., 2015) and COVID19(Reusch et al., 2021).

1.2.8 Adaptive immune system

The adaptive immune system is composed of cells that recognise specific pathogen derived antigens via a set of specialized cell surface receptors. The two types of cells that makes up the adaptive immune system are lymphocytes known as T cells and B cells. In adaptive immunity there is a first/primary response to the pathogenic antigen which causes a large number of effector lymphocytes to undergo clonal expansion to respond against the specific pathogen. Once the pathogen is cleared the effector populations undergo contraction phase characterised by the loss of effector cells and the emergence and maintenance of long-lived memory cells which are capable of mounting a stronger and more rapid response should there be another encounter with the same pathogen(Cantrell, 2015).

1.2.9 T cells

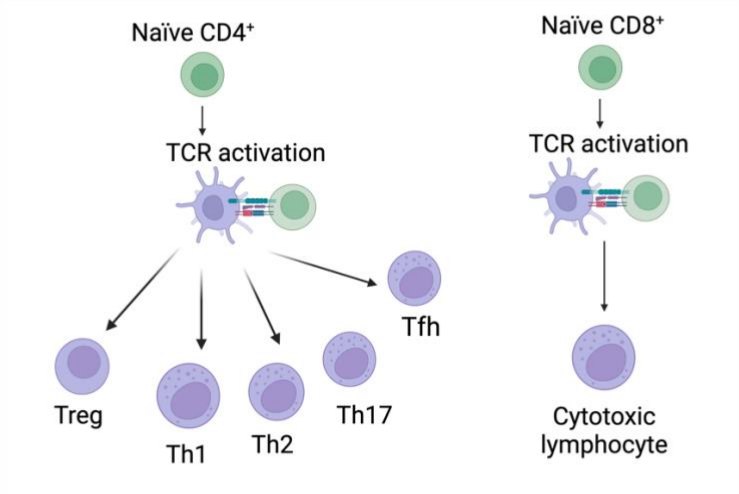

Figure 1.7: Schematic showing the main types of T cells, with the most prominent division being based on the expression of the CD4 or CD8 proteins. CD4$^+$ naïve T cells give rise to helper T cell lineages, and CD8$^+$ naïve T cells give rise to cytotoxic T cells.

T cells originally derive from HSCs in the bone marrow, but complete their development and maturation in the thymus, which gives rise to their name. All T cells express specific type of surface receptors, the T-cell antigen receptor (TCR). The majority of T cells express $\alpha\beta$ chain of the TCR, however there is also a subset that expresses $\gamma\delta$ chains. The $\gamma\delta$ T cells are more abundant in the barrier sites such as the intestines(Kabelitz et al., 2020). Conventional $\alpha\beta$ T cells recognise their ligands and are stratified by the expression of co-receptors with the capacity to recognise class I major histocompatibility complex (MHC), CD8, or class II MHC, CD4 (Fig. 1.7). $\gamma\delta$ T cells typically recognise their ligands independent of antigen processing and MHC complexes.

CD8$^+$ T cells respond to pathogens to differentiate into cytolytic effectors that directly kill infected or transformed cancer cells. CD4$^+$ T cells are known to produce an array of chemokines and different cytokines that can excerpt effects on T and/or B cell differentiation as well as innate immune cells like neutrophils. Because of these functions that are referred to as “helper” T cells(Cantrell, 2015), though the existence of cytolytic effector CD4$^+$ cells has also been reported(Hua et al., 2013). CD4$^+$ T cells are further stratified by their specific role within the adaptive immune system; Th1 cells are defined by the production of IFNγ and are involved in responses against intracellular pathogens, Th2 cells are defined by the production of interleukin (IL) 4 and IL13 and are typically involved in responses against extracellular pathogens, Th17 are defined by the production of IL17 and IL22 with pro-inflammatory functions. Regulatory T (Treg) cells are characterised by immunosuppression and preventing autoimmunity and finally follicular helper T (Tfh) cells regulate the development of B cells.

In T cell mediated immune responses there is a strong clonal expansion of antigen specific T cells followed by their differentiation into effector cells. At the end of the primary immune response and pathogen clearance a proportion of naïve T cells differentiate into long-lived memory T cells(Sprent and Surh, 2001). These memory T cells are then capable of mounting a more vigorous secondary response to antigens by destroying the infected cells, as is the role of cytotoxic T cells (CTLs) or controlling B cells responses (CD4$^+$ cells).

T cells also have important tissue residency characteristics, as a proportion of memory cells have been found to establish residency in a variety of tissues, like the lung(Teijaro et al., 2011), skin(Gebhardt et al., 2009, Mackay et al., 2013), intestinal epithelium(Masopust et al., 2010), liver(Pallett et al., 2017) among others. These tissue resident T cells lack the expression of homing receptors that direct them back towards the lymph nodes, like CD62L and CCR7, and also display specific adaptations to the tissue they reside in.

1.3 Web applications, programming languages and databases

An important part of this thesis revolves around building tools or applications that enable the visualisation and sharing of proteomics data. As the size and complexity of the datasets increases, the more important it becomes to provide effective ways to interact and extract information from such data. To achieve this goal, we created two web applications (web apps) described in chapter 5, the Encyclopedia of Proteome Dynamics (EPD) and the Immunological Proteome Resource (ImmPRes), both are apps and data ecosystems. To understand their design and functionality it is important to understand some basic computational concepts which are explained in this section.

1.3.1 R programming language

R is an interpreted programming language that was specifically developed for statistical analysis and data visualisation. R is accessed via a command line interface, however development environments with a graphical interface are available such as R Studio (https://www.rstudio.com/). R lacks the support for programming paradigms and as such it is not considered a suitable language for full application development.

One of its most important features is the integration of user created packages, which have allowed scientific analysis tools to flourish. One of the most valuable resources for analysing scientific data is Bioconductor(Gentleman et al., 2004), an open source software development projects for the analysis of data, which includes important packages for the analysis of proteomic data like limma(Ritchie et al., 2015), MSstats(Choi et al., 2014) and pRoloc(Gatto et al., 2014).

1.3.2 Python

Python is a programming language that was developed over 30 years ago. It is a language created to emphasise readability while supporting a wide array of programming paradigms, making it particularly useful to the development of web applications. Python has consistently ranked among the most popular programming languages over the last 10 years.

Python is frequently used for full application development, however implementations of a data driven packages like NumPy, SciPy and plotting packages like Matplotlib have also made python an effective tool to use for scientific computing and scripting. To date however, Python lacks the breadth of scientific data analysis and visualisation packages that are found in R.

1.3.3 Web applications (web app)

A web application (web app) is a software application that runs on a remote server that is connected to a network and is accessible via an online interface, therefore they require an internet connection to be accessed. They have been used in areas such as online retail for more than two decades now. The web app code is hosted and ran in the remote server previously mentioned, as such it can be updated without requiring users to perform any actions or changes to their system. They are generally accessed using a web browser, implying that an active internet connection is required for their use.

Web apps have also been used for scientific purposes, ranging from analytical apps that perform tasks like enrichment or overrepresentation analysis(Mi et al., 2017, Wang et al., 2017, Sherman et al., 2022) to providing access to some of biggest and most important scientific databases like UniProt(The UniProt, 2017) and Ensembl(Zerbino et al., 2018) and providing a simple way for users from all around the world to access services or data.

Web apps have also more recently been used as a tool to share results and provide access to the data derived from large-scale scientific experiments across the different omics domains, including sequencing(Heng et al., 2008, Immunological Genome, 2020, Uhlen et al., 2019) and proteomics(Brenes et al., 2018a, Rieckmann et al., 2017, Go et al., 2021, Tinti and Ferguson, 2022, Wolf et al., 2020).

1.3.4 Databases

Databases are defined as highly organised collections of electronic data that are stored in one or multiple computers/servers. Different types of databases have different modelling, storage and querying language properties. The first prominent database system that was created was the relational database(Codd, 1970), which models the data storage as a collection of tables consisting of rows and columns. It unified data and metadata and defined an access language based on algebra, SQL (structured query language), while providing a layer of abstraction from how the data was physically stored in the servers.

The main limitation of relational databases is that they require a comprehensive understanding of all data requirements from the modelling and design stage. They are therefore considered inflexible and difficult to redesign when the data requirements change, especially if the third normal form design approach is followed. Relational databases are generally not distributed; hence it is normally not possible to store subsets of the database on different servers or computers. Thus relational databases were limited to single servers, which made it vulnerable to hardware failure and limited in capacity by the hardware that could be installed within a single machine.

The advent of Big Data, i.e. datasets that are too large, varied or of a complex structure to be stored in traditional relational databases(Sagiroglu and Sinanc, 2013), gave rise to a wide variety of new databases with more flexible data models that collectively referred to as noSQL (not only structured query language) databases. These databases covered different implementation methods, modelling techniques and storage strategies, but in general shared a distributed characteristic. These could be stored, frequently with redundant copies of a subset of the data, on multiple computers or servers.

The distributed nature reduced the risk related to hardware failure and provided a considerable performance increase, as data could be partitioned and written to different servers at the same time, and data could be read from multiple servers with large numbers of disks at the same time too. NoSQL databases include graph databases like Neo4j, document databases like MongoDB, key-value wide-column databases like Apache Cassandra among others.

Acknowledgements

Figures in this chapter were generated using biorender.com

Chapter 2 – Technical challenges and issues related to large-scale proteomics using TMT

2.1 Publications relating to this chapter

- Multi-batch reveals false positives, batch effects and missing values (Brenes et al., 2019)
- This manuscript has had minor updates and adaptions to fit the thesis

2.1.1 Contributions to the publications on this chapter

- I proposed and conceptualized the study
- I analysed the dataset
- I interpreted the data
- I produced all the figures used in the study
- I wrote the paper with Angus Lamond and with feedback from all authors

2.1.2 Main contributions of the other authors to the publications on this chapter

- The hiPSC cells lines were grown and reprogrammed in the Sanger Centre as part of the HipSci Consortium(Kilpinen et al., 2017)
- The mass spectrometry raw files were generated In the Lamond Laboratory by Dalila Bensaddek
- Jens Hukelmann assisted in the interpretation of the data

2.2 Introduction

High-throughput, shotgun proteomics, using data dependent acquisition (DDA), now enables the comprehensive study of proteomes, allowing the identification of 10,000 or more proteins from cells and tissues(Bekker-Jensen et al., 2017, Beck et al., 2011, Meier et al., 2018, Kawashima et al., 2022, Muntel et al., 2019a). However, to achieve such deep proteome coverage using DDA, extensive prefractionation of extracts prior to mass spectrometry (MS) analysis used to be frequently required(Camerini and Mauri, 2015, Bekker-Jensen et al., 2017). Furthermore, to evaluate statistically the significance of the resulting data, a minimum of 3 replicates for each sample/condition is also necessary(Rost et al., 2015, Turck et al., 2007), therefore also increasing acquisition time involved is increased still further for experiments that analyse the multi-dimensional characteristics of the proteome; e.g. studying differences in PTMs like phosphorylation (Zhou et al., 2001, Beausoleil et al., 2006, Villen et al., 2007, Dephoure et al., 2008, Bekker-Jensen et al., 2020, Collins et al., 2007, Martinez-Val et al., 2021, Huang et al., 2019, Searle et al., 2019, Ochoa et al., 2020, Urbaniak et al., 2013, Nett et al., 2009), protein-protein interactions (PPI) (Fleming et al., 2016, Sullivan et al., 2013, Trinkle-Mulcahy et al., 2008, Mellacheruvu et al., 2013, Ji et al., 2021, Leitner et al., 2012, Kirkwood et al., 2013, Crozier et al., 2017, Larance et al., 2016) , protein subcellular localisation (Thul et al., 2017, Dunkley et al., 2004, Christoforou et al., 2016, Dunkley et al., 2006, Sadowski et al., 2006, Mulvey et al., 2017, Christopher et al., 2021, Itzhak et al., 2016, Guther et al., 2014, Mardakheh et al., 2016, Geladaki et al., 2019) or protein synthesis and turnover (Boisvert et al., 2012, Ly et al., 2018, Tinti et al., 2019, Doherty et al., 2009, Schwanhausser et al., 2011, Welle et al., 2016, Martin-Perez and Villen, 2017, Zecha et al., 2018, Rolfs et al., 2021).

To cope with the challenges of large-scale DDA based proteomics analyses, strategies have been developed to allow multiple samples to be analysed in parallel, through multiplexing isotopically tagged peptides. Two of the main examples are, TMT(Thompson et al., 2003) and iTRAQ(Ross et al., 2004), which both use isobaric tags for simultaneous peptide identification and quantification. TMT in particular is widely used(Isasa et al., 2015, McAlister et al., 2014), reflecting the ability of multiplexed TMT to increase sample throughput in proteomics studies and reduce the “missing values” problem that arises from the stochastic sampling inherent in DDA proteomics(Lazar et al., 2016, Webb-Robertson et al., 2015). Thus, within a single multiplex TMT batch, the number of missing values at the protein level is low, frequently <2%(Isasa et al., 2015). Furthermore, the precision of the quantification within a multiplex TMT batch is high(O'Connell et al., 2018). However, it is less clear how well multiplexed TMT performs for very large-scale analyses, involving numerous TMT batches.

Using a proteomic dataset of human hiPSC cells, involving 24 separate 10-plex TMT batches(Brenes et al., 2018b). We compare the quantitation of data both within and between 10-plex batches and focus our analysis on 4 main issues: (i) missing values, (ii) accuracy of quantification, (iii) false positives and (iv) the effect of both reporter ion interference (RII) and co-isolation interference (CII).

We show there is an inflationary effect on missing values as data from multiple batches are integrated both at the protein and peptide level. We evaluated reproducibility both by studying the coefficient of variation (CV) within each 10-plex TMT batch, and by comparing a reference line (technical replicates of the hiPSC line "bubh_3") that were common to every batch. Furthermore, the incidence of false positives was studied by using Y chromosome peptides as an internal control. The hiPSC lines quantified in this dataset were derived from 163 different donors including both male and female, hence the peptides mapped to the Y chromosome should be absent from female lines. Nonetheless, we confirm that these Y chromosome-specific peptides were consistently detected in the female channels of all TMT batches. Finally, by using these Y chromosome peptides, we quantified the effect of ion co-isolation and reporter ion interference upon TMT quantification accuracy.

2.3 Methods

2.3.1 Experimental design and statistics rationale

The study consists of 240 hiPSC replicates, 217 biological replicates and 24 technical replicates, derived from 163 different donors. The study comprises twenty-four 10-plex TMT batches. Each batch consisted of 1 common reference line (technical replicates of hiPSC cell line "bubh_3") and 9 different hiPSC cell lines. The technical replicates were used for the data normalisation strategy described below. Out of the 240 replicates analysed, 142 were derived from female donors and from 98 male donors.

2.3.2 TMT Sample preparation

For protein extraction, hiPSC cell pellets were washed with ice cold PBS and redissolved immediately in 200 µL of lysis buffer (8 M urea in 100 mM triethyl ammonium bicarbonate (TEAB)) and mixed at room temperature for 15 minutes. Cellular DNA was sheared using ultrasonication (6 X 20 s on ice). The proteins were reduced using tris-carboxyethylphosphine TCEP (25 mM) for 30 minutes at room temperature, then alkylated in the dark for 30 minutes using iodoacetamide (50 mM). Total protein was quantified using the EZQ assay (Life Technologies). For the first digestion with mass spectrometry grade lysyl endopeptidase, Lys-C (Wako, Japan), the lysates were diluted 4-fold with 100 mM TEAB then further diluted 2.5-fold before a second digestion with trypsin. Lys-C and trypsin were used at an enzyme to substrate ratio of 1:50 (w/w). The digestions were carried out overnight at 37ºC, then stopped by acidification with trifluoroacetic acid (TFA) to a final concentration of 1% (v:v). Peptides were desalted using C18 Sep-Pak cartridges (Waters) following manufacturer's instructions.

For tandem mass tag (TMT)-based quantification, the dried peptides were re-dissolved in 100 mM TEAB (50 μl) and their concentration was measured using a fluorescent assay (CBQCA, Life Technologies). For each 10-plex TMT batch 100 μg of peptides from each cell line to be compared, in 100 μl of TEAB, were labelled with a different TMT tag (20 μg ml^{-1} in 40 μl acetonitrile) (Thermo Scientific), for 2 h at room temperature. After incubation, the labelling reaction was quenched using 8 μl of 5% hydroxylamine (Pierce) for 30 min and the different cell lines/tags were mixed and dried *in vacuo*.

The TMT samples were fractionated using off-line, high-pH reverse-phase (RP) chromatography: samples were loaded onto a 4.6 × 250 mm Xbridge BEH130 C18 column with 3.5-μm particles (Waters). Using a Dionex bioRS system, the samples were separated using a 25-min multistep gradient of solvents A (10 mM formate at pH 9) and B (10 mM ammonium formate pH 9 in 80% acetonitrile), at a flow rate of 1 ml min^{-1}. Peptides were separated into 48 fractions, which were consolidated into 24 fractions. The fractions were subsequently dried and the peptides re-dissolved in 5% formic acid and analysed by LC–MS/MS.

2.3.4 TMT LC–MS/MS

TMT-based analysis. Samples were analysed using an Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific), equipped with a Dionex ultra-high-pressure liquid-chromatography system (RSLCnano). RPLC was performed using a Dionex RSLCnano HPLC (Thermo Scientific). Peptides were injected onto a 75 µm × 2 cm PepMap-C18 pre-column and resolved on a 75 µm × 50 cm RP- C18 EASY-Spray temperature-controlled integrated column-emitter (Thermo Scientific), using a four-hour multistep gradient from 5% B to 35% B with a constant flow rate of 200 nl min^{-1}. The mobile phases were: 2% ACN incorporating 0.1% FA (solvent A) and 80% ACN incorporating 0.1% FA (solvent

B). The spray was initiated by applying 2.5 kV to the EASY-Spray emitter and the data were acquired under the control of Xcalibur software in a data-dependent mode using top speed and 4 s duration per cycle. The survey scan is acquired in the orbitrap covering the *m*/*z* range from 400 to 1,400 Thomson with a mass resolution of 120,000 and an automatic gain control (AGC) target of 2.0 x 10^5 ions. The most intense ions were selected for fragmentation using CID in the ion trap with 30% CID collision energy and an isolation window of 1.6 Th. The AGC target was set to 1.0 x 10^4 with a maximum injection time of 70 ms and a dynamic exclusion of 80 s, the scan rate was set to 'Rapid'. During the MS3 analysis for more accurate TMT quantifications, 5 fragment ions were co-isolated using synchronous precursor selection with a window of 2 Th and further fragmented using HCD collision energy of 55%. The fragments were then analysed in the orbitrap with a resolution of 60,000. The AGC target was set to 1.0 x 10^5 and the maximum injection time was set to 105 ms.

2.3.5 Machine, blanks and standards

All of the TMT batches were analysed on the same Orbitrap Fusion MS instrument. Between each individual TMT experiment, one blank was run, followed by analysis of a 15 peptide Retention Time Calibration (RTC) standard, to evaluate retention time drift. This was followed by analysis of an MCF10a total cell digest standard to evaluate peptide and protein identifications. The last step consisted of analysis of two blanks, one with an oscillating gradient and one with the gradient matching the samples to be run.

2.3.6 Identification & Quantification

The data from all twenty-four 10-plex TMT batches were collected and analysed simultaneously, using Maxquant(Cox and Mann, 2008, Tyanova et al., 2016a) v. 1.6.3.3. The FDR threshold was set to 1% for each of the respective Peptide Spectrum Match (PSM) and Protein levels. The data was searched with the following parameters; type was set to Reporter ion MS3 with 10plex TMT, stable modification of carbamidomethyl (C), variable modifications, oxidation (M), acetylation (protein N terminus), deamidation (NQ), Glutamine to pyro-glutamate (N-terminus), with a 2 missed tryptic cleavages threshold, reporter mass tolerance set to 0.03 ppm. Minimum peptide length was set to 7 amino acids. Proteins and peptides were identified using UniProt (SwissProt December 2018). The run parameters are accessible at ProteomeXchange(Vizcaino et al., 2014) via the PRIDE repository(Vizcaino et al., 2016, Perez-Riverol et al., 2018), along with the full MaxQuant(Cox and Mann, 2008) quantification output (PDX010557).

2.3.7 Filtering

All proteins that were marked as ‘Reverse’, ‘Potential Contaminants’ or ‘Only identified by site’ were discarded. The final subset comprised 9,640 proteins. Peptides marked as ‘Potential contaminants’ or ‘Reverse’ were also filtered from the analysis. The final peptide dataset comprised 178,491 peptides.

2.3.8 Copy number generation

Protein copy numbers were calculated following the proteomic ruler approach(Wisniewski et al., 2014). For protein, p, $uCN_{b,c,p}$ is the uncorrected protein copy number:

$$uCN_{b,c,p} = protein\ MS3\ signal_{b,c,p} \times \frac{A}{M_p} \times \frac{6.85 * 10^{-12}}{\sum_{h \in b,c} histones\ MS3\ signal_h}$$

$$for\ batch\ b\ \in \{1,2, \dots 24\}\ and\ channel\ c\ \in \{126C, 127N, \dots, 131N\}$$

where A is Avogadro’s constant, M_p is the molar mass of the protein p, protein MS3 signal is the protein MS3 intensity and histones MS3 signal is the MS3 intensity for all histones, h.

These uncorrected copy numbers, which will be referred to here as “raw copy numbers”, were used to study the coefficient of variation (CV). To control for technical variation between the 24 different 10-plex batches, a correction factor, cf, was applied to every protein, p, in every batch, b, to adjust the protein copy numbers.

$$cf_{b,p} = \frac{uCN_{b,126C,p}}{\sum_b uCN_{b,126C,p}/24}$$

$$for\ b\ \in \{1,2, \dots 24\}$$

where $uCN_{b,126C,p}$ is the protein copy number derived from reporter channel 126C (the reference channel). The normalised copy number, $normCN$, is calculated for protein, p in all batches, b, and all channels c:

$$normCN_{b,c,p} = \frac{uCN_{b,c,p}}{cf_{b,p}}$$

$$for\ batch\ b\ \in \{1,2, \dots 24\}\ and\ channel\ c\ \in \{126C, 127N, \dots ,131N\}$$

2.3.9 Missing value calculations

First, to estimate missing values within this DDA analysis, a list of unique proteins/peptides that were detected with at least 1 reporter intensity greater than zero were calculated for each batch. To determine the number of missing values within each 10-plex TMT batch, the number of unique proteins/peptides per reporter channel was compared to the number of unique proteins/peptides identified within the batch. This approach was applied to generate the missing value calculations for each of the 24 individual 10-plex TMT batches.

To assess the effect of integrating multiple TMT batches, random sampling was performed to estimate how missing values are affected by a progressive increase in the number of 10-plex TMT batches analysed. This was performed in an incremental fashion starting from 2 and finishing with 22 batches (PT6388 was not used for this analysis), with 500 iterations per level. At the first level 2 batches would be selected at random with no replacement, and at the last level 22 batches would be selected at random, again with no replacement. This was performed with the R function “sample()” part of the base R-core package.

At each level a new list of proteins/peptides detected with at least 1 reporter ion intensity greater than zero within any of the integrated TMT batches was calculated, and the number of proteins/peptides with intensity greater than 0 per reporter channel was evaluated against the new list.

2.3.10 Coefficient of variation

The coefficient of variation (CV) in protein abundance levels was calculated using the log10 transformed protein copy numbers.

$$CV = \frac{S}{X} \times 100$$

For each protein the CV is equal to the copy number standard deviation (S) divided by the mean copy number (X) times 100. The protein CV within each 10-plex TMT batch was calculated for all 10 cell lines within the same batch, using all proteins detected in every reporter channel. The reference line CV was calculated using proteins that were detected in the TMT^{10} -126C (reference line) channel across all of the 24 10-plex TMT batches.

2.3.11 Correlation Clustering

For each 10-plex TMT batch, a concordance correlation value was calculated for all cell lines within the same batch. The calculations were performed using “correlation()” function from the R package “agricolae” version 1.2.8.

The same process was applied to calculate the concordance correlation values for the reference lines, i.e. using reporter channel 126C in all TMT batches.

2.3.12 Peptide intensity normalisation

The replicate normalised intensity, rni, was calculated per peptide, q:

$$rni_q = log_{10}(\frac{peptide\ MS3\ signal_{b,c,q}}{median(I_{b,c})})$$

$$I_{b,c} = \{peptide\ MS3\ signal_{b,c,q}\ :\ \forall q\}\ given\ batch, b\ and\ channel, c$$
$$for\ batch\ b\ \in \{1,2, \dots 24\}\ and\ channel\ c\ \in \{126C, 127N, \dots, 131N\}$$

The median normalised intensity, mni, for peptide, q, is the median of all batches, b, and channels, c:

$$mni_q = median(rni_{b,c,q})$$

$$for\ all\ batches\ b\ \in \{1,2, \dots 24\}\ and\ channels\ c\ \in \{126C, 127N, \dots, 131N\}$$

The $global\ median$ is the median of mni for all peptides, q:

$$global\ median = median(mni_q)$$

2.3.13 Reporter ion interference classification

The reporter ion interference (RII) targets are based on a typical product data sheet for 10-plex TMT Label Reagents from ThermoFisher Scientific, as summarised in the table below:

Mass Tag	Reporter channel	-1Da (secondary RII)	+1Da (primary RII)
TMT^{10} -126	0	--	127C
TMT^{10} -127N	1	--	128N
TMT^{10} -127C	2	126	128C
TMT^{10} -128N	3	127N	129N
TMT^{10} -128C	4	127C	129C
TMT^{10} -129N	5	128N	130N
TMT^{10} -129C	6	128C	130C
TMT^{10} -130N	7	129N	131
TMT^{10} -130C	8	129C	--
TMT^{10} -131	9	130N	--

2.3.14 Analysis of reporter ion interference

To study the effect of reporter ion interference across different TMT channels, we selected a subset of 69 peptides that were specific to the following list of protein coding genes uniquely located on the Y chromosome; “CDY1”, “CDY2A”, “DDX3Y”, “EIF1AY”, “KDM5D”, “NLGN4Y”, “PCDH11Y”, “RPS4Y1”, “TBL1Y”, “USP9Y” & “UTY”.

This approach of using peptide values from Y chromosome specific genes depends upon there being a diverse mixture of male and female donor-derived hiPSC lines in each 10-plex TMT batch. However, two of the 24 TMT batches comprised exclusively female donor-derived hiPSCs, which had been shown not to have Y chromosome derived DNA in QC analyses[(Kilpinen et al., 2017)]. For these female donor-specific batches, any peptide assigned to Y chromosome specific genes was excluded from the analysis. An additional batch, PT6388, displayed an irregular behaviour, and was hence also discarded from the posterior analysis. A final subset of 65 Y chromosome-specific peptides were used for this analysis (supplemental data(Brenes et al., 2019) for list).

2.3.15 Peptide male vs reporter ion interference ratios

The peptide ratios comparing male channels vs female channels, mpr, subjected to different reporter ion interference conditions, $cond$, were calculated per 10-plex TMT batch, b, for peptide, q, utilising the replicate normalised intensities:

$$mpr_{b,q} = \frac{median(RNI_{b,male,q})}{median(RNI_{b,cond,q})}$$

$$RNI_{b,male,q} = \{rni\ _{b,c,q}: \forall\ male\ channels, c\},$$

$$RNI_{b,cond,q} = \{rni_{b,c,q}: \forall\ cond\ channels, c\},$$

$$for\ b\ \in \{1,2, \dots 24\}\ and\ cond\ \in \{primary\ RII, secondary\ RII, double\ RII, no\ RII\}.$$

The box plot comparing male replicates to the different reporter ion interference conditions used these peptide batch ratios and was plotted using “ggplot2” version 3.0.0(Wickham, 2016).

2.3.16 Peptide reporter ion interference ratios

The peptide ratios, npr, comparing different reporter ion interference (RII) conditions, $cond$, in female channels to female channels with no reporter ion interference, $noRII$, were calculated within each 10-plex TMT batch, b, for peptide, q, utilising the replicate normalised intensities.

$$npr_{b,q} = \frac{median(RNI_{b,cond,q})}{median(RNI_{b,noRII,q})}$$

$$RNI_{b,cond,q} = \{rni_{b,c,q} : \forall\ cond\ channels, c\},$$

$$RNI_{b,noRII,q} = \{rni\ _{b,c,q} : \forall\ noRII\ channels, c\},$$

$$for\ b\ \in \{1,2, \ldots 24\}\ and\ cond\ \in \{primary\ RII, secondary\ RII, double\ RII\}.$$

These results were stratified by the global median, where peptides with median normalised either intensity greater than or equal to the global median were considered 'High intensity' and those lower than the global median were considered 'Low intensity'. The box plot comparing different reporter ion interference conditions to the replicates not affected by reporter ion interference used these peptide batch ratios and was plotted using "ggplot2" version 3.0.0(Wickham, 2016).

2.4 Results

2.4.1 Inflation of missing values in multi-batch TMT

A known advantage of using TMT is the low index of missing values that are present within a single TMT batch. Recent studies report as low as <1% missing values at the protein level(O'Connell et al., 2018), albeit data are usually not reported at the peptide level.

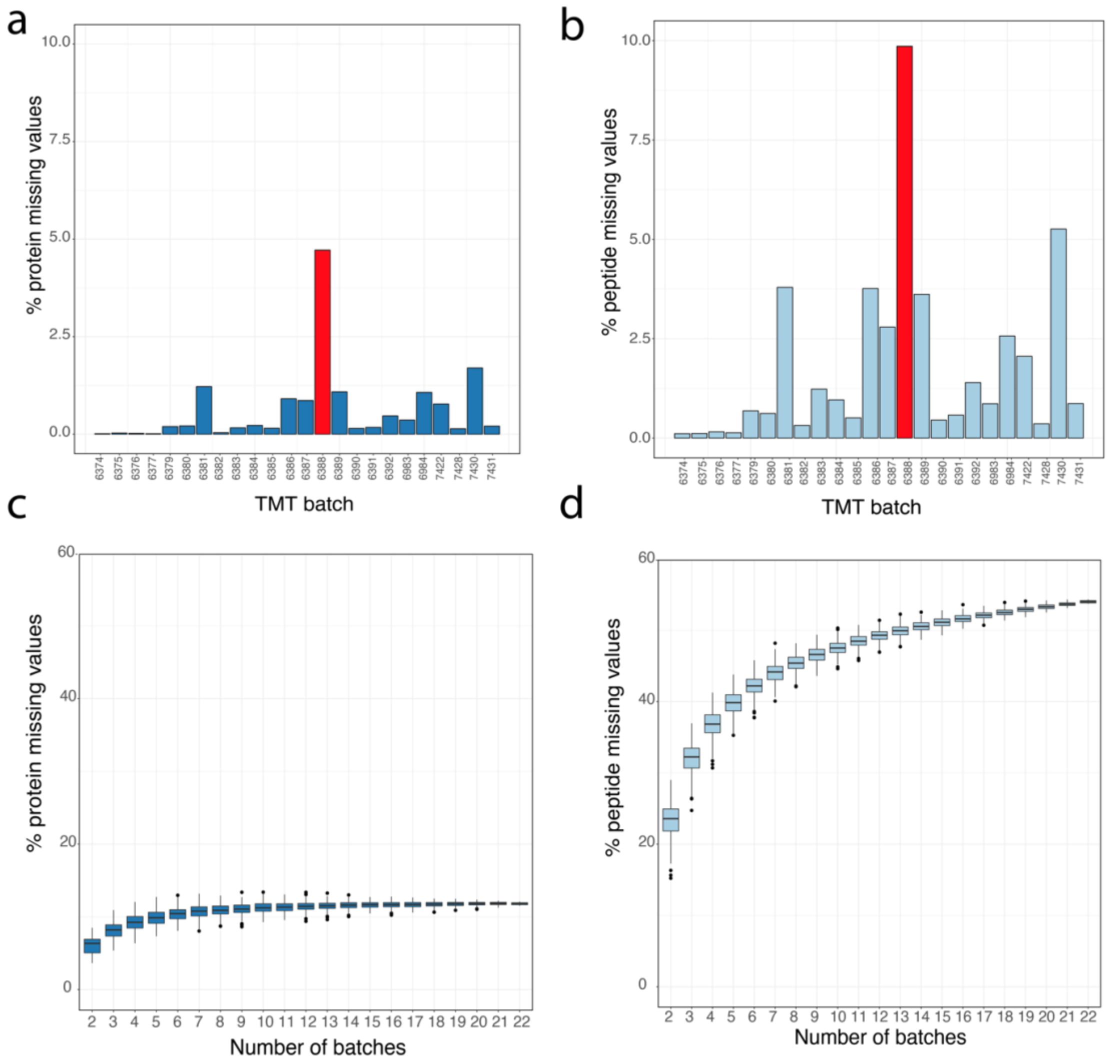

Fig. 2.1- Protein and peptide missing values: (*a*) Bar plot showing the percentage of missing values for each TMT batch calculated at the protein level. **(*b*)** Bar plot showing the percentage of missing values for each TMT batch calculated at the peptide level. **(*c*)** Box plot showing the results for protein missing values as a function of the number of 10-plex TMT batches (see methods). **(*d*)** Box plot showing the results for peptide missing values as a function of the number of 10-plex TMT batches (see methods). For all boxplots the lower and upper hinges represent the 1st and 3rd quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge, the lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge.

We searched our hiPSCs data against the smaller SwissProt database instead of TrEMBL and started by analysing the hiPSC 10-plex TMT data for the number of missing values at the protein level within each TMT batch (Fig. 2.1a). It is important to note this dataset uses extensive fractionation, 24 fractions with 4-hour gradient per TMT-batch, to maximise depth of coverage and minimise missing values. With reduced fractionation the results would be considerably more pronounced than what is shown here. The preliminary results are consistent with previous reports, i.e. 92% of the 24 different 10-plex TMT batches show <1% missing values at the protein level, with only 1 outlier with missing protein values >1.5%. This was experiment PT6388, which is highlighted in red

(Fig. 2.1a&b). Furthermore, when we analyse the data at the peptide level, there is very close agreement to the protein data, with 92% of the 24 different 10-plex TMT batches having <5% missing peptide values, however the outlier batch has an exacerbated effect and displays 9% missing values at the peptide level. We therefore excluded PT6388 from the rest of the analysis.

These previous results do not address the effect of integrating data from multiple, independent 10-plex TMT batches into a single analysis. To study the effect of data integration, we increased the number of batches selected, from 2 to 22 and recalculated the number of missing values that were present (Fig. 2.1c&d; see methods). At the protein level, the median number of missing values increases from 0.19% with one 10-plex TMT batch, to 6.35% when data from a second 10-plex TMT batch were integrated (Fig. 2.1c). When we integrate data from 5 different 10-plex TMT batches, the median number of missing values at the protein level escalated to ~10%. This situation was exacerbated when the analysis was performed at the peptide level (Fig. 2.1d). When integrating data from just two 10-plex TMT batches, the median number of missing peptide values was >23%. Even more striking, it only required integrating data from 5 different 10-plex TMT batches to produce ~40% missing values at the peptide level. The data suggest peptides are not reproducibly detected among batches, but whether this effect was limited to low abundance peptides remained unknown.

Based on the previous results, we decided to perform a more in-depth analysis on the inflation rate of peptide missing values. We observed that the number of peptides identified within each 10-plex TMT batch is relatively constant (Fig. 2.2a), but found it variable across the different TMT-batches. The median number of peptides identified per batch was 84,046 with a standard deviation of 11,354. To further analyse these peptide level data, we first median-normalised the MS3 intensities for all peptides in all cell lines (see methods). The $\log_{10}$ median normalised MS3 intensities spanned 6 orders of magnitude (see methods; Fig. 2.2b).

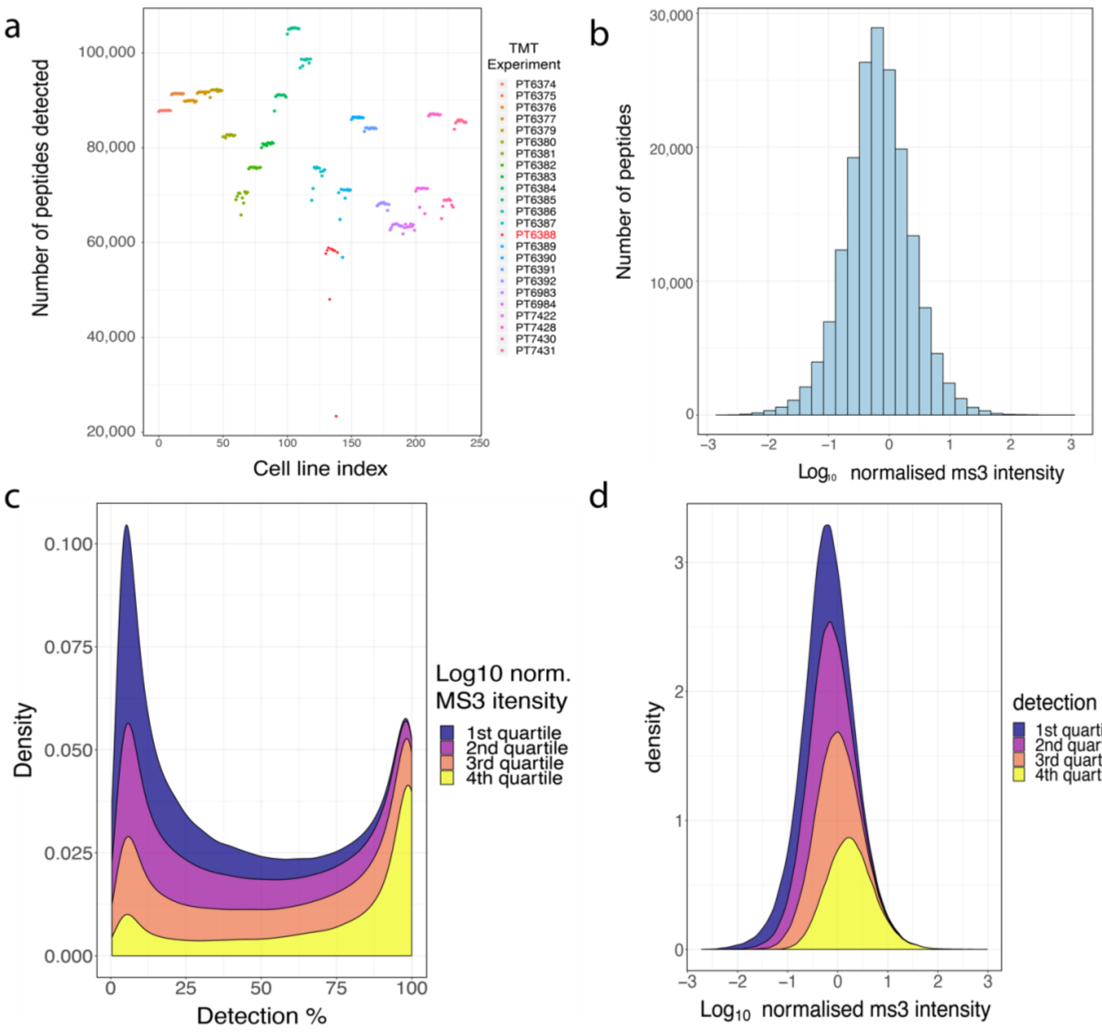

Fig. 2.2 - Peptide identifications and intensities: ***(a)*** Number of peptides identified with MS3 intensity greater than zero in all TMT channels, coloured by TMT batch. ***(b)*** Histogram of the median normalized peptide intensity (see methods). ***(c)*** Stacked density plot showing peptides grouped by normalized peptide intensity quartiles and their percentage of detection across all TMT channels. ***(d)*** Stacked density plot showing quartiles of identification rates for each peptide and their corresponding $\log_{10}$ normalized MS3 intensity.

We next analysed the peptide dataset by quartiles, based on the $\log_{10}$ median normalised intensity values (Fig. 2.2c). The first quartile represented the 25% least abundant peptides and the fourth quartile the 25% most abundant peptides. There are only 11 peptides among the 25% least abundant peptides that are detected in all TMT replicates and only 603 peptides that are seen in > 90% of the replicates. As DDA selects the n most abundant ions reaching the mass-spectrometer during a MS1 scan(Hu et al., 2016) (where n typically is 10-30), this bias is predictable. However, when we analyse the results from the 25% most abundant peptides (which have a median normalised intensity of 0.32, representing the 84th percentile of abundance), we see that 26% of these peptides are still only detected in less than half of all replicates. Furthermore only 24% of these peptides were detected in all the samples. In total there are 12,140 peptides that were detected in all replicates, regardless of intensity classification, which represents a mere 6.81% of all peptides identified.

Next, we analysed the data by comparing the identification quartiles, organised by the percentage of replicates in which they were detected (Fig. 2.2d). The first quartile represented the 25% of peptides that were detected least frequently, i.e. in less than 29 TMT replicates. Of these peptides ~1/3 of them had an intensity higher than the global median (see methods; -0.184), highlighting that even relatively abundant peptides are not identified consistently. Overall, about half of the total number of peptides identified are detected in <40% of all replicates, showing that like other DDA methodologies, multi-batch TMT also suffers from an inflation missing values.

2.4.2 Variation between 10-plex TMT batches

Multiple studies have documented TMT as a method producing precise quantitation, in some cases having a coefficient of variation (CV) ~3x lower than comparable label free data(O'Connell et al., 2018). Most of these studies have focused on analysing quantitative precision within a single TMT batch, and do not explore the effect of integrating data from multiple TMT batches into one analysis. However, large-scale DDA based proteomic experiments looking at hundreds of cell lines or patients need to employ multiple TMT batches in a single experiment(Hennrich et al., 2018, Nusinow et al., 2020), hence understanding the multi batch effect is vital.

We calculated protein copy numbers for 230 hiPSC replicates, which included 208 biological replicates and 23 technical replicates of a control hiPSC line (bubh_3), across 23 10-plex TMT batches (PT6388 is once again removed from the analysis). We then proceeded to calculate Lin's concordance correlation coefficient, which measures the agreement between two variables and is proposed to evaluate reproducibility(Lin, 1989). This was done for every hiPSC line within each TMT 10-plex batch, and for all the technical replicates of the control line, bubh_3, which was always allocated to channel TMT^{10} 126, across the 23 10-plex TMT batches.

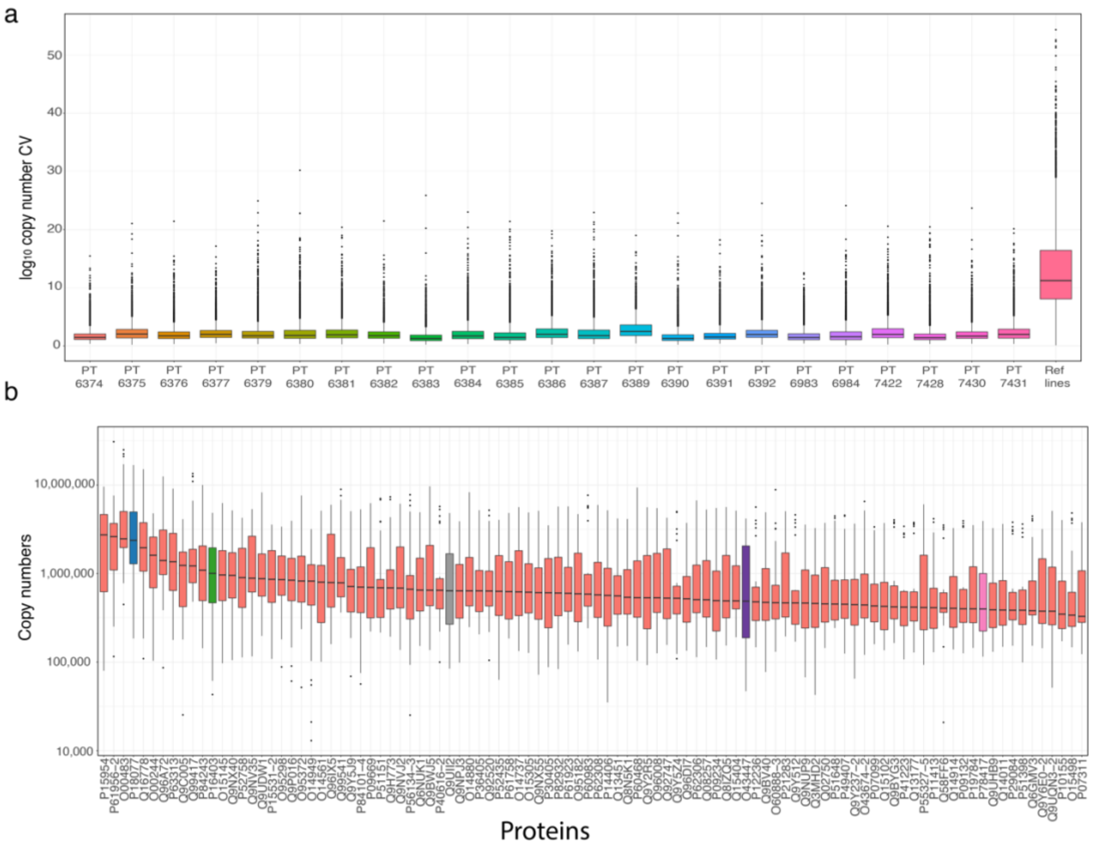

Fig. 2.3 - Variation: ***(a)*** Box plots showing the protein copy number coefficient of variation for all proteins detected in each 10-plex TMT batch as well as all proteins detected in all the reference line replicates (TMT channel 126C in all batches). ***(b)*** Box plots showing the protein copy numbers of the 100 most abundant proteins with a coefficient of variation greater than or equal to 7.5 across all reference line replicates (TMT channel 126C in all batches). For both boxplots the lower and upper hinges represent the 1st and 3rd quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge, the lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge.

The concordance correlation coefficient within each 10-plex TMT batch is very high, (median value 98% concordance), highlighting the precision of the quantitation within each single batch. However, when the same calculation is applied to the technical replicates of the control hiPSC line across 23 respective batches, the median concordance coefficient drops to 81%. Highlighting clear batch effects.

To explore this situation further, we calculated the CV for the $\log_{10}$ transformed protein copy numbers$^{\text{(Limpert et al., 2001)}}$ (see methods), both within each 10-plex TMT batch, and across 23 controls (Fig. 2.3a). When we calculated the protein CV exclusively within each 10-plex TMT batch, the median was 1.72, and with all 10-plexes showing a median protein CV <2.5. Accordingly, the data show that for every batch, proteins with a CV >7.5 were considered outliers (Fig. 2.3a). These data show high precision of quantitation within each individual multiplexed experiment. Similarly, to evaluate accuracy we calculated the CV for all technical replicates of the reference hiPSC line, bubh_3, which were analysed in channel TMT^{10} 126 in every 10-plex. The median CV of all the proteins detected in the technical replicates, was 11.03, i.e. 6.4-fold higher than the median within-batch CV. The CV of more than 80% of all proteins quantified in the technical replicates would be considered outliers in all the within batch 10-plex TMT analyses.

It is commonly assumed in proteomics studies that variation predominantly affects low intensity proteins and peptides. We decided to test this assumption and analysed the variation of extremely abundant proteins. We focused the analysis on the 23 technical replicates of bubh_3, and selected the top 100 most abundant proteins where the CV was greater than 7.5 and created box plots based on their copy numbers (Fig. 2.3b).

We chose to highlight 5 cases; 60S ribosomal protein L35a (RPL35A; highlighted in blue), Histone H1.2 (HIST1H1C; highlighted in green), ATPase inhibitor, mitochondrial (ATP5IF1; highlighted in grey), Peptydyl-prolyl cis-trans isomerase H (PPIH; highlighted in purple) and Glutathione S-transferase omega-1 (GSTO1; highlighted in pink). RPL35A ranges in expression from ~16,785,000 to ~185,000 copies and was identified with 10 unique + razor peptides (URP). HIST1H1C ranges from ~4,847,000 to ~43,000 copies and was identified with 6 URP, ATP5IF1 from ~4,825,000 to ~84,000 copies and was identified with 6 URP, PPIH from ~4,180,000 to ~45,000 copies and was identified with 14 URP and GSTO1 from ~3,920,000 to ~115,000 and was identified with 16 URP. All of these 5 proteins are highly abundant, with a median copy number > 500,000 and have been identified with >6 peptides that were used for the quantitation, yet their expression levels vary drastically between the different technical replicates of bubh_3, highlighting that the variation is not limited to low abundance proteins.

We also note that while the majority of hiPSC lines in this study come from healthy donors, some of these TMT batches, e.g. PT6390, contain mixtures of hiPSC lines derived from both healthy donors and donors with rare genetic diseases, including “Usher syndrome”, “Monogenic Diabetes” and “Bardet-Biedl syndrome”. Nonetheless, the median protein CV within PT6390 is still ~10 fold lower than the CV obtained from analysing the 23 technical replicates of bubh_3, indicating that TMT batch effects have a bigger influence on the proteomics data than a healthy vs diseased physiology.

Our results highlight that while multiplex TMT is a useful and precise methodology for quantitative proteomics, it is important to be aware also of its potential limitations, particularly when attempting large-scale experiments which will require multiple TMT batches to be integrated. It is worth noting that copy numbers already provide a layer of normalisation(Wisniewski et al., 2014), however it proves ineffective at dealing with batch variation. These findings underline it is essential to understand the potential for batch variation to affect data quality. To reduce the batch effect several methods have been proposed, from protein expression inference(Schwacke et al., 2009), to a standard reference line(Herbrich et al., 2013). Here we used the technical replicates of a reference hiPSC line as an internal reference standard to control for variation between batches (see methods). Using this normalisation method provided a median CV of 2.96% across all cell lines and technical replicates, making the results closer to the metrics obtained within each TMT batch.

2.4.3 Incidence of false positives

The hiPSC dataset(Brenes et al., 2018b, Kilpinen et al., 2017) provided us with an excellent opportunity to study the incidence of false positives within a multi-plex TMT batch. The study utilised hiPSC lines derived from both male and female donors within twenty-two of the twenty-four 10-plex TMT batches analysed here. Since only the lines from male donors should include proteins encoded by genes exclusively on the Y chromosome, this effectively provided a set of endogenous "spike-in" peptides, which we could use to monitor the presence of false positives as well as exploring the effects of reporter ion interference (RII) between TMT channels and co-isolation interference (CII).

The dataset detected 11 proteins that were mapped to the Y chromosome. Correspondingly, all unique peptides derived from these proteins should only be present in the TMT channels with male cell lines and, in theory, should be absent in the TMT channels with female cell lines. To avoid mismatches arising from shared peptides, we focussed our analysis on a subset of 69 peptides that mapped uniquely to the following Y chromosome specific genes; "CDY1", "CDY2A", "DDX3Y", "EIF1AY", "KDM5D", "NLGN4Y", "PCDH11Y", "RPS4Y1", "TBL1Y", "USP9Y", "UTY". Additionally, since two of the 10-plex TMT batches analysed (PT6384 and PT7422) had only female cell lines, any potential Y chromosome-specific peptides that were detected in these batches were treated as incorrect identifications and discarded from further analysis. Furthermore, batch PT6388 was also considered an outlier and not included for further analysis. As a result, we focussed on 65 unique, Y chromosome encoded peptides that were used as "male-specific" spike-in references for the analysis of false positives within the previously mentioned 21 TMT batches.

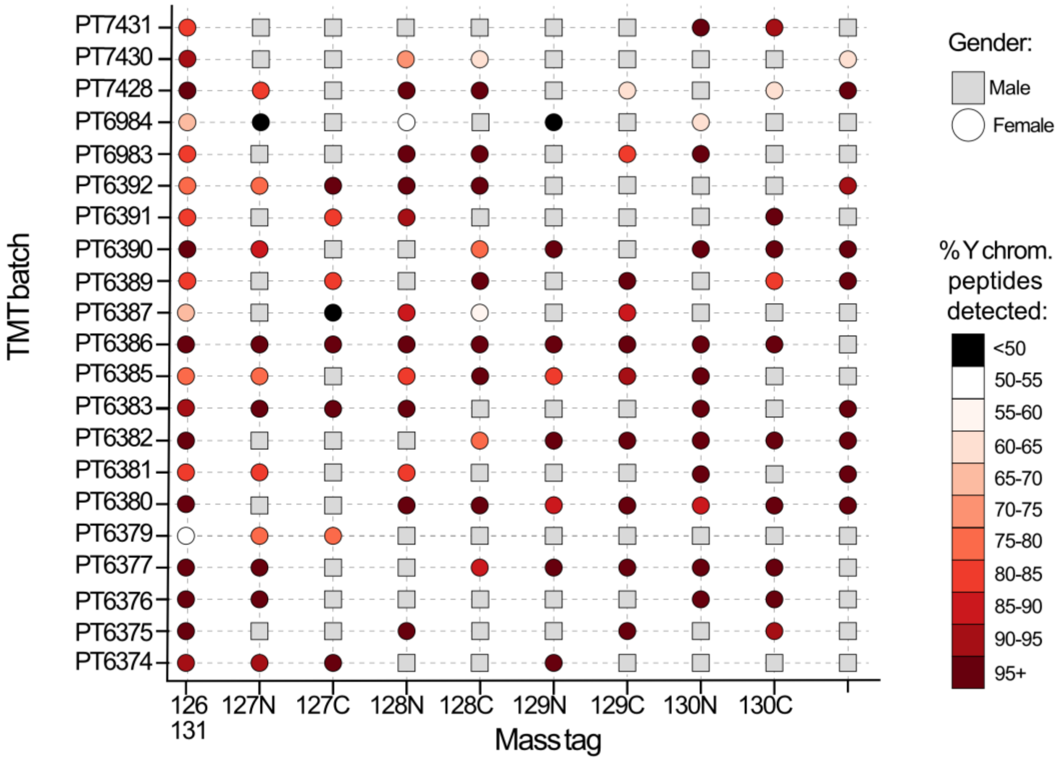

Fig. 2.4 - Y chromosome peptides in female channels: Schematic showing the incidence of false positives across 21 TMT batches and their reporter ion mass tags. Male cell lines are shown as a grey square, female cell lines are represented by a circle. The female lines (circles) are shaded to indicate the percentage of Y chromosome specific peptides that were detected in their channel within each TMT batch.

We evaluated false positives by exploring how frequently the respective female TMT channels were quantifying signal from Y chromosome-specific peptides (Fig. 2.4). Surprisingly, this showed that in all TMT channels containing a female cell line a minimum of 40% of the Y chromosome-specific peptides identified within the batch also had signal in the female channels. Remarkably, across all these batches, a median of 89% of Y chromosome-specific peptides quantified in each batch were also quantified in TMT channels that contained female cell lines. The Y chromosome peptides should not be present in any female line, hence the level of detection mentioned earlier is unexpected. But not entirely unpredictable as within a single TMT batch there are consistent reports of almost no missing values being present, which is particularly problematic if said TMT batch contains heterogenous populations. We infer that the appearance of signal for Y chromosome-specific peptides in the channels containing female cell lines likely results from a combination of co-isolation and reporter ion interference.

2.4.4 Reporter ion interference and co-isolation interference

Reporter ion interference, also known as cross-label isotopic impurity, can arise from manufacture level impurities and experimental error(Ow et al., 2009). Co-isolation interference is the effect caused by multiple labelled peptides being selected within the isolation window(Muntel et al., 2019b). To study both of these conditions we focused on the previously mentioned Y chromosome-specific peptides, as these should only be present in the male channels and absent in the female channels, therefore any signal detected in female channels should be artificial. We used the Y chromosome peptides to evaluate the difference in replicate normalised peptide intensities (for details on the normalisation see methods) between male and female lines, across all the twenty-one 10-plex TMT batches (Fig. 2.5a).

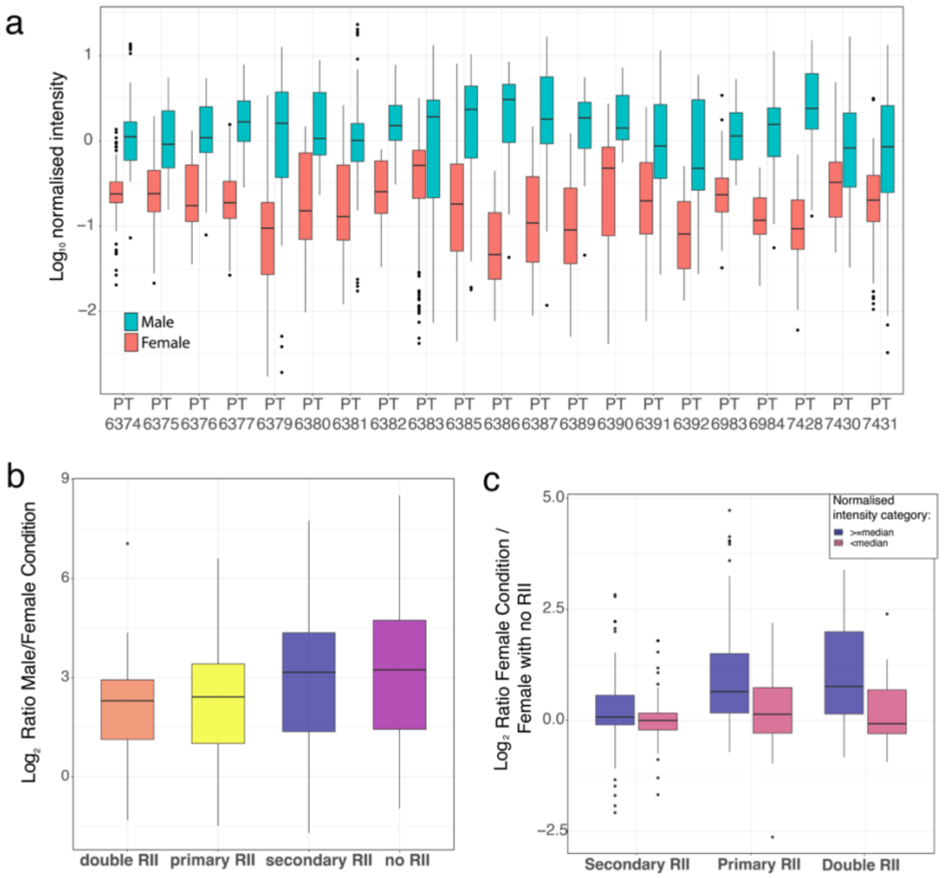

Fig. 2.5 - TMT channel leakage analysis: ***(a)*** Box plot showing the median normalized intensity of Y chromosome specific peptides for both female and male cell lines across 21 TMT batches. ***(b)*** Box plot of ratios for Y chromosome specific peptides, comparing male channels *versus* female channels affected by different reporter ion interference type. ***(c)*** Box plot of ratios for Y chromosome specific peptides, stratified by the median normalized intensity, comparing female channels affected by different types of reporter ion interference *versus* female channels not subjected to reporter ion interference. For all 3 boxplots the lower and upper hinges represent the 1st and 3rd quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge, the lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge.

The results of the analysis revealed significant variation between TMT batches. For example, some batches, such as PT6379 and PT6386, have 17 and 65-fold difference in intensity between male and female channels, simplifying the detection of false positives due to co-isolation interference. However, other batches, e.g. PT7430 and PT6391, only show a 2.5 and 4.4-fold difference, respectively making the detection of the false positives problematic. We note both of the previously mentioned batches display low intensity peptides and hence low signal-to-noise ratios, making female channels more vulnerable to co-isolation interference.

To further evaluate the interference, we selected female channels with no primary (+1 Da) or secondary (-1 Da) reporter ion interference (no cross-label isotopic impurity; see methods), as likely examples of co-isolation interference(Paulo et al., 2016). Peptides in male channels show a median of 9.43-fold higher normalised intensity (see methods) compared to female channels not affected by reporter ion interference. However, the effects vary depending on the peptide intensity thresholds. High intensity peptides, where the median normalised peptide intensity across male lines was greater than or equal to the global median (-0.184; see methods), displayed 12.8-fold higher intensities. Low intensity peptides, where the median normalised peptide intensity of male lines was lower than the global median, only displayed 2.14-fold increased intensity.

We also examined the potential effects of reporter ion interference (RII). For this analysis, we calculated a peptide specific ratio for each condition (see methods), firstly comparing the male channels vs the different types of reporter ion interference present in female channels. These conditions were, 'primary RII' when a male channel affects by isotopic impurity the +1Da female channel of the same isotope, for example a hypothetical contamination of a male line in channel 126C to female line in channel 127C. 'secondary RII' is when a male channel affects by isotopic impurity the -1Da female channel of the same isotope, for example a male line in channel 127C to female line in channel 126C. 'double RII' is when a female channel is affected by both 'primary RII' and 'secondary RII' from male channels. Channels not affected by either primary, or secondary RII were labelled as 'no RII'. The ratios comparing males to the previous conditions were used to generate the box plot (Fig 5b).

The male lines were a median of 9.4 fold higher than female channels not affected by reporter ion interference ("no RII"), but only 4.9 fold higher than female channels subjected to primary and secondary reporter ion interference ("double RII"). The smallest effect was caused by “secondary RII”, where the male lines were 8.94 fold higher than the female lines, suggesting that “primary RII” is the main source of isotopic impurities within a 10-plex (it is important to note that this conclusion is different for a 16 or 18-plex experiment) and explaining why we see little difference between the “primary RII” and “double RII” conditions. We note that for both “primary RII” and “double RII” the false positives are within the 8-fold increase/decrease range for *bona fide* changes in protein/peptide expression levels often detected within proteomic datasets(Hukelmann et al., 2016, Ly et al., 2015).

To quantify the differences between primary and secondary reporter ion interference across peptide abundance categories, we now compared female channels affected by the different types of reporter ion interference (“primary RII”, “secondary RII” and “double RII”) to female channels not affected by reporter ion interference (Fig. 5c; see methods). We also stratified this analysis by the peptide intensity, with high intensity values being either greater than or equal the global median, and low intensity being lower than the global median. Within the analysis we confirmed that the smallest effect was caused by “secondary RII”; for high intensity Y chromosome-specific peptides it displayed only a 5% increase compared to the channels not affected by reporter ion interference and virtually no change in low abundance peptides (Fig 5c). The “primary RII” displayed a more pronounced effect, with a median increase of 57% in the high intensity peptides and 10% in the low intensity ones. The combination of primary and secondary RII produced a median of 70% increase in the high intensity peptides and a small reduction in the median for the low intensity peptides.

These results suggest low intensity peptides are mostly affected by co-isolation interference, as reporter ion interference has little to no effect on their ratios, but that reporter ion interference can have profound effect in quantification of high intensity peptides. This provides important practical information that aids the design of TMT experiment to help minimise the potential effects on data quantification of cross condition/population reporter ion interference.

2.4.5 Optimising the experimental design

For all studies based on more than a single multi-plex TMT batch, we advocate at least one relevant internal reference sample should be included in each batch, as previously described(Herbrich et al., 2013, Plubell et al., 2017), and suggest they are assigned to either channel

126C, or 127N (134C and 135N are also good alternative if using an 18-plex experiment). For a 10 or 11-plex experiment these channels avoid “primary RII”, the main cause of isotopic impurities, and are only affected by “secondary RII”, which only causes a median increase of 2.2% in intensity. In contrast, placing the reference line at channel 131N, or 131C within a 10-plex or 11-plex increases the impact of reporter ion interference by exposing them to “primary RII”.

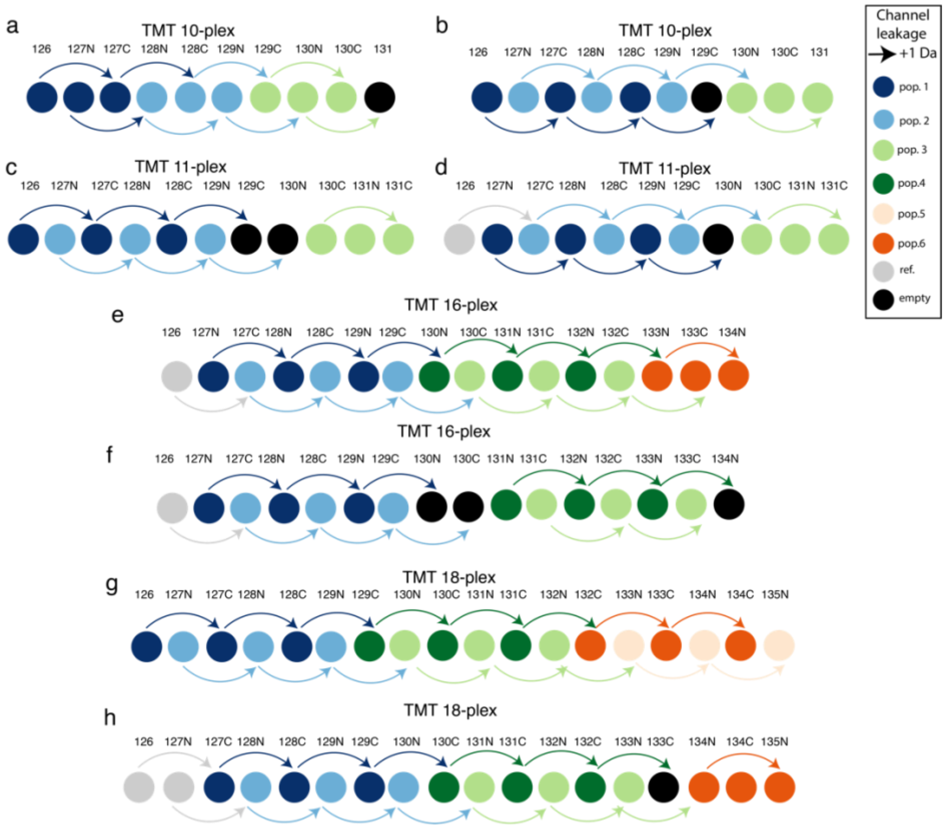
a
TMT 10-plex
126 127N 127C 128N 128C 129N 129C 130N 130C 131
b
TMT 10-plex
126 127N 127C 128N 128C 129N 129C 130N 130C 131
c
TMT 11-plex
126 127N 127C 128N 128C 129N 129C 130N 130C 131N 131C
d
TMT 11-plex
126 127N 127C 128N 128C 129N 129C 130N 130C 131N 131C
e
TMT 16-plex
126 127N 127C 128N 128C 129N 129C 130N 130C 131N 131C 132N 132C 133N 133C 134N
f
TMT 16-plex
126 127N 127C 128N 128C 129N 129C 130N 130C 131N 131C 132N 132C 133N 133C 134N
g
TMT 18-plex
126 127N 127C 128N 128C 129N 129C 130N 130C 131N 131C 132N 132C 133N 133C 134N 134C 135N
h
TMT 18-plex
126 127N 127C 128N 128C 129N 129C 130N 130C 131N 131C 132N 132C 133N 133C 134N 134C 135N
Channel leakage
+1 Da
pop. 1
pop. 2
pop. 3
pop.4
pop.5
pop.6
ref.
empty

Fig. 2.6 - TMT experimental design from reporter ion interference analysis: Schematics representing potential sample layouts showing: ***(a)*** 3-3-3 grouped layout for a 10-plex TMT batch with 2 populations and 3 replicates each. 4 channels are being affected by cross population primary reporter ion interference. ***(b)*** Alternative layout for a 10-plex TMT batch with 2 populations and 3 replicates each. 1 channel is being affected by cross population primary reporter ion interference ***(c)*** optimal 11-plex configuration for 3 populations with 3 replicates each. By leaving two empty channels, it eliminates cross population reporter ion interference. ***(d)*** optimal 11-plex configuration for 3 populations with 3 replicates each, with one empty channel and one reference line channel. Only 1 channel suffers primary reporter ion interference. ***(e)*** Layout for a 16-plex TMT with 5 populations and 3 replicates each. 4 channels are being affected by cross population primary reporter ion interference. ***(f)*** Layout for a single batch design 18-plex TMT with 6 populations and 3 replicates each. 4 channels are being affected by cross population primary reporter ion interference. ***(g)*** Layout for a multi batch design 18-plex TMT with 6 populations and 3 replicates each. 5 channels are being affected by cross population primary reporter ion interference.

Our results also show TMT experimental designs that can help to minimise the effects of primary and secondary reporter ion interference between the different populations/conditions. For example, in a 10-plex TMT study, when three conditions are being analysed, each with three biological replicates, a 3-3-3 grouped layout would cause multiple channels to be affected by cross population/condition reporter ion interference (Fig. 2.6a). The optimal design would involve alternating the first two conditions while leaving an empty channel before the last (Fig. 2.6b).

However, there is little incentive to use a 10-plex, as all 10-plex TMT setups involve cross population/condition reporter ion interference. An 11-plex TMT set up enables a design without reporter ion interference between the three conditions/populations but requires two empty channels at 129C and 130N to achieve this (Fig. 2.6c). If a control/reference channel is included, as advocated, then it should be placed in channel 126C, while allocating the empty channel to position TMT^{11} 130N, between the alternating experimental conditions and the final replicates of the third condition (Fig. 2.6d). All the suggested setups aim to reduce cross condition/population RII (+1 and -1Da), thereby avoiding decreases in quantification accuracy by isotopic impurities.

The newer 16-plex(Li et al., 2020) and 18-plex(Li et al., 2021) considerably expand the multiplexing capacities, but also complicates the channel allocation as well as the primary/secondary RII balance. There are currently virtually no public large-scale datasets using the 16 and 18-plex, but according to their product sheet, they display a more prominent effect of secondary RII (-1 Da), as channel 135N is reported to produce twice as much secondary RII compared to 131N. The product sheet also reports limited primary RII leakage into 133N-135N, this can be exploited when designing templates for the 16 and 18-plex studies.

A 16-plex would have a spare channel when a five condition three replicate setup is used, therefore the single and multi-batch allocations would be identical. We recommend pairing the populations and alternating the channels between the first two pairs (Fig. 2.6e) as this will reduce the number of channels affected by cross condition/population reporter ion interference. The final population would have replicates allocated sequentially but would be located in the channels least affected by primary RII (Fig. 2.6e). If the study uses less than five conditions, then the empty channels can be allocated to prevent all cross condition/population primary and secondary RII (Fig. 2.6f). The 18-plex allows for a complete allocation of channels, allowing a six-population analysis when the experiment consists of a single TMT 18-plex batch (Fig. 2.6g), again we recommend pairing two populations and alternating them. When organising a multi-batch 18-plex study, we recommend using two control channels and an empty channel at 133C, this would insure only 1 replicate of each condition was affected by cross condition/population reporter ion interference (Fig. 2.6h).

2.5 Discussion

Quantitative proteomic analysis using TMT labelling has become one of the most popular DDA methods currently used, thanks to its multiplexing capabilities, scalability, low number of missing values and precision when a single multiplexed batch is analysed. However, when large studies that require the use of multiple parallel TMT batches are performed(Plubell et al., 2017, Maes et al., 2016, Nusinow et al., 2020), the situation becomes more complicated. Here, we have used the analysis of data integrated from over 20, 10-plex TMT batches to investigate accuracy, missing values, false positives, co-isolation interference, reporter ion interference and experimental design within very large-scale proteomics experiments. We have focussed on a model data set derived from the analysis of human iPS cell lines, derived from both male and female donors(Kilpinen et al., 2017).

The resulting data confirms that a single batch of multi-plex TMT experiments minimises the typical missing values issue associated in proteomics with Data Dependent Acquisition (DDA), both at the protein and peptide levels. However, this situation changes as data from two or more separate multi-plex TMT batches are integrated. As multiple batches are combined the missing values index inflates rapidly. This effect is particularly striking at the peptide level, where integrating data from only two different batches causes the missing values index to increase from <2% to ~24%. Even though the inflation rate at the protein level is lower, the integration of the second batch pushes missing protein values from <0.5% to >6%. This inflationary effect can decrease the accuracy of results derived from large-scale experiments that compare data generated from multiple TMT batches and is similar to what is seen in label free workflows. One potential solution would be to utilise MS2-based TMT quantitation, as this has been reported to produce more total peptide identifications(Myers et al., 2018), however there is no guarantee this will detect peptides/proteins more reproducibly across batches. Furthermore MS2-based TMT quantitation will intensify the disruptive effect of the co-isolation interference, which has been recently shown to be particularly problematic for low abundance reporter ions while using older versions of tune (Smith et al., 2022).

While single TMT batches can provide remarkably precise results within the multiplexed experiment, we found that this is often not reproducible across multiple batches. To study reproducibility, we normalised the data using the proteomic ruler(Wisniewski et al., 2014) and for every protein we calculated the coefficient of variation (CV), both for the 23 technical replicates of the reference line, and within each of 10-plex TMT batches. The median copy number CV of the technical replicates was ~6.4-fold higher than data from different donors analysed within the same 10-plex TMT batch. This also underlines the importance of normalising batch effects, in our case via a common control sample within each TMT batch which allows for objective data normalisation to minimise the batch effects, as has been reported(Liu et al., 2017, Ping et al., 2018). We showed that by introducing at least one suitable internal reference standard (IRS), though preferably multiple IRS channels would be preferred, within each TMT experiment the batch effects can be normalised. The challenge lies in identifying a suitable control that is truly representative for most proteins being compared within the experiment and in creating a control that is highly reproducible across all the TMT batches.

This study has also highlighted the issue of false positives, reporter ion interference and co-isolation interference. The dataset we selected provided an ideal set up to analyse these factors, as it contained hiPSC lines derived from both male and female donors. Thus, by identifying a set of peptides uniquely mapped to the male-specific Y chromosome, these provided a convenient set of internal controls to monitor the expression of false positives, defined here as correctly identified peptides which are not meant to be expressed in specific conditions or populations. The data showed that even for a 10-plex TMT batch with only two male channels (PT6380), the remaining 8 female channels still quantified 97.5% of all the Y chromosome-specific peptides that were detected in that batch. This means there are false positives being consistently detected in the female channels within the multiplexed experiments, which suggests there are severe limitations when analysing heterogenous populations within the same TMT batch. This issue in particular of additional relevance to single cell proteomics, where it is important to understand the subset of proteins expressed by the cells, and where the false positives caused by TMT can cause important biological misinterpretations. This issue we attribute mostly to co-isolation interference and we note that the issue has been reduced, though not completely eliminated, with newer generation Orbitrap MS instruments, where the improved source has enhanced the signal to noise ratio(Paulo et al., 2016). Furthermore newer isobaric tagging methods have been developed which claim to be co-isolation free(Virreira Winter et al., 2018).

Furthermore, we used the previously mentioned Y chromosome peptides to study the different reporter ion interference (RII) conditions across 21 different 10-plex TMT batches with different number of male and female derived cell lines, as well as different channel combinations. The data highlighted the effects of primary (male channel isotopic contamination into a +1Da female channel) and secondary (male channel isotopic contamination into a -1Da female channel). Reporter channels affected by both primary and secondary reporter ion interference showed a median signal increase in high intensity peptides of 1.7-fold compared to channels not subjected to reporter ion interference. This was found to be primarily caused by "primary RII" as the data also showed that "secondary RII" had the smallest effect with only a median 1.02-fold increase compared to channels with no reporter ion interference. It is important to note the work was carried out before the release of the 18-plex, where the effect of "secondary RII" is reported to be much more prevalent.

To best avoid the effects of reporter ion interference, we have used these data to propose optimised experimental set ups for assigning samples to specific channels that can either minimize, or eliminate (when possible), the effect of primary and secondary reporter ion interference between conditions/populations. Nonetheless, we highlight again that mixing significantly different populations within a TMT batch, for example hiPSCs and terminally differentiated somatic cells, will introduce false positives within the data, as illustrated here by the Y chromosome-specific peptides detected within all female cell lines. For such large-scale experiments it is also vital to have strict quality control (QC) procedures in place to evaluate and maintain a constant performance within the instrument/s. In our case one of the TMT batches (PT6388) revealed poor performance in the QC run. The failed QC run was not detected until after the samples were run, producing poor results within that batch. We therefore recommend that to execute large-scale TMT a rigorous QC procedure should be set in place before the start of the experiment.

In conclusion, as proteomics continues to move into large-scale work, we find that TMT is a valuable methodology for DDA analysis and its potential to increase scalability and produce precise quantitation have made it a justifiably popular approach for high-throughput proteomic studies. Here, we have provided an in-depth, practical evaluation of parameters affecting the generation of high-quality quantitative data from very large-scale TMT-based proteomics analyses, and we highlight some of the limitations which should be carefully considered when planning these experiments. We hope the resulting information will prove useful for improving experimental design and resulting data quality for many future proteomics projects.

Chapter 3 – TMT-based large-scale characterisation of human induced pluripotent stem cells

3.1 Publications relating to this chapter

- Erosion of human X chromosome inactivation causes major remodelling of the iPSC proteome (Brenes et al., 2021b)
- This manuscript has had minor updates and adaptions to fit within this thesis
- The supplemental tables are available as supplemental data in the publication

3.1.1 Contributions to the publications on this chapter

- I proposed and conceptualized the study
- I processed and analysed all the proteomic dataset
- I integrated all proteomic and transcriptomic data
- I conceptualized the allele specific analysis
- I generated all figures used in this study
- I wrote the paper with input from the remaining authors

3.1.2 Main contributions of the other authors to the publications on this chapter

- The hiPSC cells lines were grown and reprogrammed in the Sanger Centre as part of the HipSci Consortium(Kilpinen et al., 2017)
- The transcriptomic data was processed in the Stegle Laboratory by Daniel Seaton and Bogdan Mirauta
- The allele specific analysis was performed in the Stegle Laboratory by Daniel Seaton
- The mass spectrometry raw files were generated in the Lamond Laboratory by Dalila Bensaddek
- The Ribo Mega-SEC experiments were performed in the Lamond Laboratory by Harunori Yoshikawa

3.2 Introduction

In humans and other mammalian species, female cells have two copies of the X chromosome, while males have a single X chromosome and a much smaller Y chromosome that is not present in females. In females, one of the two X chromosomes undergoes silencing, causing repression of transcription and thereby inactivating expression of alleles located on this second copy of the X. This process is termed X chromosome inactivation (XCI). The XCI process in female cells is considered to be a critical dosage compensation mechanism that evolved in mammals as a way to equalize X-linked gene expression between males and females (Graves, 2016, Livernois et al., 2012). XCI is vital for embryonic development and failure to induce XCI has been shown to cause embryonic lethality(Borensztein et al., 2017, Takagi and Abe, 1990). Furthermore, skewed XCI has also been shown to have major clinical consequences, with the emergence of numerous sex-specific genetic disorders, such as Rett's Syndrome(Lyst and Bird, 2015).

The initiation of XCI is controlled by a specific locus, termed the X-inactivation centre (Xic)(Augui et al., 2011). The mechanism of XCI involves a profound structural reorganization of the inactivated copy of the X chromosome, which becomes heterochromatic and visibly compacted (Augui et al., 2011, Giorgetti et al., 2016). Within the Xic, a long, non-coding RNA, called 'XIST', has been shown to be an important component of the XCI process(Marahrens et al., 1997, Penny et al., 1996). Accumulation of XIST RNA across the inactive copy of the X chromosome triggers the changes that produce the transcriptionally inactive state (Dossin et al., 2020, Galupa and Heard, 2018).

Over a decade ago, breakthrough studies reported that terminally differentiated somatic cells could be reprogrammed back into a pluripotent state via the exogenous expression of a small set of transcription factors(Takahashi and Yamanaka, 2006, Takahashi et al., 2007, Yu et al., 2007). The resulting human induced pluripotent stem cells (hiPSCs) were shown to share the hallmarks of their embryonic counterparts, including the induction of XCI(Wutz, 2012). However, for these cells, as well as for human embryonic stem cells (hESC), XCI has been shown to be unstable in culture. Thus, some human primed iPSCs exhibit erosion of XCI, where the X chromosome loses H3K27me3 marks, as well as losing expression of XIST RNA expression (Anguera et al., 2012, Dandulakis et al., 2016, Mekhoubad et al., 2012). Although the role of XIST in relation to escape from XCI remains unclear, it has been reported that loss of XIST expression is characteristic of class III hESCs that display eroded XCI (Geens and Chuva De Sousa Lopes, 2017).

We first explore the global consequences for human gene expression when XCI is eroded, using a collection of iPSCs derived from healthy female donors that were all reprogrammed from primary skin fibroblasts (Kilpinen et al., 2017). Specifically, we have analysed the impact of XCI erosion using 74 independent HipSci (www.hipsci.org) hiPSC lines derived from female donors and also compared these to 46 lines derived from male donors, using both RNAseq and proteomic data(Mirauta et al., 2020). The data show that for our collection of iPSCs a decrease in the expression of the lncRNA XIST was correlated with significantly higher biallelic expression, reflecting increased erosion of XCI.

We also report a global analysis comparing in parallel RNA and protein expression levels for lines that were stratified based upon having either High, or Low, expression levels of XIST RNA. This provides the first in depth analysis of how erosion of XCI in human cells affects gene expression at the protein level. The data show that erosion of XCI increases both transcription and protein production from genes on the inactive X chromosome, while comparisons to the male lines show this erosion significantly affects dosage compensation at the protein level. Remarkably, we also uncover a widespread increase in the abundance of many proteins encoded by genes on the autosomes, independent of a parallel increase in transcription. Female cells with low XIST RNA show a median ~13% increase in total protein levels, along with higher levels of polysomes and components of the translational machinery. These data indicate that erosion of XCI can affect the expression of a much wider range of proteins and disease-linked gene loci than previously realised based on RNA analysis alone.

All of the raw and processed mass spectrometry (MS) files are available within PRIDE (Vizcaino et al., 2016, Perez-Riverol et al., 2018) [PXD010557], while the RNAseq data are available within ENA (Amid et al., 2020) [PRJEB7388].

3.3 Methods

3.3.1 Experimental Model and Subject Details:

All lines included in this study are part of the HipSci resource and were reprogrammed from primary fibroblasts as previously described (Kilpinen et al., 2017). Out of the total of more than 800 hiPSC lines available within the HipSci resource (hipsci.org), 120 derived from healthy donors and with proteomic analysis were used in this study. All lines derived from healthy female donors (subset of 74 hiPSC lines) were then used for the XCI analysis and included all lines derived from healthy male donors (subset of 46 hiPSC lines) for the dosage compensation analysis.

3.3.1 TMT Sample preparation & LC–MS

The sample preparation and LC-MS are as described in the previous chapter.

3.3.2 Ribo Mega-SEC hiPSC lines cell culture

For the Ribo Mega-SEC analyses 4 hiPSC lines with High XIST RNA levels ('iiyk_2', 'iiyk_4', 'nufh_3' and nufh_4') and 3 lines with Low XIST RNA levels ('fawm_4', 'bawa_1' and 'aizi_3') were used. The lines were maintained in TESR medium (Ludwig et al., 2006), supplemented with FGF2 (Peprotech, 30 ng/ml) and noggin (Peprotech, 10 ng/ml), on growth factor reduced geltrex basement membrane extract (Life Technologies, 10 μg/cm^2) coated dishes at 37^{o}C in a humidified atmosphere of 5% CO_2 in air.

Cells were routinely passaged twice a week as single cells, using TrypLE select (Life Technologies) and replated in TESR medium that was further supplemented with the Rho kinase inhibitor Y27632 (Tocris, 10 μM), to enhance single cell survival. Twenty-four hours after replating, Y27632 was removed from the culture medium. For proteomic analyses, cells were plated in 100 mm geltrex coated dishes at a density of 5x10^4 cells cm^{-2} and allowed to grow to for 3 days, until confluent, with daily medium changes.

3.3.3 Ribo Mega-SEC

Ribo Mega-SEC for the separation of polysomes and ribosomal subunits using size exclusion chromatography was performed as previously reported(Yoshikawa et al., 2018), with a slight modification. Briefly, 2.5 x 10^6 cells were washed once with ice-cold PBS, scraped in ice-cold PBS and collected by centrifugation at 500 g for 5 min (all centrifugations at 4°C). The cells were lysed by vortexing for 10 sec in 250 μL of polysome extraction buffer (20 mM Hepes-NaOH (pH 7.4), 130 mM NaCl, 10 mM $MgCl_2$, 5% glycerol, 1% CHAPS, 0.2 mg/ml heparin, 2.5 mM DTT, 20 U SUPERase In RNase inhibitor, cOmplete EDTA-free Protease inhibitor), incubated for 15 min on ice, and centrifuged at 17,000 g for 10 min. Supernatants were filtered through 0.45 μm Ultrafree-MC HV centrifugal filter units (Millipore).

Using a Dionex Ultimate 3,000 Bio-RS uHPLC system (Thermo Fisher Scientific), a SEC column (Agilent Bio SEC-5, 2,000 Å pore size, 7.8 × 300 mm with 5 μm particles) was equilibrated with three column volumes of filtered SEC buffer (20 mM Hepes-NaOH (pH 7.4), 60 mM NaCl, 10 mM $MgCl_2$, 0.3% CHAPS, 0.2 mg/ml heparin, 2.5 mM DTT, 5% glycerol) (all column conditioning and separation at 5°C) and 100 µl of 10 mg/ml of filtered bovine serum albumin (BSA) solution diluted by PBS was injected once to block the sites for non-specific interactions. After monitoring the column condition by injecting standards, including 10 µL of 10 mg/mL BSA solution and 5 µL of HyperLadder 1 kb (BIOLINE), 200 µL of the filtered cell lysates was injected onto the pre-equilibrated SEC column. The flow rate was 0.4 mL/min and the chromatogram was monitored by measuring UV absorbance at 215, 260 and 280 nm with a 1 Hz data collection rate by the Diode Array Detector.

3.3.4 RNA-seq data processing

Raw RNA-seq data were obtained from the ENA project: PRJEB7388. CRAM files were merged on a sample level and converted to a single FASTQ file per sample. Sequencing reads were trimmed to remove adapters and low-quality bases (Trim Galore!), followed by read alignment using STAR (v.020201) (Dobin et al., 2013), using the two-pass alignment mode and the default parameters as proposed by ENCODE (c.f. STAR manual). All alignments were relative to the GRCh37 reference genome, using ENSEMBL 75 as transcript annotation (Zerbino et al., 2018).

Samples with low quality RNA-seq were discarded if they had either less than 2 billion bases aligned, had less than 30% coding bases, or had a duplication rate higher than 75%. Gene-level RNA expression was quantified from the STAR alignments using featureCounts (Liao et al., 2014) (v1.6.0), which was applied to the primary alignments using the "-B" and "-C" options in stranded mode, using the ENSEMBL 75 GTF file. Quantifications per sample were merged into an expression table using the following normalization steps. First, gene counts were normalized by gene length. Second, the counts for each sample were normalized by sequencing depth using the edgeR (Robinson et al., 2010) adjustment. Transcript isoform expression was quantified directly from the (unaligned) trimmed reads using Salmon (Patro et al., 2017) (v0.8.2), using the '--seqBias', '--gcBias' and 'VBOpt' options in 'ISR' mode to match our inward stranded sequencing reads. The transcript database was built on transcripts derived from ENSEMBL 75. The TPM values as returned by Salmon were combined into an expression table.

3.3.5 Allele specific analysis

Allele-specific quantification of expression from the X chromosome was calculates using RNA-Seq reads mapping to the X chromosome. Allele-specific counts were obtained from SNPs present in DBSNP using GATK ReadCounter with the command 'GenomeAnalysisTk.jar -T ASEReadCounter -U ALLOW_N_CIGAR_READS –minMappingQuality 10 –minBaseQuality 2', restricted to SNPs which were known to be heterozygous in the analysed sample. The allele-specific fraction of expression was defined as the fraction of transcript reads mapping to the less expressed allele, restricting to heterozygous X chromosome SNPs with at least 20 overlapping reads. These fractions were then averaged across SNPs at a whole-chromosome level (Fig. 3.2a,c), and for individual genes (Fig. 3.6a,d). Note that, for a given gene in a given sample, this quantification could only be performed when the donor for that sample carries a heterozygous common variant in that gene. This reduced the number of samples for which allele-specific expression could be computed for each gene.

3.3.6 Primary vs secondary allele

The primary allele for each hiPSC line is defined as the allele with the highest number of transcript reads mapping to it. Conversely the secondary allele is defined as the allele with the lowest number of transcript reads mapped to it.

3.3.7 Proteomics identification and quantification

The identification and quantification strategy used is the same as described in the previous chapter.

3.3.8 Razor + unique peptides

Peptides that are unique to a single protein sequence are known as "unique peptides" and peptides that are shared between multiple protein sequences are known as "shared peptides". Within MaxQuant, shared peptides are assigned to a single protein group, following Occam's Razor. The number of unique peptides, plus the number of shared peptides used for the quantification of a protein group, is referred to as Razor + unique peptides.

3.3.9 hiPSC Copy number generation

Protein copy numbers were calculated using the proteomic ruler (Wisniewski et al., 2014) and using the MS3 reporter intensity. An additional batch correction factor for each TMT experiment was applied as previously described (Brenes et al., 2019).

3.3.10 Protein content

The protein content for cell lines in both the High and Low XIST RNA populations was calculated based on the copy numbers. The molecular weight of each protein (converted to picograms) was multiplied by the number of copies for the corresponding protein and this was then summed for all proteins across all lines to calculate the protein content.

3.3.11 Chromosome mapping

To map gene products to their specific chromosomes, we utilised the UniProt (The UniProt, 2017) protein-chromosome mapping service. We used their output to produce a list of unique proteins for each specific chromosome. Subsequently, we mapped the proteins detected in our hiPSC dataset to their corresponding chromosomes, based on the UniProt mapping file.

3.3.12 X-inactivation stratification and analysis

Based on the RNAseq data, 74 hiPSC lines were classified into 3 distinct categories, based on XIST expression. 30 hiPSC lines where XIST expression was <1 Log_2 TPM were classified as 'Low XIST'. 35 hiPSC lines where XIST expression was higher than 2.75 Log_2 TPM were classified as High XIST and the remaining 9 lines were classified as 'Medium' XIST.

3.3.13 High XIST filtering

The High XIST population contained two proteomic experiments, PT7422 and PT6386, contributing a large number of High XIST replicates within their 10-plex TMT experiment. As the maximum number of replicates per 10-plex within the Low XIST group was 4, we performed hierarchical clustering to reduce the number of lines contributed by PT7422 and PT6386 to a maximum of 4, in order to minimise batch effects. The final number of lines with High XIST, post filtering, was 26.

3.3.14 GO Enrichment analysis

All of the GO enrichment analyses were done using Panther (Mi et al., 2017) and used the 8,511 proteins that were detected in both the RNAseq and proteomics datasets as background. We performed a biological process overrepresentation analysis for all proteins that were significantly increased (q-value <0.05), where the corresponding transcript was not significantly increased in expression. Furthermore, an additional biological process overrepresentation of significantly increased (q-value<0.05) ribosome biogenesis proteins was carried out.

3.3.15 Hypergeometric analysis

The hypergeometric analyses were all done in R using the phyper function from the stats package (v.3.6.0). For this analysis, a subset of proteins (1,825), which were significantly increased in expression without a corresponding mRNA increase, were selected. We first looked at the number of proteins within the previous subset with a peptide count either greater than, or equal to, the 75th percentile (26 peptides) and used phypher to determine hypergeometric p-values. We then compared this result to the one produced by randomly selecting 1,825 proteins and repeating the previous process. This was done 100,000 times. We also looked at the number of proteins with copy numbers either greater than, or equal to, the 75th percentile (407,724 copies) and used phypher to determine hypergeometric p-values. We then compared this result to the one produced by randomly selecting 1,825 proteins and repeating the previous process. This was done 100,000 times.

3.3.16 60S/40S ratio

The ratios were calculated by summing the copy numbers from all proteins of the 60S ribosomal subunit, divided by the sum of all copy numbers from the 40S ribosomal subunit, for each individual hiPSC line.

3.3.17 UniProt to Ensembl mapping

Mapping of UniProt accessions to Ensembl gene identifiers was done in R using the “UniProt.ws” package version 2.24.1

3.3.18 Kinase map

The kinase map was generated within the Encyclopedia of Proteome Dynamics (Brenes et al., 2018a) using the KinoViewer (Brenes and Lamond, 2019).

3.3.19 Sequence coverage maps

The sequence coverage maps for EIF1AX, EIF2S3 and RPS6KA3 were generated using Jalview (Waterhouse et al., 2009) version 2.11.1.3

3.3.20 Quantification and statistical analysis

The proteomics data used for the analysis were obtained from the ProteinGroups.txt output of MaxQuant (Cox and Mann, 2008, Tyanova et al., 2016a) v. 1.6.3.3. Contaminants, reverse hits and 'only identified by site' proteins were excluded from analysis. Overall, we quantified 9,631 protein groups in at least one of the samples. For additional stringency and to reduce batch variation, only proteins with 3 or more 'Razor + unique peptides' were considered.

Fold changes and p-values were calculated in R utilising the bioconductor package LIMMA(Ritchie et al., 2015) version 3.7. FDR calculations were performed in R with the "qvalue" package version 2.10.0. For both the RNAseq and proteomics differential expression analyses, gene products with a q-value<=0.05 were considered significant. For comparisons looking at individual gene products or aggregated gene families, Welch's t-test was used. In this case results with a p-value<=0.05 were considered significant.

3.3.21 Data and Code availability:

The mass-spectrometry dataset, PXD010557, supporting the current study is available in PRIDE (https://www.ebi.ac.uk/pride/archive/projects/PXD010557). The RNAseq dataset, PRJEB7388, supporting the current study is available in the ENA project (https://www.ebi.ac.uk/ena/browser/view/PRJEB7388).

3.4 Results

3.4.1 Overview of the coverage of the HipSci RNAseq and proteomic datasets

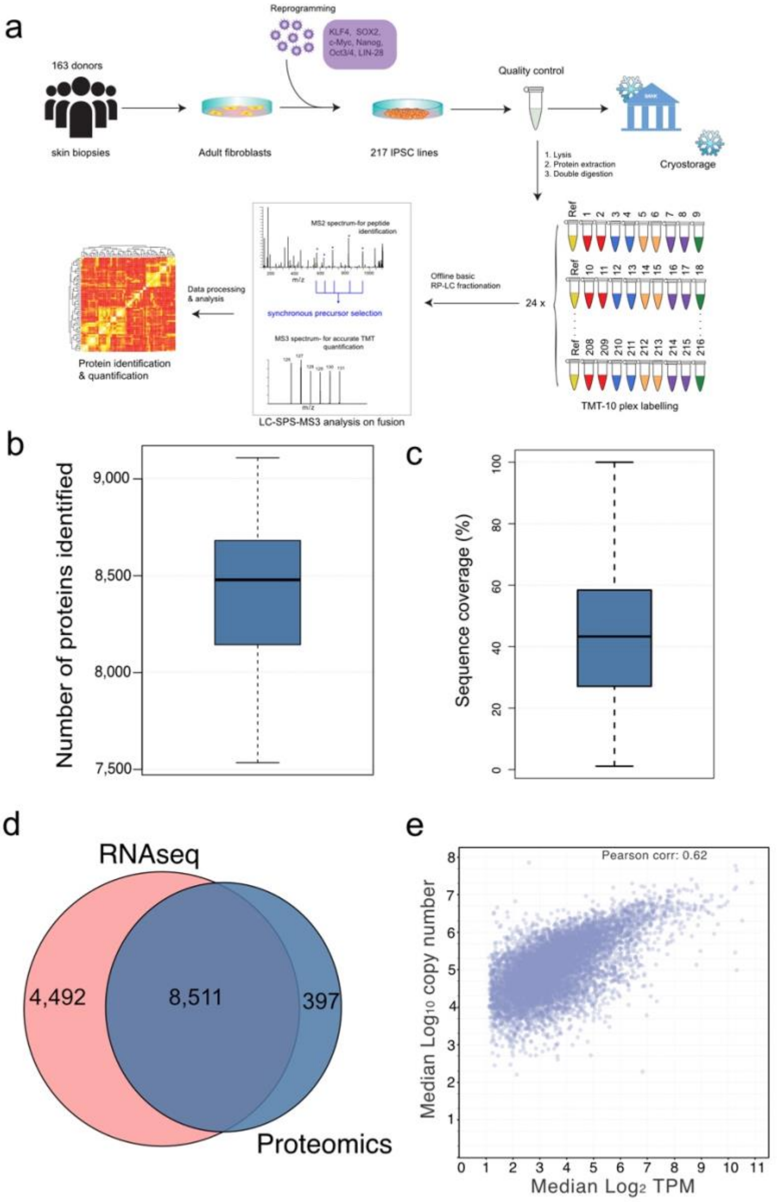

Figure 3.1- Comprehensive coverage: ***(a)*** The HipSci proteomics workflow starting with the reprogramming of fibroblasts obtained from skin biopsies(Kilpinen et al., 2017) and leading to the TMT-based proteomic workflow. In total 24 10-plex TMT batches were used, each with an internal reference standard present in channel 126. ***(b)*** Box plot showing the number of proteins identified per line across the 56 filtered (see methods) female hiPSC lines. ***(c)*** Box plot showing the sequence coverage for all proteins detected within the dataset. ***(d)*** Pie chart showing the overlap between quantified gene products in the proteomics and RNAseq datasets. ***(e)*** Scatter plot showing the median $\log_2$ transcripts per million (TPM) vs the median $\log_{10}$ copy number for all gene products. For all Box plots the lower and upper hinges represent the 1st and 3rd quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge, the lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge

All of the hiPSC lines used for this study were generated by the HipSci project (www.hipsci.org). Skin biopsies were taken from donors and the primary skin fibroblasts were isolated, these were then reprogrammed into primed hiPSCs and were subjected to rigorous quality control procedures, which included array-based genotyping and gene expression profiling, as well as an evaluation of their pluripotency and differentiation properties (Kilpinen et al., 2017). This study analyses gene expression data (Mirauta et al., 2020, Brenes et al., 2019) generated from 74 independent hiPSC lines derived from healthy female donors and 46 lines derived from healthy male donors. The lines were grown using identical culture conditions and aliquots were divided for parallel RNAseq and proteomic analyses. The MS-based proteomic data were acquired using a tandem mass tag (TMT) workflow (see Methods; Fig. 3.1a).

The proteomic data for this study were processed using MaxQuant (Cox and Mann, 2008, Tyanova et al., 2016a) and searched against the manually curated SwissProt database (The UniProt, 2017) with a 1% FDR threshold at the peptide spectrum match (PSM) and protein level (for more details see methods). Overall, it detected the expression of >9,500 protein groups (i.e., proteins/protein isoforms without discriminating peptides; hereon termed ‘proteins’), with a median of 8,479 proteins identified across all lines (Fig. 3.1b) and median protein sequence coverage of 42% across all proteins (Fig. 3.1c). All downstream analyses were performed on a subset of 8,908 proteins, which were each identified with at least 3 ‘Razor + unique peptides’ (see methods). To compare protein expression levels between the respective hiPSC lines, protein copy numbers were estimated using the ‘proteomic ruler’ (Wisniewski et al., 2014) approach, and using the batch correction method previously described (Brenes et al., 2019, Plubell et al., 2017). This is well suited for the analysis of HipSci lines, which have been shown to have near identical DNA content (Kilpinen et al., 2017). From the RNAseq data, after filtering, a total of 12,798 transcripts were quantified, (see methods), with matching protein level data for 65% of these (Fig. 3.1d). To explore the relationship between RNA and protein abundance levels in this set of hiPSC lines, we calculated the Pearson correlation of transcript abundance vs protein abundance (Fig. 3.1e). This resulted in a Pearson correlation coefficient of 0.62, which is similar to what has been reported by multiple previous studies comparing mRNA and protein expression levels, both in different human cell types and for other mammalian species (Edfors et al., 2016, Lundberg et al., 2010, Ly et al., 2014).

3.4.2 Erosion of X-chromosome inactivation

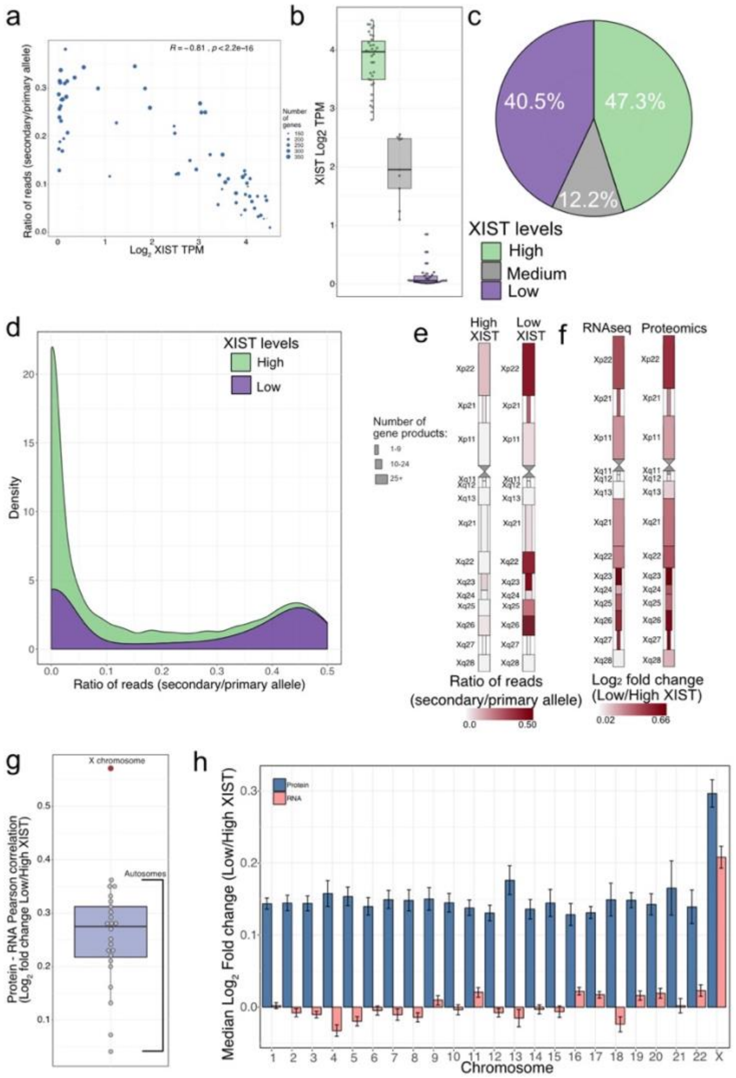

Figure 3.2 – XIST and XCI: ***(a)*** Scatter plot showing the ratio of reads derived from the secondary allele (lowest expressed allele) compared to the primary allele (highest expressed allele) for all X-linked transcripts vs the log$_2$ XIST TPM for all healthy female lines. The size of the circle is determined by the number of transcripts used for the analysis ***(b)*** Box plots showing the Log$_2$ Transcripts Per Million (TPM) for the long non-coding RNA XIST across all 3 populations; Low, Medium and High XIST. ***(c)*** Pie chart showing the percentage of healthy female lines within each XIST stratified population. ***(d)*** Stacked density plot for all X-linked gene products across all lines showing the ratio of reads mapped to the secondary allele compared to the primary allele for both the High and Low XIST populations. ***(e)*** X chromosome map showing the ratio of reads derived from the secondary allele compared to the primary allele across chromosomal bands for both the High and Low XIST populations. The size of the rectangles represents the number of gene products per band. ***(f)*** X chromosome map showing the log$_2$ fold change (Low/High XIST) across chromosomal bands for both the RNAseq and proteomic datasets. The size of the rectangles represents the number of gene products per band. ***(g)*** Box plot showing the Pearson correlation coefficient comparing log$_2$ fold change (Low/High XIST) at the RNAseq and proteomics level for all chromosomes. Autosomes are coloured in grey; the X chromosome is coloured in red. ***(h)*** Bar plot showing the median Log$_2$ fold change (Low/High XIST) for all gene products aggregated at the chromosome level for both RNAseq and proteomics. The error bars represent the SEM. For all Box plots the lower and upper hinges represent the 1st and 3rd quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge, the lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge

As it had been previously reported that there was a correlation between the loss of XIST RNA coating the inactive X chromosome and erosion of XCI, we first evaluated the relationship between the levels of XIST RNA expression and the erosion of XCI within the HipSci hiPSC lines, as measured by an allele specific expression (ASE) analysis on X-linked genes (Fig. 3.2a). This showed a clear correlation between hiPSC lines with low XIST expression and increased levels of biallelic expression for X-linked genes. Interestingly, a parallel analysis focussed on XACT RNA, another long non-coding RNA implicated in the mechanism of XCI in human (Vallot et al., 2017), showed little to no correlation with altered biallelic expression in these hiPSC lines (Fig. 3.3).

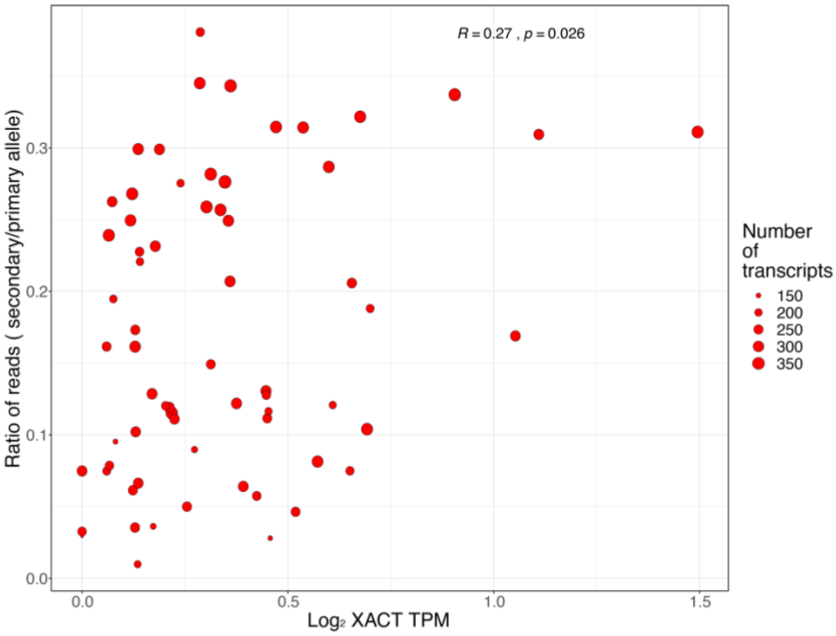

Figure 3.3 – XACT and XCI: Scatter plot showing the ratio of reads derived from the secondary allele (lowest expressed allele) compared to the primary allele (highest expressed allele) for all X-linked transcripts vs the log$_2$ XACT TPM for all healthy female lines. The size of the circle is determined by the number of transcripts used for the analysis

Next, the 74 hiPSC lines were stratified into Low, Medium and High XIST RNA populations, based on the RNAseq expression data (Fig. 3.2b; Table S2). This identified two main populations, with 47.3% showing the expected high levels of XIST RNA and a surprisingly elevated proportion, 40.5%, of the hiPSC lines having very low levels of XIST RNA expression (Fig. 3.2c). A minor population (12.2%) that showed an intermediate level of XIST expression was also identified. However, as this population represented a low number of hiPSC lines it wasn’t used for the main downstream analysis. The rest of the study focussed on comparing gene expression specifically between the populations stratified by High and Low levels of XIST RNA expression.

We next set out to examine the relationship between XIST expression levels and allelic expression within X-linked gene products (Fig. 3.2d). This analysis showed that hiPSC lines with High levels of XIST expression had a significantly lower fraction of genes with reads derived from the lowest expressed allele (hereon termed secondary allele; see methods), compared to the Low XIST population. Gene products within hiPSC lines displaying High levels of XIST had a median of 99.5% of reads originating from the highest expressed allele (hereon termed primary allele; see methods), with only 0.5% from the secondary allele. hiPSC lines with Low levels of XIST showed an increase in the proportion of reads derived from the secondary allele, with a median of 22.6% of reads being derived from the secondary allele and only 77.4% from the primary allele. We therefore conclude that XIST expression levels provide a suitable marker for detecting erosion of XCI within the hiPSC lines analysed.

We next mapped all the X-linked genes to their respective bands within the X chromosome and studied allele expression for genes across all bands, for both the High and Low XIST populations (Fig. 3.2e). These data again emphasise that the population with High XIST expression has much lower biallelic expression than the Low XIST population. However, it was apparent that the level of biallelic expression is not uniform across the X chromosome. Even within the High XIST population, certain bands, such as Xp22, Xq23 and Xq26, are more prone to displaying increased expression from the secondary allele (Fig. 3.2e). These same bands also displayed higher biallelic expression within the Low XIST population.

We expanded the chromosomal band analysis by calculating the median $\log_2$ fold change between the High and Low XIST populations, for all gene products and across all bands, at both the RNAseq and proteomics levels (Fig. 3.2f). As seen for the allelic expression, the fold change across chromosomal bands was not uniform. Interestingly, some of the hotspots highlighted by the allelic analysis (e.g. bands Xp22, Xq23 and Xq26), were also among the sites showing highest levels of change in RNA and protein expression. This is consistent with previous observations showing that there are specific loci that can preferentially escape XCI(Balaton and Brown, 2016, Tukiainen et al., 2017). We note that the independent transcriptomic and proteomic data sets both displayed very similar patterns of gene expression variation in response to XIST levels across the X chromosome. This concordance in RNA and protein-level data are consistent with a predominantly transcription-driven regulation of X-linked gene expression.

We also wanted to understand how the changes in gene expression between the High and Low XIST populations behaved for genes across all other chromosomes. Hence, we used all gene products that were detected both in both the RNAseq and proteomics datasets, aggregated them at the chromosome level and compared their respective RNA and protein fold changes. This chromosome-specific view showed that the highest fold change concordance is observed within X-linked genes, with a Pearson correlation of 0.56 (Fig. 3.2g). However, this same level of concordance was not observed across all other chromosomes, as each of the autosomes had a much lower correlation coefficient than was seen for the X chromosome, with the second highest being chromosome 10 with 0.36 and the median being 0.27. The data thus indicate a difference between X-linked genes and genes on all the autosomes.

To quantify these differences, we compared the median fold change for RNAs and proteins across all chromosomes (Fig. 3.2h). Unsurprisingly, the highest median increase observed within the Low XIST, compared to the High XIST population, at both the RNA and protein levels, occurs for genes on the X chromosome. However, the proteomics data, unlike the RNAseq data, also displayed increased median fold changes across all other chromosomes as well.

In summary, the RNA expression data show that hiPSC lines with High levels of XIST RNA display significantly lower biallelic expression than the hiPSC lines with Low XIST, with 99.5% of the reads derived from the primary allele and only 0.5% from the secondary allele. The lines with low expression of XIST showed a higher proportion of reads derived from the secondary allele (22.6%), consistent with erosion of XCI. The data also show that erosion of XCI in iPSCs results in both increased transcription and protein expression for X-linked genes. However, the same was not observed for genes on the autosomes, where the increased median fold change seen across all chromosomes was only detected at the protein level.

3.4.3 Impact of XCI erosion on the autosomes and dosage compensation

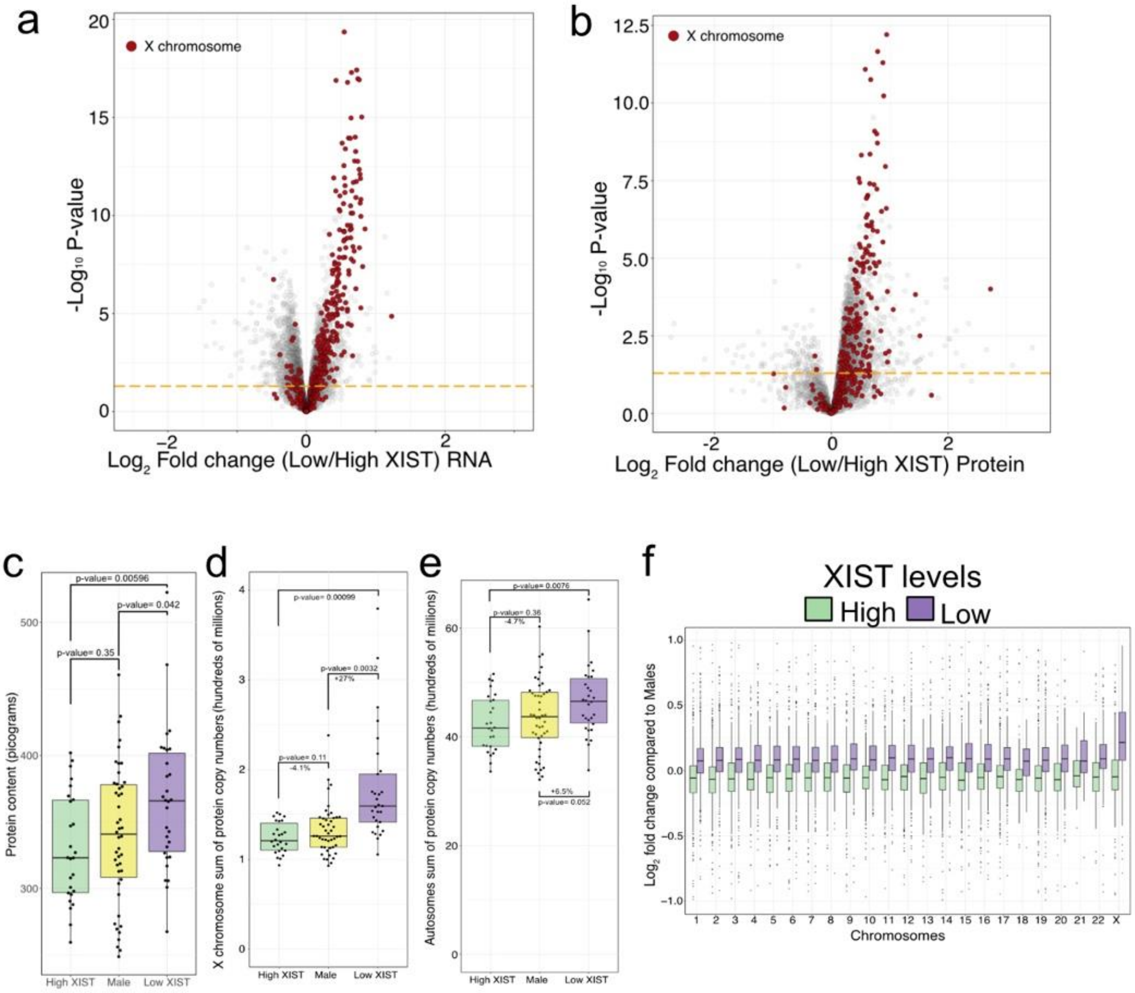

Figure 3.4 – Multi-omic overview: ***(a)*** Volcano plot showing the $\log_2$ fold change (Low/High XIST) on the X axis, with the -$\log_{10}$ p-value on the y axis for the RNAseq dataset. X chromosome transcripts are highlighted in red; autosome transcripts are coloured grey. All transcripts above the orange line have a p-value lower than 0.05 ***(b)*** Volcano plot showing the $\log_2$ fold change (Low/High XIST) on the X axis, with the -$\log_{10}$ p-value on the y axis for the proteomic dataset. X chromosome proteins are highlighted in red; autosomal proteins are coloured grey. All proteins above the orange line have a p-value lower than 0.05 ***(c)*** Box plots showing the estimated protein content (see methods) for the High XIST, Low XIST and Male populations. ***(d)*** Box plots showing the estimated protein copy numbers across the X chromosome for the High XIST, Low XIST and Male lines. ***(e)*** Box plots showing the estimated protein copy numbers across all autosomes for the High XIST, Low XIST and Male lines. ***(f)*** Box plot showing the protein $\log_2$ fold change (High XIST/Males and Low XIST/Males) across all chromosomes. For all Box plots the lower and upper hinges represent the 1st and 3rd quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge, the lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge

Next, we focussed on a differential expression analysis, comparing the High vs Low XIST-stratified populations at both the RNA and protein levels. This showed that ~55% of X-linked genes in the Low XIST population exhibited significantly increased expression of both RNA and protein, as compared to the High XIST population (Fig. 3.4a&b). However, when comparing gene expression from autosomes, once again we detected differences between the RNA and protein datasets. Thus, 9% of autosomal transcripts (1,087 out of 12,042), were significantly increased in expression and 11.2% (1,344 out of 12,042), significantly decreased in expression, in the Low compared to the High XIST population. In contrast, the proteomics data showed that 27.8% of the autosome-encoded proteins quantified (2,383 out of 8,593), were significantly increased in expression in the Low XIST population, while only 1.2% of autosome-encoded proteins (107 out of 8,593), were significantly decreased (Fig. 3.4b). These results show that there is a clear effect within the proteomics data that is not recapitulated by the RNAseq data.

Based on the increased fold change across all chromosomes and significantly increased expression of over 2,300 autosomal proteins, we suspected that the Low XIST population of hiPSC lines may have a higher average protein content per cell, as compared to the High XIST lines. To test this hypothesis, we used the MS data to estimate the total protein content and compare both XIST-stratified populations to each other. Furthermore, this dataset also provided the first opportunity to study the impact of human dosage compensation at the protein level and thus to determine how the global proteome may respond to erosion of XCI. Hence, we also compared both XIST stratified populations to 46 hiPSC lines derived from healthy male donors (see methods).

No significant differences in total protein content were detected between the High XIST female lines and the male lines (Fig. 3.4c). However, the Low XIST population had a significant increase of 7.3% in total protein content, when compared to the male lines, and an even more pronounced increase of 13.2%, when compared to the High XIST lines (Fig. 3.4c). To check if these changes in protein content were related to potential cell cycle differences between the respective stratified cell lines, we analysed the expression of a panel of genes previously characterised as being cell cycle regulated (Ly et al., 2014). Gene expression at both the RNAseq and protein levels showed no significant differences for these known cell cycle regulated genes between the High and Low XIST populations (Fig. 3.5). Hence, we conclude that the observed differences in protein content linked with erosion of XCI are likely not the result of altered kinetics of cell cycle progression.

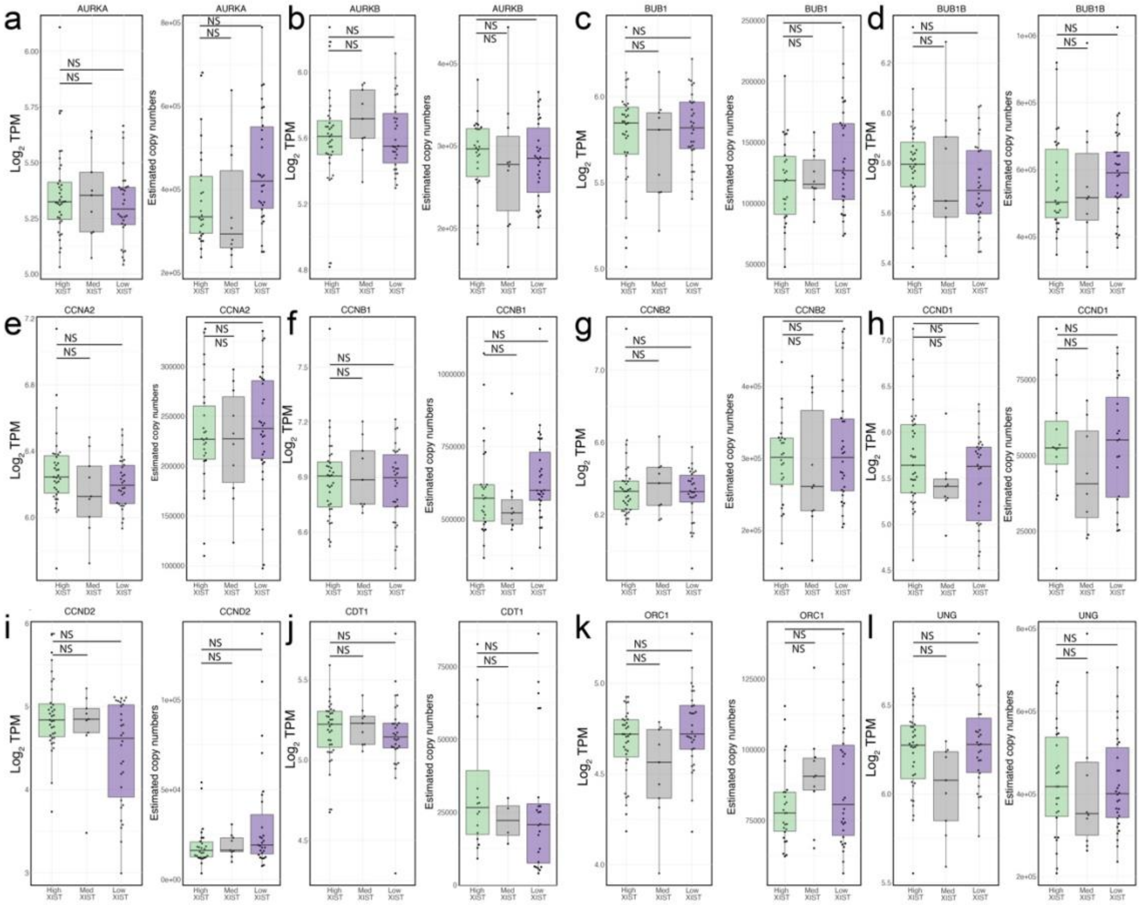

Figure 3.5 – Cell cycle markers: All box plots show the log2 TPM and estimated protein copy numbers within the High and Low XIST populations for a gene product. ***(a)*** AURKA. ***(b)*** AURKB. ***(c)*** BUB1. ***(d)*** BUB1B. ***(e)*** CCNA2. ***(f)*** CCNB1. ***(g)*** CCNB2. ***(h)*** CCND1. ***(i)*** CCND2. ***(j)*** CDT1. ***(k)*** ORC1. ***(l)*** UNG. For all box plots the upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge, the lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge.

We drilled down on these comparisons further and focussed on the total copy numbers for all X-linked proteins across both XIST-stratified female populations and the male lines. Once again, this comparison revealed no significant differences between the High XIST female population and the males (Fig. 3.4d). In contrast, the Low XIST population saw a dramatic increase in protein copy numbers of 27%, compared to the males (Fig. 3d). When we repeated this analysis for autosomal proteins (Fig. 3.4e), the Low XIST female population displayed 6.5% higher protein copy numbers than the males and 11.8% higher than the High XIST female population. Hence, these data suggested there was little or no significant difference in total protein levels between the High XIST female population when compared to hiPSC lines from male donors, while the Low XIST female population was significantly different to both. The Medium XIST population appeared to be more closely aligned to the High XIST than to the Low XIST population (Fig. 3.6a-d)

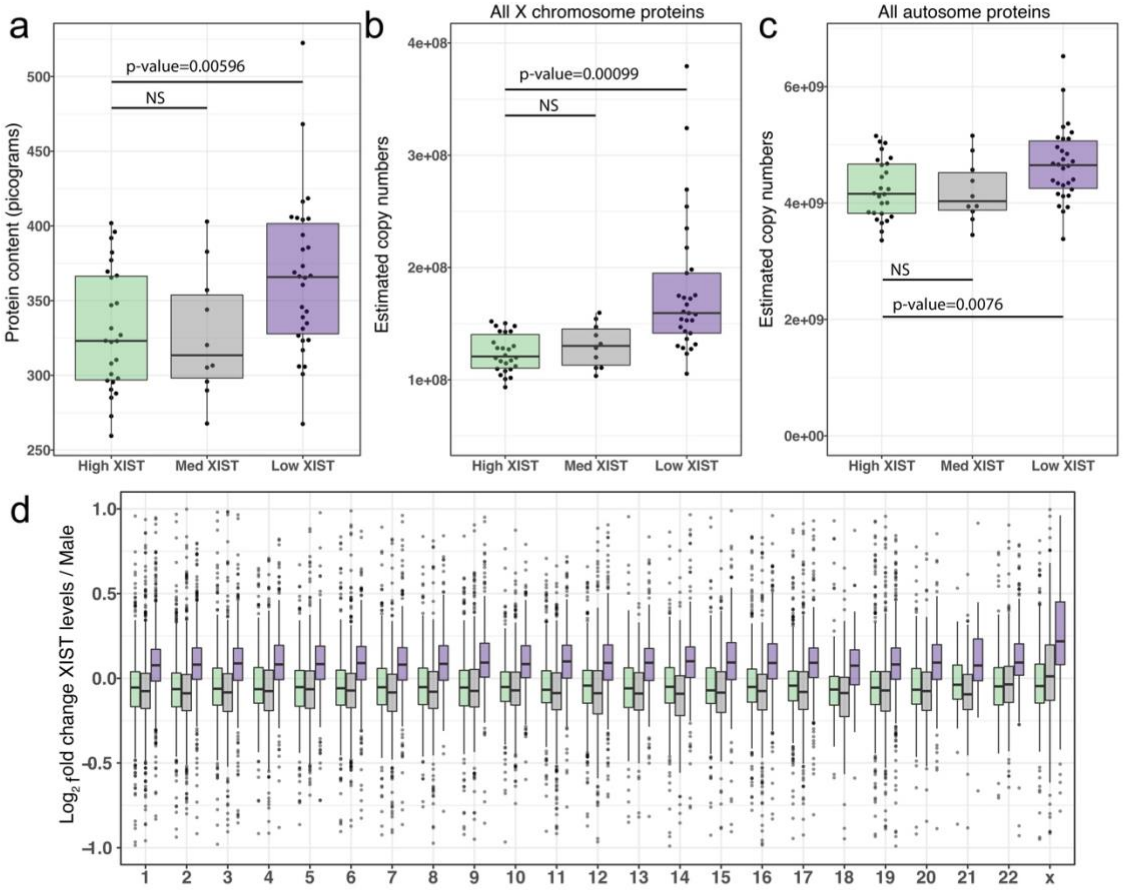

Figure 3.6 – High, Medium and Low XIST protein level data: ***(a)*** Box plot showing the estimated protein content (see methods) for the High, Medium and Low XIST. ***(b)*** Box plot showing the sum of protein copy numbers across the X chromosome for the High XIST, Medium and Low XIST lines. ***(c)*** Box plot showing the sum of protein copy numbers across all autosomes for the High XIST, Medium and Low XIST lines. ***(d)*** Box plot showing the protein log2 fold change for the High, Medium and Low XIST populations when compared to the male lines across all chromosomes. For all box plots the upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge, the lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge.

To obtain a more granular view, we compared the changes in expression for each protein within both the High and Low XIST female populations, in comparison to the male population, for each of their respective chromosomes (Fig. 3.4f & 3.6d). When comparing the male to High XIST female lines, the data showed no significant fold change difference between the X chromosome and any other chromosome. This demonstrates that XCI is effective in ensuring similar expression levels of X linked proteins between males and females with robust XCI. However, the situation is different when comparing male-derived lines with Low XIST female lines, where the median fold change for proteins encoded by genes on the X chromosome is significantly higher than for genes on all other chromosomes.

3.4.4 Proteome specific response to XCI

We next investigated in more detail how the proteome was altered in the Low XIST population and examined which types of proteins and which protein functions showed changes. We focussed on proteins that showed significantly increased expression (q-value<0.05) in the Low XIST population, as compared to the High XIST population, but without a corresponding significant increase in the RNAseq data. We discovered that this group of proteins were robust identifications, as they were enriched in high abundance proteins (abundance greater than the 75th percentile; see methods), with high numbers of Razor + unique peptides detected (RUP greater than the 75th percentile; Fig. 3.7a-b).

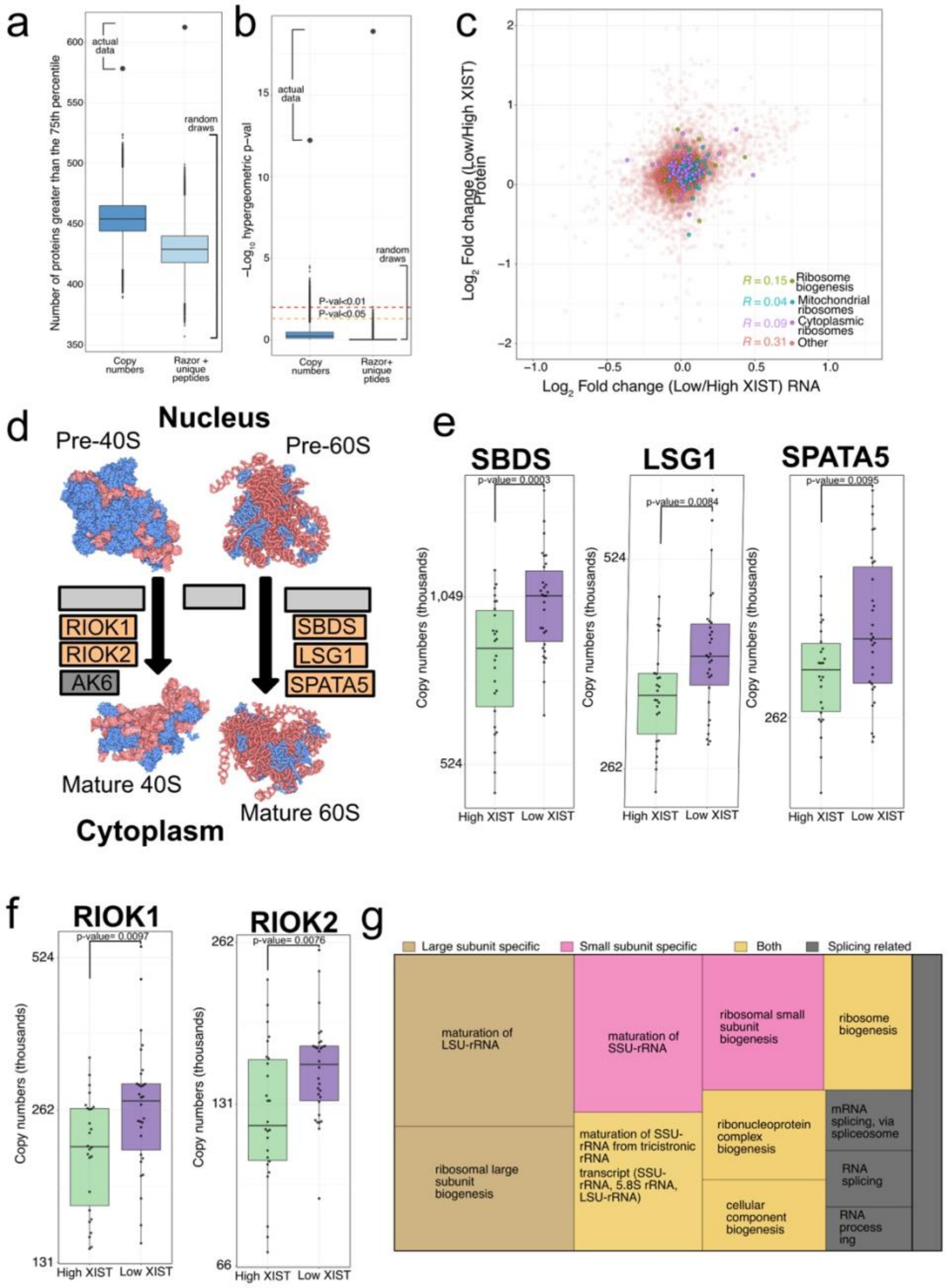

Figure 3.7 – Ribosome biogenesis: ***(a)*** Box plot showing the number of proteins with copy numbers greater than the 75th percentile and Razor + unique peptides greater than the 75th percentile for the simulations and the actual experimental data (see methods). ***(b)*** Box plot showing the hypergeometric p-value for proteins with copy numbers greater than the 75th percentile and Razor + unique peptides greater than the 75th percentile for the simulations and the actual experimental data (see methods). ***(c)*** Scatter plot showing the $\log_2$ fold change (Low/High XIST) at the protein and RNA level. Ribosome biogenesis, cytoplasmic and mitochondrial ribosomal proteins are highlighted, and Pearson correlation coefficients provided. ***(d)*** Schematic showing the cytoplasmic ribosome biogenesis proteins with proteins significantly increased in expression highlighted in orange ***(e)*** Box plots showing the estimated protein copy numbers for SBDS, LSG1 and SPATA5 within the Low and High XIST populations. ***(f)*** Box plots showing the estimated protein copy numbers for RIOK1 and RIOK2 within the Low and High XIST populations. ***(g)*** Treemap plot showing the results of a Biological Process overrepresentation test focussed on ribosome biogenesis proteins. The rectangle size is proportional to the enrichment level of the specific terms. For all Box plots the lower and upper hinges represent the 1st and 3rd quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge, the lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge

A follow-up analysis, focussed on biological processes, was carried out via a Gene Ontology (GO) overrepresentation test, using Panther[Mi et al., 2017]. This showed that these proteins were enriched specifically for the GO terms ‘ribonucleoprotein complex biogenesis’ and ‘mRNA metabolic process’. We note that these enriched GO terms are associated with many proteins involved in post-transcriptional mechanisms that could increase total protein expression from a constant amount of mRNA. This includes genes related to processes including ribosome subunit biogenesis, ribosome function and the control of protein translation. Consistent with this, when comparing changes in expression at the respective RNA and protein levels between the XIST RNA-stratified populations, Pearson correlation coefficients were particularly low for the ribosomal (0.09) and ribosome subunit biogenesis (0.15) proteins (Fig. 3.7c). This indicates a potentially important role for post-transcriptional mechanisms in regulating protein expression from these genes.

Overall, > 36% of all proteins involved in ribosome subunit biogenesis, as described in KEGG (Kanehisa and Goto, 2000), showed significantly (Q-value<0.05) increased expression within the Low XIST RNA population and of those, >70% of genes encoding proteins involved in cytoplasmic stages of ribosome subunit biogenesis showed increased protein expression (Fig. 3.7d). For example, SPATA5 (a human homologue of yeast Drg1), SBDS and LSG1 (Fig. 3.7e), which are all involved in the final step of 60S maturation and the atypical RIO kinases RIOK1 and RIOK2 (Fig. 3.7f), involved in the final step of 40S maturation (Vanrobays et al., 2003), all showed significantly increased protein expression within the Low XIST RNA population. Interestingly, all of these proteins are all involved in the cytoplasmic quality control of ribosomes [Pena et al., 2017, Karbstein, 2013, Cerezo et al., 2019). To drill down further into the process of ribosome subunit biogenesis, we performed a more granular enrichment analysis on the proteins involved within this pathway and found the highest enrichment was on terms related to large subunit biogenesis and ribosomal RNA (Fig. 3.7g)

3.4.5 Proteome specific changes affecting ribosomes and translation initiation

When we focused on the total estimated copy numbers for the cytoplasmic ribosomes, we noticed they mirrored the protein content closely, with a mean increase of 13% (p-value 0.0019) in the Low XIST population (Fig. 3.8a). Interestingly, the changes were not uniform between the large subunit (60S) and the small subunit (40S). The largest increase in the Low XIST population affected proteins belonging to the 60S subunit (Fig. 3.8b), resulting in a significant change in the ratio of protein copy numbers between 60S and 40S ribosomal proteins (Fig. 3.8c). Overall, ~42% of the ribosomal proteins and ribosomal protein S6 kinases detected were significantly increased in expression in the Low XIST RNA population (Fig. 3.8d). Of these proteins, the one with the highest p-value (p-value <6.86e-^{07}) was p90 ribosomal S6 kinase (RPS6KA3), which is encoded on the X chromosome and is linked to cell growth via increased cap-dependent translation through phosphorylation of RPS6 (Roux et al., 2007) and RPTOR (Carriere et al., 2008) (for all significantly increased X-linked kinases see Fig. 3.9).

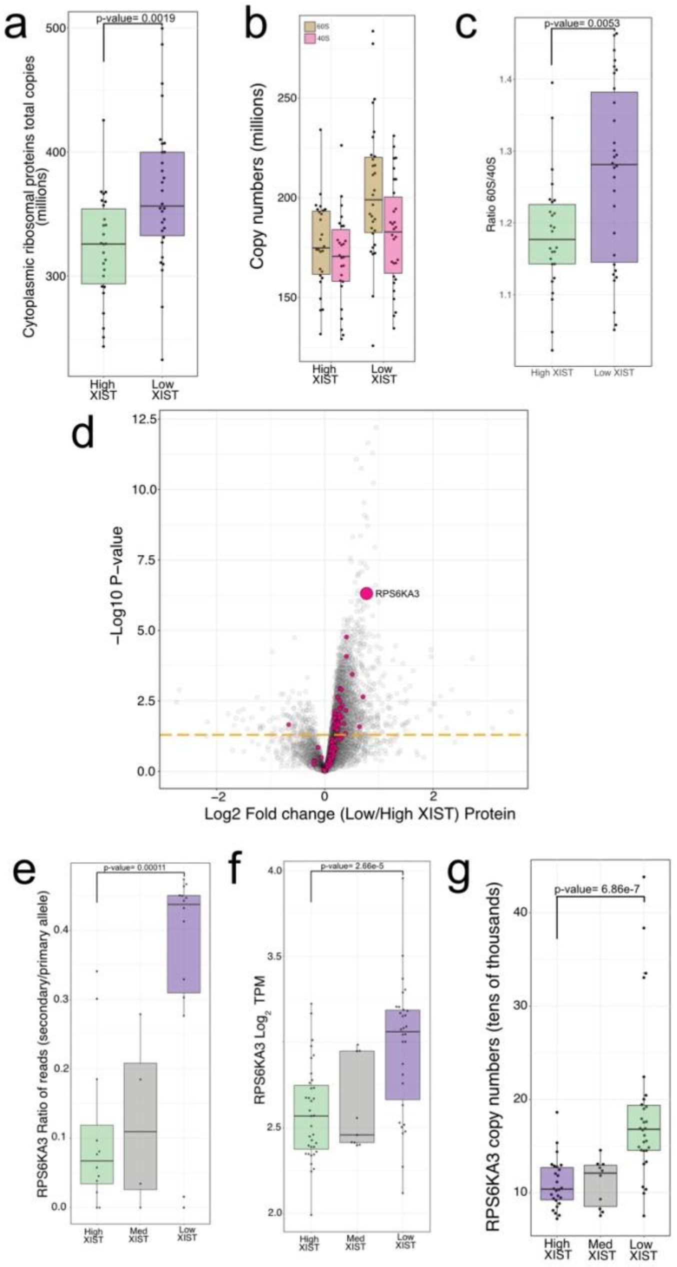

Figure 3.8 – Ribosomes and translational initiation: ***(a)*** Box plot showing the estimated copy numbers for the sum of all cytoplasmic ribosomal proteins within the High and Low XIST populations. ***(b)*** Box plot showing the estimated copy numbers for the sum of all 60S (large ribosomal subunit) and 40S (small ribosomal subunit) proteins within the High and Low XIST populations. ***(c)*** Box plot showing the ratio of the sum of 60S to 40S ribosomal proteins within the High and Low XIST populations. ***(d)*** Volcano plot showing the protein $\log_2$ fold change (Low/High XIST) on the X axis, with the $-\log_{10}$ p-value on the y axis. Ribosomal proteins and ribosomal S6 kinases are highlighted in pink; all other proteins are coloured grey. All proteins above the orange line have a p-value lower than 0.05 ***(e)*** Box plot showing the ratio of reads mapped to the secondary allele compared to the primary allele for RPS6KA3 within the High, Medium and Low XIST populations. ***(f)*** Box plot showing the $\log_2$ TPM of RPS6KA3 within the High, Medium and Low XIST populations. ***(g)*** Box plot showing the estimated protein copy numbers of RPS6KA3 within the High, Medium and Low XIST populations. For all Box plots the lower and upper hinges represent the 1st and 3rd quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge, the lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge

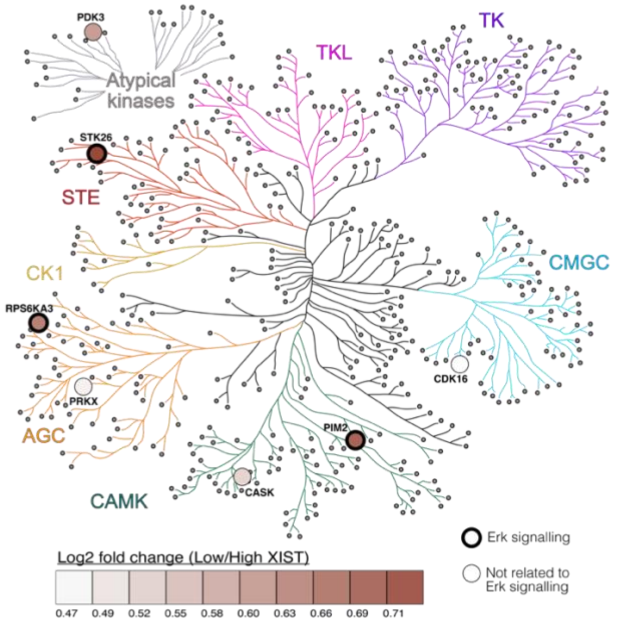

Figure 3.9 – X-linked Kinome: Protein level kinase map for all X-linked kinases showing the log2 fold change (Low/High XIST) for the kinases that were significantly in expression within the Low XIST population compared to the High XIST population.

As RPS6KA3 is an X-linked kinase, we checked the ratio of reads mapped to the secondary allele, compared to the primary allele and detected significant differences between the High and Low XIST lines. Thus, the median ratio for the High XIST population was 0.07, whereas the median ratio in the Low XIST population was 0.44 (Fig. 3.8e). We also detected significantly higher expression of RPS6KA3 within the Low XIST, as compared to the High XIST population, in both the RNAseq (Fig. 3.8f) and proteomics (Fig. 3.8g) data sets. These data support a model in which transcriptional de-repression of the inactive X chromosome in low XIST lines increases the expression of proteins encoded on the X, which in turn can increase mRNA translation and protein expression for autosomal genes.

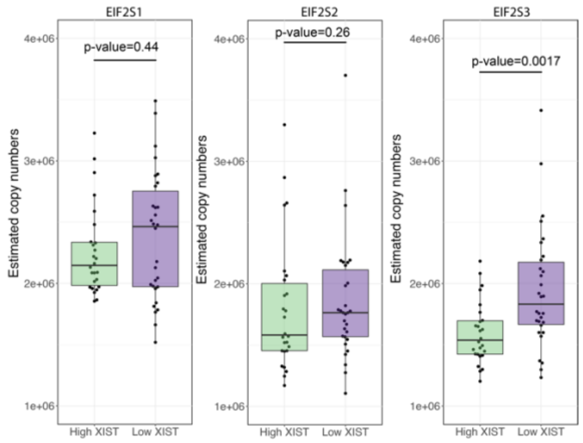

Figure 3.10 – eIF2 copy numbers: Box plots showing the estimated copy numbers for the High and Low XIST populations for EIF2S1, EIF2S2 and EIF2S3. For all boxplots the lower and upper hinges represent the 1st and 3rd quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge, the lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge.

We next looked for additional X-linked genes with the potential to affect mRNA translation, focussing on the translation initiation factors EIF2S3 and EIF1AX. These proteins both form part of the 43S translation preinitiation complex (Jackson et al., 2010). EIF2S3 is a member of the heterotrimeric eIF2 complex, which delivers an initiator methionyl transfer RNA to the ribosome. Based on the protein expression data in this study, EIF2S3 appears to be the rate limiting subunit of the eIF2 complex within the hiPSC lines (Fig. 3.10). As seen with RPS6KA3, a significantly higher ratio of reads from the secondary allele of EIF2S3 are seen within the Low XIST, compared to the High XIST population (Fig. 3.11a), together with significantly higher expression at the RNAseq (Fig. 3.11b) and proteomics levels (Fig. 3.11c). Similarly, EIF1AX is involved in virtually all the steps in mRNA translation initiation, from pre-initiation to ribosomal subunit joining (Nag et al., 2016) and shows a dramatic median increase of over 2.3 million protein copies in the low XIST RNA population. A higher median fraction of reads was seen from the EIF1AX secondary allele (0.31 in the Low XIST vs 0.13 in the High XIST; Fig. 3.11d). However, this difference did not meet a statistically significant threshold, likely due to outlier lines within the High XIST population with a high number of reads from the secondary allele. Nonetheless, the expression levels of EIF1AX at both the RNA (Fig. 3.11e) and protein (Fig. 3.11f) levels were both significantly increased in the low XIST lines. To validate further the MS-based identifications and quantifications, we mapped the peptide coverage for all 3 X-linked translational modulators (i.e., RPS6KA3, EIF2S3 and EIF1AX), which revealed robust results with high numbers of Razor + unique peptides (all over 10 RUP) detected, as well as high sequence coverage (all over 53%; Fig. 3.12).

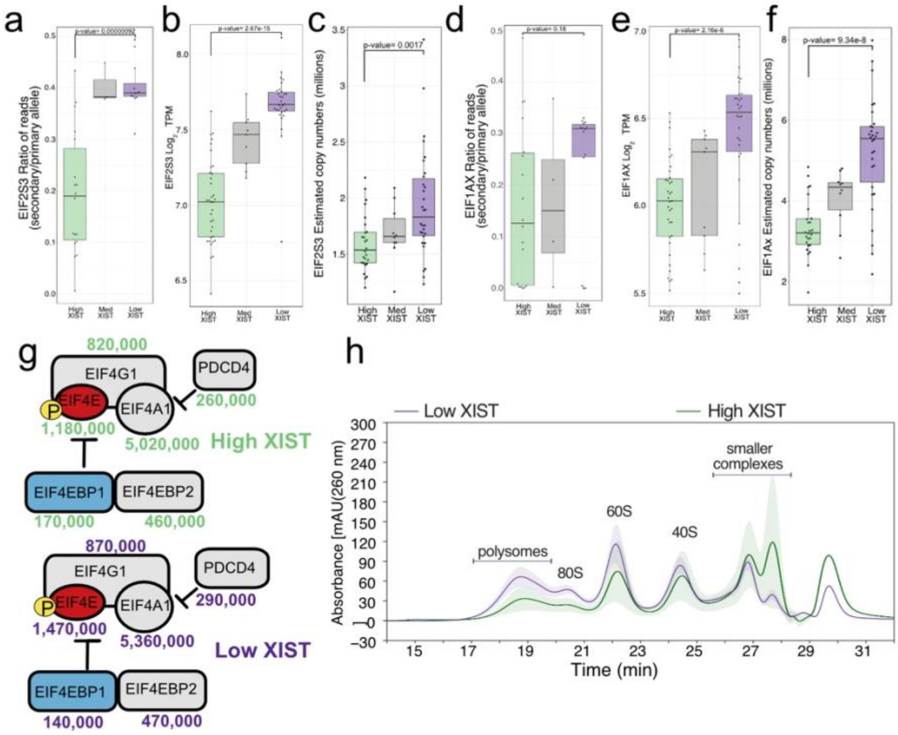

Figure 3.11 – Translational machinery: For all Box plots the lower and upper hinges represent the 1st and 3rd quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge, the lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge. **(a)** Box plot showing the ratio of reads mapped to the secondary allele (lowest expressed allele) compared to the primary allele (highest expressed allele) of EIF2S3 within the High, Medium and Low XIST populations. **(b)** Box plot showing the Log_2 TPM for EIF2S3 within the High, Medium and Low XIST populations. **(c)** Box plot showing the protein copy numbers for EIF2S3 within the High, Medium and Low XIST populations. **(d)** Box plot showing the ratio of reads mapped to the secondary allele compared to the primary allele for EIF1AX within the High, Medium and Low XIST populations. **(e)** Box plot showing the Log_2 TPM of EIF1AX within the High, Medium and Low XIST populations. The lower and upper hinges represent the 1st and 3rd quartiles. **(f)** Box plot showing the protein copy numbers of EIF1AX within the High, Medium and Low XIST populations. **(g)** Schematic showing the protein copy numbers for the eIF4F complex and its inhibitors displayed for both the High and Low XIST populations. Proteins represented via red boxes are significantly increased, light blue boxes are significantly decreased and elements in grey boxes remain unchanged. **(h)** Line plot showing the Ribo Mega-SEC derived line plot showing the mean High and Low XIST polysome profile with the coloured ribbon representing the standard deviation.

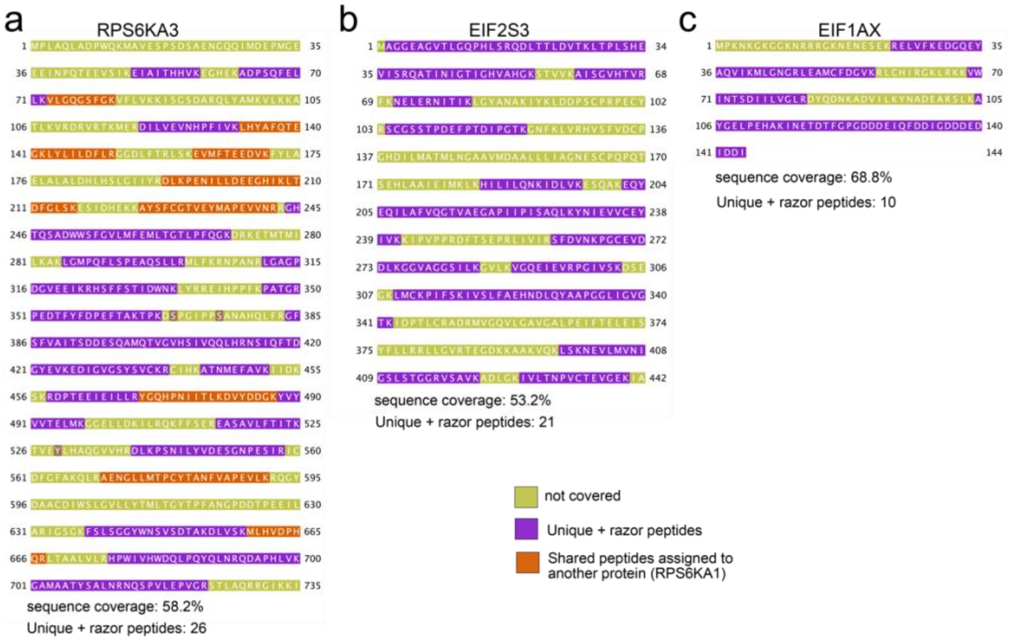
a
RPS6KA3
sequence coverage: 58.2%
Unique + razor peptides: 26
b
EIF2S3
sequence coverage: 53.2%
Unique + razor peptides: 21
c
EIF1AX
sequence coverage: 68.8%
Unique + razor peptides: 10
not covered
Unique + razor peptides
Shared peptides assigned to another protein (RPS6KA1)

Figure 3.12 – Peptides and sequence coverage: Peptide maps showing the amino acid sequence and MS derived peptide coverage for **(a)** RPS6KA3. **(b)** EIF2S3. **(c)** EIF1AX.

The potential impact of eroded XCI on factors affecting the efficiency of translation initiation was not limited to upregulation of proteins encoded by genes on the X chromosome. For example, eIF4F is another complex that is vital for translation initiation(Merrick, 2015). It is composed of the alpha (EIF4A1 & EIF4A3), epsilon (EIF4E) and gamma (EIF4G1 & EIF4G3) subunits. The stoichiometry of EIF4E, the cap binding subunit (located on chromosome 4), to its inhibitors EIF4EBP1 (located on chromosome 8) and EIF4EBP2 (located on chromosome 10), has been proposed as one of the regulatory mechanisms for eIF4F translational control (Richter and Sonenberg, 2005). Interestingly, in the Low XIST population, there is a significant increase in EIF4E levels compared to the High XIST population, alongside a parallel significant decrease in levels of EIF4EBP1 (Fig. 3.11g). The decrease in EIF4EBP1 is particularly relevant in this case, as only ~1.2% of all proteins are significantly decreased in expression in the Low XIST population.

In summary, the data are consistent with erosion of XCI in the Low XIST RNA population causing a major change in the proteome that results in a global increase in total protein levels, which is mediated, at least in part, via increased levels of translation. To test the hypothesis that the hiPSC lines with low levels of XIST RNA have increased levels of translation, we compared polysome and ribosome levels between hiPSC lines showing either High, or Low levels of XIST RNA, using the Ribo Mega-SEC method(Yoshikawa et al., 2018).

This size exclusion chromatography (SEC)-based method can separate large protein complexes, including polysomes, 80S monosomes and ribosome subunits.

The Ribo Mega-SEC data were remarkably consistent with the MS-based proteomics data. This showed that both the polysome containing fractions and 80S monosome containing fractions are increased significantly in extracts from hiPSC lines with Low levels of XIST RNA, compared to extracts from hiPSC lines with High XIST RNA levels (Fig. 3.11h). Moreover, the SEC data show a more pronounced increase within the Low XIST RNA lines in the ratio of 60S to 40S ribosome subunits, consistent with larger increase in 60S ribosomal protein levels in the MS data (Fig. 3.8b & c). The polysome fractionation data are thus consistent with a model in which erosion of XCI in the hiPSC lines expressing low levels of XIST RNA upregulates protein translation capacity and leads to a global increase in protein levels from multiple genes on both autosomes and the X chromosome.

3.5 Discussion

This study provides the first in depth global analysis of how erosion of X chromosome inactivation (XCI) affects gene expression and dosage compensation at the protein level in human cells and compares this to matching RNAseq data. We analysed RNA expression in 74 independent human iPSC lines derived from healthy female donors that were generated by the HipSci consortium(Kilpinen et al., 2017). This showed that ~40% of these lines expressed very low levels of the long, non-coding XIST RNA. Further analysis of bi-allelic expression from X chromosome encoded genes showed that low levels of XIST RNA strongly correlated with erosion of XCI, as reflected by a significantly increased fraction of reads being mapped to the lowest expressed (secondary) allele. We therefore characterised how the erosion of XCI remodels gene expression and the human proteome, by comparing in detail the levels of RNA and protein expression across 56 female hiPSC lines, stratified according to high vs low XIST RNA levels, as well as comparing these to 46 lines derived from healthy male donors, which have only one X chromosome.

First, the data show that hiPSC lines derived from healthy female donors with low expression of XIST significantly upregulated levels of both RNA and protein from many genes on the X chromosome. Including the evidence of increased biallelic expression in the Low XIST lines, these data indicate that erosion of XCI causes increased protein expression from X-linked genes primarily via a transcriptional mechanism.

We also leveraged the unique characteristics of our dataset to compare how protein expression levels are affected by dosage compensation in the respective High and Low XIST-stratified female lines vs the male lines. The data indicate that when compared to the males, the High XIST lines, which exhibit robust XCI, show similar patterns of expression for genes derived from the X chromosome. This suggests that when XIST levels are high, there is effective dosage compensation for the extra X chromosome copy, acting at the protein level. However, for the lines with Low XIST expression levels, which exhibit higher erosion of XCI, a very different situation is evident. In this case the X chromosome encoded proteins are increased in expression by 27% compared to males, now making gene expression from the X chromosome significantly different between males and females.

Second, we also detected a significant increase in protein expression levels from 26% of autosomal genes within the Low XIST female lines. In contrast with genes on the X chromosome, 21% of all autosomal gene products were only increased at the protein level and not RNA. Thus, autosome encoded genes showed a low overall RNA:protein fold change correlation (median Pearson correlation for autosomes of 0.27, compared with 0.56 for the X chromosome).

It should be noted that, unlike the RNAseq data, protein copy numbers can be estimated from the MS data without the need for spike-ins, via the proteomic ruler (Wisniewski et al., 2014). The estimated copy numbers allow us to calculate and explore differences in absolute protein content. With this RNAseq data set and normalisation approach, which did not include spike-in controls, it was not possible to detect potential transcriptional amplification effects (Loven et al., 2012). Furthermore, recent reports have suggested that in mESCs, changes in ERK signalling can cause hypomethylation(Choi et al., 2017, Song et al., 2019) potentially also affecting global transcription. Our data reveal that multiple X-linked kinases associated to ERK signalling are significantly increased in expression within the Low XIST population (Fig. S4). Their contribution to cell phenotypes herein remains to be determined. Therefore, we cannot completely exclude the possibility that absolute changes at the transcript level may also occur for autosomal genes, which are masked due to technical issues in the detection of transcriptional amplification. Moreover, while erosion of XCI is generally thought to specifically affect the expression of X-linked genes, we note that increased transcription from autosomal genes has been reported in murine trophoblasts when they failed to induce XCI (Sakata et al., 2017). Nonetheless, the data from our study indicate that erosion of XCI in human cells can affect protein levels encoded by a much wider range of genes than was previously shown by RNAseq data alone, including autosome-linked genes and disease loci.

Third, the comparison of High vs Low XIST hiPSC populations showed that the Low XIST expressing cell lines had a median increase of ~13% in total protein content, which was also significantly higher than the protein content of the lines derived from healthy male donors. In considering potential mechanisms that could cause this increased protein content, several lines of evidence suggest it may result, at least in part, from post-transcriptional regulation affecting the translation efficiency of a subset of mRNAs. It has been proposed that translation rates positively correlate with protein abundance(Brockmann et al., 2007, Liu et al., 2016, Marguerat et al., 2012). Congruently our data show the autosome-encoded proteins that are significantly increased in expression, but without a corresponding mRNA increase, are enriched in high abundance proteins with high peptide counts.

Focussing on the proteins that show statistically significant, RNA-independent increases in expression in Low XIST lines, revealed an enrichment for GO terms associated with ribonucleoprotein complex biogenesis. Furthermore, independent polysome profiling analyses comparing High and Low XIST hiPSC lines, using the Ribo Mega-SEC method (Yoshikawa et al., 2018), also showed a significant increase in polysomes and 80S ribosomes within the Low XIST lines. Thus, analyses using both the separate MS proteomics and polysome methods, support the view that iPS cells derived from healthy female donors showing XCI erosion have increased protein translation activity, resulting in a global increase in total protein levels. It will be interesting to analyse in the future whether mRNAs encoding the subset of autosomal proteins that show increased abundance share some common features or sequence motifs that promote efficient translation.

In light of the elevated polysome levels and increased protein expression observed in the Low XIST hiPSC population, it is interesting that two X-linked genes that encode important regulators of translational initiation (EIF1AX and EIF2S3), as well as a kinase known to modulate translation (RPS6KA3), all show highly increased expression at both the RNA and protein levels. It has been proposed that protein synthesis is principally regulated at the initiation stage(Jackson et al., 2010). The EIF1AX and EIF2S3 proteins are thus candidates for mediating, at least in part, the mechanism whereby erosion of XCI causes an increase in protein translational capacity. Interestingly, both the EIF1AX and EIF2S3 genes have been categorized previously as facultative XCI escapees(Zhang et al., 2013a, Belling et al., 2017), meaning they are amongst a subset of X-linked genes that can escape transcriptional repression, despite the global repressed XCI state. This local increased gene dosage effect suggests that even female lines with normal XCI may differ in translational capacity from male cells with only a single X chromosome.

Our data show that erosion of XCI in primed human induced pluripotent stem cells has the potential to cause major changes at the level of protein expression, which in turn could have important implications for disease progression and response to therapy in females. Many autosomal genes which control cell growth, survival and proliferation like ERK2(Papa et al., 2019), FYN(Saito et al., 2010) and CDK6(Yang et al., 2017) were shown to be significantly increased in abundance in eroded hiPSCs, suggesting increased oncogenic potential. Though it is unlikely that undifferentiated PSCs will be injected into human patients, the differentiated cells derived from these hiPSCs could be. And these differentiated cells would share the XCI state, and likely also the increased abundance of autosomal proteins seen in the hiPSC they were derived from. Thus, it is important to consider potential solutions to the problem. It has been proposed that the erosion of XCI in primed pluripotent stem cells occurs under specific growth conditions under prolonged culture(Cloutier et al., 2022). This would fit with the results seen within our data, where multiple hiPSC lines derived from the same donor can have completely different XCI states. If female hiPSCs are to be used in clinical treatments or regenerative medicine, some mitigation to prevent the erosion of XCI should be considered. One potential solution is to modify the culture conditions to use a specific xenofree medium, which was proposed to significantly reduce the incidence of erosion of XCI(Cloutier et al., 2022). Another alternative solution is to either revert the primed hiPSCs back into a naïve pluripotency state, or to directly reprogram somatic cells back into a naïve pluripotency state. Naïve pluripotent stem cells would have both X chromosomes active and would theoretically reinstate X inactivation upon exit of the naïve pluripotency stage, thus minimising the issues of erosion of XCI.

Finally, it would be of great interest to explore if the changes seen in hiPSCs that were caused by the loss of XIST expression and erosion of XCI suggest might occur *in-vivo* in other cell types that are known to undergo prolonged expansion. It has been reported that in rapidly proliferating cells, such as T and B lymphocytes undergoing clonal expansion, there is a loss of expression of XIST on the X chromosome (Wang et al., 2016, Syrett et al., 2017). This suggests there is potential for increased erosion of XCI and consequently increased abundance of autosomal proteins in females as seen with the hiPSCs. T and B cells, as well as other hematopoietic populations, are of particular importance because changes in the abundance of proteins that function as activating or inhibitory receptors, immunosuppressive molecules or metabolic enzymes could have direct implications in health, disease propensities and auto immune diseases, conditions in which we know there are sex specific biases.

Chapter 4 – Large-scale DIA-based neutrophil proteomics identifies temporal changes and hallmarks of delayed recovery in COVID19 patients

4.1 Pre-print relating to this chapter

- Neutrophil proteomics identifies temporal changes and hallmarks of delayed recovery in COVID19 (Long et al., 2022)
- All supplemental tables are available at: https://dmail-my.sharepoint.com/:f:/g/personal/ajbrenesmurillo_dundee_ac_uk/EmRtmTwhHu5GojFV3yF5P20BDqbhxR4S3T9BWRj-a2bbkQ?e=cLpfcK

4.1.1 Contributions to the pre-print on this chapter

- I helped to design and refine the data independent acquisition (DIA) method used in the study
- I assisted in planning the proteomic pipeline for the study
- I analysed the proteomic data and performed the main interpretation of the results
- I generated all figures used in this study
- I wrote the paper in collaboration with the first and corresponding authors and with input from the remaining authors

4.1.2 Main contributions of the other authors to the pre-print on this chapter

- The blood samples were collected in Ninewells Hospital by a team from the Chalmers Laboratory lead by Merete Long and Holly Keir
- The neutrophils and PBMCs were isolated by a team from the Chalmers Laboratory lead by Merete Long and Holly Keir
- The neutrophil and PBMC samples were prepared for proteomic analysis in the Cantrell Laboratory by Andrew Howden, Christina Rollings and Amy Lloyd
- Merete Long performed the PRG2 ELISA
- The mass spectrometry raw files were generated by the FingerPrints Proteomics Facility in the School of Life Sciences
- Andrew Howden, Gabriel Sollberger, Merete Long, Doreen Cantrell and James Chalmers assisted in the interpretation of the data

4.2 Introduction

Coronavirus disease 2019 (COVID19) caused by SARS-CoV-2 infection has a diverse spectrum of presentations; from asymptomatic or pre-symptomatic, to mild illness with flu-like symptoms, moderate illness demonstrating lower respiratory disease and severe or critical illness involving development of pneumonia and acute respiratory distress syndrome (ARDS)(Wang et al., 2020, Huang et al., 2020, Xu et al., 2020). The immune response plays a key role in determining the outcome of SARS-CoV-2 infection, with an excessive inflammatory response accounting for the majority of morbidity and mortality. Major immune events following COVID19 include generation of a type I interferon response (IFN-I)(Wilk et al., 2020, Lee and Shin, 2020), reduced numbers of T cells(Lucas et al., 2020, Bergamaschi et al., 2021, Mathew et al., 2020), cytokine storm(Del Valle et al., 2020, Lucas et al., 2020, Bergamaschi et al., 2021, Jose and Manuel, 2020), emergency myelopoiesis and myeloid compartment dysregulation(Schulte-Schrepping et al., 2020, Wilk et al., 2020, Kuri-Cervantes et al., 2020, Mathew et al., 2020), as well as procoagulant pathway activation(Nicolai et al., 2020, Jose and Manuel, 2020). A hallmark of severe disease is T-cell lymphopenia in combination with increased blood neutrophil counts(Mann et al., 2020, Tan et al., 2020, Reusch et al., 2021, Huang et al., 2020) . Neutrophils are key effector cells of the innate immune response that play a role in responses to pathogens and tumours(Burn et al., 2021, Quail et al., 2022) and in severe COVID19 there is evidence of emergency myelopoiesis and the appearance of circulating immature and dysfunctional neutrophils in peripheral blood(Aschenbrenner et al., 2021, Silvin et al., 2020, Carissimo et al., 2020, Schulte-Schrepping et al., 2020, Morrissey et al., 2021). There is also a signature of neutrophil activation in patients with severe COVID19 that predicts disease trajectory(Meizlish et al., 2021, Thwaites et al., 2021), as well as evidence of increased NETosis(Dowey et al., 2022). Neutrophil populations following SARS-CoV-2 infection have been characterized by combinations of high-resolution mass cytometry(Panda et al., 2022, Morrissey et al., 2021, Rendeiro et al., 2021), flow cytometry(Carissimo et al., 2020) and by single cell RNA sequencing(Wilk et al., 2020, Stephenson et al., 2021, Shaath et al., 2020, Melms et al., 2021). However, it is recognised that in-depth quantitative analysis of cellular proteomes can provide insights unobtainable from transcriptomes, particularly relating to neutrophil biology(Hoogendijk et al., 2019). Proteomic analysis of neutrophils has thus identified a type 1 interferon and prothrombotic hyperinflammatory signature in neutrophils isolated from COVID19 infected individuals with acute respiratory distress syndrome (ARDS)(Reyes et al., 2021). However, there has been no systematic analysis of neutrophil proteomes in patients with COVID19 of differing severity and no exploration of the persistence of changes in neutrophil phenotypes following SARS-CoV-2 infection. Accordingly, in the current study mass spectrometry was used to map the protein signatures of peripheral blood neutrophils from a cohort of patients who were admitted to hospital with COVID19 from May 2020 to December 2020, sampled up to 29 days following hospitalization, with matched control populations. This extensive patient cohort and sampling strategy has provided an in-depth analysis of the neutrophil proteomes in health and disease and provided novel insights into changes in neutrophils during acute SARS-CoV-2 infection and during the disease recovery phase.

4.3 Methods

The PREDICT-COVID19 study was a prospective observational case-control study conducted at Ninewells Hospital, Dundee, UK. Patients with suspected or confirmed COVID19 were enrolled within 96 hours of hospital admission with SARS-CoV-2 infection confirmed by RT-PCR performed on combined oropharyngeal and nasopharyngeal swabs. Written informed consent was provided by all participants. Two control populations were included; (i) patients presenting with community-acquired lower respiratory tract infections (LRTI) not due to SARS-CoV-2 infection and (ii) age and sex matched, non-infected controls.

Blood sampling was performed at enrolment (study day 1), and for the COVID19 cohort additional blood sampling was performed on day 7, 15 and 29 while in hospital. The timing of blood sampling was standardised (between 0900-1100h each day). A subset of participants who had been discharged returned as outpatients for sampling at day 29. Full inclusion and exclusion criteria are shown online. Inclusion criteria for controls were age ≥16 years, absence of an infection-related diagnosis, judged as clinically stable by the investigator and able to give informed consent. Exclusion criteria were known or past SARS-CoV-2 infection in the past 3 months, known contact with a COVID19 positive case in the preceding 14 days, any current infection, and any contraindication to venepuncture or participation in the study as judged by the investigator.

4.3.1 Clinical variables

Baseline severity was classified according to the WHO scale as WHO3= Hospitalized, not requiring supplementary oxygen, WHO4= Hospitalized, requiring oxygen through facemask or nasal prongs, WHO5/6= requiring high flow nasal oxygen, continuous positive airway pressure or invasive mechanical ventilation. Patients were categorised as either recovered (WHO=1) or not recovered (WHO2-3) at day 29, as determined on the basis of symptoms reported at follow-up, with patients still hospitalized categorized as non-recovered.

4.3.2 Clinical outcomes

Key clinical outcomes against which proteomic data were to be compared were mortality, requirement for mechanical ventilation, requirement for non-invasive ventilation, clinical severity on a 7-point ordinal scale as well as clinical outcome at day 29 on a 7-point ordinal scale.

4.3.3 Neutrophil isolation

20 ml of peripheral venous blood was drawn into 2x10ml EDTA(K2) vacutainers using a 21G butterfly needle for cell isolation. Within a maximum of 2h of venepuncture, neutrophils were isolated using EasySep™ Direct Human Neutrophil Isolation Kit (STEMCELL Technologies #19666) utilising negative immunomagnetic selection, as per the manufacturer’s instructions.

In brief, 50µl Selection Cocktail and Direct RapidSpheres™ were added per ml of blood and incubated for 5 min. DPBS (without Mg2$^+$/Ca2$^+$) containing 1mM EDTA was added to a final volume of 50ml and the tube inserted into an Easy50 EasySep™ magnet (STEMCELL Technologies #18002). After incubation for 10 min, the neutrophil-rich upper layer was transferred to a new tube and the same volume of RapidSpheres™ as in the first step was added. Following 5 min incubation, the tube was inserted into the magnet for a 5 then 10 min incubation, transferring the neutrophil-rich suspension to a new tube each time. 10µl of the final neutrophil suspension was removed for cell counting using a disposable haemocytometer, sealed using 65µl Gene Frames (ThermoScientific #AB0577). The remaining cell suspension was centrifuged (300g, 6 min) to pellet neutrophils. Plasma supernatant was removed and discarded, and the pellet was washed by gentle resuspension in 10ml DPBS followed by further centrifugation (300g, 6 min).

4.3.4 PBMC isolation

10ml of venous blood drawn into EDTA-coated tubes as described above was diluted 1:1 with DPBS, then gently layered onto 15ml of Lymphoprep^{TM} solution (STEMCELL #07811) in a 50ml SepMate^{TM} column (STEMCELL #85450), and centrifuged for 20 min (1200*g*). Subsequently, the resulting layer of PBMCs was decanted and 10µl of the cell suspension was removed for cell quantitation as detailed above. The PBMC suspension was centrifuged at 300*g* for 8 mins, the cell pellet was then resuspended in 10 ml of DPBS as a washing step to remove contaminating free proteins, and centrifuged again (300*g*, 8 min).

4.3.5 Cell sample storage

The washed neutrophil or PBMC pellet was resuspended in DPBS to achieve a cell concentration of 5×10^6 cells/ml. 1ml of the suspension was transferred into an Eppendorf Protein LoBind tube and cells were pelleted (300 g, 5 min). The supernatant was discarded and cell pellet immediately frozen by inserting the tube into a pre-cooled metal block at -80°C. Pellets were stored at -80°C until lysis for liquid-chromatography mass spectrometry (LC-MS).

4.3.6 Neutrophil and PBMC isolation and sample preparation for LC-MS

Within two hours of venepuncture, neutrophils were isolated using the EasySep™ Direct Human Neutrophil Isolation Kit (STEMCELL Technologies #19666) utilising negative immunomagnetic selection and peripheral blood mononuclear cell (PBMCs) isolated by density-gradient separation using Lymphoprep^{TM} and SepMate^{TM} columns. Isolated cells were pelleted and stored at -80°C until analysis. To minimise batch effects of lysis buffers and conditions, stored cell pellets ($5\text{x}10^6$ cells) were lysed in batches of 50–100 samples.

4.3.7 Sample preparation for LC-MS

To minimise batch effects of lysis buffers and conditions, stored cell pellets ($5\text{x}10^6$ cells) were lysed in batches of 50–100 samples. For this, 400µl of freshly-prepared lysis buffer was added (5% sodium dodecyl sulphate (SDS), 10mM tris(2-carboxyethyl) phosphine hydrochloride (TCEP) and 50mM triethylammonium bicarbonate (TEAB; pH8.5) in dH2O) at RT. Samples were vortexed at 20,000 rpm for 10s, twice, at 5 min intervals. Samples were shaken at 500 rpm at room temperature for 5 minutes and then boiled at 98 C for

5 minutes. Samples were allowed to cool before sonicating for 30 cycles (30 seconds on and 30 seconds off). 1 µl of benzonase (Millipore) was added to each sample and samples incubated at 37 C for 15 minutes. Samples were alkylated with the addition of iodoacetamide to a final concentration of 20 mM and incubated for 1 hour in the dark. During the alkylation step, protein was quantified using the EZQ assay (Thermo Fisher). 100 µg of protein was loaded onto S-Trap mini columns(HaileMariam et al., 2018) (Protifi) and processed following the manufacturer's instructions. Digests were performed using 5 µg trypsin/sample at 47 C for 2 hours. Peptides were eluted from S-Trap columns firstly with 50mM ammonium bicarbonate, followed by 0.2% aqueous formic acid and lastly with 50% aqueous acetonitrile containing 0.2% formic acid. Eluted peptides were dried and suspended in 1 % formic acid and quantified using the CBQCA assay (Thermo Fisher).

4.3.8 Liquid chromatography mass spectrometry (LC-MS) analysis

1.5 µg of peptide for each sample was analysed on a Q-Exactive-HF-X (Thermo Scientific) mass spectrometer coupled with a Dionex Ultimate 3000 RS (Thermo Scientific) as described previously(Reyes et al., 2021). The following LC buffers were used: buffer A (0.1% formic acid in Milli-Q water (v/v)) and buffer B (80% acetonitrile and 0.1% formic acid in Milli-Q water (v/v)). 1 µg aliquot of each sample was loaded at 15 µL/min onto a trap column (100 µm × 2 cm, PepMap nanoViper C18 column, 5 µm, 100 Å, Thermo Scientific) equilibrated in 0.1% trifluoroacetic acid (TFA). The trap column was washed for 3 min at the same flow rate with 0.1% TFA then switched in-line with a Thermo Scientific, resolving C18 column (75 µm × 50 cm, PepMap RSLC C18 column, 2 µm, 100 Å). Peptides were eluted from the column at a constant flow rate of 300 nl/min with a linear gradient from 3% buffer B to 6% buffer B in 5 min, then from 6% buffer B to 35% buffer B in 115 min, and finally to 80% buffer B within 7 min. The column was then washed with 80% buffer B for 4 min and re-equilibrated in 3% buffer B for 15 min. Two blanks were run between each sample to reduce carry-over. The column was kept at a constant temperature of 50^{o} C at all times.

The data was acquired using an easy spray source operated in positive mode with spray voltage at 1.9 kV, the capillary temperature at 250° C and the funnel RF at 60° C. The MS was operated in DIA mode using parameters previously described (Muntel et al, 2019) with some modifications. A scan cycle comprised a full MS scan (m/z range from

350-1650, with a maximum ion injection time of 20 ms, a resolution of 120 000 and automatic gain control (AGC) value of 5 × 10^6). MS survey scan was followed by MS/MS DIA scan events using the following parameters: default charge state of 3, resolution 30.000, maximum ion injection time 55 ms, AGC 3 x 10^6, stepped normalized collision energy 25.5, 27 and 30, fixed first mass 200 m/z. The inclusion list (DIA windows) and windows widths are shown below. Data for both MS and MS/MS scans were acquired in profile mode. Mass accuracy was checked before the start of samples analysis.

4.3.9 DIA isolation windows

Window	**Window start (m/z)**	**Window width**	**Window overlap**	**Window**	**Window start (m/z)**	**Isolation window**	**Window Overlap**
1	349.975	66.8	0.525	**24**	663.25	14.5	0.5
2	416.25	13.5	0.5	**25**	677.25	13.5	0.5
3	429.25	11.5	0.5	**26**	690.25	13.5	0.5
4	440.25	12.5	0.5	**27**	703.25	14.5	0.5
5	452.25	11.5	0.5	**28**	717.25	16.5	0.5
6	463.25	11.5	0.5	**29**	733.25	15.5	0.5
7	474.25	11.5	0.5	**30**	748.25	16.5	0.5
8	485.25	10.5	0.5	**31**	764.25	18.5	0.5
9	495.25	11.5	0.5	**32**	782.25	17.5	0.5
10	506.25	11.5	0.5	**33**	799.25	18.5	0.5
11	517.25	11.5	0.5	**34**	817.25	19.5	0.5
12	528.25	10.5	0.5	**35**	836.25	20.5	0.5
13	538.25	11.5	0.5	**36**	856.25	20.5	0.5
14	549.25	10.5	0.5	**37**	876.25	22.5	0.5
15	559.25	11.5	0.5	**38**	898.25	24.5	0.5
16	570.25	10.5	0.5	**39**	922.25	26.5	0.5
17	580.25	11.5	0.5	**40**	948.25	28.5	0.5
18	591.25	12.5	0.5	**41**	976.25	31.5	0.5
19	603.25	12.5	0.5	**42**	1007.25	35.5	0.5
20	615.25	12.5	0.5	**43**	1042.25	41.5	0.5
21	627.25	11.5	0.5	**44**	1083.25	50.5	0.525
22	638.25	13.5	0.5	**45**	1133.225	516.8	
23	651.25	12.5	0.5				

4.3.10 Spectronaut 14 processing

The raw DIA data was processed using Spectronaut(Bruderer et al., 2015) 14. As the whole dataset was too large to analyse in a single cell batch, the data was processed in multiple batches. Firstly, a DIA library was generated from 106 DIA raw files using Pulsar, this library is included in the PRIDE submisison. Protein inference was performed using Trypsin/P as Digest Rule, and Specific as Digest Type. PSM, Peptide and Protein FDR were set to 0.01. The library consisted of 126,065 precursors, 70,531 Peptides and 5,901 Protein Groups.

The data was subsequently searched against this library with the parameters personalised as follows: the decoy method was set to 'Inverse', the Precursor and Protein q-value Cutoff was set to 0.01, major grouping was set to 'Protein Group Id' and minor grouping to 'Stripped Sequence'. The major group quantity was set to 'Sum peptide quantity' and the minor group quantity to 'Sum precursor quantity'. The Major Group Top N and Minor Group Top N were unselected. The quantity MS-Level was set to 'MS2' and the quantity type set to 'Area', data filtering was set to 'Qvalue'. Cross Run Normalization was disabled as was the imputation strategy. The data search was split into 6 batches of less than 75 raw files, and the results were then combine using the 'SNE combine' feature. The individual SNE files and the spectral library are included in the PRIDE(Perez-Riverol et al., 2018) submission under identifies PXD036082 and PXD036089.

4.3.11 Statistical methods

The differential expression analyses were performed in R (v. 4.0.3) and the global p-values and fold changes were calculated via the Bioconductor package Limma(Ritchie et al., 2015) (v 3.46.0). The estimated copy numbers were $\log_2$ converted before being fed to the linear model function lmFit() which used the method='robust' parameter. The linear model was the evaluated using the empirical Bayes statistics for differential expression function eBayes() with the parameter set to robust= TRUE. The q-value was calculated with the Bioconductor package qvalue (v 2.22.0). Differences in global protein levels were considered significant when the q-value <= 0.05. P-values for protein families and the PBMC proteins were calculated in R using Welches T-test, with differences considered significant when the p-value <= 0.05. Overrepresentation analyses were performed using webgestalt(Wang et al., 2017).

4.3.12 Protein copy numbers

Protein copy numbers were calculated in R based on the proteomic ruler(Wisniewski et al., 2014). The copy numbers for all patients are included in Supplemental Table 14.

4.3.13 Protein content

The estimated protein content was calculated based on the protein copy numbers and converted into picograms.

4.3.14 Protein filtering

Proteins identified by a single razor peptide were not included in the analysis. Furthermore, proteins were also filtered using the contaminant list provided by MaxQuant(Cox and Mann, 2008). For the differential expression analysis, proteins needed to be detected in 2 samples or more.

4.3.15 Data batches

Neutrophil samples collected in the study were lysed for proteomic analyses in 4 main batches, therefore statistical analyses were performed in the interests of avoiding bias from batched processing. Batches 1, 2 and 4 contain data from control, LRTI and COVID19 patients. Batch 3 contained data derived exclusively from COVID19 patients. Comparisons of day 1 COVID19 and LRTI patients to controls used data from batches 1,2 and 4. Comparisons done between different day 1 COVID19 patients stratified by WHO score were done with batches 1,2,3 and 4. All analysis done on longitudinal data (Day 7 and Day 29) compared to controls was done on samples from batch 4. Finally, the comparison of Day 1 vs Day 7 for COVID19 patients used batches 3 and 4.

4.3.16 Overrepresentation analysis

Overrepresentation analyses (ORA) were performed using webgestalt(Wang et al., 2017). The background for all analyses was defined as all the proteins that were detected within the dataset. Only proteins that were detected with a q-value<0.05 were used for the ORAs. The analyses included ‘Biological Process’ and ‘Cellular Component’ as the gene ontology functional databases, and 'CORUM' as the network database. A minimum number of 5 genes per category was selected and the significant level was changed from Top 10 to using FDR<0.05

4.3.17 Figure generation

The boxplots and raincloud plots were all generated in R using ggplot2 (v 3.3.5), ggdist (v. 3.1.1), ggbeeswarm (v 0.6.0). Sankey diagrams were generated using sankeymatic (https://sankeymatic.com/build/), while the schematic diagrams were generated using biorender (https://biorender.com/). The heatmaps for this manuscript were generated using Morpheus (https://software.broadinstitute.org/morpheus/). The input was based on the median copy numbers and for each protein the data was normalised based on the maximum copy number displayed, leading to a scale from 0 to 1.

4.3.18 Granule proteins

Proteins were classified into the different granule subsets based on the work previously described by multiple proteomics experiments(Hoogendijk et al., 2019, Rorvig et al., 2013). The full list of proteins is available at Supplemental Table 15.

4.3.19 MHC class II proteins

Boxplots showing the estimated copy numbers for all MHC-II proteins include are based on following proteins; HLA-DMB, HLA-DOA, HLA-DOB, HLA-DPA1, HLA-DRA, HLA-DRB1, HLA-DRB3 and HLA-DRB5.

4.3.20 Data availability

All the mass spectrometry files, as well as the processed search result files are available at PRIDE(Perez-Riverol et al., 2018, Perez-Riverol et al., 2022) under the identifiers PXD036082 and PXD036089.

4.4 Results

217 patients were enrolled between May 2020 and December 2020. The cohort consisted of 84 individuals with confirmed COVID-19, 91 with acute LRTI and 42 non-infected controls (Table 4.1). The COVID19 and control populations were well matched at the baseline (day 1), with slightly more females in the uninfected control population.

	SARS-CoV-2 positive	Non-COVID19 LRTI	Non-infected controls
Number of patients	84	91	42
Patient Age (mean-SD)	65.5 (50-80) 69 (53-78)	65.7 (49 - 82) 71 (57-78)	58.5 (41-76) 62 (43-75)
Number of males (%)	44 (52.4%)	45 (49.5%)	16(40.0%)
Patients with Chronic cardiac disease	36 (42.9%)	37 (40.7%)	14 (35.0%)
Patients with Chronic respiratory disease	24 (28.6%)	42 (46.2%)	11 (27.5%)
Patients with Diabetes	10 (11.9%)	13 (14.3%)	7 (17.5%)
Patients with BMI >30kg/m2	24 (28.6%)	29 (31.9%)	11 (27.5%)
Severity at enrolment			N/A
Hospitalized, not requiring oxygen	32 (38.1%)	51 (56.0%)	
Hospitalized requiring oxygen	33 (39.3%)	36 (39.6%)	
Requiring ventilatory support	19 (22.6%)	4 (4.4%)	
X-ray findings			N/A
No changes	19 (22.6%)	37 (40.7%)	
Unilateral pneumonia	13 (15.5%)	40 (44.0%)	
Bilateral pneumonia	50 (59.5%)	14 (15.4%)	
Not done	2 (2.4%)		
Time from symptom onset to hospitalization (days)	9.7 (12.4) 7 (4-11)	Not recorded	N/A
Length of hospital stay (days)	11.8 (10.3) 8 (4-17)	7.1 (6.1) 6 (1-8)	N/A
Blood neutrophil counts (mean-SD)	5.6 (2.8)	8.2 (3.7)	N/A
Blood eosinophil count, cells/ul (mean-SD)	40 (73)	126 (181)	N/A

Table 4.1 - Patient characteristics at enrolment

At baseline 32 (38.1%) patients with COVID19 did not require supplementary oxygen, 33 (39.3%) required supplementary oxygen but did not require ventilatory support and 19 (22.6%) required mechanical ventilation or non-invasive ventilatory support. 46 patients received 6mg dexamethasone daily (54.8%).

4.4.1 The protein landscape of neutrophils from hospitalised COVID19 patients

Peripheral blood neutrophil proteomes from patients with COVID19, non-COVID19 LRTI and control groups were analysed by mass spectrometry (Fig 4.1a). The data showed that over 5,800 unique proteins were identified with a median of 4,923 proteins/sample (Fig 4.1b) allowing both quantitative and qualitative comparisons of neutrophil proteomes. There were no significant differences in the total protein content of neutrophils from control, LRTI or COVID19 patients (Fig. 4.1c) but there were differences in their protein landscape composition. There were 300 proteins significantly increased and 123 proteins significantly decreased in abundance in the neutrophil proteomes of LTRI patients compared to controls (Fig. 4.1d; Supplemental Table 4.1). These changes include decreases in abundance of CD10 (MME), a marker of immature neutrophils, and components of azurophilic granules including elastase and myeloperoxidase. Proteins increased in abundance control vesicular trafficking and mRNA processing (Supplemental Table 4.2). When the proteomes of neutrophils from COVID19 patients were compared to controls, changes in expression of 1,748 proteins were detected (Fig. 4.1e; Supplemental Table 4.3). There was increased expression of 1,008 proteins including a striking signature of interferon-regulated proteins as well as changes in proteins controlling metabolic processes including glycolysis and fatty acid metabolism. Proteins decreased in expression in neutrophils from COVID19 patients included CD10, components of endosomal sorting complexes and key enzymes controlling glycogenolysis.

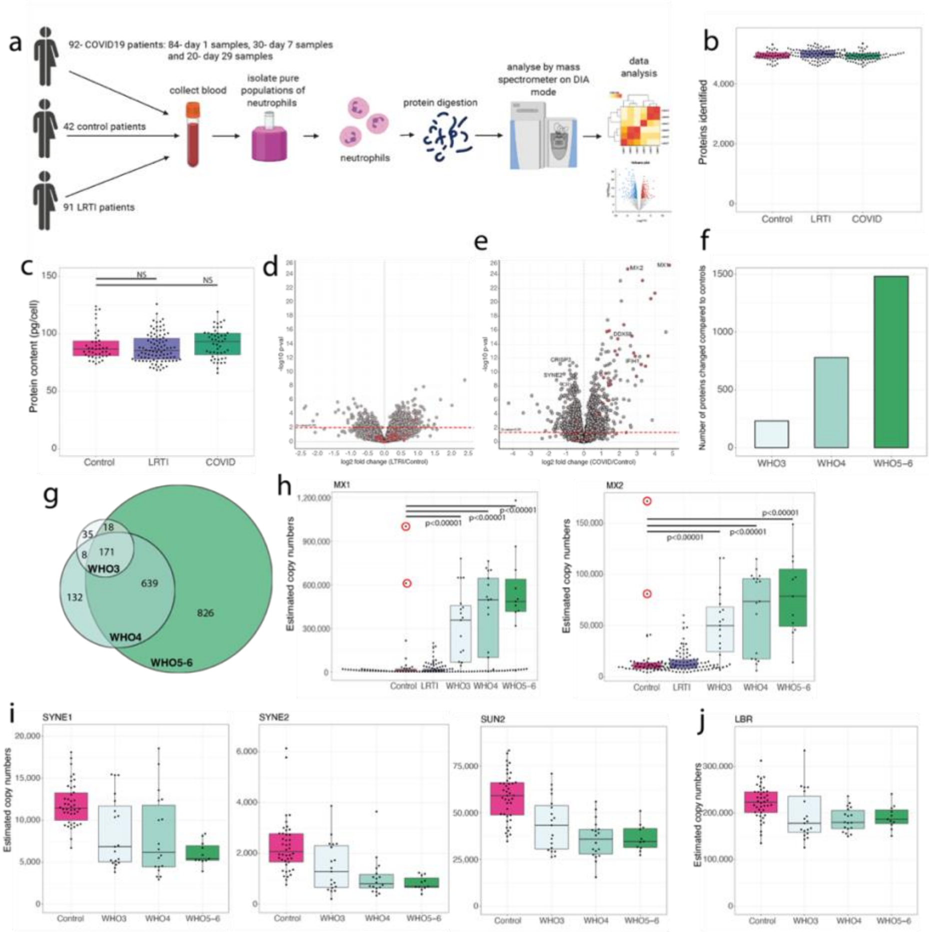

Figure 4.1- Core COVID19 neutrophil proteomic signature: ***(a)*** Schematic showing the sample collection, neutrophil isolation and mass spectrometry workflow ***(b)*** Box plot showing the number of proteins identified across all samples for Control, LRTI and COVID19. ***(c)*** Box plot showing the estimated protein content for all samples for Control, LRTI and COVID19. ***(d)*** Volcano plot showing the fold change and p-value comparing the neutrophil proteomes of LRTI patients to the controls. IFN-induced proteins are coloured in red. The red dotted line represents Q-value=0.05. ***(e)*** Volcano plot showing the fold change and p-value comparing the neutrophil proteomes of COVID19 patients to the controls. IFN induced proteins are coloured in red. The red dotted line represents Q-value=0.05. ***(f)*** Bar plot showing the number of proteins that are significantly changed in abundance when comparing the neutrophil proteomes of WHO3 (moderate), WHO4 (severe) and WHO5-6 (critically severe) COVID19 patients to the controls. ***(g)*** Venn diagram showing the overlap of significantly altered proteins across the stratified COVID19 patient cohorts. ***(h)*** Bos plots showing the estimated protein copy numbers for MX1 and MX2 across control, LRTI, WHO3, WHO4 and WHO5-6 COVID19 patients. Patients circled in red were pre-symptomatic and later tested SARS-CoV-2 positive. Box plots showing the estimated protein copy numbers for ***(i)*** SYNE1, SYNE2 and SUN2 and ***(j)*** laminin B receptor (LBR) across Control, WHO3, WHO4 and WHO5-6 patients. For all boxplots the whiskers extend from the hinge to the largest and smaller values no further than 1.5 x interquartile range.

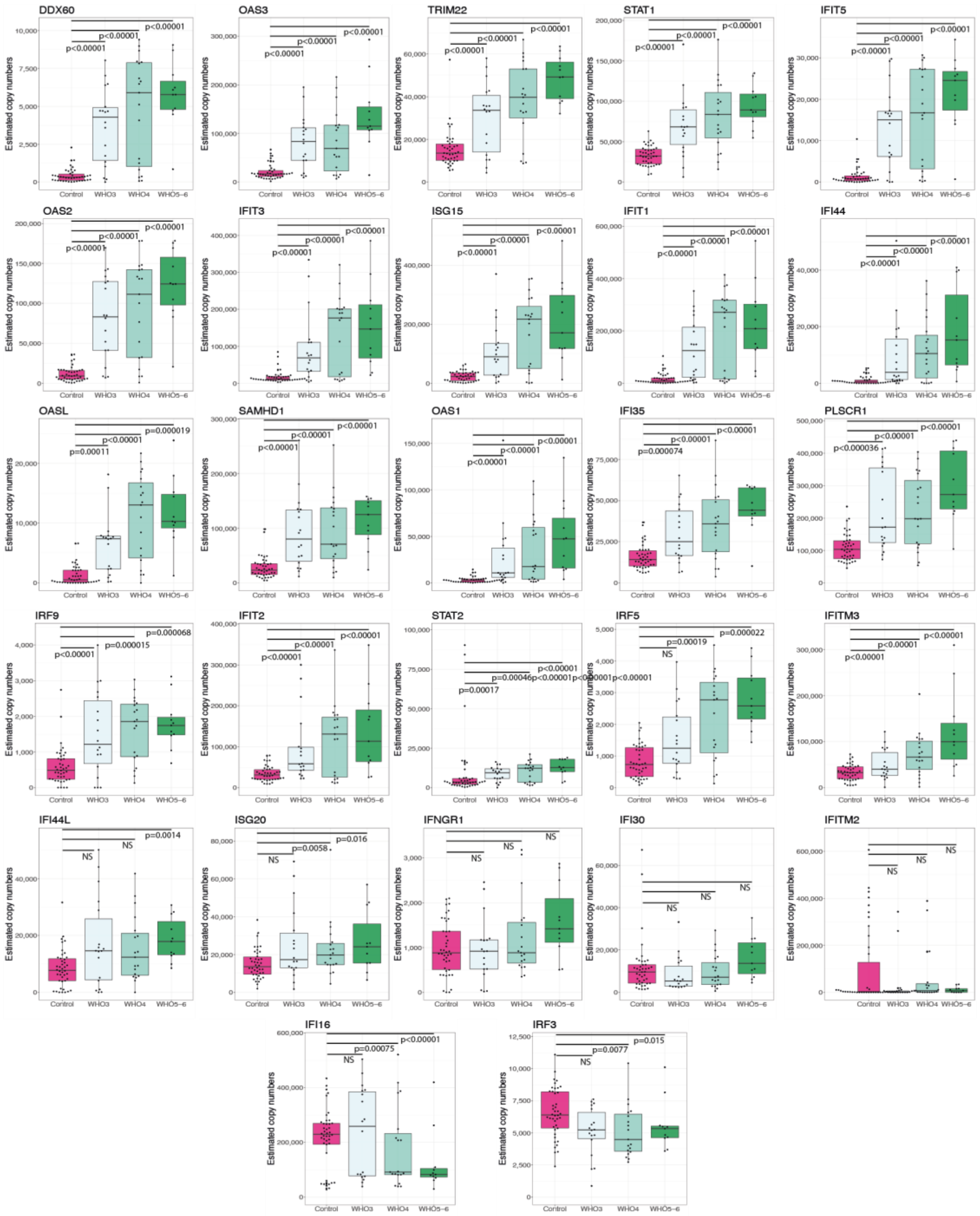

Figure 4.2: Box plots showing the estimated copy numbers for the main type I interferon response proteins at day 1 in the neutrophil proteomes across control participants and WHO3 (moderate), WHO4 (severe) and WHO5-6 (critically ill) COVID19 patients. For all boxplots the whiskers extend from the hinge to the largest and smaller values no further than 1.5 x interquartile range.

In a complementary analysis, the COVID19 patients were stratified based on the World Health Organisation (WHO) severity scale, reflecting patient severity upon hospital admission. WHO3 patients were considered of moderate severity, as they were hospitalised but did not require supplemental oxygen. WHO4 patients were considered severe patients, they were hospitalised and required supplemental oxygen. WHO5-6 patients were considered critically severe, they were hospitalised and required mechanical ventilation. We compared the neutrophil proteomes of the stratified patient groups to the control proteomes and found that the number of proteins significantly changed in abundance increased with disease severity: 221 proteins were significantly changed in neutrophils from WHO3 patients (Supplemental Table 4.4), 779 in WHO4 (Supplemental Table 4.5) and 1,483 in WHO5-6 patient neutrophils (Fig. 4.1f; Supplemental Table 4.6).

The data also revealed a core signature of 171 proteins that were significantly changed across the neutrophil proteomes of all 3 stratified COVID19 patient groups (Fig. 4.1g; Supplemental Table 4.7). This signature included 101 proteins that were significantly increased in abundance in all patient groups, and > 50% of these represented interferon (IFN) induced proteins, suggestive of an IFN response present in the majority of neutrophils from WHO3, WHO4 and WHO5-6 patients (Fig. 4.1h; Fig. 4.2). Furthermore, this IFN response clearly distinguished the proteomes of COVID19 patients from those of the LTRI and control cohorts, even identifying two pre-symptomatic COVID19 patients who were originally enrolled in the control group (Fig. 4.1h). The core signature also included 70 proteins that were decreased in abundance across all stratified COVID19 patient cohorts. Prominently among these were proteins linked to regulation of the structure and rigidity of the nucleus, e.g., SYNE1, SYNE2 and SUN2 (Fig. 4.1i) and the laminin b receptor (LBR; Fig. 4.1j).

4.4.2 Neutrophil phenotypes linked to COVID19 disease severity

We next studied potential changes in the neutrophil proteomes of COVID19 patients that associated with disease severity. One well established consequence of COVID19 is emergency myelopoiesis and the release of immature neutrophils into circulation(Reusch et al., 2021), particularly in cases of severe COVID19. In this respect, the proteomic data reflected a similar signature. It highlighted reduced abundance of CD10 (Fig. 4.3a) and cytosolic components of the NADPH oxidase complex (Fig. 4.3b), which are known to have higher expression in fully mature neutrophils. It also displayed increased expression of the proliferation marker PCNA in neutrophils from WHO4 and WHO5-6 patients (Fig. 4.3c), which is characteristic of immature neutrophils(Hsu et al., 2019, Hoogendijk et al., 2019). Immature neutrophils are also known for their reduced density, thus can be found in the PBMC layer upon density-gradient separation(Denny et al., 2010, Pavon et al., 2012).

We used mass spectrometry to analyse PBMC proteomes derived from these patients. Many previous studies have shown PBMC abnormalities in patients with COVID19 disease(Stephenson et al., 2021, Schulte-Schrepping et al., 2020, Diao et al., 2020, Xie et al., 2021, Shaath and Alajez, 2021, Arunachalam et al., 2020, Kuri-Cervantes et al., 2020, Laing et al., 2020) . Accordingly, the PBMC proteomic data showed a skewed composition with evidence for decreased levels of T cells in WHO5-6 patients and increases in MZB1^{+} cells in all COVID19 patients (Fig. 4.4), but more relevantly it also showed increased expression of key neutrophil proteins such as neutrophil elastase in the PBMCs of WHO5-6 patients (Fig. 4.3d). The increased presence of more low-density neutrophils is consistent with an increased frequency of immature neutrophils.

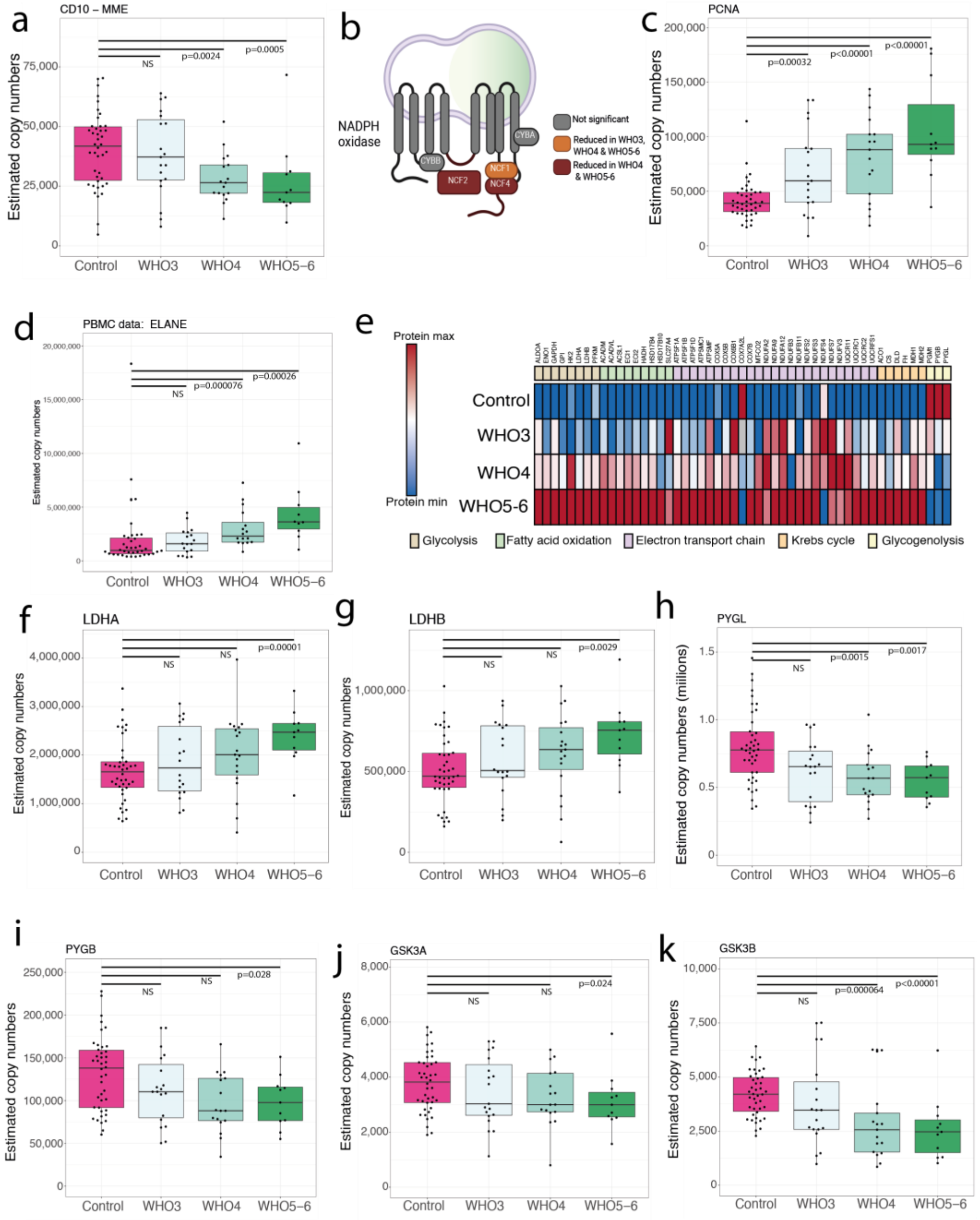

Figure 4.3 – Immature neutrophils and metabolic changes in COVID19: ***(a)*** Box plot showing the estimated protein copy numbers for CD10 (MME) across control, WHO3 (moderate), WHO4 (severe) and WHO5-6 (critically severe) COVID19 patients. ***(b)*** Schematic showing the NAPDH oxidase complex and the protein subunits that are significantly changed in abundance compared to the control group. ***(c)*** Box plot showing the estimated protein copy numbers for PCNA across Control, WHO3, WHO4 and WHO5-6 patients. ***(d)*** Box plot showing the estimated protein copy numbers for ELANE in the PBMC proteomes across Control, WHO3, WHO4 and WHO5-6 patients. ***(e)*** Heatmap showing the metabolic proteins that were significantly changed in abundance in WHO5-6 COVID19 patients. The data are normalised to the maximum value of each protein. Box plot showing the estimated protein copy numbers for ***(f)*** LDHA, ***(g)*** LDHB, ***(h)*** PYGL, ***(i)*** PYGB, ***(j)*** GSK3A, ***(k)*** GSK3B in the neutrophil proteomes across control, WHO3, WHO4 and WHO5-6 patients. For all boxplots the whiskers extend from the hinge to the largest and smaller values no further than 1.5 x interquartile range.

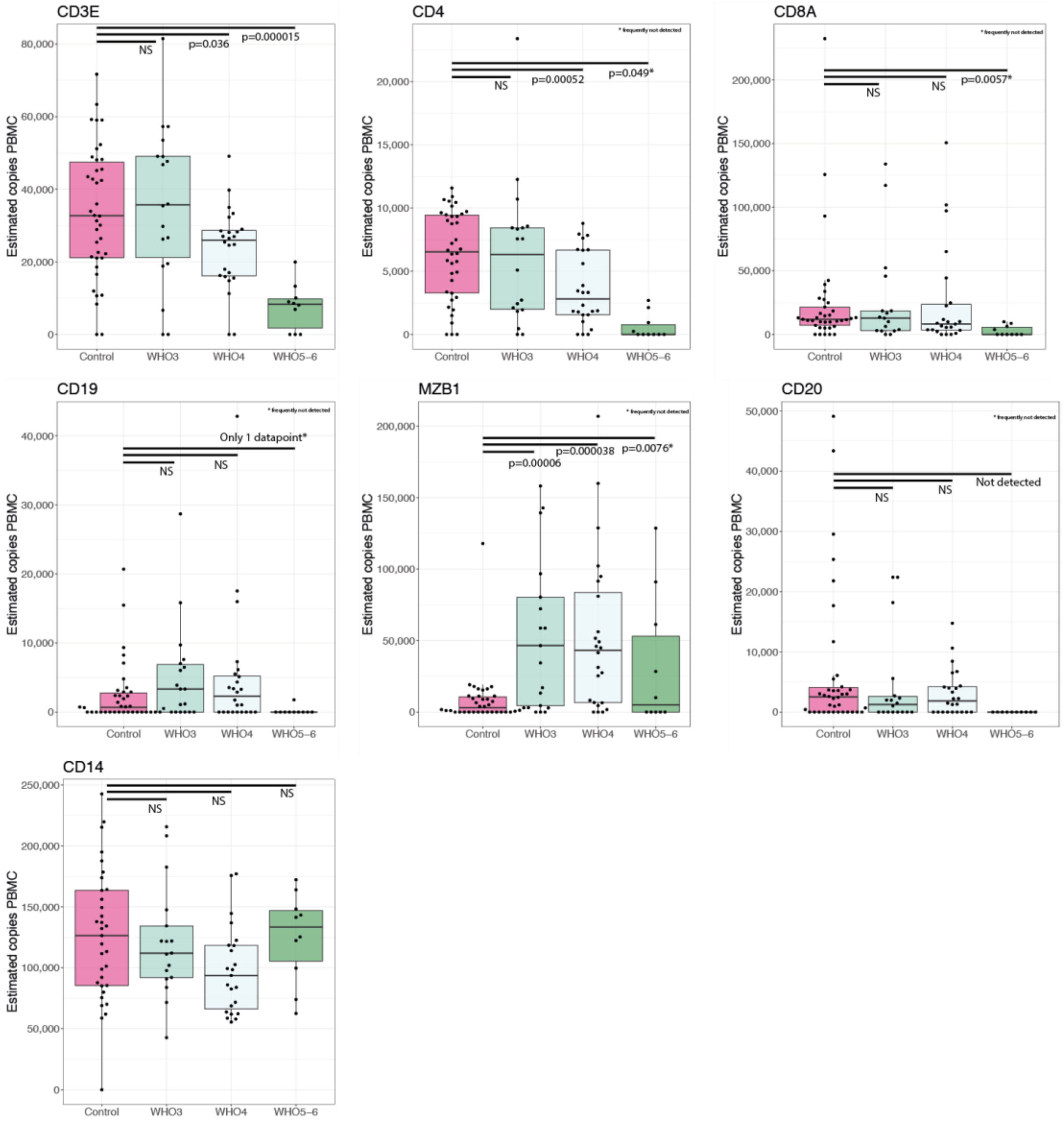

Figure 4.4: Box plots showing the estimated protein copy numbers for the key cell type markers in the PBMCs proteomes at day 1 across Control, WHO3 (moderate), WHO4 (severe) and WHO5-6 (critically ill) patients. For all boxplots the whiskers extend from the hinge to the largest and smaller values no further than 1.5 x interquartile range.

We also observed changes in the metabolic profile of neutrophils from WHO5-6 patients, some linked to neutrophil maturation stage and some to hypoxia(Watts et al., 2021).

The neutrophil proteomes from WHO5-6 patients had increased expression of enzymes involved in fatty acid oxidation, electron transport chain and the TCA (Fig. 4.3e), all shown to be more highly expressed in immature neutrophils(Riffelmacher et al., 2017, Injarabian et al., 2019). However, neutrophil proteomes from WHO5-6 patients also displayed significantly higher abundance of vital glycolytic enzymes that convert pyruvate into lactate like lactate dehydrogenase A (LDHA; Fig. 4.3f)) and B (LDHB; Fig. 4.3g), which have been reported to be increased in abundance with hypoxia(McGettrick and O'Neill, 2020). This suggests that not all metabolic changes are related to neutrophil maturity. The metabolic profiling also highlighted changes related to glycogen synthesis and glycogen breakdown, where rate limiting enzymes of glycogenolysis(Zhang et al., 2012), glycogen phosphorylases PYGL (Fig. 4.2h) and PYGB (Fig. 4.3i) and the negative regulators of glycogen synthesis the glycogen synthase kinases(Embi et al., 1980) (Fig. 4.3j-k) were significantly decreased in abundance WHO5-6 patients compared to controls.

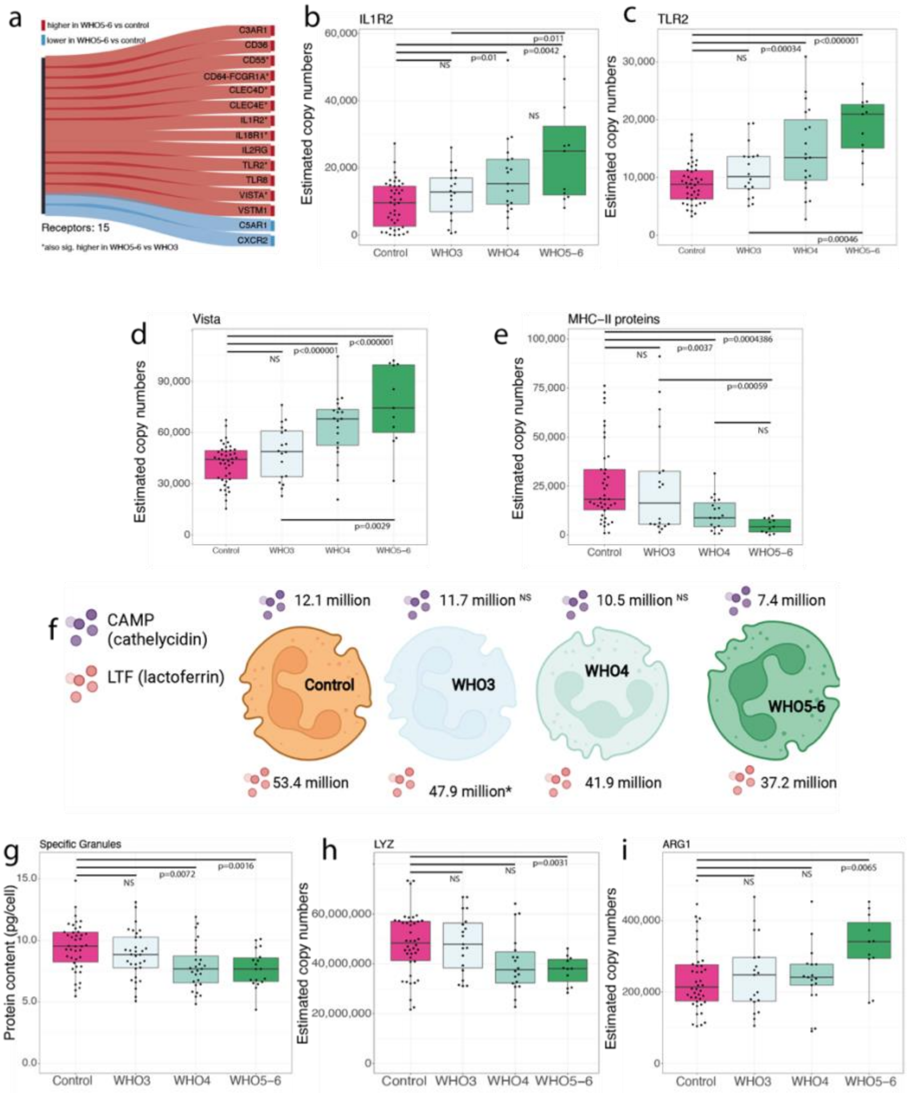

Figure 4.5 – Markers of severity in the neutrophil proteomes of COVID19 patients: ***(a)*** Sankey diagram showing the immunomodulatory receptors that are significantly changed in abundance in the neutrophil proteomes WHO5-6 compared to controls. Proteins in red are significantly increased in abundance in the neutrophil proteomes of WHO5-6, proteins in blue were significantly decreased in abundance. Box plots showing the estimated protein copy numbers for ***(b)*** IL1R2, ***(c)*** TLR2, ***(d)*** Vista (VSIR) ***(e)*** all MHC-II class proteins across control, WHO3 (moderate), WHO4 (severe) and WHO5-6 (critically severe) COVID19 patients. ***(f)*** Schematic showing the estimated median copy numbers of LTF and CAMP across control, WHO3, WHO4 and WHO5-6 patients. ***(g)*** Box plot showing the estimated protein content for proteins primarily contained within Specific Granules across control, WHO3, WHO4 and WHO5-6 patients. Box plots showing the estimated protein copy numbers for ***(h)*** LYZ and ***(i)*** ARG1 across control, WHO3, WHO4 and WHO5-6 COVID19 patients. For all boxplots the whiskers extend from the hinge to the largest and smaller values no further than 1.5 x interquartile range.

Furthermore the neutrophils from COVID19 patients also displayed changes in surface receptors (Fig. 4.5a), with decreased expression of migratory receptors C5AR1 and CXCR2, and increased abundance of CD64 (FCGR1A), a known neutrophil activation marker(Hoffmann, 2009, Fjaertoft et al., 1999) across all patient groups (Fig. 4.5a). Some of the changes in these receptors associated with disease severity. With increased abundance in the neutrophil proteomes of WHO5-6 patients of IL1R2, a decoy receptor for IL1 (Fig. 4.5b), the pattern recognition receptor TLR2 (Fig. 4.5c) and V-domain Ig suppressor of T cell activation (VISTA, VSIR) (Fig 4.5d), a negative immune checkpoint regulator. The data also highlighted reduced abundance of MHC-class II proteins that also associated with disease severity (Fig. 4.5e).

Neutrophils derived from WHO5-6 patients also had differences in granule protein composition. Notably they displayed a reduction in LTF (lactoferrin) and CAMP (cathelicidin; Fig. 4.3f), two of the three most abundant specific (secondary) granule proteins. This translated into a significant reduction in the protein content of specific granules in WHO4 and WHO5-6 patients compared to controls (Fig. 4.5g). Granule production is a hierarchical process linked to neutrophil maturity. The first granules to be produced are azurophilic granules, followed by specific and tertiary granules and lastly secretory vesicles (SV). Our data showed no significant changes in the protein content of azurophilic granules, ficolin granules, gelatinase granules or secretory vesicles (Fig. 4.6), suggesting no direct link to neutrophil maturity, but suggesting a potential link to neutrophil degranulation. We did however see altered expression of some azurophilic granule proteins linked with immunomodulation namely a reduction in LYZ (lysozyme) (Fig. 4.5h) and an increase in ARG1 (arginase 1) (Fig. 4.5i). Other immunosuppressive proteins like TGFB1 were also increased in abundance. These data highlight how SARS-CoV-2 infection results in neutrophils with distinct potential immunoregulatory capacities and distinct capacities to respond to cues from other immune cells, pathogens or cytokines.

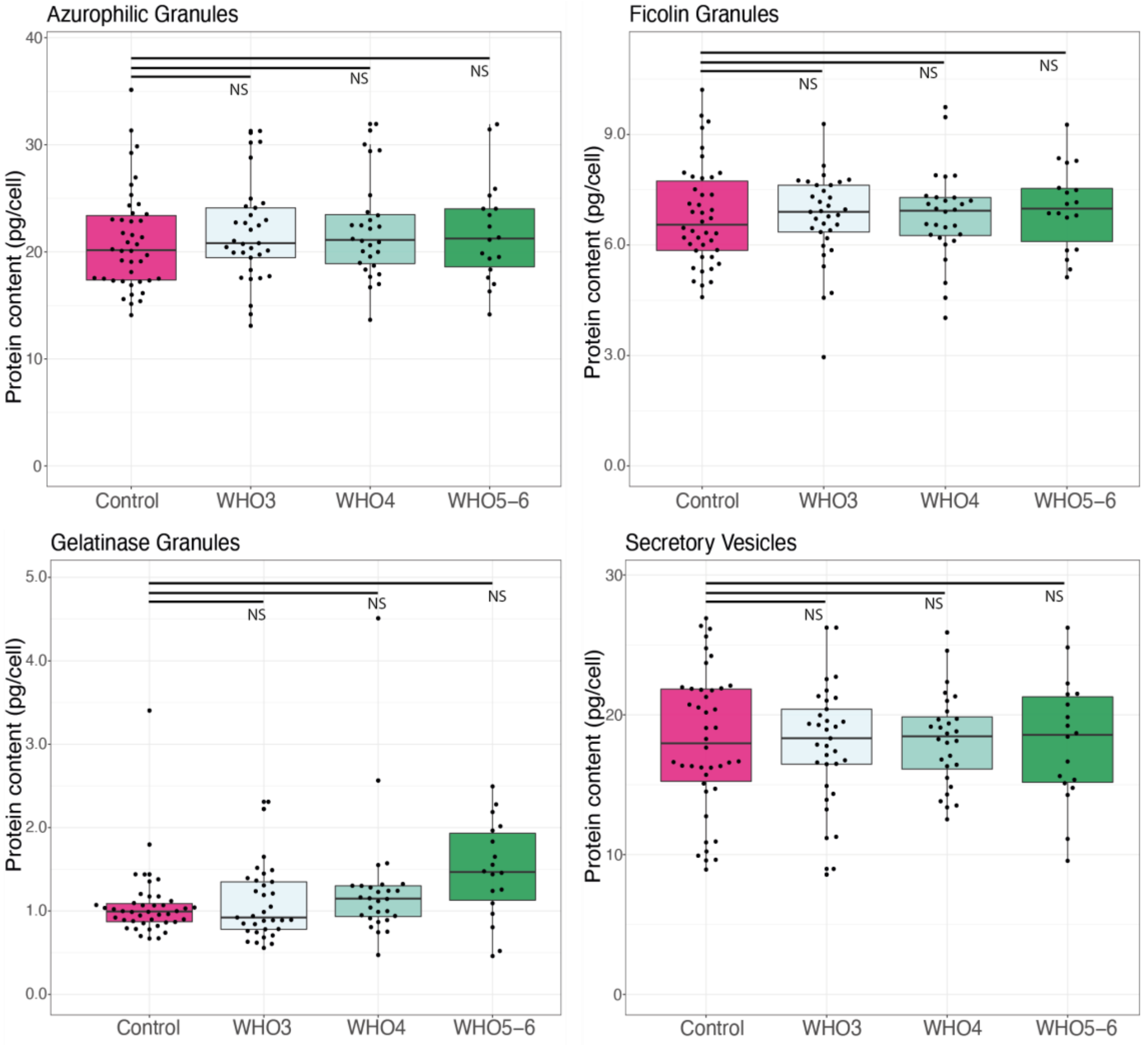

Figure 4.6: Box plots showing the estimated protein content of Azurophilic (primary), ficolin and gelatinase (tertiary) granules and secretory vesicles at day 1 in neutrophil proteomes across control, WHO3 (moderate), WHO4 (severe) and WHO5-6 (critically ill) patients. For all boxplots the whiskers extend from the hinge to the largest and smaller values no further than 1.5 x interquartile range.

4.4.3 Transient interferon and altered metabolic signatures in neutrophils from SARS-CoV-2-infected individuals

To gain insight into changes in neutrophil proteomes related to the temporal disease dynamics, we examined neutrophils from COVID19 patients 7 days after recruitment into the study. At this timepoint there were still major differences between the neutrophil proteomes from COVID19 patients versus controls, with 2,081 proteins significantly changed in abundance (Supplemental Table 4.8). Using the same patient stratification strategy as described for the day 1 data, the magnitude of the changes in protein signatures in the day 7 samples also associate with disease severity: 239 proteins changed in WHO3 (Supplemental Table 4.9) patients compared to controls, 1,373 in WHO4 (Supplemental Table 4.10) and 828 proteins changed in WHO5-6 patients (Fig. 4.7a, Supplemental Table 4.11).

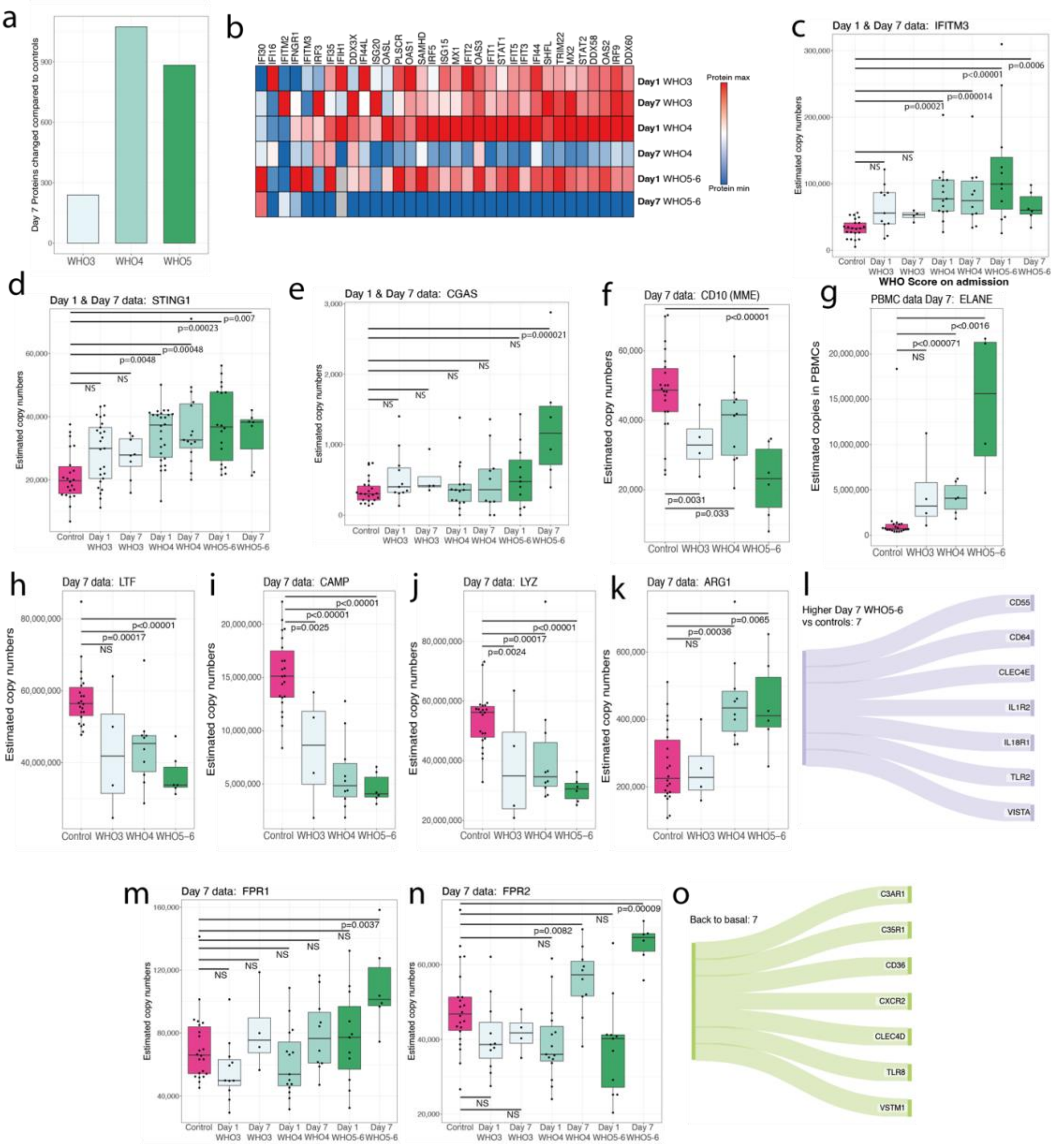

Figure 4.7 – Transient and prolonged effects in the day 7 proteome: ***(a)*** Bar plot showing the number of proteins significantly changed at day 7 in the WHO3 (moderate), WHO4 (severe) and WHO5-6 (critically severe) proteomes compared to controls. ***(b)*** Heatmap showing the main type I Interferon induced proteins across WHO3, WHO4 WHO5-6 patients for both day 1 and day 7. The data are normalised to the maximum value of each protein. Box plots showing the estimated protein copy numbers for ***(c)*** IFITM3, ***(d)*** STING1 and ***(e)*** CGAS across controls and WHO3, WHO4 and WHO5-6 patients in both day 1 and day 7. ***(f)*** Box plot showing the estimated protein copy numbers for CD10 across Control, WHO3, WHO4 and WHO5-6 patients at day 7. ***(g)*** Box plot showing the estimated protein copy numbers in the PBMC proteomes for ELANE across Control, WHO3, WHO4 and WHO5-6 patients at day 7. Box plots showing the estimated protein copy numbers for ***(h)*** LTF, ***(i)*** CAMP, ***(j)*** LYZ and ***(k)*** ARG1 across controls, WHO3, WHO4 and WHO5-6 patients at day 7. ***(l)*** Sankey diagrams showing immunomodulatory receptors which are significantly higher in the WHO5-6 patients at day 7 compared to controls. Box plots showing the Estimated protein copy numbers for ***(m)*** FPR1 and ***(n)*** FPR2 across controls and WHO3, WHO4 and WHO5-6 patients in both day 1 and day 7. **(o)** Sankey diagrams showing immunomodulatory receptors which

returned to basal levels (i.e. control levels) in the WHO5-6 patients at day 7 compared to controls. For all boxplots the whiskers extend from the hinge to the largest and smaller values no further than 1.5 x interquartile range.

We explored the abundance of important neutrophil responses to COVID19, like the IFN response, to explore if there were any changes across the two timepoints. A robust IFN-I response was still evident in WHO3 patients at day 7, showing no difference in the magnitude of the interferon signature compared to day 1 (Fig 4.7b). In contrast, the proteomes of WHO4 and WHO5-6 patients showed a prominent reduction of the IFN-I signature at day 7 (Fig. 4.7b), with WHO5-6 showing almost no significant differences compared to controls, except for IFITM3 (Fig. 4.7c), IFNGR1, STING (Fig. 4.4d) and cGAS (Fig. 4.7e) which remained elevated at day 7 compared with controls.

In contrast to the reduced interferon signature, the neutrophils and PBMCs from the WHO4 and WHO5-6 patients still showed hallmarks of immature/low-density neutrophils (Fig. 4.7f&g), and their neutrophils demonstrated persistent differences in granule proteins, with LTF, CAMP and LYZ still significantly reduced and the immunosuppressive proteins ARG1 and TGFB1 significantly increased (Fig. 4.7h-k). Furthermore, multiple signalling receptors that had been identified as markers of disease severity at day 1 remained elevated in the day 7 analyses such as TLR2, VISTA, IL1R2 and IL18R1 (Fig. 4.7l), suggesting therapeutic value across multiple stages of disease progression.

The day 7 data also showed some very prominent changes that were not seen at day 1, such as the increased abundance of the migratory and activating formyl peptide receptors FPR1 (Fig. 4.7m) and FPR2 (Fig. 4.7n). Moreover, half of the receptors that were increased at day 1 were no longer different to the control neutrophils (Fig. 4.7o), and this included other vital migratory receptors like C5AR1 and CXCR2, suggesting the neutrophils at day 7 maintain an immunosuppressive capacity but increase their potential to migrate from the blood into the tissues.

4.4.4 Neutrophil proteomic signatures of recovered versus non recovered COVID19 patients

To explore longer term effects of SARS-CoV-2 infection on neutrophils we then focussed on the proteomic analysis of neutrophils from COVID-19 patients 29 days after enrolment. For these analyses, a new patient stratification strategy was used, this was based on patient outcome at day. The neutrophil proteomes of control patients were compared to patients who were at home and fully recovered (R-WHO1) and to patients at home with persistent symptoms and limitations or still hospitalized (R-WHO2-3). Neutrophils from R-WHO2-3 patients showed significant changes in 1,111 proteins compared to neutrophils from control population (Supplemental table 4.12), which is almost three times the number of proteins significantly changed in R-WHO1 patients versus controls (Supplemental table 4.13, Fig. 4.8a). There were 268 proteins that were changed in neutrophils from both R-WHO1 and R-WHO2-3 patients, when compared to controls (Fig. 4.8b). These included changes in prominent receptors and adaptor proteins, such as increased abundance of the activating and pro-survival CSF receptor CSF2RA (Fig. 4.8c), reduced abundance of main immunoreceptor tyrosine-based activation motif (ITAM) adaptor FcRγ (Fig. 4.8d), reduced abundance of the glycogen phosphorylase PYGL (Fig. 4.8e) and the migratory receptor LTB4R (Fig. 4.8f). In particular, the reduced abundance of FcRγ suggests a potential deficiency in responding to immunoglobulin-opsonized pathogens.

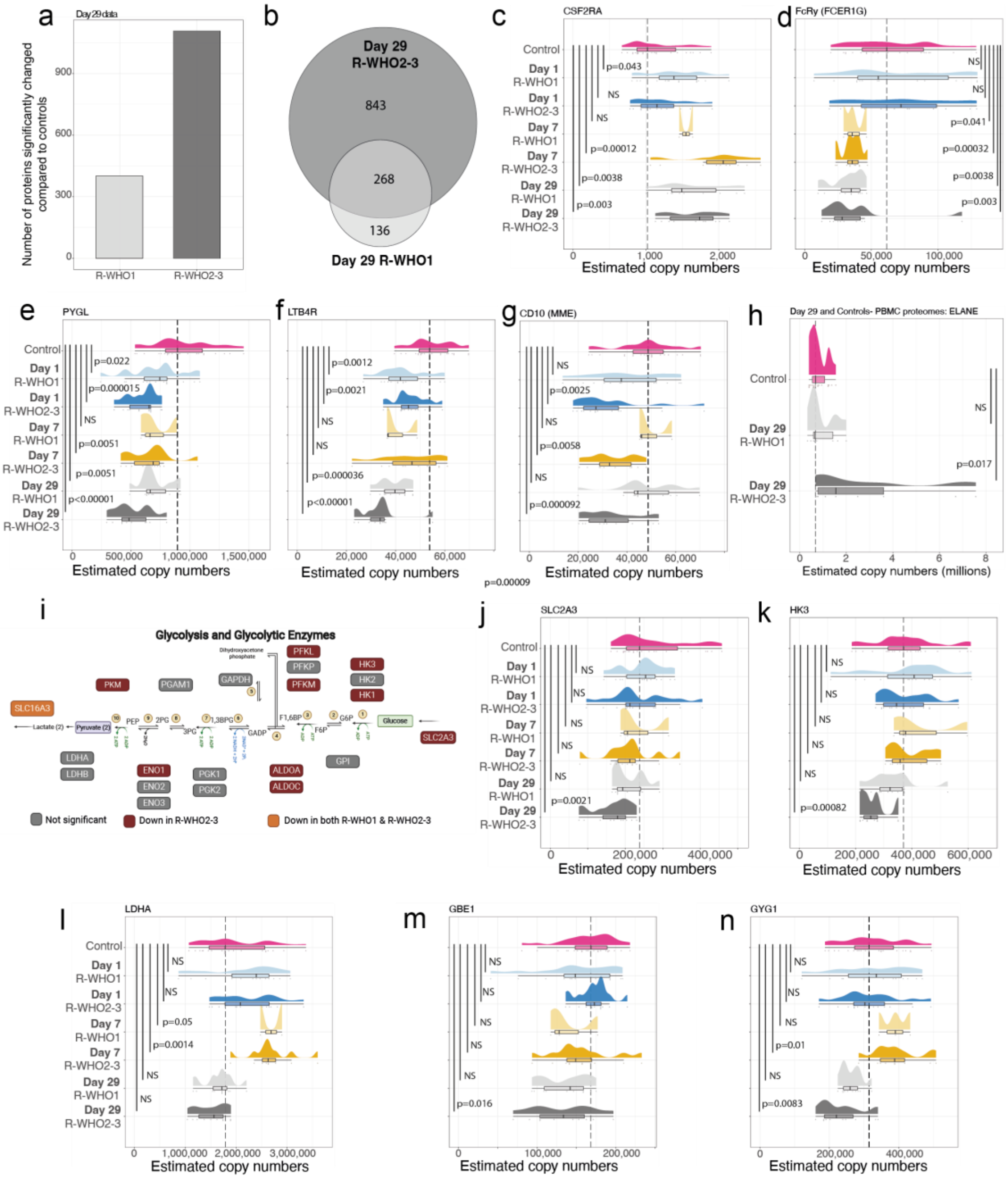

Figure 4.8 – Metabolic changes in the neutrophil proteomes at day 29: ***(a)*** Bar plot showing the number of proteins significantly changed in the neutrophil proteomes of R-WHO1 (recovered) or R-WHO2-3 (not recovered) patients at day 29 compared to controls. ***(b)*** Venn diagram showing the overlap of significantly changed proteins in the R-WHO1 and R-WHO2-3 day 29 proteomes. Rain cloud plots showing the estimated protein copy numbers for ***(c)*** CSF2RA, ***(d)*** FcRy (FCER1G), ***(e)*** PYGL, ***(f)*** LTB4R and ***(g)*** CD10 across controls, Day 1, Day 7 and Day 29 COVID19 patients stratified in R-WHO1 and R-WHO2-3 ***(h)*** Rain cloud plot showing the estimated protein copy numbers in the PBMC proteomes for ELANE across controls as well as Day 29 stratified in R-WHO1 and R-WHO2-3. ***(i)*** Schematic showing the glycolytic pathway highlighting proteins that were significantly reduced in abundance in the R-WHO2-3 patients. Rain cloud plot showing the estimated protein copy numbers for ***(j)*** SLC2A3, ***(k)*** HK3, ***(l)*** LDHA, ***(m)*** GBE1 and ***(n)*** GYG1 across controls, Day 1, Day 7 and Day 29 COVID19 patients stratified into R-WHO1 and R-WHO2-3. All Raincloud plots include density plots as well as boxplots. For all boxplots the whiskers extend from the hinge to the largest and smaller values no further than 1.5 x interquartile range.

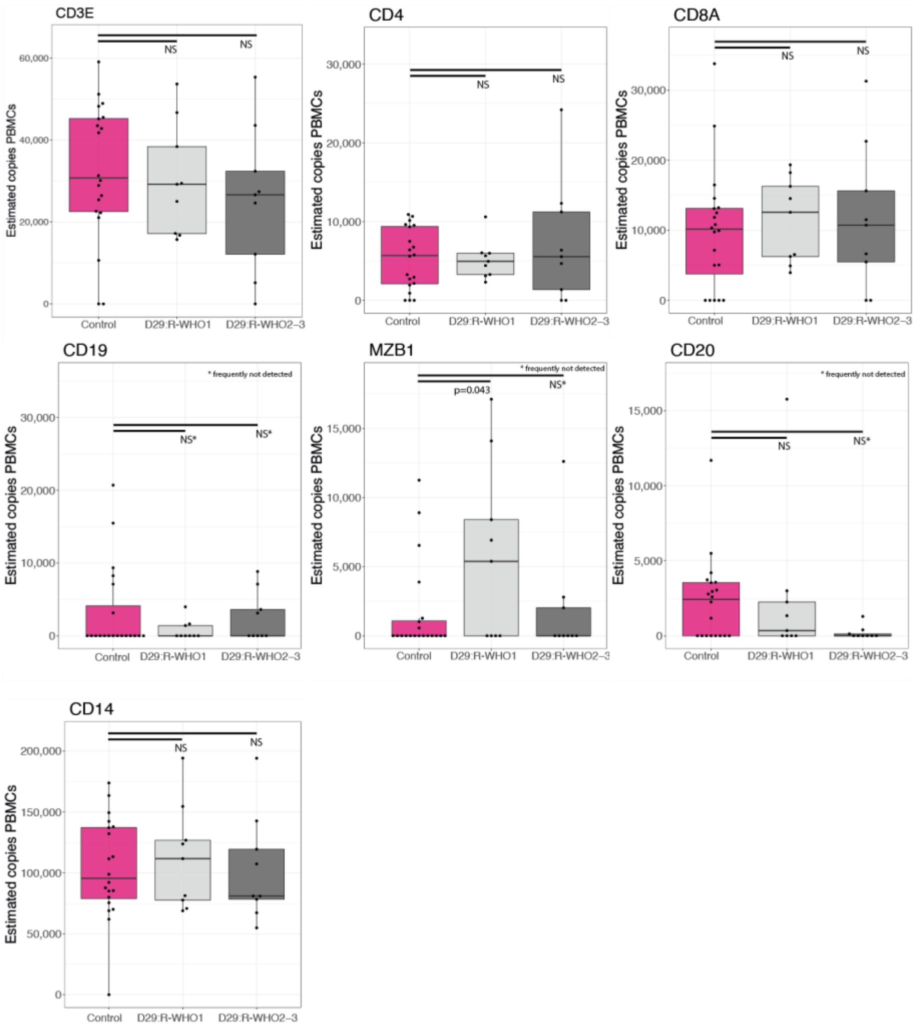

Figure 4.9: Box plots showing the estimated protein copy numbers for key cell type markers derived from PBMC proteomes across control participants, Day 29 R-WHO1 (recovered) and R-WHO2-3 (not recovered) COVID19 patients. For all boxplots the whiskers extend from the hinge to the largest and smaller values no further than 1.5 x interquartile range.

The data also revealed 843 proteins that were only changed in R-WHO2-3 neutrophils; these included the previously mentioned proteomic markers for immature neutrophils (Fig. 4.8g&h), with data from both the PBMC and neutrophil proteomes suggesting the presence of immature neutrophils at day 29 was linked to delayed recovery. The PBMC proteome no longer indicated alterations in B cells in R-WHO2-3 patients, nor persisting in T cells as described in the literature(Shuwa et al., 2021, Wilson et al., 2022, Mann et al., 2020) (Fig. 4.9). Other characteristics of neutrophils from R-WHO2-3 patients related to changes in key metabolic proteins that control glycolysis, glycogenolysis and glycogen synthesis.

Neutrophils have been shown to have a dependence on two metabolic pathways, glycolysis when nutrients are abundant, and glycogenolysis in nutrient deprived conditions(Sadiku et al., 2021). R-WHO2-3 patients displayed reduced abundance of rate limiting regulators of glycolysis like SLC2A3 (GLUT3), Hexokinase 3 and the lactate transporter SLC16A3 (Fig. 4.8i). Strikingly, the changes to glycolytic proteins were different to those seen in neutrophils at day1 and 7. For example, day1 and day7 neutrophils did not show decreased expression of GLUT3 (Fig. 4.8j) or Hexokinase 3 (Fig. 4.8k). Additionally, glycolytic enzymes that were significantly increased in abundance in day 7 neutrophils, were now unchanged compared to the control proteomes (e.g. lactate dehydrogenase) (Fig. 4.8l). One other difference in the metabolic proteome in the day29 neutrophils was reduced expression in R-WHO2-3 patients of enzymes that control glycogen synthesis such as the glycogen branching enzyme (GBE1; Fig. 4.8m) and glycogenin 1 (GYG1; Fig. 4.8n). The data suggest that the neutrophils derived from R-WHO2-3 patients display a metabolic proteome with reduced glycolysis and glycogenolysis, similar to what has previously been described in neutrophils from individuals with COPD, which associated with reduced energy production leading to impaired neutrophil survival and antimicrobial capacity(Sadiku et al., 2021).

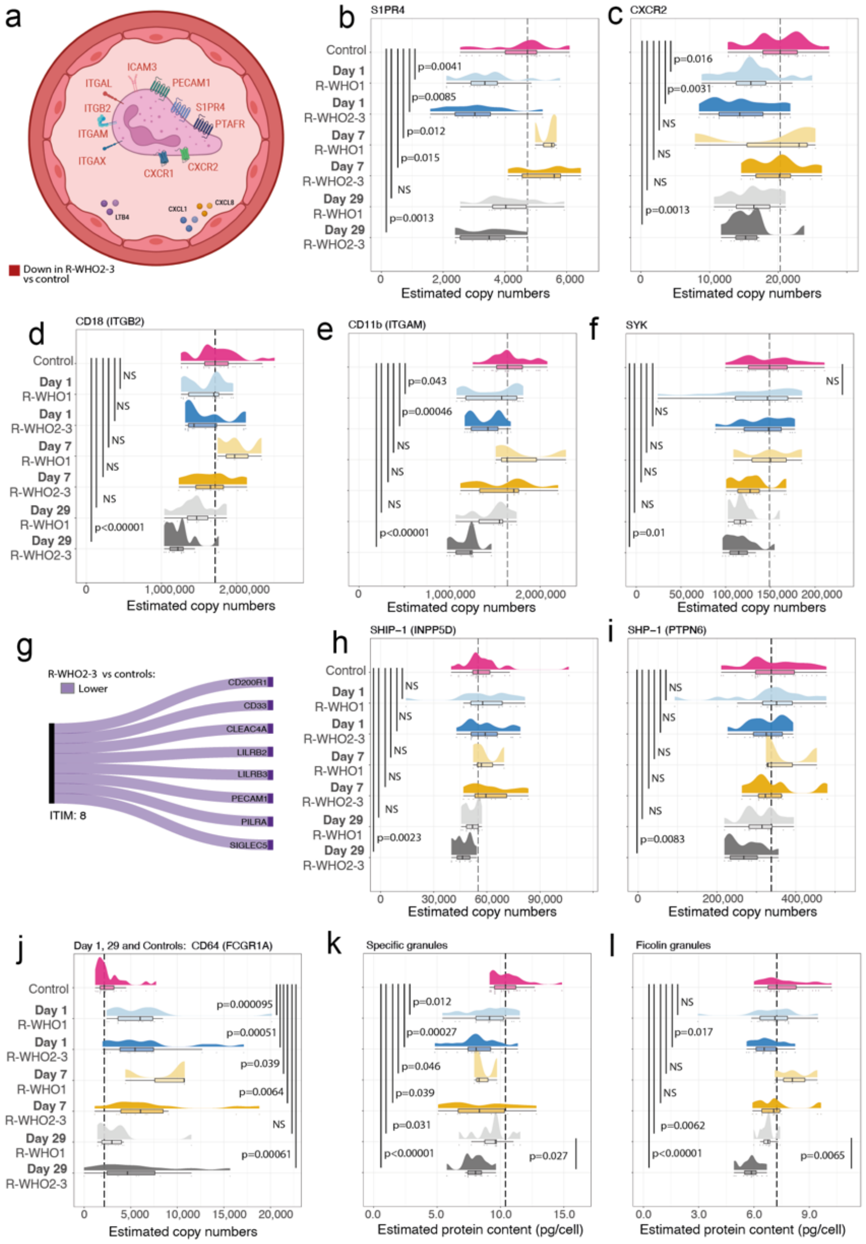

Figure 4.10 – Neutrophil migratory and inhibitory machinery at day 29: ***(a)*** Schematic showing the migratory receptors and integrins that are exclusively reduced in abundance in the R-WHO2-3 patients compared to Controls. Rain cloud plot showing the estimated protein copy numbers for ***(b)*** S1PR4, ***(c)*** CXCR2, ***(d)*** CD18 (ITGB2), ***(e)*** CD11b (ITGAM) and ***(f)*** SYK across controls, Day 1, Day 7 and Day 29 COVID19 patients stratified into R-WHO1 (recovered) or R-WHO2-3 (not recovered). ***(g)*** Sankey diagram showing the inhibitory receptors that are significantly decreased in abundance in the neutrophil proteomes of R-WHO2-3 patients. Rain cloud plot showing the estimated protein copy numbers for ***(h)*** SHIP-1 (INPP5D), ***(i)*** SHP-1 (PTPN6) and ***(j)*** CD64 (FCGR1A) across controls, Day 1, Day 7 and Day 29 COVID19 patients stratified in R-WHO1 and R-WHO2-3. Rain cloud plot showing the estimated protein content for all proteins contained primarily within ***(k)*** Specific Granules and ***(l)*** Ficolin Granules across controls, Day 1, Day 7 and Day 29 COVID19 patients stratified in R-WHO1 and R-WHO2-3. All Raincloud plots include density plots as well as boxplots. For all boxplots the whiskers extend from the hinge to the largest and smaller values no further than 1.5 x interquartile range.

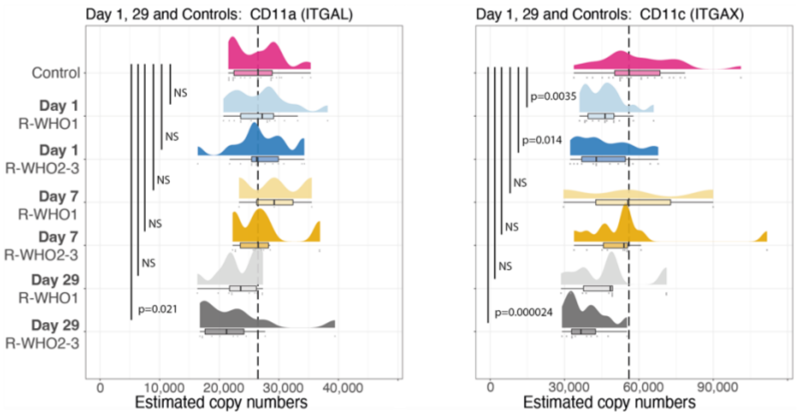

Figure 4.11: Rain cloud plots showing the estimated copy numbers of CD11a & CD11c in neutrophil proteomes across control participants, Day 1 COVID19, Day 7 COVID19 and Day29 all stratified by R-WHO1 (recovered) and R-WHO2-3 (not recovered) patients. All Raincloud plots include density plots as well as boxplots. For all boxplots the whiskers extend from the hinge to the largest and smaller values no further than 1.5 x interquartile range.

One other striking observation was that neutrophils from R-WHO2-3 patients displayed a systematic reduction in the abundance of receptors that control neutrophil migration from the blood into sites of inflammation (Fig. 4.10a), including the sphingosine-1-phosphate receptor S1PR4 (Fig. 4.10b) and chemokine receptor CXCR2 (Fig. 4.10c). They also displayed reductions in the most abundant integrin subunits, which are vital for the extravasation process. There were reductions in CD18 (Fig. 4.10d) and CD11b (Fig. 4.10e), as well as CD11a and CD11c (Fig. 4.11) along with SYK (Fig. 4.10f) a kinase that mediates integrin signalling. With the exception of CD11c, this reduced abundance was not present in neutrophil proteomes at day 1 or 7 (Fig. 4.10d-f), suggestive of reduced migratory capacity.

The neutrophil proteomes of R-WHO2-3 patients also showed reduced abundance of the inhibitory machinery which plays a prominent role in restricting unwanted neutrophil activation(Favier, 2016, van Rees et al., 2016). Like the C-Type-Lectin family and leukocyte immunoglobulin-like family proteins (Fig. 4.10g). Similarly they also displayed reduced expression of the SH-2-containing inositol 5'phosphatase, SHIP-1 (INPP5D; Fig. 4.10h) and the SH-2 containing tyrosine phosphatase SHP-1 (PTPN6; Fig. 4.10i), which also mediate inhibitory signals(Ono et al., 1996, Ono et al., 1997) and in which mutations have been linked to severe autoimmune disease(Green and Shultz, 1975). This would suggest the neutrophils display a reduced activation threshold, which is concordant with the increased abundance of the activation marker CD64 (Fig. 4.10j).

In the proteomes of neutrophils from R-WHO2-3 patients at day29 some of the changes observed at day 1 and 7, like the increased abundance of ARG1 and TGFB1, were no longer present (Fig. 4.12), however the reductions in specific granules (Fig. 10k) driven by LTF and CAMP persisted (Fig. 4.12). Furthermore, the proteomes of R-WHO2-3 patients displayed significant reductions in protein content of ficolin granules (Fig. 4.10l), a subset of highly mobilizable granules, distinct from the gelatinase granules(Rorvig et al., 2009). As granule abundance is linked to maturity, we again considered whether these observed changes reflected increased prevalence of immature neutrophils in the R-WHO2-3 individuals. As previously mentioned, the last granules to be produced as neutrophils mature are the tertiary granules and secretory vesicles and these set of granules displayed no significant changes in protein content in the neutrophil proteomes from R-WHO2-3 patients (Fig. 4.13). The selective reductions in specific and ficolin granules would thus appear to be maturity independent.

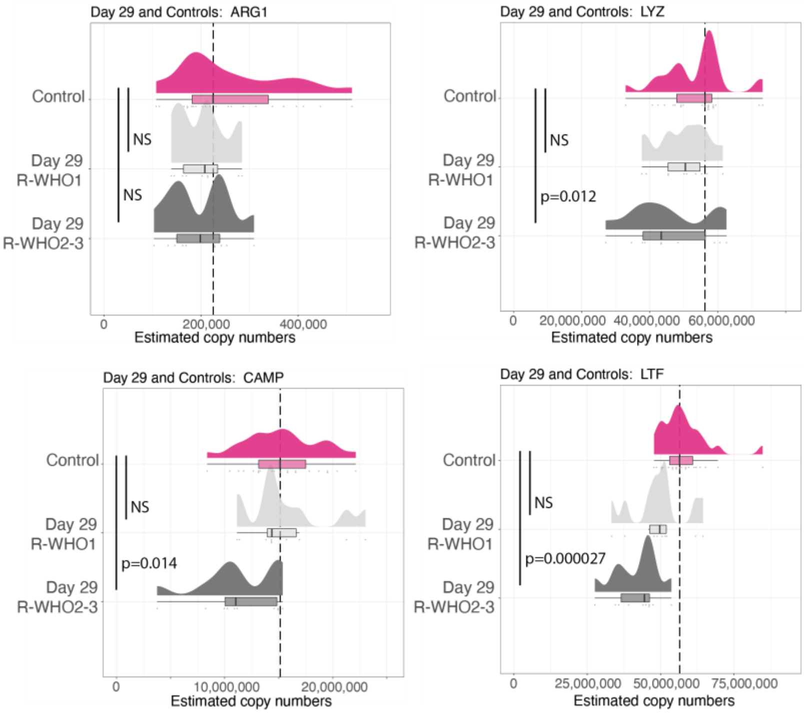

Figure 4.12: Rain cloud plots showing the estimated protein copy numbers of LYZ, ARG1, CAMP & LTF at day 29 in the neutrophil proteomes of control group participants, and day 29 R-WHO1 (recovered) and R-WHO2-3 (not recovered) COVID19 patients. All Raincloud plots include density plots as well as boxplots. For all boxplots the whiskers extend from the hinge to the largest and smaller values no further than 1.5 x interquartile range.

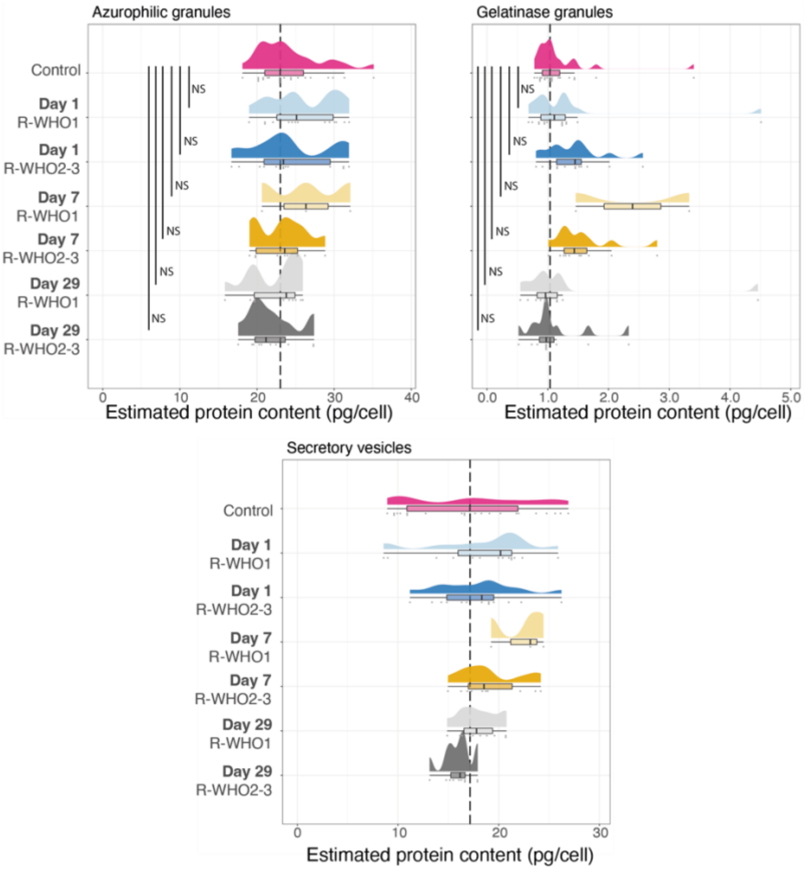

Figure 4.13: Rain cloud plots showing the estimated protein content of azurophilic and gelatinase granules and secretory vesicles in the neutrophil proteomes of control participants at baseline, and Day 1 COVID19, Day 7 COVID19 and Day29 all stratified by R-WHO1 (recovered) and R-WHO2-3 (not recovered) patients. All Raincloud plots include density plots as well as boxplots. For all boxplots the whiskers extend from the hinge to the largest and smaller values no further than 1.5 x interquartile range.

4.4.5 Contaminants and biomarkers: eosinophil proteins as predictors of ICU admission

Another important objective of this this project was to determine if protein abundance could be predictive of disease severity, thus having the potential to be used as a disease marker in the clinic. This biomarker analysis revealed that when comparing the proteomes of COVID19 patients to controls in the early infection, there were systematic differences in abundance of a subset of proteins that were classified as eosinophil granules, such as eosinophil peroxidase (EPX; Fig. 4.14a), galectin-10 (CLC; Fig. 4.14b), eosinophil major basic protein (PRG2; Fig. 4.14c) and eosinophil major basic protein 2 (PRG3; Fig. 4.14d). All four of these proteins displayed significantly reduced abundance in the neutrophil proteomes of COVID19 patients. The neutrophils in this study had been isolated using negative immunomagnetic selection which is an efficient isolation method that has been shown to provide a highly pure (95%+) population of neutrophils. However it has also been shown that eosinophils can be the main contaminants present in these preparations(Thomas et al., 2015). As such we became interested in understanding if these 4 proteins were eosinophil specific or if they were also expressed in neutrophils.

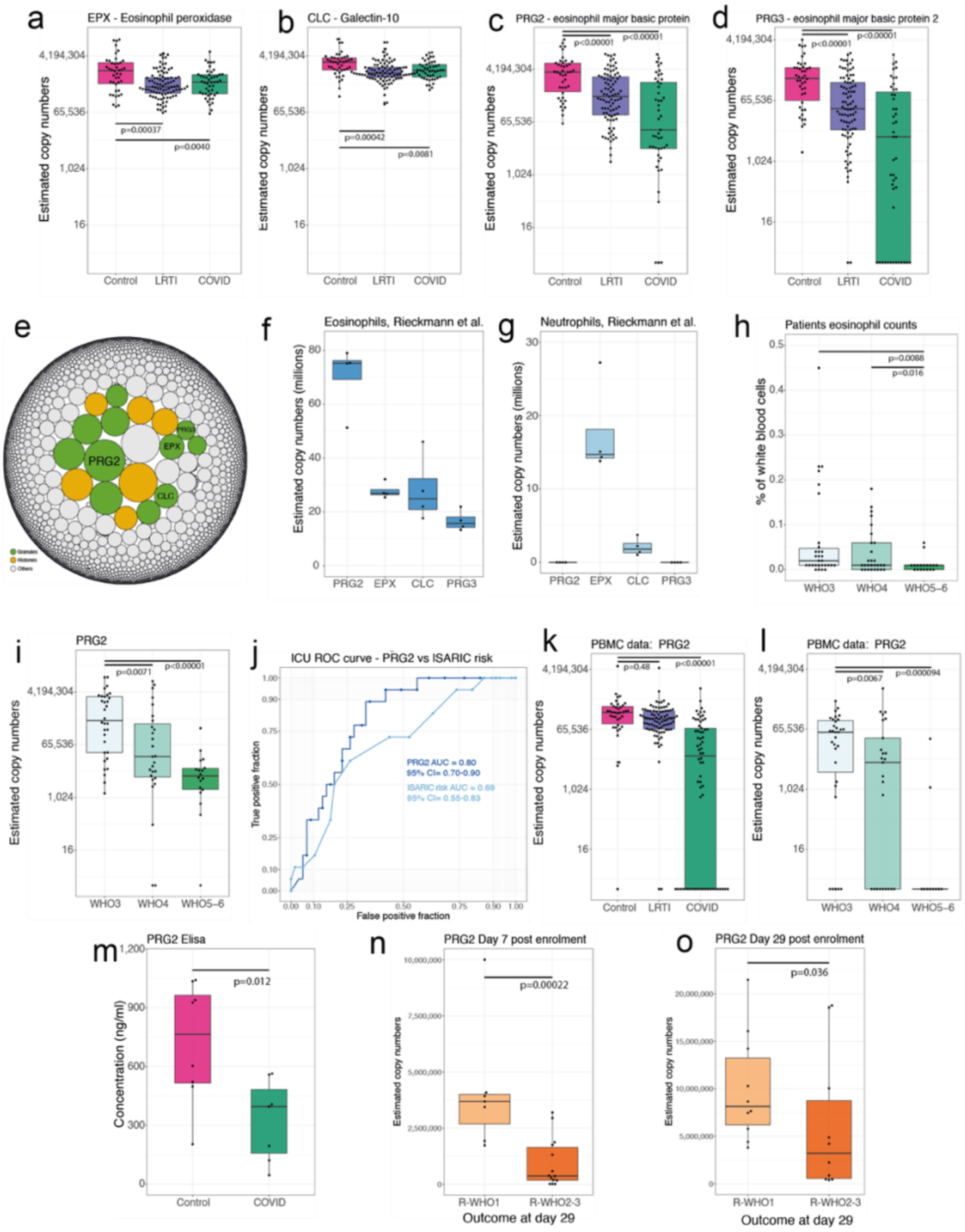

Figure 4.14 – PRG2 abundance in the neutrophil and PBMC proteomes: Box plot showing the estimated protein copy numbers for across control, LRTI and COVID19 patients for ***(a)*** EPX, ***(b)*** CLC, ***(c)*** PRG2, ***(d)*** PRG3. **(e)** Bubble plot showing the estimated copy numbers of eosinophil proteins in the Rieckmann data(Rieckmann et al., 2017). Box plot swhong the estimated protein copy numbers for PRG2, EPX, CLC and PRG3 derived from the Rieckmann data in ***(f)*** the eosinophil proteomes ***(g)*** the neutrophil proteomes. ***(h)*** Box plot showing the eosinophil counts across WHO3 (moderate), WHO4 (severe) and WHO5-6 (critically severe) COVID19 patients at day 1 of enrolment. ***(i)*** Box plot showing the estimated protein copy numbers for PRG2 at day 1 of enrolment across WHO3, WHO4 and WHO5-6 COVID19 patients. ***(j)*** ROC curve for the estimated copy numbers of PRG2 and the ISARIC risk score predicting ICU admission. ***(k)*** Box plot showing the estimated protein copy numbers for across control, LRTI and COVID19 patients for PRG2 in the PBMC proteomes. ***(l)*** Box plot showing the estimated protein copy numbers for PRG2 across WHO3, WHO4 and WHO5-6 COVID19 patients in the PBMC proteomes. ***(m)*** Box plot showing the concentration of PRG2 within the blood of control and COVID19 patients as measure by an ELISA. Box plots showing the estimated protein copy numbers for PRG2 across R-WHO1 (recovered at day 29) and R-WHO2-3 (not recovered at day 29) COVID19 patients in the neutrophil proteomes at ***(n)*** day 7 post enrolment and ***(o)*** day 29post enrolment. For all boxplots the whiskers extend from the hinge to the largest and smaller values no further than 1.5 x interquartile range.

For this purpose, we mined a publicly available proteomic dataset that characterised the proteomes of neutrophils and eosinophils after using fluorescence assisted cell sorting (FACS) to isolate both populations$^{\text{(Rieckmann et al., 2017)}}$. FACS has been shown to be an effective method to exploit the differences between neutrophils and eosinophils to generate more pure populations$^{\text{(Dorward et al., 2013)}}$. Within the Rieckmann dataset all 4 of the previous proteins were among the 50 most abundant proteins in the eosinophil proteomes (Fig. 4.14e), with PRG2 being the most abundant protein (Fig. 4.14f). The neutrophil proteomes in the Rieckmann dataset highlighted that both PRG2 and PRG3 display very low abundance in neutrophils, both having less than 5,000 protein copies per cell, while EPX and CLC were highly abundant, both displaying over 1.8 million copies per cell (Fig. 4.14g). This suggests that EPX and CLC are likely to also be expressed in some neutrophil populations, unlike PRG2 and PRG3.

It had been reported that COVID19 patients frequently displayed eosinopenia$^{\text{(Wang et al., 2020)}}$, which is when there is a prominent reduction in the number of eosinophils in the blood. Hence, we checked the peripheral blood eosinophil counts for COVID19 patients at day 1 of enrolment and stratified by patient severity upon admission (as previously described). This analysis found that WHO5-6 patients had significantly lower number of eosinophils compared to both WHO3 and WHO4 patients (Fig. 4.14h). However, the value of the eosinophil counts as a predictive marker was limited due to lack of sensitivity, with multiple WHO3 and WHO4 patients having no detectable eosinophils.

We hypothesised that the estimated copy numbers of PRG2 could provide a more sensitive proxy measure for eosinophil abundance, and thus for critically severe COVID19. We compared PRG2 abundance across the stratified COVID19 patient groups and found its abundance was significantly reduced in both WHO4 and WHO5-6 patients compared to WHO3 patients (Fig. 4.14i). To evaluate the value of PRG2 abundance to predict critically severe illness, a receiver operating characteristic (ROC) curve analysis was performed. This revealed that PRG2 copy numbers had an area under the curve (AUC) of 0.8 (0.70-0.90). The AUC is a measure an effective measure of sensitivity and specificity that is summarised the whole ROC curve. The AUC of the estimated copy numbers of PRG2 was higher than the AUC for a clinically recognised metric like ISARIC risk score$^{\text{(Knight et al., 2020)}}$ (Fig. 4.14j).

As orthogonal validation we also studied the abundance of PRG2 across the PBMC proteomes in a similar analysis as with the neutrophil data. PRG2 was significantly reduced in abundance in the PBMC proteomes of COVID19 patients compared to controls (Fig. 4.14k), and was also significantly lower in the PBMC proteomes of WHO5-6 patients compared to WHO3 and WHO4 patients (Fig. 4.14l). This suggested PRG2 could be a robust biomarker to predict critically severe disease. However, mass spectrometry-based methods are not routinely used in this clinic. We therefore wanted to explore PRG2 abundance with an assay that could be easily translated into a clinical assay. Consequently, an enzyme-linked immunosorbent assay (ELISA) was used to measured PRG2 concentrations in the patients' blood, and it also found PRG2 to be significantly reduced in COVID19 patients compared to controls (Fig. 4.14m).

PRG2 appears to be a robust biomarker, with abundance not only having a correlation with severe acute disease, but also having a correlation to delayed recovery from COVID19. With patients stratified by their outcome at day 29 as previously described, the neutrophil proteomes of R-WHO1 (recovered) patients at day 7 displayed significantly higher abundance of PRG2 (Fig. 4.14n) compared to those not fully recovered (R-WHO2-3), suggesting a potential links between eosinophil levels, using PRG2 as a proxy, and recovery from COVID19. This pattern was still significant at day 29, where R-WHO1 patients still had significantly higher abundance of PRG2 compared to R-WHO2-3 patients (Fig. 4.14o).

4.5 Discussion

This study represents a large-scale and comprehensive proteomic characterisation of hundreds of neutrophil proteomes studied in both health and disease. It provides a valuable resource to understand the phenotypes of peripheral blood neutrophils derived from healthy individuals, in addition to an in-depth proteomic map of the changes in neutrophil proteomes that were seen during acute COVID19 and the immediate recovery period. The present data show IFN-I is a core COVID19 signature present in majority of neutrophil proteomes derived from COVID19 patients at day 1 of enrolment, regardless of disease severity. IFN is a vital anti-viral pathway, hence its engagement in the acute infection would be expected. However, the current data also showed that this signature can be transient with kinetics dependent on disease severity. In the neutrophils of COVID19 patients with moderate disease there was a sustained IFN induced protein signature, whereas in patients with critically severe COVID19 this signature was lost, making them no different to controls. This data suggests that either severe and critically severe COVID19 patients had already successfully cleared the virus and no longer required an anti-viral response, or that the premature reduction in IFN signalling could have a detrimental effect by not obstructing viral replication. If the latter holds true, it suggests stratified IFNβ treatment could prove to be beneficial for these patients.

In our data we also identified novel neutrophil markers of disease severity, including multiple receptors with great relevance to neutrophil responses like the pattern recognition receptor TLR2 and the inhibitory receptor VISTA. TLR2 has been recently shown to recognise, bind to and engage inflammatory signalling pathways in response to the spike protein of SARS-CoV-2(Khan et al., 2021). It's persistent upregulation and capacity to induce NFKB suggests it can be involved in promoting neutrophil inflammatory pathways, in which case antagonistic antibodies against TLR2 could be a potential mechanism to limit excessive neutrophil activation. Similarly, the increased abundance of Vista in the neutrophils of patients with critically severe COVID19 is a potential target of interest. Vista has been described as a checkpoint inhibitor in T cells, where loss of function lowers the T-cell activation threshold(Wang et al., 2014). It is also highly expressed in myeloid cells, as this datasets highlights. It has been proposed that agonistic antibodies against Vista could help reduce inflammation caused by macrophages(ElTanbouly et al., 2019, ElTanbouly et al., 2020), and its sustained upregulation in neutrophils suggests it could also be a promising target. Vista had been shown to reduce the abundance of CXCR2 in neutrophils and thus could also help manage excessive innate immune activation(ElTanbouly et al., 2019, ElTanbouly et al., 2020). In this case in particular, agonistic antibodies to Vista could be another important alternative to help manage the innate cytokine storm.

The current study also noted metabolic changes in neutrophils correlated with severity and recovery status. Critically ill patients display changes in abundance of proteins across a wide array of metabolic pathways some of which are explained by the increased proportion of immature neutrophils in patients with severe disease, while others also like LDHA, are likely linked to hypoxia(McGettrick and O'Neill, 2020, Watts et al., 2021) and neutrophil extracellular traps (NETs) (Awasthi et al., 2019). It was also interesting to find that rate limiting enzymes of glycogenolysis were significantly decreased across all timepoints, suggesting an attempt to maintain glycogen stores. When looking at the delayed recovery COVID19 patients, their metabolic profiles were distinct from the changes seen in neutrophils isolated from patients on day 1 and 7. Neutrophils have been shown to depend on glycolysis for energy production when the environment is nutrient rich(Jeon et al., 2020, Borregaard and Herlin, 1982, Pearce and Pearce, 2013, Burn et al., 2021), and to default to glycogen breakdown when it is not(Sadiku et al., 2021). These two metabolic pathways are vital to neutrophil functions, and our data show significantly lower abundance of key rate-limiting proteins of both glycolysis and glycogenolysis in the proteomes of non-recovered patients. Reductions in both pathways impair the bioenergetic capacity of neutrophils and have been shown to lead to impaired killing and impaired survival capacities in chronic disease(Sadiku et al., 2021).

The dysfunctional phenotype of post-COVID19 neutrophils was not limited to metabolism. Peripheral blood neutrophils depend on signalling receptors and integrins to recognise migratory signals and perform the extravasation process. The neutrophil proteomes of non-recovered patients displayed a systematic reduction in migratory receptors, from the chemokine receptors to spingosine-1-phosphate receptors. They also showed significantly reduced abundance of subunits of the Mac-1 and LFA-1 complexes that mediate leucocyte extravasation(Anderson and Springer, 1987). These changes suggest impaired capacity of neutrophils to migrate from the blood into the sites of inflammation, which may theoretically make patients post-COVID19 vulnerable to secondary infections; as mutations that affect function or abundance of CD18 in human neutrophils cause Leucocyte Adhesion Deficiency (LAD) and result in increased susceptibility to bacterial infections(Etzioni et al., 1992).

Non-recovered COVID19 patients also displayed a systematic reduction in the abundance of inhibitory receptors and phosphatases which are required to limit neutrophil activation(Ono et al., 1996, Ono et al., 1997, Shultz et al., 1997, Favier, 2016, van Rees et al., 2016). This reduction in the inhibitory machinery might explain the persistent degranulating phenotype present in peripheral blood neutrophils, which can have deleterious consequences as activated neutrophils in the bloodstream can cause considerable tissue damage and have been described as potentially lethal(Burn et al., 2021, Yang et al., 2021). However, additional work is required to understand these results, as the programmed disarming of neutrophils has also been described as a protective mechanism(Adrover et al., 2020).

In conclusion, COVID19 remains a risk to global health, and we believe our data has identified a core neutrophil proteomic signature associated to acute disease. The longitudinal analysis also identified potential stratified treatment options related to transient IFN response as well as persistent upregulation of neutrophil receptors linked to excessive inflammation. Furthermore, the investigation characterised a molecular phenotype linked to delayed recovery which potentially represents a mechanism relating to the extended symptoms COVID19 patients can experience.

Chapter 5 – Data sharing and visualisation for large-scale proteomics

5.1 Publications relating to this chapter

- The Encyclopedia of Proteome Dynamics: a big data ecosystem for (prote) omics (Brenes et al., 2018a)
- The Encyclopedia of Proteome Dynamics: The KinoViewer (Brenes and Lamond, 2019)
- Tissue environment, not ontogeny, defines murine intestinal intraepithelial T lymphocytes (T-IEL) (Brenes et al., 2021a)

5.1.1 Pre-prints relating to this chapter

- The Immunological Proteome Resource(Brenes et al., 2022)

5.1.2 Contributions to the publications on this chapter

- I designed and built the Encyclopedia of Proteome Dynamics, the KinoViewer and Immunological Proteome Resource
- I wrote all the EPD and ImmPRes papers with input from all authors
- I analysed the data for the T-IEL paper
- Maud Vandereyken, Mahima Swamy and I interpreted the data and co-wrote the T-IEL paper with input from all other authors
- Maud Vandereyken and I generated all the figures in the T-IEL paper

5.1.3 Main contributions of other authors to the publications on this chapter

- Vackar Afzal, Robert Kent, Christopher Martin, Emma Arandjelovic, Barry Carr, Yasmen Ahmad and Scott Greig developed the core peptracker suite which provided the user authentication and the base HTML template to the EPD
- The data contained within the EPD were generated in the Lamond Laboratory by Tony Ly, Mark Larance, Michelle Boisvert, Kathryn Kirkwood, Thomas Crozier, Armel Nicholas, Dalila Bensaddek and Jens Hukelmann
- The data contained within ImmPRes were generated by Jens Hukelmann, Laura Spinelli, Andrew Howden, Julia Marchingo, Linda Sinclair, Christina Rollings, Iain Phair, Stephen Matthews, Sarah Ross, David Finlay and Mahima Swamy
- Mahima Swamy generated the T-IEL samples and supervised the T-IEL paper
- Jens Hukelmann generated the mass spectrometry files for the T-IEL paper
- Maud Vanderekyen, Olivia James & Harriet Watt performed the validation experiments in the T-IEL paper

5.2 Introduction

Mass spectrometry (MS) - based proteomics is now the predominant approach used for large-scale detection and quantitation of proteins in cells and tissues(Larance and Lamond, 2015, Aebersold and Mann, 2016, Aebersold and Mann, 2003). As such there has been a rapid growth in the number and size of proteomics datasets generated, with over 500 new proteomic datasets being deposited into PRIDE each month(Perez-Riverol et al., 2022). This has been aided by the continual improvement of mass spectrometers, featuring increased speed and resolution, combined with improvements in sample handling workflows and peptide chromatography. The growing trend for large-scale projects is to use DIA and have short gradients to maximise throughput(Messner et al., 2020).

For many years it has been recognised that visualising proteomics data is a complicated but vital aspect of generating datasets(Oveland et al., 2015). With more datasets being created and with the increased size and complexity of each of them it also becomes particularly important to consider how those datasets will be made available to the scientific community. Web-based apps are particularly well suited for this task and their value to proteomics data is now more recognised(Rieckmann et al., 2017, Go et al., 2021, Wolf et al., 2020, Brenes et al., 2022, Brenes and Lamond, 2018). These apps support the efforts to share the proteomics data with the wider community that has biological interest in the data that has been generated.

With this goal in mind, we created the two distinct web applications, The Encyclopedia of Proteome Dynamics(Brenes et al., 2018a) (EPD) and The Immunological Proteome Resource (Brenes et al., 2022) (ImmPRes), to provide simple graphical access to large-scale mass spectrometry-based proteomics data to the research community. Both the EPD and ImmPRes have integrated multiple datasets and provide the option to explore these data via predefined interactive visualisations, like volcano plots and bar plots, aimed at providing simple and quick visuals to explore large datasets. Both tools were created on the premise of open access to the data and as such have made all processed data contained within the tool readily available, as well as providing direct links to PRIDE(Perez-Riverol et al., 2018, Perez-Riverol et al., 2022) for the raw files produced by the different experiments.

5.3 Methods

5.3.1 EPD design

The EPD was designed as a big data solution, using a noSQL ecosystem. It combines a graph database and a key-store columnar database hybrid. The graph is built in Neo4j (https://neo4j.com/) and the key-store columnar hybrid is apache Cassandra (http://cassandra.apache.org/). It operates by dividing all data evenly across all the nodes in a ring architecture with an odd number of nodes. The replication factor of the data was set to n+1/2 to provide resilience against hardware failure. The EPD also uses D3.js (https://d3js.org/) as the client-side java script (JS) library used to generate the interactive visualisations.

The EPD is part of the peptracker suite (https://peptracker.com/) and as such the application has been programmed in python and is implemented in an ecosystem that uses Django (https://www.djangoproject.com/) as a web framework with nginx running the webserver and load balancer (https://www.nginx.com/).

5.3.2 Neo4j design

Neo4j hosts a graph which models the interactions that are represented by a force diagram within the graphical interface. It stores a descriptive hierarchy describing the datasets, experiment types and organisms and stores associations of the different datasets to each of these descriptive elements. It also stores associations between the protein nodes and all the different datasets where they have been identified.

5.3.3 Cassandra

Cassandra hosts all the quantitative data at the protein level for all experiments contained within the EPD. It also hosts aggregated peptide data with a global false discovery rate which is available for download.

It has specific tables created for each experiment type with data redundancy, such as repeating the protein accession and gene name in all tables, in order to optimize data access and read velocity. As the data from each of the datasets are stored once and not modified after storage (except under extraordinary circumstances), the risk to data integrity via a less stringent data mode remains limited.

5.3.4 KinoViewer design

The KinoViewer was designed to operate as part of the EPD. It uses a manually drawn scalable vector graphics (SVG) file to represent the hierarchical relationships between the kinases. The hierarchical representation within the SVG file is based on the original Manning tree(Manning et al., 2002), but with the cyclin-dependent kinase hierarchy updated to match the newer philogeny(Malumbres, 2014). It uses JS to produce an interactive visualisation that is produced based on proteomics or transcriptomic data.

5.3.5 ImmPRes design

ImmPRes has been developed in python and implemented via plotly (https://plotly.com/) dash (https://plotly.com/dash/). All of the proteomics data were processed using MaxQuant**(Cox and Mann, 2008, Tyanova et al., 2016a)** (https://www.maxquant.org/), with the search parameters included in each PRIDE submission. All copy numbers were calculated using R (https://www.r-project.org/) v 4.0.3, or with Perseus**(Tyanova et al., 2016b)** and the differential expression analysis performed with the Bioconductor package LIMMA**(Ritchie et al., 2015)** v 3.50.3.

5.3.6 ImmPres data storage

ImmPRes was designed with a simplified infrastructure for easier maintenance. To minimise the storage requirements and improve app performance ImmPRes stores only summary level text file for all datasets. It stores the averaged copy numbers across the different populations present in each dataset, and for differential expression analysis it stores pre-computed p-values and fold changes. This reduced the storage requirements by 10-100 fold across the different datasets. It also implicitly improves the performance of the web app as complex calculations do not need to be computed on the fly.

5.3.7 Processing and analysis of proteomics data

All the datasets on ImmPRes and the EPD were processed, searched and quantified with MaxQuant(Tyanova et al., 2016a, Cox and Mann, 2008). The search parameters for the individual datasets differ, but are included in each PRIDE submission. For ImmPRes all datasets used a false discovery rate of 1% at both the protein and at the peptide-spectrum match level.

5.3.8 Data filtering

For both the EPD and ImmPRes any proteins that were labelled as 'Reverse', 'Pontential Contaminant' or 'Only identified by site' in the MaxQuant output were filtered out from the results.

5.3.9 ImmPRes copy number calculations

The estimated protein copy numbers were calculated using the proteomic ruler(Wisniewski et al., 2014). For MS3-based TMT datasets, the estimated copy number calculation required additional steps which involved the assignment of proportional MS1 intensity based on the ratios of the MS3 reporter ion intensity as previously described(Howden et al., 2019). Briefly, for each protein the reporter ion intensities across all channels were totalled, this total was then used to divide each reporter ion intensity to obtain a ratio. The MS1 intensity of each protein was then assigned across reporters based on the previous ratio.

5.3.10 ImmPRes mice details

For proteomics experiments cell populations were isolated from C57BL/6 wild type mice, OT-I TCR transgenic(Pircher et al., 1989) and P14 TCR transgenic mice[40] and mice with T cell selective deletions of Myc (Dose et al., 2009, Mycko et al., 2009)or *Slc7a5*(Sinclair et al., 2013) (full details are in the 'Publications & protocols' tab on immpres.co.uk). All mice for the Dundee based experiments were maintained in the Biological Resource Unit at the University of Dundee using procedures that were approved by the University Ethical Review Committee and under the authorization of the UK Home Office Animals (Scientific Procedures) Act 1986.

5.3.11 ImmPRes Label free SP3 Proteomics sample preparation

Cell pellets composing of 1 – 5 million cells (depending on the number of fractions being analysed) were lysed in 400 µl lysis buffer at room temperature in 4% SDS, 50 mM TEAB pH 8.5, 10 mM TCEP under agitation (5 min, 1200 rpm on tube shaker), boiled (5 min, 500 rpm on tube shaker), then sonicated with a BioRuptor (30s on, 30s off x30 cycles). Protein concentration was determined using EZQ protein quantitation kit (Invitrogen) as per manufacturer instructions. Lysates were alkylated with 20 mM iodoacetamide for

1 hr at room temperature in the dark, before protein clean up by SP3(Hughes et al., 2014) procedure. Briefly, 200 µg of 1:1 mixed Hydrophobic and Hydrophilic Sera-Mag SpeedBead Carboxylate-Modified Magnetic Particles were added per protein sample then acidified to ~pH 2.0 by addition 10:1 Acetonitrile: Formic Acid. Beads were immobilised on a magnetic rack and proteins washed with 2 × 70% ethanol and 1 × 100% acetonitrile. Rinsed beads were reconstituted in 0.1% SDS 50 mM TEAB pH 8.5, 1 mM $CaCl_2$ and digested overnight with LysC followed by overnight digestion with Trypsin, each at a 1:50 enzyme to protein ratio. Peptide clean up was performed as per SP3 procedure(Hughes et al., 2014). Briefly, protein-bead mixtures were resuspended and 100% acetonitrile added for 10 min (for the last 2 min of this beads were immobilised on a magnetic rack). Acetonitrile and digest buffer were removed, peptides were washed with acetonitrile and eluted in 2% DMSO. Peptide concentration was quantified using CBQCA protein quantitation kit (Invitrogen) as per manufacturer protocol. Formic acid was added to 5% final concentration.

Samples were fractionated using high pH reverse phase liquid chromatography. Samples were loaded onto a 2.1 mm x 150 mm XBridge Peptide BEH C18 column with 3.5 µm particles (Waters). Using a Dionex Ultimate3000 system, the samples were separated using a 25 min multistep gradient of solvents A (10 mM formate at pH 9 in 2% acetonitrile) and B (10 mM ammonium formate pH 9 in 80% acetonitrile), at a flow rate of 0.3 mL/min. Peptides were separated into 16 fractions which were consolidated into eight fractions. Fractionated peptides were dried in vacuo then dissolved in 5% Formic Acid for analysis by LC-ES-MS/MS.

5.3.12 Label free urea sample preparation

Cell pellets composing of 1 – 5 million cells (depending on the number of fractions being analysed) were lysed in 400 µl lysis buffer in urea buffer (8 M urea, 50 mM Tris pH 8.0, 1 mM TCEP) with protease (cOmplete mini EDTA free, Roche) and with phosphatase inhibitors (PhosStop, Roche) at room temperature for half an hour.

Samples were sonicated, and protein concentration was determined using a BCA assay according to the manufacturer's instructions. To reduce samples, 10 mM DTT was added for 30 min at room temperature. After reduction, 50 mM iodoacetimide was added to each sample and incubated for 45 min at room temperature in the dark. Samples were then diluted to 4 M urea with Tris buffer (100 mM Tris pH 8.0, 1 mM $CaCl_2$). LysC (Wako), reconstituted in Tris buffer, was then added to each sample at a ratio of 50:1 protein:LysC and the samples were incubated with LysC overnight at 30°C. Samples were then transferred to low-bind 15-ml falcon tubes (Eppendorf) and further diluted with Tris buffer to 0.8 M urea. Trypsin (Promega), reconstituted with Tris buffer, was then added to the samples at a ratio of 50:1 protein:trypsin and the samples were incubated for 8 hours at 30°C. After digestion, the samples were desalted with C18 SepPack cartridges (Waters) and dried down in a SpeedVac (Genevac). Dried peptide samples were resuspended in 10 mM sodium borate 20% (v/v) acetonitrile (pH 9.3) and fractionated into 16 fractions by strong anion exchange (SAX) chromatography. Peptide samples for proteomic analysis were fractionated with an Ultimate 3000 HPLC equipped with an AS24 strong anion exchange (SAX)(Ritorto et al., 2013). For the separation, the buffers used were 10 mM sodium borate (pH 9.3) (SAX buffer A) and 10 mM sodium borate (pH 9.3), 500 mM NaCl (SAX buffer B). Peptide samples were resuspended in 210 µl of 10 mM sodium borate 20% (v/v) acetonitrile (pH 9.3) and injected onto the SAX column and separated using an exponential elution gradient starting with Buffer A. In total, 16 peptide fractions were collected and desalted with Sep-pack C18 96 well desalting plates (Waters). Desalted peptides were dried down with a SpeedVac (Genevac).

5.3.13 Label Free Liquid chromatography tandem mass spectrometry analysis

For each fraction, 1 µg was injected onto a nanoscale C18 reverse-phase chromatography system (UltiMate 3000 RSLC nano, Thermo Scientific) then electrosprayed into an Orbitrap mass spectrometer (LTQ Orbitrap Velos Pro; Thermo Scientific). For chromatography buffers were as follows: HPLC buffer A (0.1% formic acid), HPLC buffer B (80% acetonitrile and 0.08% formic acid) and HPLC buffer C (0.1% formic acid). Peptides were loaded onto an Acclaim PepMap100 nanoViper C18 trap column (100 µm inner diameter, 2 cm; Thermo Scientific) in HPLC buffer C with a constant flow of 10 µl/min. After trap enrichment, peptides were eluted onto an EASY-Spray PepMap RSLC nanoViper, C18, 2 µm, 100 Å column (75 µm, 50 cm; Thermo Scientific) using the buffer gradient: 2% B (0 to 6 min), 2% to 35% B (6 to 130 min), 35% to 98% B (130 to 132 min), 98% B (132 to 152 min), 98% to 2% B (152 to 153 min), and equilibrated in 2% B (153 to 170 min) at a flow rate of 0.3 µl/min. The eluting peptide solution was automatically electrosprayed using an EASY-Spray nanoelectrospray ion source at 50° and a source voltage of 1.9 kV (Thermo Scientific) into the Orbitrap mass spectrometer (LTQ Orbitrap Velos Pro; Thermo Scientific). The mass spectrometer was operated in positive ion mode. Full-scan MS survey spectra (mass/charge ratio, 335 to 1800) in profile mode were acquired in the Orbitrap with a resolution of 60,000. Data were collected using data- dependent acquisition: the 15 most intense peptide ions from the preview scan in the Orbitrap were fragmented by collision-induced dissociation (normalized collision energy, 35%; activation Q, 0.250; activation time, 10 ms) in the LTQ after the accumulation of 5000 ions. Precursor ion charge state screening was enabled, and all unassigned charge states as well as singly charged species were rejected. The lock mass option was enabled for survey scans to improve mass accuracy. (Using Lock Mass of 445.120024).

5.3.14 TMT Proteomics sample preparation

Cell pellets composing of 1 – 5 million cells (depending on the number of fractions being analysed) were lysed in 400 µl lysis buffer (4% sodium dodecyl sulfate, 50 mM TCEP (pH 8.5) and 10 mM tris(2-carboxyethyl)phosphine hydrochloride). Lysates were boiled and sonicated with a BioRuptor (30 cycles: 30 s on and 30 s off) before alkylation with 20 mM iodoacetamide for 1 h at 22 °C in the dark. The lysates were subjected to the SP3(Hughes et al., 2014) procedure for protein clean-up(Hughes et al., 2014) before elution into digest buffer (0.1% sodium dodecyl sulfate, 50 mM TEAB (pH 8.5) and 1 mM $CaCl_2$) and digested with LysC and Trypsin, each at a 1:50 (enzyme:protein) ratio. TMT labeling and peptide clean-up were performed according to the SP3 protocol.

The TMT samples were fractionated using off-line high-pH reverse-phase chromatography: samples were loaded onto a 4.6 mm × 250 mm XbridgeTM BEH130 C18 column with 3.5 µm particles (Waters). Using a Dionex BioRS system, the samples were separated using a 25-min multistep gradient of solvents A (10 mM formate at pH 9 in 2% acetonitrile) and B (10 mM ammonium formate at pH 9 in 80% acetonitrile), at a flow rate of 1 ml min^{-1}. Peptides were separated into 48 fractions, which were consolidated into 24 fractions. The fractions were subsequently dried, and the peptides were dissolved in 5% formic acid and analyzed by liquid chromatography–mass spectrometry.

5.3.15 TMT Liquid chromatography electrospray–tandem mass spectrometry analysis

For each fraction, 1 µg was analysed using an Orbitrap Fusion Tribrid mass spectrometer (Thermo Fisher Scientific) equipped with a Dionex ultra-high-pressure liquid chromatography system (RSLCnano). Reversed-phase liquid chromatography was performed using a Dionex RSLCnano high-performance liquid chromatography system (Thermo Fisher Scientific). Peptides were injected onto a 75 µm × 2 cm PepMap-C18 pre-column and resolved on a 75 µm × 50 cm RP C18 EASY-Spray temperature-controlled integrated column-emitter (Thermo Fisher Scientific) using a 4-h multistep gradient from 5% B to 35% B with a constant flow of 200 nl min^{-1}. The mobile phases were: 2% acetonitrile incorporating 0.1% formic acid (solvent A) and 80% acetonitrile incorporating 0.1% formic acid (solvent B). The spray was initiated by applying 2.5 kV to the EASY-Spray emitter, and the data were acquired under the control of Xcalibur software in a data-dependent mode using the top speed and 4 s duration per cycle. The survey scan was acquired in the Orbitrap covering the *m/z* range from 400–1,400 Thomson units (Th), with a mass resolution of 120,000 and an automatic gain control (AGC) target of 2.0×10^5 ions. The most intense ions were selected for fragmentation using collision-induced dissociation in the ion trap with 30% collision-induced dissociation energy and an isolation window of 1.6 Th. The AGC target was set to 1.0×10^4, with a maximum injection time of 70 ms and a dynamic exclusion of 80 s. During the MS3 analysis for more accurate TMT quantifications, ten fragment ions were co-isolated using synchronous precursor selection, a window of 2 Th and further fragmented using a higher-energy collisional dissociation energy of 55%. The fragments were then analyzed in the Orbitrap with a resolution of 60,000. The AGC target was set to 1.0×10^5 and the maximum injection time was set to 300 ms.

5.4 Results

5.4.1 The Encyclopedia of Proteome Dynamics (EPD)

5.4.1.1 Introduction

The Encyclopedia of Proteome Dynamics (EPD) is a web app created to share and explore high-throughput, curated proteomics datasets generated within the Lamond laboratory and close collaborators and annotated with consistent metadata, using carefully controlled vocabulary. It is a data ecosystem that has been available at https://peptracker.com/epd/ . Its goal was to provide a simple user interface to share the proteomic data with the scientific community in an accessible format.

The EPD's data ecosystem is composed of a series of noSQL (not only structured query language) databases, Cassandra and Neo4j, with a web application and server (see methods) and a graphical interface for users to interact with. The EPD is a unified online resource that combines processed MS-based proteomics data from many heterogenous large-scale studies on multiple human cell types(Mirauta et al., 2020), across multiple model organisms(Larance et al., 2015, Hukelmann et al., 2016) and a wide array of research areas like cell cycle analysis(Ly et al., 2015, Ly et al., 2017), protein turnover(Boisvert et al., 2012) and protein-protein interactions (PPI)(Kirkwood et al., 2013, Larance et al., 2016).

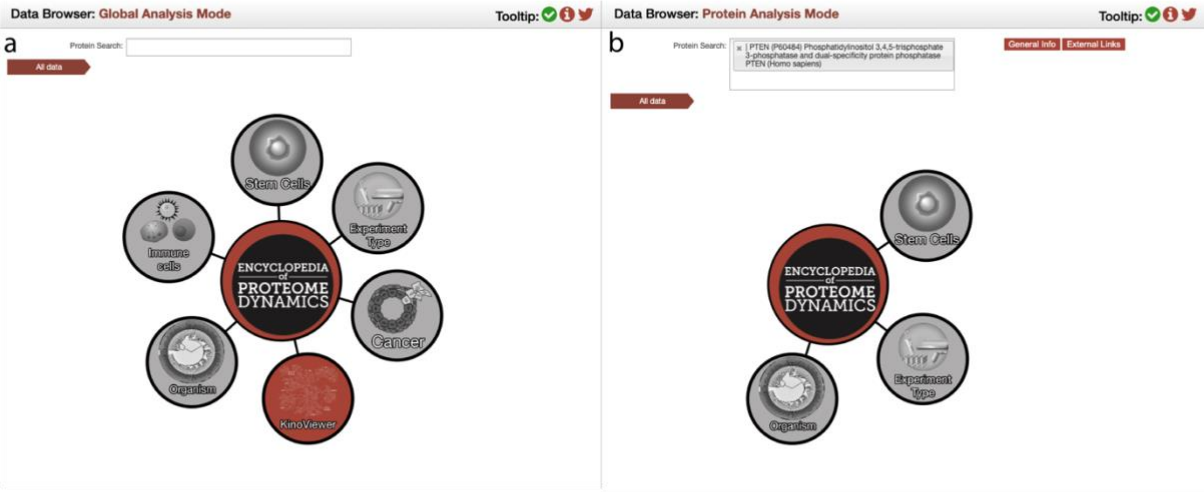
Data Browser: Global Analysis Mode
Tooltip:
a
Protein Search:
All data
Stem Cells
Experiment Type
Immune cells
ENCYCLOPEDIA of PROTEOME DYNAMICS
Cancer
Organism
KinoViewer
Data Browser: Protein Analysis Mode
Tooltip:
b
Protein Search:
× | PTEN (P60484) Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN (Homo sapiens)
General Info
External Links
All data
Stem Cells
ENCYCLOPEDIA of PROTEOME DYNAMICS
Experiment Type
Organism

Figure 5.1 – EPD navigation: ***(a)*** Schematic showing the global analysis navigation. This mode is enabled when no protein has been selected. ***(b)*** Schematic showing the Protein Analysis navigation, this mode is enabled after a protein is selected on the search box.

For users to interact with the EPD, the main interface dealing with data access and navigation is implemented as a force diagram which is a graphical representation of the graph structure stored in Neo4j. It allows users to interact with the system and datasets without requiring any technical expertise. The navigation has two main functionality modes, a global search (Fig. 5.1a) across all datasets even when no proteins are selected or a protein search and a protein search to drill down on specific results for a predetermined protein (Fig. 5.1b). The 'Global Analysis Mode', henceforth referred to as Global mode, provides access to overview plots which visualise aggregated elements within a dataset, like volcano plots and histograms. The 'Protein Analysis Mode', henceforth referred to as Protein mode, allows users to perform a search for a specific protein of interest using the search box. The protein search is implemented to work with the protein accession, gene name or protein description. Once a protein is selected, the Protein mode filters out any datasets where the protein of interest has not been detected (Fig. 5.1b).

5.4.1.2 Types of experiments and visualisations

The EPD hosts an array of experiments that have different methodologies and biological aims, ranging from proteomic characterisations of different cell cycle stages(Ly et al., 2015), protein complex studies using Size Exclusion Chromatography(Kirkwood et al., 2013, Larance et al., 2016) (SEC) to large-scale analysis of characterisation of human induced pluripotent stem cells(Mirauta et al., 2020). As such there are a wide array of different visualisations aimed at best representing the data from these heterogenous experiments.

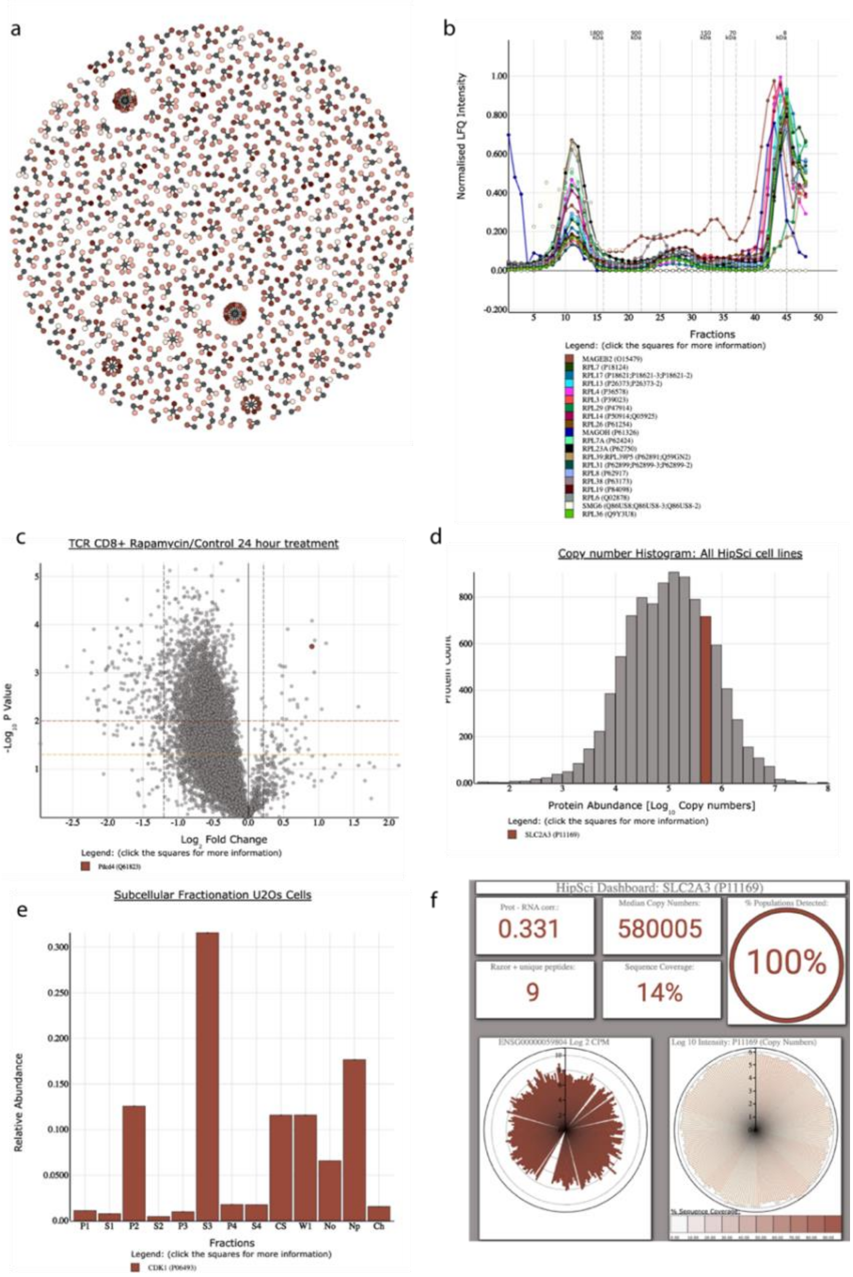

Figure 5.2 – EPD data visualisation: ***(a)*** Force diagram used in the Global Analysis Mode to show the results of a clustering analysis for the SEC-based protein complexes study. All the linked circles represent elements within the same cluster. ***(b)*** Line plot (available in both Global and Protein Analysis modes) used to show all the members of a predicted protein complex across the elution profile of a size exclusion chromatography-based mass spectrometry experiment. ***(c)*** Volcano plot (available in both Global and Protein Analysis modes) used to show the results of a differential expression analysis. In this case showing the estimated copy numbers in cytotoxic T cells with and without a 24-hour rapamycin treatment. ***(d)*** Histogram (available in both Global and Protein Analysis modes) used to show the comparative abundance of SLC2A3 compared to all proteins detected in human induced pluripotent stem cells. ***(e)*** Bar plot (available in both Global and Protein Analysis modes) used to show the abundance of CDK1 across different subcellular compartments. ***(f)*** Dashboard plot, only available in Protein Analysis Mode, showing the full data summary for SLC2A3 in the large-scale human induced pluripotent stem cell datasets.

Some plots within the EPD are only available in specific data modes either Global or Protein analysis mode. The force diagrams plots (not the ones used in the navigation) are used to represent the output of clustering analyses (Fig. 5.2a). This plot type is only available in Global mode and in this case shows the results of a dataset looking at formaldehyde crosslinked protein complexes(Larance et al., 2016). The plot represents a proteome wide clustering analysis of normalised SEC elution profiles, where clicking on the central black node on any of the clusters allows you to plot all proteins contained in the cluster on a line plot and display their normalised intensity across all the SEC fractions (Fig. 5.2b). Within this line plot, all elements can be removed from the plot by clicking on the colour box next to their name and additional proteins can be added by using the search box.

Volcano plots (Fig. 5.2c) are also a classical Global analysis plot, as they allow users to explore the differential expression analysis of all proteins detected in at least two replicated per condition within a dataset. Volcano plots are versatile and are also available when doing 'Protein Analysis Mode', in this case the selected protein would be highlighted in red as shown (Fig. 5.2c). Similarly, histograms are used in both Global and Protein modes, as they allow a general comparison of protein abundance and dynamic range when in Global mode, and a comparison of the abundance of a specific protein when on Protein mode (Fig. 5.2d). Bar plots are frequently used to represent the abundance of individual proteins in the Protein mode, in this case showing CDK1 abundance across subcellular compartments in U2OS cells (Fig. 5e). Finally, the most complex custom plot created on the EPD is the dashboard, which was developed for large-scale datasets like the HipSci dataset(Mirauta et al., 2020). The custom dashboard shows summary data, QC data as well as abundance across all replicates. In this case it is showing data for the hiPSC protein copy numbers and integrated Counts per million (CPM) from the transcriptomic data, while also displaying QC statistics like the sequence coverage and then number of unique + razor peptides. (Fig. 5.2f).

All plots contained within the EPD use interactive behaviour based on 'on-click events' by either displaying additional information or additional functionality. In some plots types, they allow you to highlight or add new elements to the plots. On click is also used to remove elements from the plots or obtain additional information about the selected protein, with interactive links into useful databases like UniProt(The UniProt, 2017).

5.4.1.3 The KinoViewer

The KinoViewer(Brenes and Lamond, 2019) is an extension of the EPD that was created to visualise kinase expression from proteomic or transcriptomic data. Protein kinases are a class of enzymes that catalyse the transfer of the gamma phosphate group from ATP onto specific hydroxyl groups on amino acid sidechains. The site-specific phosphorylation of protein substrates can drastically alter their function, by changing, for example, either their activity and/or stability. The protein kinase family is also of major clinical relevance, with over 240 kinase inhibitors that are either already approved drugs, or else involved in clinical trials(Klaeger et al., 2017).

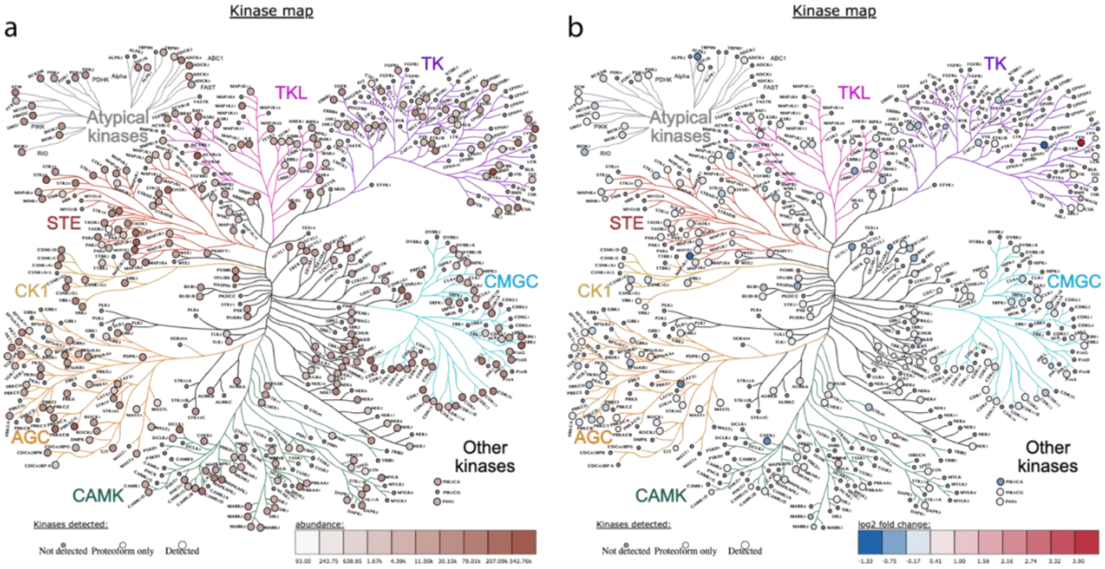

Figure 5.3 – The KinoViewer: ***(a)*** KinoViewer schematic showing the abundance map of all detected kinases derived from a proteomics dataset. The abundance measure is dependent on the input from the user. ***(b)*** KinoViewer schematic showing the differential expression map for all detected kinases derived from a proteomics dataset.

The KinoViewer is the only section within the EPD that allowed users to upload external data to create a custom visualisation. It is integrated into the EPD and has been available online at https://peptracker.com/epd/analytics/ by clicking on the red KinoViewer node. This opens a specific graphical interface with a text box that allows users to paste quantitative proteomic, or transcriptomic data. The KinoViewer then uses this quantitative data to produce either a single colour gradient abundance map (Fig. 5.3a) based on the intensities, protein copy numbers or transcript-based counts or a diverging colour scheme differential expression map (Fig. 5.3b) based on fold changes. The type of plot depends on the formatting and type of data that is provided, with additional information about how to use The KinoViewer provided when clicking on the “i” button on the KinoViewer graphical interface.

5.4.2 The Immunological Proteome Resource (ImmPRes)

5.4.2.1 Introduction

The immune system is a vast network of interacting cells that function to protect organisms against foreign pathogens. Understanding the roles of different cells, signalling pathways and proteins within this system is vital to understanding health and disease. In this context, high quality RNA based resources(Heng et al., 2008, Immunological Genome, 2020, Uhlen et al., 2019) have played vital roles in defining expression levels of RNAs across bulk and single cell immune populations. However, proteins are the molecules that form the backbone of cell structure and control virtually all metabolic processes and regulatory mechanisms(Larance and Lamond, 2015). The proteome of a cell is a dynamic system that is constantly modulated by changes in rates of protein synthesis and degradation(Marchingo and Cantrell, 2022). mRNA levels are thus not effective predictors of protein abundance(Liu et al., 2016, Buccitelli and Selbach, 2020), and quantitative characterisation of cellular proteomes versus transcriptomes has been shown to provide invaluable information about the shaping of lymphocyte identity(Hukelmann et al., 2016).

Mass spectrometry instrumentation, sample processing and software level breakthroughs have enabled the comprehensive annotation of the proteome in a scalable and cost-effective manner(Bekker-Jensen et al., 2017). Thus, defining the proteome of hematopoietic cells in steady state, as well as in response to different conditions such as T cell receptor activation or specific cytokines has become an attainable project of great potential benefit to the immunological community. A resource looking at a select number of human hematopoietic populations at the protein level was recently developed(Rieckmann et al., 2017), but there is no current resource exploring the proteomes of murine leukocyte populations. Mice are widely accepted to be excellent models to study the mammalian immune system(Phifer-Rixey and Nachman, 2015) and have played a vital role in immunology research. The tractability of mouse genomes means that mice have been used extensively to explore the importance of different immune cell populations and different immune regulatory molecules. As such, resources mapping the proteomic profile of murine hematopoietic cells and in particular mouse T cell populations would be of great value to the community.

To achieve this goal, we created The Immunological Proteome Resource (ImmPRes; http://immpres.co.uk/), an open access public resource integrating proteomic data generated by large-scale mass-spectrometry analysis of murine hematopoietic populations, with a future aim to expand into human populations. It is a multidisciplinary collaborative effort between immunology and mass spectrometry-based labs with the objective to help define an in-depth high-quality map of the immune proteome.

One of the main goals of ImmPRes is to provide open access to all proteomic data, ranging from the raw MS files to the processed copy number summaries. It does so by providing access to a simple graphical interface designed to interact with the different proteomic datasets, and to download the raw data for large-scale reanalysis. All MS raw files are uploaded to Proteomics Identifications Database(Perez-Riverol et al., 2018, Perez-Riverol et al., 2022) (PRIDE) where they can be downloaded and reanalysed, while the processed data are made available for download directly on ImmPRes.

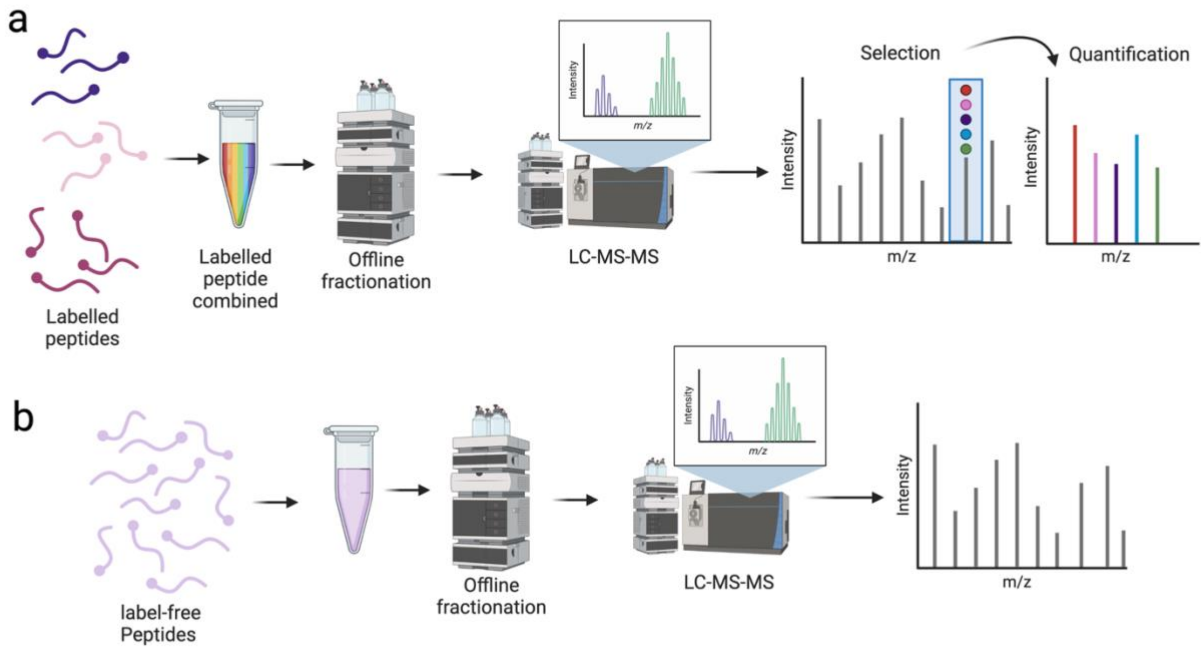

Figure 5.4- Mass spectrometry workflows: ***(a)*** Schematic showing the proteomic TMT-based proteomic workflow which used SP3 for sample processing, intensive offline fractionation in the HPLC and synchronous precursor selection at the MS3 within the orbitrap Fusion. ***(b)*** Schematic showing the proteomic label free proteomic workflow which used SP3 for sample processing, intensive offline fractionation in the HPLC and synchronous precursor selection at the MS3 within the orbitrap Fusion

Data reproducibility and integrity are a priority within the resource. As such there is an in-depth protocols section which explains in detail both the sample processing as well as the mass spectrometry analysis. Furthermore, the sample preparation steps for each of the different hematopoietic populations are exhaustively documented within the protocols section. Specific details on the mouse strains, tissue preparation protocols, growth media, cytokines added, purification protocols and activation details (where applicable) are available within the protocols section at http://immpres.co.uk/ .

The current mass spectrometry-based datasets have been acquired in Data Dependent Acquisition (DDA) mode using either an SP3(Hughes et al., 2014) or urea based sample preparation coupled with isobaric labelling using Tandem Mass Tags(Thompson et al., 2003) (TMT) (Fig. 5.4a), or a label-free strategy (Fig. 5.4b). All the DDA data were acquired with extensive fractionating of the samples for an in-depth overview and were analysed with a rigorous false discovery rate and with data imputation disabled. Furthermore, all datasets containing heterogenous populations were searched without using "match between runs" to avoid false identifications being propagated(Lim et al., 2019). For the analytic side, the "proteomic ruler"(Wisniewski et al., 2014) method was used for the normalisation of all current datasets. This method uses the mass spectrometry signal of histones as an internal standard which avoids the error prone steps of cell counting and enables the estimation of protein copy numbers per cell. Knowledge of protein copy number is invaluable for a full understanding of cell phenotypes. For example, it allows exploration of the stoichiometry of proteins within protein complexes and provides an abundance measure with direct biological meaning to the end user.

5.4.2.2 ImmPRes datasets overview

ImmPRes contains multiple large-scale proteomics datasets containing hundreds of raw files and identifying thousands of proteins. The initial resource has integrated numerous datasets (Fig. 5.5), across different hematopoietic populations and across different conditions. The data are readily available to be browsed and explored using the graphical interface.

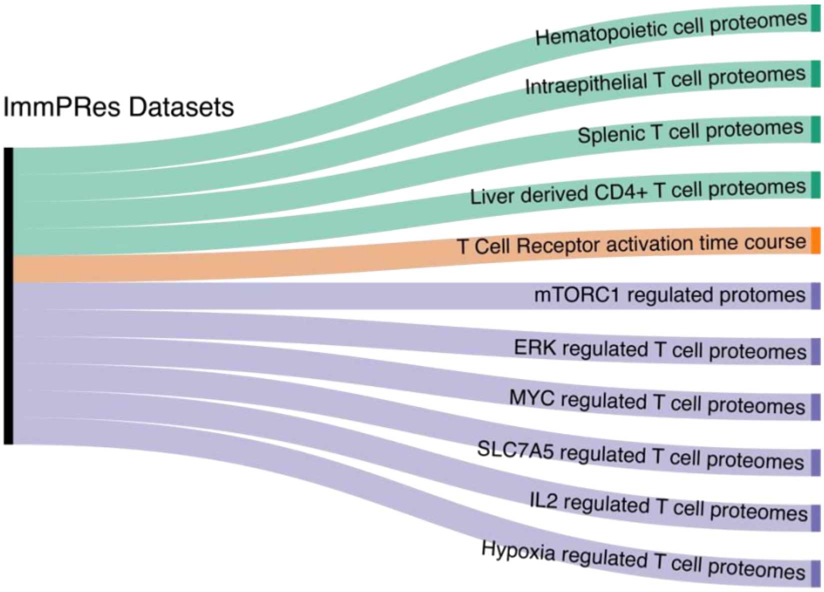
ImmPRes Datasets
Hematopoietic cell proteomes
Intraepithelial T cell proteomes
Splenic T cell proteomes
Liver derived CD4+ T cell proteomes
T Cell Receptor activation time course
mTORC1 regulated protomes
ERK regulated T cell proteomes
MYC regulated T cell proteomes
SLC7A5 regulated T cell proteomes
IL2 regulated T cell proteomes
Hypoxia regulated T cell proteomes

Figure 5.5- Sankey diagram showing all the different mass spectrometry-based datasets that have been integrated into ImmPRes.

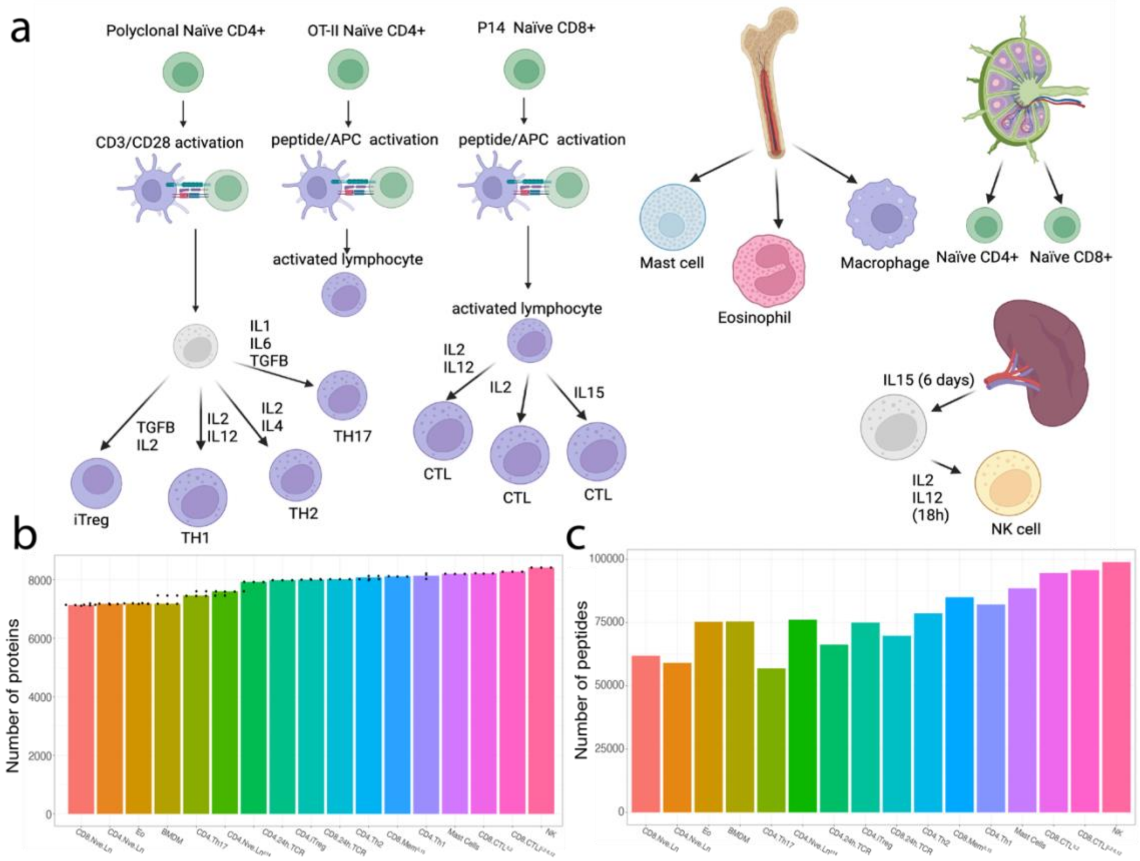

Figure 5.6 - Primary data record: ***(a)*** Schematic showing the different populations which are contained within the 'Hematopoietic cell proteomes' data record. ***(b)*** Bar plot showing the number of proteins identified for all populations. ***(c)*** Bar plot showing the number of peptides identified for all populations.

Hematopoietic cell proteomes

The **'Hematopoietic cell proteomes'** is the largest and most in-depth dataset hosted on ImmPRes. It is a TMT based characterization of 16 different immune cell populations, including 8 previously unpublished populations (Fig. 5.6a). This large dataset contains naïve CD4+ and CD8+ T cell subpopulations isolated from lymph nodes, multiple in vitro generated effector T cell subsets, in vitro generated regulatory T cells as well as innate leukocytes such as mast cells and macrophages(Howden et al., 2019). The in-depth coverage meant that over 10,000 proteins were identified across all populations, without using matching or imputation (Fig. 5.6b). Furthermore, all populations had a peptide coverage greater than 55,000 peptides and ranging up to almost 99,000 peptides (Fig. 3c). The raw files and MaxQuant output for this dataset are available in PRIDE with identifier PXD012058(Brenes and Cantrell, 2018) and PXD020091(Brenes and Cantrell, 2020) while the processed copy numbers are available in the 'Downloads' tab in ImmPRes.

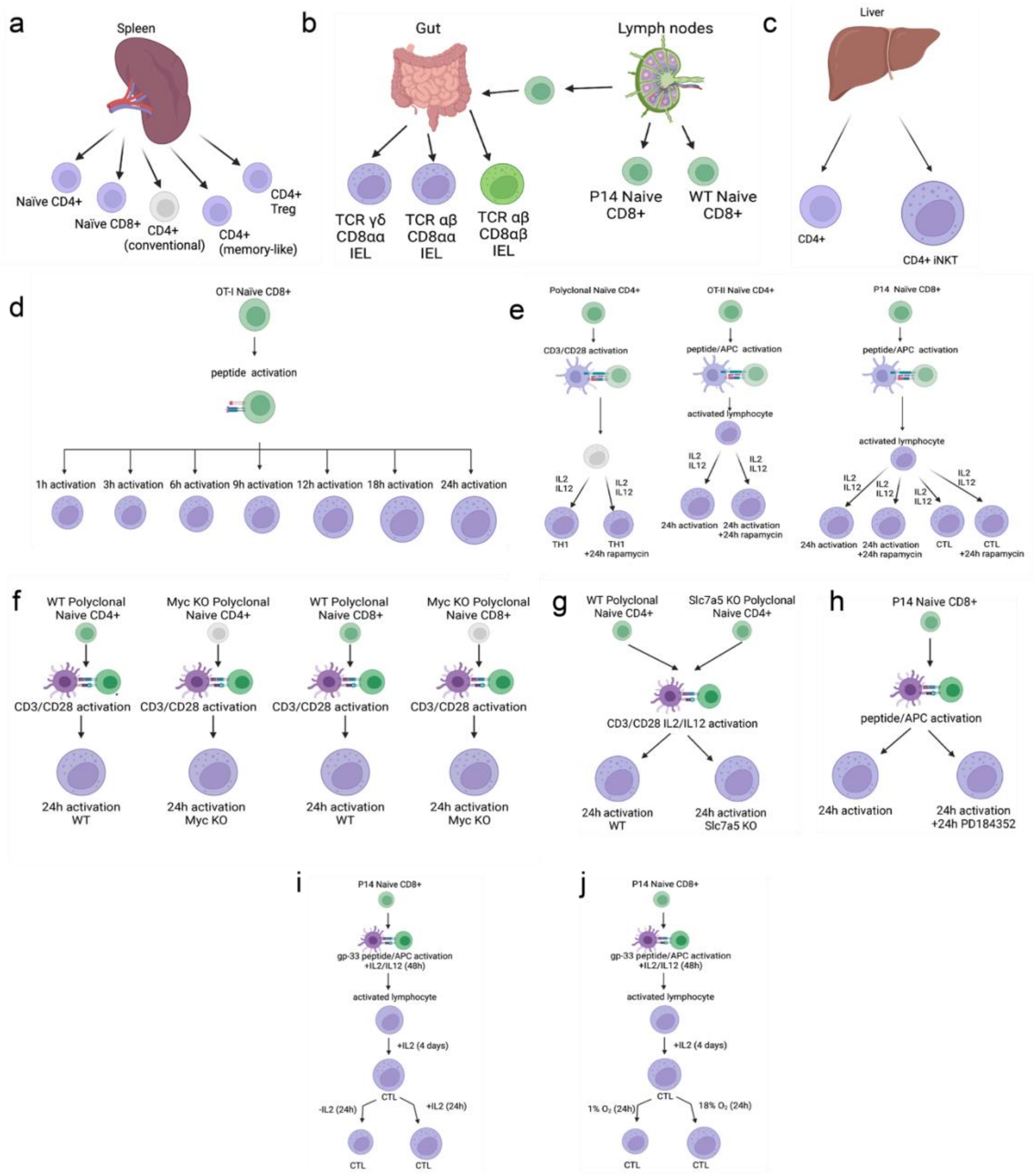

Figure 5.7- Data records overview: Schematics representing: ***(a)*** Populations contained in the 'Splenic T cell proteomes' data record**.** ***(b)*** Populations which are contained within the 'Intraepithelial T lymphocyte proteomes.' data record. ***(c)*** Populations contained within the 'Liver derived CD4+ T cell proteomes.' data record. ***(d)*** Populations contained within the 'T Cell Receptor activation time course.' data record. ***(e)*** Populations and treatments contained within the 'Mtorc1 regulated proteome' data record**.** ***(f)*** Populations contained within the 'Myc regulated proteomes' data record. ***(g)*** Populations contained within the 'Slc7a5 regulated proteomes' data record. ***(h)*** Populations and treatments contained within the 'Erk regulated proteomes' data record. ***(i)*** Populations and treatments contained within the 'IL2 regulated proteomes' data record. ***(j)*** Populations and treatments contained within the 'Hypoxia regulated proteomes' data record.

Splenic T cell proteomes

T cells in the spleen are frequently used to study T cell biology. ImmPRes includes a TMT-based **LC-MS** dataset**,** labelled as **'Splenic T cell proteomes'** on the Data Browser, characterising the proteomes of ex-vivo splenic T cells, including naïve CD8+ T cells, naïve CD4+ T cells, memory like CD4+ T cells and CD4+ regulatory T cells (Fig. 5.7a). The raw files and MaxQuant output for this dataset are available in PRIDE with identifier PXD020091(Brenes and Cantrell, 2020), while the processed copy numbers are available in the 'Downloads' tab in ImmPRes.

Intraepithelial T lymphocyte proteomes

Intestinal intraepithelial lymphocytes (IEL) comprise a distinct group of innate-like and memory T cells that collectively form one of the largest T cell compartments in the body. ImmPRes includes a TMT-based **LC-MS** dataset exploring the proteomes of intraepithelial T lymphocytes (T-IELs)**,** labelled as **'Intraepithelial T cell proteomes'** on the Data Browser (Fig. 5.7b). The dataset characterises the proteome of TCRαβ CD8αβ, TCRαβ CD8αα and TCRγδ CD8αα T-IELS along with two conventional TCRαβ CD8αβ lymph node derived naïve CD8+ T cell populations, wild type (WT) and P14**(Brenes et al., 2021a)**. The raw files and MaxQuant output for this dataset are available in PRIDE with identifier PXD023140**(Brenes and Lamond, 2020)**, while the processed copy numbers are available in the 'Downloads' tab in ImmPRes.

Liver derived CD4+ T cell proteomes

T cells play a critical role in liver immunity and take part both in the initiation and in the resolution of intrahepatic inflammation(Ficht and Iannacone, 2020). The liver contains conventional CD4 T cells, and Natural Killer T (NKT) cells that express an invariant Vα14 T cell receptor that recognizes glycolipid/CD1d antigen complexes (iNKTs) and play a role in immune surveillance and immune homeostasis**(Ficht and Iannacone, 2020)**. ImmPRes includes a TMT-based **LC-MS** dataset**,** labelled as **'Liver derived CD4+ T cell proteomes'** on the Data Browser, characterising the proteomes of ex-vivo liver derived CD4+ T cells along with iNKT cells (Fig. 5.7c). The raw files and MaxQuant output for this dataset are available in PRIDE with identifier PXD036319(Brenes and Cantrell, 2022), while the processed copy numbers are available in the 'Downloads' tab in ImmPRes.

Time course analysis of CD8+ T cell proteome remodelling in response to T Cell Receptor (TCR) engagement

A label-free **LC-MS** dataset, labelled as **'T Cell Receptor activation timecourse'** on the Data Browser, analysing the proteome of CD8+ naïve T cells responding to cognate antigen over a time course (Fig. 5.7d). The T cells were CD8+ cells expressing a TCR complex that recognises the ovalbumin peptide SIINFEKL in the context of H2K^{b} (OT1-T cells). The time course data was collected at 0, 1, 3, 6, 9, 12, 18 and 24 hours of activation of OT-1 T cells with SIINFEKL. The raw files and MaxQuant output for this dataset are available in PRIDE with identifier PXD016443(Marchingo and Cantrell, 2020), while the processed copy numbers are available in the 'Downloads' tab in ImmPRes.

mTORC1 regulated lymphocyte proteomes

One key signalling molecule that controls protein turnover in mammalian cells is the nutrient sensing protein kinase mammalian target of rapamycin complex 1 (mTORC1)**(Kim et al., 2002)**. ImmPRes contains a TMT-based **LC-MS** dataset, labelled as **'mTORC1 regulated proteomes'** on the Data Browser, which explores the effect of how rapamycin, an inhibitor of mTORC1 reshapes T cell proteomes (Fig. 5.7e). The data compares how mTORC1 inhibition impacts the immediate response of naïve CD4^{+} and CD8^{+} T cells to antigen, versus how 24 hours of mTORC1 inhibition reshapes differentiated effector CD4^{+} TH1cells and CD8^{+} T cytotoxic T lymphocytes (CTL)**(Howden et al., 2019)**. The impact of mTORC1 inhibition on six additional populations are set to be released in the near future. The raw files and MaxQuant output for this dataset are available in PRIDE with identifier PXD012058**(Brenes and Cantrell, 2018)** and PXD020091**(Brenes and Cantrell, 2020)**, while the processed copy numbers are available in the 'Downloads' tab in ImmPRes.

Myc control of the proteomic landscape of immune activated CD4+ and CD8+ T cells

T cell expansion and differentiation are critically dependent on the transcription factor Myc(Marchingo et al., 2020). ImmPRes contains a label-free **LC-MS** dataset, labelled as **'Myc regulated T cell proteomes'** on the Data Browser, studying the effects of Myc-deletion on the proteomes of immune activated CD4+ and CD8+ T cells (Fig. 5.7f). The dataset comprises wild type CD4+ and CD8+ naïve T cells, along with wild type and Myc null CD4+ and CD8+ T cells polyclonally activated with CD3 and CD28 antibodies(Marchingo et al., 2020). The raw files and MaxQuant output for this dataset are available in PRIDE with identifier PXD016105(Marchingo and Cantrell, 2019), while the processed copy numbers are available in the 'Downloads' tab in ImmPRes.

Analysis of how the amino acid transporter SLC7A5 fuels CD4+ T cell proteomes

T lymphocytes regulate nutrient uptake to meet the metabolic demands of an immune response, thus they respond to antigen by upregulating expression of many amino-acid transporters including SLC7A5, the System L ('leucine-preferring system') transporter. Slc7a5 transports large neutral amino acids into T cells and is the dominant methionine transporter(Sinclair et al., 2013), accordingly SLC7A5null T cells cannot proliferate or differentiate in response to antigen. ImmPRes includes a label-free **LC-MS** dataset, labelled as **'SLC7A5 regulated T cell proteomes'** on the Data Browser, showing the effects of Slc7a5-deficiency on the proteomes of immune activated CD4+ T cells (Fig. 5.7g). The dataset compares WT naïve CD4+ T cells and wild type and SLC7a5 null CD4+ T cells polyclonally activated with CD3 and CD28 antibodies in the presence of cytokines interleukin 2 (IL2) and IL12(Marchingo et al., 2020). The raw files and MaxQuant output for this dataset are available in PRIDE with identifier PXD016105(Marchingo and Cantrell, 2019), while the processed copy numbers are available in the 'Downloads' tab in ImmPRes.

Analysis of how antigen receptor driven proteome restructuring in CD8+ T cells is regulated by extracellular signal-regulated kinases (ERKs)

A key T cell signalling module is mediated by ERK1/2 serine/threonine kinases which are activated in response to antigen receptor engagement. ImmPRes contains a label-free **LC-MS** dataset, labelled as **'ERK regulated T cell proteomes'** on the Data Browser, showing the effect of inhibiting ERK activity on antigen receptor driven proteome restructuring in CD8+ T cells (Fig. 5.7h). The dataset compares naïve CD8+ T cells and $CD8^+$ T cells expressing antigen receptors specific for lymphocytic choriomeningitis virus glycoprotein peptide gp33-41 activated for 24 hours with the gp33 peptide in the presence or absence of the kinase inhibitor PD184352 which prevents activation of ERK1/2(Damasio et al., 2021). The raw files and MaxQuant output for this dataset are available in PRIDE with identifier PXD023256(Howden and Cantrell, 2020), while the processed copy numbers are available in the 'Downloads' tab in ImmPRes.

Analysis of how Interleukin 2 controls the proteome of CD8 cytotoxic T cells

Interleukin-2 (IL-2) regulates transcriptional programs and protein synthesis to promote the differentiation of effector $CD8^+$ cytotoxic T lymphocytes (CTLs)(Ross and Cantrell, 2018). ImmPRes contains a label-free **LC-MS** dataset, labelled as **'IL2 regulated T cell**

proteomes' on the Data Browser, showing the effect of IL2 on CTL expressing T cell receptor specific for the gp-33 peptide from lymphocytic choriomeningitis virus (Fig. 5.7h). P14 naïve CD8+ T cells were activated with the gp-33 peptide for 48 hours and cultured in IL2 for 4 days to generate CTL. Label-free high-resolution mass spectrometry was then used to compare the proteome of CTLs maintained in IL-2 and CTLs deprived of IL-2 for 24 hours(Rollings et al., 2018). The raw files and MaxQuant output for this dataset are available in PRIDE with identifier PXD008112(Ross and Cantrell, 2017), while the processed copy numbers are available in the 'Downloads' tab in ImmPRes.

Analysis of hypoxia induced remodelling of cytotoxic T cell proteomes

During immune responses T cells must function in oxygen-deficient, or hypoxic, environments. ImmPRes contains a label-free **LC-MS** dataset, labelled as 'Hypoxia regulated T cell proteomes' on the Data Browser, showing the effect of 24 hours of hypoxia (1% oxygen saturation) on the proteomes of CTL (Fig. 5.7i). P14 naïve CD8+ T cells were activated with the gp-33 peptide for 48 hours and cultured in IL2 for 4 days to generate CTL. A population of CTLs was then maintained in hypoxic conditions (1% oxygen saturation) for 24 hours, while another was maintained in normoxic conditions (18% oxygen saturation) for 24 hours(Ross et al., 2021). The raw files and MaxQuant output for this dataset are available in PRIDE with identifier PXD026223(Ross and Cantrell, 2021), while the processed copy numbers are available in the 'Downloads' tab in ImmPRes.

5.4.3 Functionality and usage notes

All data hosted on ImmPRes are open access under CC BY 4.0 terms. As a web resource with a user-friendly graphical interface, users can easily generate interactive plots. This functionality is present for all datasets and can be found under the 'Data browser' tab. Within the Data browser the first sub tab labelled 'Datasets' provides a graphical representation of the different collections of data that have been integrated into ImmPRes, these are also shown in Figures 5.5-5.7.

The 'Copy numbers' tab (Fig. 5.8a) is a simple way to explore the estimated copy numbers for a particular protein of interest in a specific dataset. This tab has multiple dropdowns with important functionality, the first dropdown is labelled 'Dataset selection;' and it allows users to select which datasets they want to visualise from all available options in ImmPRes. This field is particularly informative as it will display the dataset name, i.e. ‘Hematopoietic cell proteomes’, it will also state the acquisition method used for the dataset, DDA (data dependent acquisition) or DIA (data independent acquisition), and it will show the labelling strategy used, be it Tandem Mass Tags or Label free. The ‘Gene search:’ dropdown allows users to search for their protein of interest using the standard Gene name, the default option shows the results for ‘Pten’. Additionally, some datasets allow the user to filter the results by selecting specific populations, i.e. within the ‘Hematopoietic cell proteomes’ dataset users can filter the results to show only ‘CD8+ T cells’.

The ’Concentrations’ tab (Fig. 5.8b) mimics the previously described behaviour, but instead of showing the estimated copy numbers, it will show a concentration-like measure (copy numbers of a protein divided by the total copy numbers), providing a measurement that normalises differences in cell size, it is an alternative way to visualise the expression levels of proteins across the datasets. The next tab within the Data Browser is the ‘Differential expression’ tab (Fig. 5.8c). This tab provides access to a volcano plot; a plot which shows the fold change and p-value when comparing two specific conditions. The volcano plots are available for some datasets and for some specific comparisons. Like the two previous tabs, the Differential expression tab has a ‘Dataset selection’ dropdown, a ‘Gene search’ dropdown and a ‘Population search’ dropdown. Users can download the results shown on the volcano plot by clicking on the ‘Download Selected Data’ button. Users can also select a segment of the volcano by pressing and holding the left click over the desired section in the plot and can download this subset of proteins by pressing the ‘Download Selected Data’ button.

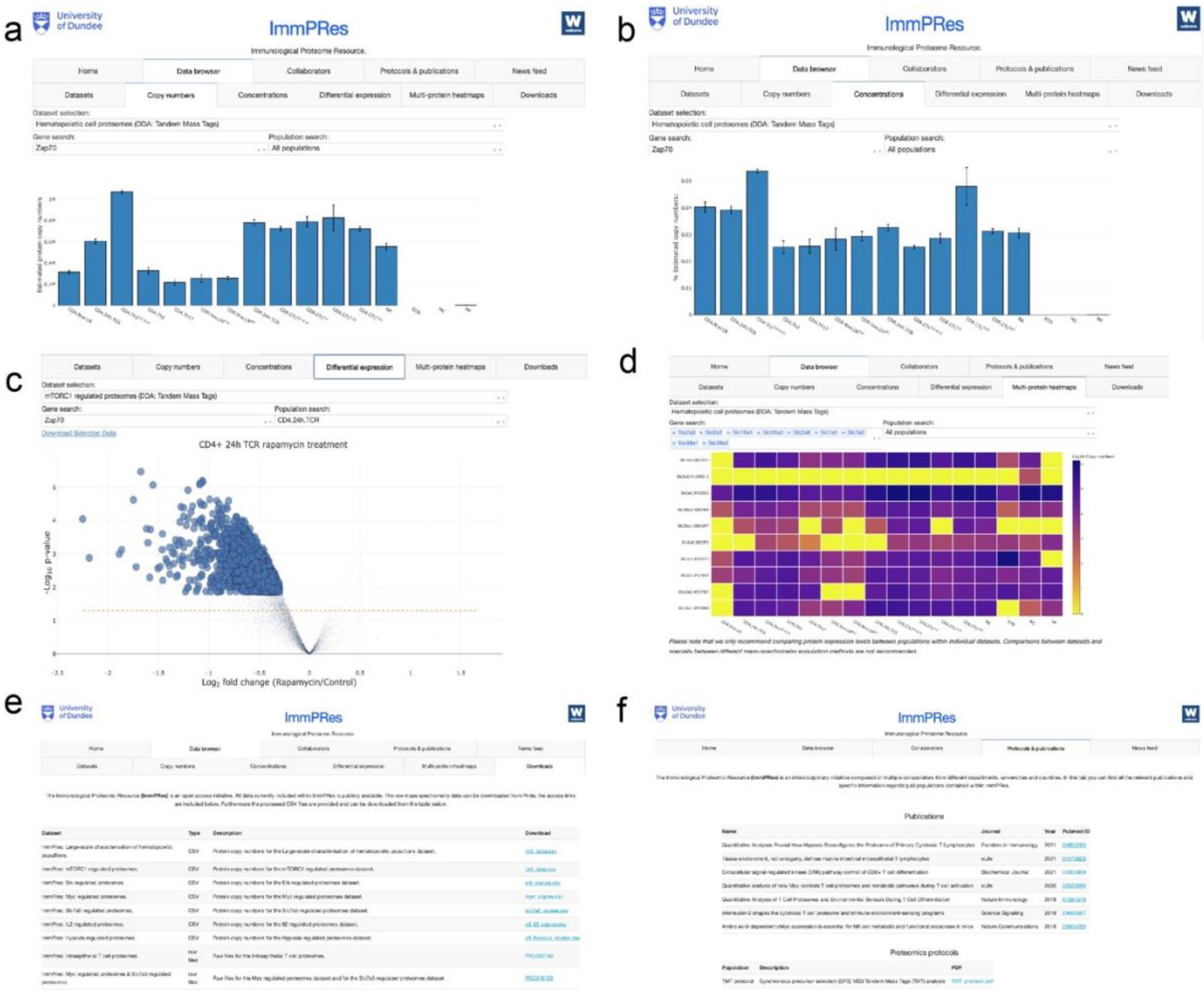

Figure 5.8- ImmPres usage and functionality: ***(a)*** Snapshot showing the 'Copy numbers' tab within ImmPRes. Different proteins can be searched by typing their gene name within the 'Gene Search' box. ***(b)*** Snapshot showing the concentrations tab within ImmPRes. Different proteins can be searched by typing their gene name within the 'Gene Search' box. ***(c)*** Snapshot showing the 'Differential expression' tab within ImmPRes. Different proteins can be searched by typing their gene name within the 'Gene Search' box. ***(d)*** Snapshot showing the 'Multi-protein heatmaps' tab within ImmPRes. Different proteins can be searched by typing their gene name within the 'Gene Search' box. ***(e)*** Snapshot showing the 'Downloads tab within ImmPRes ***(f)*** Snapshot showing the 'Protocols & publications' tab within ImmPRes

The final interactive tab is the 'Multi-protein heatmaps' (Fig. 5.8d). The functionality of the tab is the same as described above, except for the 'Gene search' dropdown, which unlike previous tabs allows the users to search for multiple genes at the same time. Every time a new protein is selected it will update the heatmap to also include the new selection. The heatmap itself shows the $\log_{10}$ converted protein copy numbers for the selected proteins and populations within the selected dataset.

The Data browser also provides access to the data should users wish to explore any dataset outside of ImmPRes. Within the 'Downloads' tab (Fig. 5.8e) there is a table that lists all the processed data, with comma separated value (CSV) files to download the estimated copy numbers for all datasets, as well as links to PRIDE(Perez-Riverol et al., 2018) which allow users to download the raw MS files for reanalysis. The final tab of relevance to usage is the ‘Protocols & publications’ tab (Fig. 5.8f). Here the users can find links to all publications linked to datasets currently contained within ImmPRes. In this same tab all the important details related the processing of the cells, the mass spectrometry and the analysis are also available for download as a PDF document.

Overarching discussion: contributions, limitations, challenges and opportunities for large and population-scale proteomics

Proteomic characterisations of different cell types, tissues and organisms have enabled scientific breakthroughs for over two decades. However, until a few years ago proteomic studies struggled with scalability and throughput to handle truly large population-scale datasets. We are now at a stage where this obstacle is no longer insurmountable. Improvements in instrument speed and sensitivity, acquisition methods, sample processing, software and labelling techniques have empowered large-scale studies comprising of hundreds or thousands of samples from patients or cell lines(Long et al., 2022, Brenes et al., 2021b, Mirauta et al., 2020, Nusinow et al., 2020, Goncalves et al., 2022, Poulos et al., 2020) and will likely move into the tens of thousands of samples per study in the short run. In this thesis we present two large-scale studies, one analysing data from over 200 hiPSC lines(Brenes et al., 2021b) and the other neutrophils from over 200 patients(Long et al., 2022) and show that by leveraging large-scale datasets we can extract new and exciting biological insights ranging from new molecular phenotypes, to biomarkers and potential new treatments. However, the first thing this work underscored are the challenges and difficulties of large-scale proteomics, as well as highlighting the importance of meticulous planning and understanding the strengths and limitations of the different proteomic methodologies that could be leveraged within these studies. Large-scale studies can suffer from batch effects at sample collection, processing and/or at the mass spectrometry stage and this holds true for both DDA and DIA, labelled or label free. Starting with labelled DDA, this work was the first, to the best of our knowledge, to highlight the issues relating to large-scale multi-batch TMT. The study showed that multi-batch TMT suffered from increased prevalence of missing values, as is seen in label free DDA studies. As the number of TMT batches increases, so do the number of peptide and protein missing values. This was one of the limitations of DDA that TMT seemed to resolve, however it does not hold true if multiple batches are needed.

It also confirmed clear and pronounced batch effects in the quantification that were specific to the TMT experiments. These effects can have severe implications, where different samples would cluster by TMT batch, irrespective of cell type, tissue or condition. These effects emphasise the importance of having an internal reference sample (IRS) within every TMT batch for data normalisation purposes. The IRS has been previously shown to be effective at providing some level of normalisation for TMT data (Plubell et al., 2017, Liu et al., 2017, Herbrich et al., 2013), as it enables the correction of the abundance levels for each protein across the different batches. Therefore, it is vitally important to ensure that the IRS is representative for all samples that will be analysed. As such it is recommended that the IRS be composed of a mixture of all samples that will be included in the project, which implies a requirement to have access to all samples before the first TMT batch is analysed. This is not always possible, especially with clinical data, and it makes adding unplanned extensions, such as adding new tissues at a later stage, incompatible with the normalisation strategy.

Furthermore, it is not trivial to create an IRS that is suitable for large heterogenous projects looking at different cell lines, cell types or tissues. This type of project would require a pooled IRS that would contain proteins from all the different samples in the study. Due to the lack of missing values within a TMT batch, this pooled heterogenous IRS would introduce false positives (understood as correct peptide identifications that are detected in populations where they shouldn't be) within the individual batches. This is particularly challenging for single cell data, where heterogenous populations are the norm and where false positives can have profound implications for the biological interpretation of the data. Furthermore, creating an effective IRS for single cell analysis has additional complications, as it is not possible to combine material from the single cells that will be analysed, and creating a combined IRS created from bulk samples can contain 10x if not >100x the loading material of the individual cells, which can then have a profound effect in quantitative accuracy(Ye et al., 2022).

Multi-batch TMT also requires previous knowledge of all the analytic goals and specific comparisons that will be made, as it will require the samples to be allocated uniformly across batches to enable these comparisons. Even when using IRS based normalisation, if the samples are not properly distributed across the different batches it makes disentangling biological differences from the technical artifacts challenging. Our study(Mirauta et al., 2020) inappropriately overrepresented disease samples, hiPSCs derived from patients with Bardet-Biedl syndrome and monogenic diabetes, in a small number of TMT batches sometimes using 8/10 replicates in a 10-plex for the disease state, instead of spreading them uniformly across batches. This made it virtually impossible to disentangle the biological signature from the technical noise.

In summary, despite all the previously mentioned challenges and limitations, multi-batch TMT, especially with the implementation of TMTpro 18-plex, is a viable DDA based alternative for large-scale projects and is currently one of the main methods used to analyse single cell proteomes(Specht et al., 2021). It was precisely with large-scale TMT that we revealed new biological insights in healthy human iPSCs. By analysing a large number of both male and female hiPSC, we were able to explore patterns in the data which revealed that the female populations could be effectively stratified, and it highlighted important differences between the two stratified populations. The stratification axis was centred around the X chromosome inactivation (XCI) state of the cells, where the cells in one population maintained robust XCI and the other suffered widespread erosion of XCI. Our in-depth large-scale proteomic data facilitated the direct comparison of these two populations, which would not have been effective with a small-scale study. The data showed that the proteome and transcriptome have good concordance when analysing the effects of erosion of X chromosome inactivation on X-linked gene products, suggesting a transcriptionally driven mechanism. However, the proteomic data, enabled by the estimated copy numbers calculation derived from the proteomic ruler(Wisniewski et al., 2014), highlighted that the proteomes of the eroded female lines also had significantly higher protein content compared to lines with robust XCI and compared to males. This effectively meant there was a loss of dosage compensation. The data showed that over 2,000 proteins were significantly more abundant in lines with an eroded X chromosome and only ~1% decreased in abundance, an effect which was not mirrored in the transcriptomic data.

We also leveraged a second large-scale proteomic dataset, this time a label-free data independent acquisition-based approach exploring the neutrophil proteomes of control and COVID19 patients. These data revealed new disease phenotypes by characterising changes in the neutrophil proteomes of COVID19 patients both at the early onset of disease as well as in delayed recovery. We had published a similar small-scale dataset looking at the neutrophils of COVID19 patients (Reyes et al., 2021) and discovered important neutrophil phenotypes that were validated within our larger study, like the presence of a type I interferon response (IFN-I) in the neutrophil proteomes of severe COVID19 patients and the increased glycolytic potential seen in the neutrophils of critically severe COVID19 patients in the acute infection phase.

However, it was only via the large-scale analysis that we discovered this IFN-I signature was present in all stratified patient groups at the acute stage, irrespective of disease severity, and that the IFN-I signature displayed important temporal dynamics. Severe and critically severe patients COVID19 patients had significant reductions in the abundance of proteins that were part of this important antiviral signature 1 week after enrolment in the study. The longitudinal analysis also highlighted that some important pattern recognition and inhibitory receptors were elevated at day 1 in critically severe patients, and were also elevated at day 7, including TLR2 and Vista. Both discoveries, enabled by the larger-scale proteomic study, have the potential to facilitate new treatments of COVID19 patients. This discovery of the divergence in IFN-I signalling at day 7 between moderate and critically severe patients means that a stratified treatment strategy could be explored. The proteomics data identified MX1 as the clearest marker for a IFN-I response in neutrophils, which ties in with what has been described in multiple sclerosis (van der Voort et al., 2009) . Conveniently a lateral flow test to measure MX1 abundance already exists, and could be used to identify patients lacking an IFN-I signature(Juntunen et al., 2017). As such critically severe patients without an IFN-I signature could be offered IFNβ treatment while avoiding the treating patients with a robust IFN signature, thus avoiding excessive inflammation. Furthermore, the increased abundance of Vista and TLR2 also offers therapeutic potential. The large proteomic dataset highlighted that both markers are significantly increased in severe and critically severe patients during the acute infection phase, but that are also elevated one week post enrolment. This suggests that both receptors could be robust targets to limit neutrophils inflammation across multiple phases in disease progression.

The proteomic analysis also discovered a promising biomarker to identify severe COVID19. PRG2, an eosinophil specific protein, was significantly reduced in abundance in COVID19 patients compared to controls. The reduction was most pronounced in the critically severe patients, thus correlating significantly with disease severity. Patient groups that displayed significantly reduced PRG2 abundance, also saw parallel reductions in eosinophil counts. These results corroborate what has already been described in the literature, as severe COVID19 patients had been shown to suffer from eosinopenia(Xie et al., 2021). However, the clinical eosinophil counts proved to be an unreliable measure that lacked the sensitivity required to accurately measure the reductions in abundance of peripheral blood eosinophils. Conversely, the estimated copy numbers of PRG2 appear to provide a much more sensitive readout. To test if this biomarker could be suitable for the clinic, we validated PRG2 abundance in the blood of control and COVID19 patients using an ELISA. The results from the ELISA showed its abundance was significantly reduced abundance in COVID19 patients, validating the proteomic results and suggesting that it could be a promising biomarker for severe disease.

Furthermore, the proteomic data also revealed molecular insights into a novel neutrophil phenotype related to delayed recovery from COVID19. The neutrophils proteomes from patients not fully recovered 29 days post enrolment, displayed a dysfunctional phenotype with reduced abundance of proteins related to neutrophil migration from the bloodstream into the tissues. They also displayed proteomic changes indicative of a metabolic profile linked to impaired effector function, with defective killing and survival capacity, along with signs of a decreased activation threshold paired with increased degranulation in the periphery. These discoveries were enabled by a large cohort that could be stratified by patient outcome at day 29, thus supporting an in-depth molecular analysis of the changes in neutrophil proteomes present in patients suffering from prolonged COVID19 symptoms.

It is important to note that label-free DIA data is not exempt of the challenges related to large-scale proteomics. This DIA-based dataset also displayed clear batch effects, in this case correlating to the sample collection batches at Ninewells Hospital. The samples were collected in the middle of a pandemic, making it extremely challenging to homogenise the composition of the batches. One batch only contained samples from COVID19 patients, with the limitation of having no controls available at that time. This same batch was the only one that exhibited pronounced reductions in virtually all sets of granule proteins, potentially triggered by degranulation post isolation, or by activation of the proteases. This made comparisons to controls or LTRI proteomes highly biased whenever this batch was used and left virtually no effective normalisation method. This issue prevented multiple analytical threads from being explored further due to the overwhelming technical effects. Inevitably large-scale studies will consist of multiple batches, so our advice is that where possible it is recommended to keep balanced batches with all conditions represented, to better enable potential normalisation. It is also important to consider including controls in the sample processing and collection, as this is an additional source of variation that is frequently not captured.

Finally, the work presented on this thesis provided some use cases of how large-scale proteomics data can be shared and visualised. Inevitably as studies grow in size, there is also an implicit increase in the complexity of the data that is generated. Consequently, developing tools to ensure this data are easily available and interpretable for the wider research community becomes a vital goal. For this purpose, we created the Encyclopedia of Proteome Dynamics (EPD), a first prototype into a web application that provided open access to the proteomic data generated by the Lamond Lab. The EPD was among the first web apps used to share proteomics data, in an interactive graphical user interface. The EPD was designed using a highly sophisticated but complex eco-system with multiple specialist databases, networking strategies and development frameworks. The complex nature of the eco-system meant its upkeep became a considerable challenge, as is frequently the case with scientific tools developed by the different research groups. This is further exacerbated when the maintenance requires highly skilled individuals with expertise in multiple specialist areas such as relational database design, noSQL databases design and implementation, graph database modelling, load balancing, networking or systems administration. These roles are not easy to recruit for, nor maintain in an academic setting, where restrictions to remunerations present a less attractive alternative than similar in industry roles.

Accordingly, when developing the second web app we sought to avoid previous pitfalls and aimed to make its design simpler to maintain while fulfilling the main goals related to open access, data sharing and interpretability to non-specialist users. The Immunological Proteome Resource (ImmPRes) is a web app that operates without databases and uses a simpler web server to deliver open access to the biggest collection of mass spectrometry-based proteomic data characterising murine hematopoietic populations. ImmPRes draws inspiration from pivotal role that ImmGen(Immunological Genome, 2020) has played in democratising access to RNAseq based data for the different hematopoietic populations. It was created to fulfil a specific but complex role, to bridge the gap between proteomics, computational biology and immunology, by providing access to a simplified user interface specifically catered to how the target audience, in this case immunologists, would interact with the data. It does this by employing a combination of quantification methods with direct biological interpretation, providing a small set of curated and easily interpretable plots, and displaying detailed graphical abstracts and protocols that describe the different populations and datasets that are hosted within the resource. It is important to recognise that the end users of the data are not mass spectrometrists, and thus measures of abundance like LFQ or iBAQ have little intrinsic meaning. This is one of the reasons ImmPRes mainly operates with estimated protein copy numbers, as they quickly provide a biological context to the proteomic data, making it more relatable and easier to understand for the target audience. This is a very important consideration, because the data needs to be shared in ways that assist the understanding and re-use by the wider research community.

How we share the outputs derived from proteomics studies is undeniably a matter of importance. A set of supplemental tables embedded within a publication is slowly being considered more of a minimum requirement than an ideal solution. Innovative solutions are now more relevant than ever, as large-scale proteomic datasets produce output files containing thousands of proteins and hundreds if not thousands of quantitative columns for each protein. Thus, discovering different ways to make this data easy to explore and re-use for the scientific community is truly an important goal. And we think that web apps like ImmPRes, which reached more than 2,000 unique users over the last year alone, provide a viable model to maximise data outreach and re-use of the processed proteomic data.

As we head into the population-scale age of proteomic research, there is little doubt that proteomic will empower novel discoveries and biological insights. From the large-scale single cell proteomic characterisation of different tissues to the population level understanding of variation that will empower stratification of patients in health and disease. Here we presented multiple large-scale proteomic datasets that have contributed new biological insights into health and disease, generated technical knowledge to improve large-scale studies and built tools that facilitate the sharing and re-use of proteomic data. There still are countless biological questions we have yet to answer, insights we are yet to understand, but there is no doubt that population-scale proteomics will make important contributions to them.

References

ADROVER, J. M., AROCA-CREVILLEN, A., CRAINICIUC, G., OSTOS, F., ROJAS-VEGA, Y., RUBIO-PONCE, A., CILLONIZ, C., BONZON-KULICHENKO, E., CALVO, E., RICO, D., MORO, M. A., WEBER, C., LIZASOAIN, I., TORRES, A., RUIZ-CABELLO, J., VAZQUEZ, J. & HIDALGO, A. 2020. Programmed 'disarming' of the neutrophil proteome reduces the magnitude of inflammation. *Nat Immunol,* 21**,** 135-144.

AEBERSOLD, R., AGAR, J. N., AMSTER, I. J., BAKER, M. S., BERTOZZI, C. R., BOJA, E. S., COSTELLO, C. E., CRAVATT, B. F., FENSELAU, C., GARCIA, B. A., GE, Y., GUNAWARDENA, J., HENDRICKSON, R. C., HERGENROTHER, P. J., HUBER, C. G., IVANOV, A. R., JENSEN, O. N., JEWETT, M. C., KELLEHER, N. L., KIESSLING, L. L., KROGAN, N. J., LARSEN, M. R., LOO, J. A., OGORZALEK LOO, R. R., LUNDBERG, E., MACCOSS, M. J., MALLICK, P., MOOTHA, V. K., MRKSICH, M., MUIR, T. W., PATRIE, S. M., PESAVENTO, J. J., PITTERI, S. J., RODRIGUEZ, H., SAGHATELIAN, A., SANDOVAL, W., SCHLUTER, H., SECHI, S., SLAVOFF, S. A., SMITH, L. M., SNYDER, M. P., THOMAS, P. M., UHLEN, M., VAN EYK, J. E., VIDAL, M., WALT, D. R., WHITE, F. M., WILLIAMS, E. R., WOHLSCHLAGER, T., WYSOCKI, V. H., YATES, N. A., YOUNG, N. L. & ZHANG, B. 2018. How many human proteoforms are there? *Nat Chem Biol,* 14**,** 206-214.

AEBERSOLD, R. & MANN, M. 2003. Mass spectrometry-based proteomics. *Nature,* 422**,** 198-207.

AEBERSOLD, R. & MANN, M. 2016. Mass-spectrometric exploration of proteome structure and function. *Nature,* 537**,** 347-55.

AHMAD, Y., BOISVERT, F. M., LUNDBERG, E., UHLEN, M. & LAMOND, A. I. 2012. Systematic analysis of protein pools, isoforms, and modifications affecting turnover and subcellular localization. *Mol Cell Proteomics,* 11**,** M111 013680.

AMID, C., ALAKO, B. T. F., BALAVENKATARAMAN KADHIRVELU, V., BURDETT, T., BURGIN, J., FAN, J., HARRISON, P. W., HOLT, S., HUSSEIN, A., IVANOV, E., JAYATHILAKA, S., KAY, S., KEANE, T., LEINONEN, R., LIU, X., MARTINEZ-VILLACORTA, J., MILANO, A., PAKSERESHT, A., RAHMAN, N., RAJAN, J., REDDY, K., RICHARDS, E., SMIRNOV, D., SOKOLOV, A., VIJAYARAJA, S. & COCHRANE, G. 2020. The European Nucleotide Archive in 2019. *Nucleic Acids Res,* 48**,** D70-D76.

ANDERSON, D. C. & SPRINGER, T. A. 1987. Leukocyte adhesion deficiency: an inherited defect in the Mac-1, LFA-1, and p150,95 glycoproteins. *Annu Rev Med,* 38**,** 175-94.

ANGUERA, M. C., SADREYEV, R., ZHANG, Z., SZANTO, A., PAYER, B., SHERIDAN, S. D., KWOK, S., HAGGARTY, S. J., SUR, M., ALVAREZ, J., GIMELBRANT, A., MITALIPOVA, M., KIRBY, J. E. & LEE, J. T. 2012. Molecular signatures of human induced pluripotent stem cells highlight sex differences and cancer genes. *Cell Stem Cell,* 11**,** 75-90.

ARUNACHALAM, P. S., WIMMERS, F., MOK, C. K. P., PERERA, R., SCOTT, M., HAGAN, T., SIGAL, N., FENG, Y., BRISTOW, L., TAK-YIN TSANG, O., WAGH, D., COLLER, J., PELLEGRINI, K. L., KAZMIN, D., ALAAEDDINE, G., LEUNG, W. S., CHAN, J. M. C., CHIK, T. S. H., CHOI, C. Y. C., HUERTA, C., PAINE MCCULLOUGH, M., LV, H., ANDERSON, E., EDUPUGANTI, S., UPADHYAY, A. A., BOSINGER, S. E., MAECKER, H. T., KHATRI, P., ROUPHAEL, N., PEIRIS, M. & PULENDRAN, B. 2020. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science,* 369**,** 1210-1220.

ASCHENBRENNER, A. C., MOUKTAROUDI, M., KRAMER, B., OESTREICH, M., ANTONAKOS, N., NUESCH-GERMANO, M., GKIZELI, K., BONAGURO, L., REUSCH, N., BASSLER, K., SARIDAKI, M., KNOLL, R., PECHT, T., KAPELLOS, T. S., DOULOU, S., KROGER, C., HERBERT, M., HOLSTEN, L., HORNE, A., GEMUND, I. D., ROVINA, N., AGRAWAL, S., DAHM, K., VAN UELFT, M., DREWS, A., LENKEIT, L., BRUSE, N., GERRETSEN, J., GIERLICH, J., BECKER, M., HANDLER, K., KRAUT, M., THEIS, H., MENGISTE, S., DE DOMENICO, E., SCHULTE-SCHREPPING, J., SEEP, L., RAABE, J., HOFFMEISTER, C., TOVINH, M., KEITEL, V., RIEKE, G., TALEVI, V., SKOWASCH, D., AZIZ, N. A., PICKKERS, P., VAN DE VEERDONK, F. L., NETEA, M. G., SCHULTZE, J. L., KOX, M., BRETELER, M. M. B., NATTERMANN, J., KOUTSOUKOU, A., GIAMARELLOS-BOURBOULIS, E. J., ULAS, T. & GERMAN, C.-O. I. 2021. Disease severity-specific neutrophil signatures in blood transcriptomes stratify COVID-19 patients. *Genome Med,* 13**,** 7.

AUGUI, S., NORA, E. P. & HEARD, E. 2011. Regulation of X-chromosome inactivation by the X-inactivation centre. *Nat Rev Genet,* 12**,** 429-42.

AWASTHI, D., NAGARKOTI, S., SADAF, S., CHANDRA, T., KUMAR, S. & DIKSHIT, M. 2019. Glycolysis dependent lactate formation in neutrophils: A metabolic link between NOX-dependent and independent NETosis. *Biochim Biophys Acta Mol Basis Dis,* 1865**,** 165542.

BALATON, B. P. & BROWN, C. J. 2016. Escape Artists of the X Chromosome. *Trends Genet,* 32**,** 348-359.

BALDWIN, M. A. 2004. Protein identification by mass spectrometry: issues to be considered. *Mol Cell Proteomics,* 3**,** 1-9.

BANDURA, D. R., BARANOV, V. I., ORNATSKY, O. I., ANTONOV, A., KINACH, R., LOU, X., PAVLOV, S., VOROBIEV, S., DICK, J. E. & TANNER, S. D. 2009. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal Chem,* 81**,** 6813-22.

BARTELS, M., GOVERS, A. M., FLESKENS, V., LOURENCO, A. R., PALS, C. E., VERVOORT, S. J., VAN GENT, R., BRENKMAN, A. B., BIERINGS, M. B., ACKERMAN, S. J., VAN LOOSDREGT, J. & COFFER, P. J. 2015. Acetylation of C/EBPepsilon is a prerequisite for terminal neutrophil differentiation. *Blood,* 125**,** 1782-92.

BEAUSOLEIL, S. A., VILLEN, J., GERBER, S. A., RUSH, J. & GYGI, S. P. 2006. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol,* 24**,** 1285-92.

BECK, M., SCHMIDT, A., MALMSTROEM, J., CLAASSEN, M., ORI, A., SZYMBORSKA, A., HERZOG, F., RINNER, O., ELLENBERG, J. & AEBERSOLD, R. 2011. The quantitative proteome of a human cell line. *Mol Syst Biol,* 7**,** 549.

BEIL, W. J., WELLER, P. F., TZIZIK, D. M., GALLI, S. J. & DVORAK, A. M. 1993. Ultrastructural immunogold localization of tumor necrosis factor-alpha to the matrix compartment of eosinophil secondary granules in patients with idiopathic hypereosinophilic syndrome. *J Histochem Cytochem,* 41**,** 1611-5.

BEKKER-JENSEN, D. B., BERNHARDT, O. M., HOGREBE, A., MARTINEZ-VAL, A., VERBEKE, L., GANDHI, T., KELSTRUP, C. D., REITER, L. & OLSEN, J. V. 2020. Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. *Nat Commun,* 11**,** 787.

BEKKER-JENSEN, D. B., KELSTRUP, C. D., BATTH, T. S., LARSEN, S. C., HALDRUP, C., BRAMSEN, J. B., SORENSEN, K. D., HOYER, S., ORNTOFT, T. F., ANDERSEN, C. L., NIELSEN, M. L. & OLSEN, J. V. 2017. An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Syst,* 4**,** 587-599 e4.

BELLING, K., RUSSO, F., JENSEN, A. B., DALGAARD, M. D., WESTERGAARD, D., RAJPERT-DE MEYTS, E., SKAKKEBAEK, N. E., JUUL, A. & BRUNAK, S. 2017. Klinefelter syndrome comorbidities linked to increased X chromosome gene dosage and altered protein interactome activity. *Hum Mol Genet,* 26**,** 1219-1229.

BERGAMASCHI, L., MESCIA, F., TURNER, L., HANSON, A. L., KOTAGIRI, P., DUNMORE, B. J., RUFFIEUX, H., DE SA, A., HUHN, O., MORGAN, M. D., GERBER, P. P., WILLS, M. R., BAKER, S., CALERO-NIETO, F. J., DOFFINGER, R., DOUGAN, G., ELMER, A., GOODFELLOW, I. G., GUPTA, R. K., HOSMILLO, M., HUNTER, K., KINGSTON, N., LEHNER, P. J., MATHESON, N. J., NICHOLSON, J. K., PETRUNKINA, A. M., RICHARDSON, S., SAUNDERS, C., THAVENTHIRAN, J. E. D., TOONEN, E. J. M., WEEKES, M. P., CAMBRIDGE INSTITUTE OF THERAPEUTIC, I., INFECTIOUS DISEASE-NATIONAL INSTITUTE OF HEALTH RESEARCH, C. B. C., GOTTGENS, B., TOSHNER, M., HESS, C., BRADLEY, J. R., LYONS, P. A. & SMITH, K. G. C. 2021. Longitudinal analysis reveals that delayed bystander CD8+ T cell activation and early immune pathology distinguish severe COVID-19 from mild disease. *Immunity,* 54**,** 1257-1275 e8.

BEYNON, R. J. & PRATT, J. M. 2005. Metabolic labeling of proteins for proteomics. *Mol Cell Proteomics,* 4**,** 857-72.

BLUDAU, I., FRANK, M., DORIG, C., CAI, Y., HEUSEL, M., ROSENBERGER, G., PICOTTI, P., COLLINS, B. C., ROST, H. & AEBERSOLD, R. 2021. Systematic detection of functional proteoform groups from bottom-up proteomic datasets. *Nat Commun,* 12**,** 3810.

BOISVERT, F. M., AHMAD, Y., GIERLINSKI, M., CHARRIERE, F., LAMONT, D., SCOTT, M., BARTON, G. & LAMOND, A. I. 2012. A quantitative spatial proteomics analysis of proteome turnover in human cells. *Mol Cell Proteomics,* 11**,** M111 011429.

BORENSZTEIN, M., SYX, L., ANCELIN, K., DIABANGOUAYA, P., PICARD, C., LIU, T., LIANG, J. B., VASSILEV, I., GALUPA, R., SERVANT, N., BARILLOT, E., SURANI, A., CHEN, C. J. & HEARD, E. 2017. Xist-dependent imprinted X inactivation and the early developmental consequences of its failure. *Nat Struct Mol Biol,* 24**,** 226-233.

BORREGAARD, N. & HERLIN, T. 1982. Energy metabolism of human neutrophils during phagocytosis. *J Clin Invest,* 70**,** 550-7.

BRENES, A., AFZAL, V., KENT, R. & LAMOND, A. I. 2018a. The Encyclopedia of Proteome Dynamics: a big data ecosystem for (prote)omics. *Nucleic Acids Res,* 46**,** D1202-D1209.

BRENES, A., BENSADDEK, D., HUKELMANN, J. L., AFZAL, V. & LAMOND, A. I. 2018b. The iPSC proteomic compendium. *bioRxiv*.

BRENES, A., HUKELMANN, J., BENSADDEK, D. & LAMOND, A. I. 2019. Multibatch TMT Reveals False Positives, Batch Effects and Missing Values. *Mol Cell Proteomics,* 18**,** 1967-1980.

BRENES, A. & LAMOND, A. I. 2018. The Encyclopedia of Proteome Dynamics: The KinoViewer. *Bioinformatics*.

BRENES, A. & LAMOND, A. I. 2019. The Encyclopedia of Proteome Dynamics: the KinoViewer. *Bioinformatics,* 35**,** 1441-1442.

BRENES, A. J. & CANTRELL, D. A. 2018. Defining the T cell proteome and the role of mTORC1 during differentiation using quantitative proteomics. *Defining the T cell proteome and the role of mTORC1 during differentiation using quantitative proteomics.* https://www.ebi.ac.uk/pride/archive/projects/PXD012058.

BRENES, A. J. & CANTRELL, D. A. 2020. Immunological Proteome Resource (ImmPRes): Part two. *Immunological Proteome Resource (ImmPRes): Part two.* https://www.ebi.ac.uk/pride/archive/projects/PXD020091.

BRENES, A. J. & CANTRELL, D. A. 2022. Immunological Proteome Resource (ImmPRes): Liver derived CD4+ T cell proteomes. https://www.ebi.ac.uk/pride/archive/projects/PXD036319.

BRENES, A. J., HUKELMANN, J., SPINELLI, L., HOWDEN, A. J., MARCHINGO, J. M., SINCLAIR, L. V., ROLLINGS, C. M., JAMES, O. J., PHAIR, I. R., MATTEWS, S. P., ROSS, S. H., ARTHUR, J. S. C., SWAMY, M., FINLAY, D. K., LAMOND, A. I. & CANTRELL, D. A. 2022. The Immunological Proteome Resource. bioRxiv.

BRENES, A. J. & LAMOND, A. I. 2020. Tissue adaptation is the dominant driver of the proteomic landscape of intestinal intraepithelial lymphocytes. https://www.ebi.ac.uk/pride/archive/projects/PXD023140.

BRENES, A. J., VANDEREYKEN, M., JAMES, O. J., WATT, H., HUKELMANN, J., SPINELLI, L., DIKOVSKAYA, D., LAMOND, A. I. & SWAMY, M. 2021a. Tissue environment, not ontogeny, defines murine intestinal intraepithelial T lymphocytes. *Elife,* 10.

BRENES, A. J., YOSHIKAWA, H., BENSADDEK, D., MIRAUTA, B., SEATON, D., HUKELMANN, J. L., JIANG, H., STEGLE, O. & LAMOND, A. I. 2021b. Erosion of human X chromosome inactivation causes major remodeling of the iPSC proteome. *Cell Rep,* 35**,** 109032.

BROCKMANN, R., BEYER, A., HEINISCH, J. J. & WILHELM, T. 2007. Posttranscriptional expression regulation: what determines translation rates? *PLoS Comput Biol,* 3**,** e57.

BRUDERER, R., BERNHARDT, O. M., GANDHI, T., MILADINOVIC, S. M., CHENG, L. Y., MESSNER, S., EHRENBERGER, T., ZANOTELLI, V., BUTSCHEID, Y., ESCHER, C., VITEK, O., RINNER, O. & REITER, L. 2015. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol Cell Proteomics,* 14**,** 1400-10.

BRUDERER, R., BERNHARDT, O. M., GANDHI, T., XUAN, Y., SONDERMANN, J., SCHMIDT, M., GOMEZ-VARELA, D. & REITER, L. 2017. Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Mol Cell Proteomics,* 16**,** 2296-2309.

BUCCITELLI, C. & SELBACH, M. 2020. mRNAs, proteins and the emerging principles of gene expression control. *Nat Rev Genet,* 21**,** 630-644.

BURN, G. L., FOTI, A., MARSMAN, G., PATEL, D. F. & ZYCHLINSKY, A. 2021. The Neutrophil. *Immunity,* 54**,** 1377-1391.

CAMERINI, S. & MAURI, P. 2015. The role of protein and peptide separation before mass spectrometry analysis in clinical proteomics. *J Chromatogr A,* 1381**,** 1-12.

CANTRELL, D. 2015. Signaling in lymphocyte activation. *Cold Spring Harb Perspect Biol,* 7.

CARISSIMO, G., XU, W., KWOK, I., ABDAD, M. Y., CHAN, Y. H., FONG, S. W., PUAN, K. J., LEE, C. Y., YEO, N. K., AMRUN, S. N., CHEE, R. S., HOW, W., CHAN, S., FAN, B. E., ANDIAPPAN, A. K., LEE, B., ROTZSCHKE, O., YOUNG, B. E., LEO, Y. S., LYE, D. C., RENIA, L., NG, L. G., LARBI, A. & NG, L. F. 2020. Whole blood immunophenotyping uncovers immature neutrophil-to-VD2 T-cell ratio as an early marker for severe COVID-19. *Nat Commun,* 11**,** 5243.

CARRIERE, A., CARGNELLO, M., JULIEN, L. A., GAO, H., BONNEIL, E., THIBAULT, P. & ROUX, P. P. 2008. Oncogenic MAPK signaling stimulates mTORC1 activity by promoting RSK-mediated raptor phosphorylation. *Curr Biol,* 18**,** 1269-77.

CEREZO, E., PLISSON-CHASTANG, C., HENRAS, A. K., LEBARON, S., GLEIZES, P. E., O'DONOHUE, M. F., ROMEO, Y. & HENRY, Y. 2019. Maturation of pre-40S particles in yeast and humans. *Wiley Interdiscip Rev RNA,* 10**,** e1516.

CHOI, J., CLEMENT, K., HUEBNER, A. J., WEBSTER, J., ROSE, C. M., BRUMBAUGH, J., WALSH, R. M., LEE, S., SAVOL, A., ETCHEGARAY, J. P., GU, H., BOYLE, P., ELLING, U., MOSTOSLAVSKY, R., SADREYEV, R., PARK, P. J., GYGI, S. P., MEISSNER, A. & HOCHEDLINGER, K. 2017. DUSP9 Modulates DNA Hypomethylation in Female Mouse Pluripotent Stem Cells. *Cell Stem Cell,* 20**,** 706-719 e7.

CHOI, M., CHANG, C. Y., CLOUGH, T., BROUDY, D., KILLEEN, T., MACLEAN, B. & VITEK, O. 2014. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics,* 30**,** 2524-6.

CHRISTOFOROU, A., MULVEY, C. M., BRECKELS, L. M., GELADAKI, A., HURRELL, T., HAYWARD, P. C., NAAKE, T., GATTO, L., VINER, R., MARTINEZ ARIAS, A. & LILLEY, K. S. 2016. A draft map of the mouse pluripotent stem cell spatial proteome. *Nat Commun,* 7**,** 8992.

CHRISTOPHER, J. A., STADLER, C., MARTIN, C. E., MORGENSTERN, M., PAN, Y., BETSINGER, C. N., RATTRAY, D. G., MAHDESSIAN, D., GINGRAS, A. C., WARSCHEID, B., LEHTIO, J., CRISTEA, I. M., FOSTER, L. J., EMILI, A. & LILLEY, K. S. 2021. Subcellular proteomics. *Nat Rev Methods Primers,* 1.

CHU, V. T., BELLER, A., RAUSCH, S., STRANDMARK, J., ZANKER, M., ARBACH, O., KRUGLOV, A. & BEREK, C. 2014. Eosinophils promote generation and maintenance of immunoglobulin-A-expressing plasma cells and contribute to gut immune homeostasis. *Immunity,* 40**,** 582-93.

CLOUTIER, M., KUMAR, S., BUTTIGIEG, E., KELLER, L., LEE, B., WILLIAMS, A., MOJICA-PEREZ, S., ERLIANDRI, I., ROCHA, A. M. D., CADIGAN, K., SMITH, G. D. & KALANTRY, S. 2022. Preventing erosion of X-chromosome inactivation in human embryonic stem cells. *Nat Commun,* 13**,** 2516.

CODD, E. F. 1970. A relational model of data for large shared data banks. *Communications of the ACM,* 13**,** 377-387.

COLLINS, M. O., YU, L. & CHOUDHARY, J. S. 2007. Analysis of protein phosphorylation on a proteome-scale. *Proteomics,* 7**,** 2751-68.

COWLAND, J. B. & BORREGAARD, N. 2016. Granulopoiesis and granules of human neutrophils. *Immunol Rev,* 273**,** 11-28.

COX, J. & MANN, M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol,* 26**,** 1367-72.

CRAIG, R. & BEAVIS, R. C. 2004. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics,* 20**,** 1466-7.

CROZIER, T. W. M., TINTI, M., LARANCE, M., LAMOND, A. I. & FERGUSON, M. A. J. 2017. Prediction of Protein Complexes in Trypanosoma brucei by Protein Correlation Profiling Mass Spectrometry and Machine Learning. *Mol Cell Proteomics,* 16**,** 2254-2267.

CROZIER, T. W. M., TINTI, M., WHEELER, R. J., LY, T., FERGUSON, M. A. J. & LAMOND, A. I. 2018. Proteomic Analysis of the Cell Cycle of Procylic Form Trypanosoma brucei. *Mol Cell Proteomics,* 17**,** 1184-1195.

DAMASIO, M. P., MARCHINGO, J. M., SPINELLI, L., HUKELMANN, J. L., CANTRELL, D. A. & HOWDEN, A. J. M. 2021. Extracellular signal-regulated kinase (ERK) pathway control of CD8+ T cell differentiation. *Biochem J,* 478**,** 79-98.

DANDULAKIS, M. G., MEGANATHAN, K., KROLL, K. L., BONNI, A. & CONSTANTINO, J. N. 2016. Complexities of X chromosome inactivation status in female human induced pluripotent stem cells-a brief review and scientific update for autism research. *J Neurodev Disord,* 8**,** 22.

DAVIES, V., WANDY, J., WEIDT, S., VAN DER HOOFT, J. J. J., MILLER, A., DALY, R. & ROGERS, S. 2021. Rapid Development of Improved Data-Dependent Acquisition Strategies. *Anal Chem,* 93**,** 5676-5683.

DE KLEER, I., WILLEMS, F., LAMBRECHT, B. & GORIELY, S. 2014. Ontogeny of myeloid cells. *Front Immunol,* 5**,** 423.

DEL VALLE, D. M., KIM-SCHULZE, S., HUANG, H. H., BECKMANN, N. D., NIRENBERG, S., WANG, B., LAVIN, Y., SWARTZ, T. H., MADDURI, D., STOCK, A., MARRON, T. U., XIE, H., PATEL, M., TUBALLES, K., VAN OEKELEN, O., RAHMAN, A., KOVATCH, P., ABERG, J. A., SCHADT, E., JAGANNATH, S., MAZUMDAR, M., CHARNEY, A. W., FIRPO-BETANCOURT, A., MENDU, D. R., JHANG, J., REICH, D., SIGEL, K., CORDON-CARDO, C., FELDMANN, M., PAREKH, S., MERAD, M. & GNJATIC, S. 2020. An inflammatory cytokine signature predicts COVID-19 severity and survival. *Nat Med,* 26**,** 1636-1643.

DEMICHEV, V., MESSNER, C. B., VERNARDIS, S. I., LILLEY, K. S. & RALSER, M. 2020. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods,* 17**,** 41-44.

DENNY, M. F., YALAVARTHI, S., ZHAO, W., THACKER, S. G., ANDERSON, M., SANDY, A. R., MCCUNE, W. J. & KAPLAN, M. J. 2010. A distinct subset of proinflammatory neutrophils isolated from patients with systemic lupus erythematosus induces vascular damage and synthesizes type I IFNs. *J Immunol,* 184**,** 3284-97.

DEPHOURE, N., ZHOU, C., VILLEN, J., BEAUSOLEIL, S. A., BAKALARSKI, C. E., ELLEDGE, S. J. & GYGI, S. P. 2008. A quantitative atlas of mitotic phosphorylation. *Proc Natl Acad Sci U S A,* 105**,** 10762-7.

DEUTSCH, E. W., PEREZ-RIVEROL, Y., CHALKLEY, R. J., WILHELM, M., TATE, S., SACHSENBERG, T., WALZER, M., KALL, L., DELANGHE, B., BOCKER, S., SCHYMANSKI, E. L., WILMES, P., DORFER, V., KUSTER, B., VOLDERS, P. J., JEHMLICH, N., VISSERS, J. P. C., WOLAN, D. W., WANG, A. Y., MENDOZA, L., SHOFSTAHL, J., DOWSEY, A. W., GRISS, J., SALEK, R. M., NEUMANN, S., BINZ, P. A., LAM, H., VIZCAINO, J. A., BANDEIRA, N. & ROST, H. 2018. Expanding the Use of Spectral Libraries in Proteomics. *J Proteome Res,* 17**,** 4051-4060.

DIAO, B., WANG, C., TAN, Y., CHEN, X., LIU, Y., NING, L., CHEN, L., LI, M., WANG, G., YUAN, Z., FENG, Z., ZHANG, Y., WU, Y. & CHEN, Y. 2020. Reduction and Functional Exhaustion of T Cells in Patients With Coronavirus Disease 2019 (COVID-19). *Front Immunol,* 11**,** 827.

DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. & GINGERAS, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* 29**,** 15-21.

DOHERTY, M. K., HAMMOND, D. E., CLAGUE, M. J., GASKELL, S. J. & BEYNON, R. J. 2009. Turnover of the human proteome: determination of protein intracellular stability by dynamic SILAC. *J Proteome Res,* 8**,** 104-12.

DONG, C., FISCHER, L. A. & THEUNISSEN, T. W. 2019. Recent insights into the naive state of human pluripotency and its applications. *Exp Cell Res,* 385**,** 111645.

DORWARD, D. A., LUCAS, C. D., ALESSANDRI, A. L., MARWICK, J. A., ROSSI, F., DRANSFIELD, I., HASLETT, C., DHALIWAL, K. & ROSSI, A. G. 2013. Technical advance: autofluorescence-based sorting: rapid and nonperturbing isolation of ultrapure neutrophils to determine cytokine production. *J Leukoc Biol,* 94**,** 193-202.

DOSE, M., SLECKMAN, B. P., HAN, J., BREDEMEYER, A. L., BENDELAC, A. & GOUNARI, F. 2009. Intrathymic proliferation wave essential for Valpha14+ natural killer T cell development depends on c-Myc. *Proc Natl Acad Sci U S A,* 106**,** 8641-6.

DOSSIN, F., PINHEIRO, I., ZYLICZ, J. J., ROENSCH, J., COLLOMBET, S., LE SAUX, A., CHELMICKI, T., ATTIA, M., KAPOOR, V., ZHAN, Y., DINGLI, F., LOEW, D., MERCHER, T., DEKKER, J. & HEARD, E. 2020. SPEN integrates transcriptional and epigenetic control of X-inactivation. *Nature*.

DOWELL, J. A., WRIGHT, L. J., ARMSTRONG, E. A. & DENU, J. M. 2021. Benchmarking Quantitative Performance in Label-Free Proteomics. *ACS Omega,* 6**,** 2494-2504.

DOWEY, R., COLE, J., THOMPSON, A. A. R., HULL, R. C., HUANG, C., WHATMORE, J., IQBAL, A., BRADLEY, K. L., MCKENZIE, J., LAWRIE, A., CONDLIFFE, A. M., KISS-TOTH, E., SABROE, I. & PRINCE, L. R. 2022. Enhanced neutrophil extracellular trap formation in COVID-19 is inhibited by the protein kinase C inhibitor ruboxistaurin. *ERJ Open Res,* 8.

DRISSEN, R., BUZA-VIDAS, N., WOLL, P., THONGJUEA, S., GAMBARDELLA, A., GIUSTACCHINI, A., MANCINI, E., ZRIWIL, A., LUTTEROPP, M., GROVER, A., MEAD, A., SITNICKA, E., JACOBSEN, S. E. W. & NERLOV, C. 2016. Distinct myeloid progenitor-differentiation pathways identified through single-cell RNA sequencing. *Nat Immunol,* 17**,** 666-676.

DUNKLEY, T. P., HESTER, S., SHADFORTH, I. P., RUNIONS, J., WEIMAR, T., HANTON, S. L., GRIFFIN, J. L., BESSANT, C., BRANDIZZI, F., HAWES, C., WATSON, R. B., DUPREE, P. & LILLEY, K. S. 2006. Mapping the Arabidopsis organelle proteome. *Proc Natl Acad Sci U S A,* 103**,** 6518-23.

DUNKLEY, T. P., WATSON, R., GRIFFIN, J. L., DUPREE, P. & LILLEY, K. S. 2004. Localization of organelle proteins by isotope tagging (LOPIT). *Mol Cell Proteomics,* 3**,** 1128-34.

EBERT, A. D., YU, J., ROSE, F. F., JR., MATTIS, V. B., LORSON, C. L., THOMSON, J. A. & SVENDSEN, C. N. 2009. Induced pluripotent stem cells from a spinal muscular atrophy patient. *Nature,* 457**,** 277-80.

EDFORS, F., DANIELSSON, F., HALLSTROM, B. M., KALL, L., LUNDBERG, E., PONTEN, F., FORSSTROM, B. & UHLEN, M. 2016. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol Syst Biol,* 12**,** 883.

ELIAS, J. E. & GYGI, S. P. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods,* 4**,** 207-14.

ELTANBOULY, M. A., CROTEAU, W., NOELLE, R. J. & LINES, J. L. 2019. VISTA: a novel immunotherapy target for normalizing innate and adaptive immunity. *Semin Immunol,* 42**,** 101308.

ELTANBOULY, M. A., ZHAO, Y., SCHAAFSMA, E., BURNS, C. M., MABAERA, R., CHENG, C. & NOELLE, R. J. 2020. VISTA: A Target to Manage the Innate Cytokine Storm. *Front Immunol,* 11**,** 595950.

EMBI, N., RYLATT, D. B. & COHEN, P. 1980. Glycogen synthase kinase-3 from rabbit skeletal muscle. Separation from cyclic-AMP-dependent protein kinase and phosphorylase kinase. *Eur J Biochem,* 107**,** 519-27.

ENG, J. K., MCCORMACK, A. L. & YATES, J. R. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom,* 5**,** 976-89.

ETZIONI, A., FRYDMAN, M., POLLACK, S., AVIDOR, I., PHILLIPS, M. L., PAULSON, J. C. & GERSHONI-BARUCH, R. 1992. Brief report: recurrent severe infections caused by a novel leukocyte adhesion deficiency. *N Engl J Med,* 327**,** 1789-92.

FAURSCHOU, M. & BORREGAARD, N. 2003. Neutrophil granules and secretory vesicles in inflammation. *Microbes Infect,* 5**,** 1317-27.

FAVIER, B. 2016. Regulation of neutrophil functions through inhibitory receptors: an emerging paradigm in health and disease. *Immunol Rev,* 273**,** 140-55.

FENN, J. B., MANN, M., MENG, C. K., WONG, S. F. & WHITEHOUSE, C. M. 1989. Electrospray ionization for mass spectrometry of large biomolecules. *Science,* 246**,** 64-71.

FICHT, X. & IANNACONE, M. 2020. Immune surveillance of the liver by T cells. *Sci Immunol,* 5.

FJAERTOFT, G., HAKANSSON, L., EWALD, U., FOUCARD, T. & VENGE, P. 1999. Neutrophils from term and preterm newborn infants express the high affinity Fcgamma-receptor I (CD64) during bacterial infections. *Pediatr Res,* 45**,** 871-6.

FLEMING, J. R., SASTRY, L., WALL, S. J., SULLIVAN, L. & FERGUSON, M. A. 2016. Proteomic Identification of Immunodiagnostic Antigens for Trypanosoma vivax Infections in Cattle and Generation of a Proof-of-Concept Lateral Flow Test Diagnostic Device. *PLoS Negl Trop Dis,* 10**,** e0004977.

GALUPA, R. & HEARD, E. 2018. X-Chromosome Inactivation: A Crossroads Between Chromosome Architecture and Gene Regulation. *Annu Rev Genet,* 52**,** 535-566.

GANZ, T. 2003. Defensins: antimicrobial peptides of innate immunity. *Nat Rev Immunol,* 3**,** 710-20.

GATTO, L., BRECKELS, L. M., WIECZOREK, S., BURGER, T. & LILLEY, K. S. 2014. Mass-spectrometry-based spatial proteomics data analysis using pRoloc and pRolocdata. *Bioinformatics,* 30**,** 1322-4.

GEBHARDT, T., WAKIM, L. M., EIDSMO, L., READING, P. C., HEATH, W. R. & CARBONE, F. R. 2009. Memory T cells in nonlymphoid tissue that provide enhanced local immunity during infection with herpes simplex virus. *Nat Immunol,* 10**,** 524-30.

GEENS, M. & CHUVA DE SOUSA LOPES, S. M. 2017. X chromosome inactivation in human pluripotent stem cells as a model for human development: back to the drawing board? *Hum Reprod Update,* 23**,** 520-532.

GEIGER, T., COX, J. & MANN, M. 2010. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol Cell Proteomics,* 9**,** 2252-61.

GELADAKI, A., KOCEVAR BRITOVSEK, N., BRECKELS, L. M., SMITH, T. S., VENNARD, O. L., MULVEY, C. M., CROOK, O. M., GATTO, L. & LILLEY, K. S. 2019. Combining LOPIT with differential ultracentrifugation for high-resolution spatial proteomics. *Nat Commun,* 10**,** 331.

GENTLEMAN, R. C., CAREY, V. J., BATES, D. M., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y., GENTRY, J., HORNIK, K., HOTHORN, T., HUBER, W., IACUS, S., IRIZARRY, R., LEISCH, F., LI, C., MAECHLER, M., ROSSINI, A. J., SAWITZKI, G., SMITH, C., SMYTH, G., TIERNEY, L., YANG, J. Y. & ZHANG, J. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol,* 5**,** R80.

GILLET, L. C., NAVARRO, P., TATE, S., ROST, H., SELEVSEK, N., REITER, L., BONNER, R. & AEBERSOLD, R. 2012. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics,* 11**,** O111 016717.

GINNO, P. A., BURGER, L., SEEBACHER, J., IESMANTAVICIUS, V. & SCHUBELER, D. 2018. Cell cycle-resolved chromatin proteomics reveals the extent of mitotic preservation of the genomic regulatory landscape. *Nat Commun,* 9**,** 4048.

GIORGETTI, L., LAJOIE, B. R., CARTER, A. C., ATTIA, M., ZHAN, Y., XU, J., CHEN, C. J., KAPLAN, N., CHANG, H. Y., HEARD, E. & DEKKER, J. 2016. Structural organization of the inactive X chromosome in the mouse. *Nature,* 535**,** 575-9.

GO, C. D., KNIGHT, J. D. R., RAJASEKHARAN, A., RATHOD, B., HESKETH, G. G., ABE, K. T., YOUN, J. Y., SAMAVARCHI-TEHRANI, P., ZHANG, H., ZHU, L. Y., POPIEL, E., LAMBERT, J. P., COYAUD, E., CHEUNG, S. W. T., RAJENDRAN, D., WONG, C. J., ANTONICKA, H., PELLETIER, L., PALAZZO, A. F., SHOUBRIDGE, E. A., RAUGHT, B. & GINGRAS, A. C. 2021. A proximity-dependent biotinylation map of a human cell. *Nature,* 595**,** 120-124.

GOH, Y. P., HENDERSON, N. C., HEREDIA, J. E., RED EAGLE, A., ODEGAARD, J. I., LEHWALD, N., NGUYEN, K. D., SHEPPARD, D., MUKUNDAN, L., LOCKSLEY, R. M. & CHAWLA, A. 2013. Eosinophils secrete IL-4 to facilitate liver regeneration. *Proc Natl Acad Sci U S A,* 110**,** 9914-9.

GONCALVES, E., POULOS, R. C., CAI, Z., BARTHORPE, S., MANDA, S. S., LUCAS, N., BECK, A., BUCIO-NOBLE, D., DAUSMANN, M., HALL, C., HECKER, M., KOH, J., LIGHTFOOT, H., MAHBOOB, S., MALI, I., MORRIS, J., RICHARDSON, L., SENEVIRATNE, A. J., SHEPHERD, R., SYKES, E., THOMAS, F., VALENTINI, S., WILLIAMS, S. G., WU, Y., XAVIER, D., MACKENZIE, K. L., HAINS, P. G., TULLY, B., ROBINSON, P. J., ZHONG, Q., GARNETT, M. J. & REDDEL, R. R. 2022. Pan-cancer proteomic map of 949 human cell lines. *Cancer Cell,* 40**,** 835-849 e8.

GOUON-EVANS, V., ROTHENBERG, M. E. & POLLARD, J. W. 2000. Postnatal mammary gland development requires macrophages and eosinophils. *Development,* 127**,** 2269-82.

GRAVES, J. A. 2016. Evolution of vertebrate sex chromosomes and dosage compensation. *Nat Rev Genet,* 17**,** 33-46.

GRAVES, P. R. & HAYSTEAD, T. A. 2002. Molecular biologist's guide to proteomics. *Microbiol Mol Biol Rev,* 66**,** 39-63; table of contents.

GREEN, M. C. & SHULTZ, L. D. 1975. Motheaten, an immunodeficient mutant of the mouse. I. Genetics and pathology. *J Hered,* 66**,** 250-8.

GUTHER, M. L., URBANIAK, M. D., TAVENDALE, A., PRESCOTT, A. & FERGUSON, M. A. 2014. High-confidence glycosome proteome for procyclic form Trypanosoma brucei by epitope-tag organelle enrichment and SILAC proteomics. *J Proteome Res,* 13**,** 2796-806.

GYGI, S. P., RIST, B., GERBER, S. A., TURECEK, F., GELB, M. H. & AEBERSOLD, R. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol,* 17**,** 994-9.

HAILEMARIAM, M., EGUEZ, R. V., SINGH, H., BEKELE, S., AMENI, G., PIEPER, R. & YU, Y. 2018. S-Trap, an Ultrafast Sample-Preparation Approach for Shotgun Proteomics. *J Proteome Res,* 17**,** 2917-2924.

HANNA, J., MARKOULAKI, S., MITALIPOVA, M., CHENG, A. W., CASSADY, J. P., STAERK, J., CAREY, B. W., LENGNER, C. J., FOREMAN, R., LOVE, J., GAO, Q., KIM, J. & JAENISCH, R. 2009. Metastable pluripotent states in NOD-mouse-derived ESCs. *Cell Stem Cell,* 4**,** 513-24.

HANNA, J. H., SAHA, K. & JAENISCH, R. 2010. Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. *Cell,* 143**,** 508-25.

HENG, T. S., PAINTER, M. W. & IMMUNOLOGICAL GENOME PROJECT, C. 2008. The Immunological Genome Project: networks of gene expression in immune cells. *Nat Immunol,* 9**,** 1091-4.

HENNRICH, M. L., ROMANOV, N., HORN, P., JAEGER, S., ECKSTEIN, V., STEEPLES, V., YE, F., DING, X., POISA-BEIRO, L., LAI, M. C., LANG, B., BOULTWOOD, J., LUFT, T., ZAUGG, J. B., PELLAGATTI, A., BORK, P., ALOY, P., GAVIN, A. C. & HO, A. D. 2018. Cell-specific proteome analyses of human bone marrow reveal molecular features of age-dependent functional decline. *Nat Commun,* 9**,** 4004.

HERBRICH, S. M., COLE, R. N., WEST, K. P., JR., SCHULZE, K., YAGER, J. D., GROOPMAN, J. D., CHRISTIAN, P., WU, L., O'MEALLY, R. N., MAY, D. H., MCINTOSH, M. W. & RUCZINSKI, I. 2013. Statistical inference from multiple iTRAQ experiments without using common reference standards. *J Proteome Res,* 12**,** 594-604.

HEREDIA, J. E., MUKUNDAN, L., CHEN, F. M., MUELLER, A. A., DEO, R. C., LOCKSLEY, R. M., RANDO, T. A. & CHAWLA, A. 2013. Type 2 innate signals stimulate fibro/adipogenic progenitors to facilitate muscle regeneration. *Cell,* 153**,** 376-88.

HERR, P., BOSTROM, J., RULLMAN, E., RUDD, S. G., VESTERLUND, M., LEHTIO, J., HELLEDAY, T., MADDALO, G. & ALTUN, M. 2020. Cell Cycle Profiling Reveals Protein Oscillation, Phosphorylation, and Localization Dynamics. *Mol Cell Proteomics,* 19**,** 608-623.

HOFFMANN, J. J. 2009. Neutrophil CD64: a diagnostic marker for infection and sepsis. *Clin Chem Lab Med,* 47**,** 903-16.

HONG, C. W. 2017. Current Understanding in Neutrophil Differentiation and Heterogeneity. *Immune Netw,* 17**,** 298-306.

HOOGENDIJK, A. J., POURFARZAD, F., AARTS, C. E. M., TOOL, A. T. J., HIEMSTRA, I. H., GRASSI, L., FRONTINI, M., MEIJER, A. B., VAN DEN BIGGELAAR, M. & KUIJPERS, T. W. 2019. Dynamic Transcriptome-Proteome Correlation Networks Reveal Human Myeloid Differentiation and Neutrophil-Specific Programming. *Cell Rep,* 29**,** 2505-2519 e4.

HOWDEN, A. J. & CANTRELL, D. A. 2020. Blocking ERK activity has a selective but critical impact on activating T cells. *Blocking ERK activity has a selective but critical impact on activating T cells.* https://www.ebi.ac.uk/pride/archive/projects/PXD023256.

HOWDEN, A. J. M., HUKELMANN, J. L., BRENES, A., SPINELLI, L., SINCLAIR, L. V., LAMOND, A. I. & CANTRELL, D. A. 2019. Quantitative analysis of T cell proteomes and environmental sensors during T cell differentiation. *Nat Immunol*.

HSU, B. E., TABARIES, S., JOHNSON, R. M., ANDRZEJEWSKI, S., SENECAL, J., LEHUEDE, C., ANNIS, M. G., MA, E. H., VOLS, S., RAMSAY, L., FROMENT, R., MONAST, A., WATSON, I. R., GRANOT, Z., JONES, R. G., ST-PIERRE, J. & SIEGEL, P. M. 2019. Immature Low-Density Neutrophils Exhibit Metabolic Flexibility that Facilitates Breast Cancer Liver Metastasis. *Cell Rep,* 27**,** 3902-3915 e6.

HU, A., NOBLE, W. S. & WOLF-YADLIN, A. 2016. Technical advances in proteomics: new developments in data-independent acquisition. *F1000Res,* 5.

HUA, L., YAO, S., PHAM, D., JIANG, L., WRIGHT, J., SAWANT, D., DENT, A. L., BRACIALE, T. J., KAPLAN, M. H. & SUN, J. 2013. Cytokine-dependent induction of CD4+ T cells with cytotoxic potential during influenza virus infection. *J Virol,* 87**,** 11884-93.

HUANG, C., WANG, Y., LI, X., REN, L., ZHAO, J., HU, Y., ZHANG, L., FAN, G., XU, J., GU, X., CHENG, Z., YU, T., XIA, J., WEI, Y., WU, W., XIE, X., YIN, W., LI, H., LIU, M., XIAO, Y., GAO, H., GUO, L., XIE, J., WANG, G., JIANG, R., GAO, Z., JIN, Q., WANG, J. & CAO, B. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet,* 395**,** 497-506.

HUANG, J. X., LEE, G., CAVANAUGH, K. E., CHANG, J. W., GARDEL, M. L. & MOELLERING, R. E. 2019. High throughput discovery of functional protein modifications by Hotspot Thermal Profiling. *Nat Methods,* 16**,** 894-901.

HUANG, L., GEBRESELASSIE, N. G., GAGLIARDO, L. F., RUYECHAN, M. C., LEE, N. A., LEE, J. J. & APPLETON, J. A. 2014. Eosinophil-derived IL-10 supports chronic nematode infection. *J Immunol,* 193**,** 4178-87.

HUGHES, C. S., FOEHR, S., GARFIELD, D. A., FURLONG, E. E., STEINMETZ, L. M. & KRIJGSVELD, J. 2014. Ultrasensitive proteome analysis using paramagnetic bead technology. *Mol Syst Biol,* 10**,** 757.

HUKELMANN, J. L., ANDERSON, K. E., SINCLAIR, L. V., GRZES, K. M., MURILLO, A. B., HAWKINS, P. T., STEPHENS, L. R., LAMOND, A. I. & CANTRELL, D. A. 2016. The cytotoxic T cell proteome and its shaping by the kinase mTOR. *Nat Immunol,* 17**,** 104-12.

IMMUNOLOGICAL GENOME, P. 2020. ImmGen at 15. *Nat Immunol,* 21**,** 700-703.

INJARABIAN, L., DEVIN, A., RANSAC, S. & MARTEYN, B. S. 2019. Neutrophil Metabolic Shift during their Lifecycle: Impact on their Survival and Activation. *Int J Mol Sci,* 21.

ISASA, M., ROSE, C. M., ELSASSER, S., NAVARRETE-PEREA, J., PAULO, J. A., FINLEY, D. J. & GYGI, S. P. 2015. Multiplexed, Proteome-Wide Protein Expression Profiling: Yeast Deubiquitylating Enzyme Knockout Strains. *J Proteome Res,* 14**,** 5306-17.

ITZHAK, D. N., TYANOVA, S., COX, J. & BORNER, G. H. 2016. Global, quantitative and dynamic mapping of protein subcellular localization. *Elife,* 5.

JACKSON, R. J., HELLEN, C. U. & PESTOVA, T. V. 2010. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol,* 11**,** 113-27.

JACOBSEN, E. A., OCHKUR, S. I., PERO, R. S., TARANOVA, A. G., PROTHEROE, C. A., COLBERT, D. C., LEE, N. A. & LEE, J. J. 2008. Allergic pulmonary inflammation in mice is dependent on eosinophil-induced recruitment of effector T cells. *J Exp Med,* 205**,** 699-710.

JANEWAY, C. A., JR. & MEDZHITOV, R. 2002. Innate immune recognition. *Annu Rev Immunol,* 20**,** 197-216.

JANEWAY, C. A., JR., TRAVERS, P., WALPORT, M. & SHLOMCHIK, M. 2001. *Immunobiology, 5th edition,* New York.

JEON, J. H., HONG, C. W., KIM, E. Y. & LEE, J. M. 2020. Current Understanding on the Metabolism of Neutrophils. *Immune Netw,* 20**,** e46.

JI, Z., TINTI, M. & FERGUSON, M. A. J. 2021. Proteomic identification of the UDP-GlcNAc: PI alpha1-6 GlcNAc-transferase subunits of the glycosylphosphatidylinositol biosynthetic pathway of Trypanosoma brucei. *PLoS One,* 16**,** e0244699.

JOSE, R. J. & MANUEL, A. 2020. COVID-19 cytokine storm: the interplay between inflammation and coagulation. *Lancet Respir Med,* 8**,** e46-e47.

JUNTUNEN, E., SALMINEN, T., TALHA, S. M., MARTISKAINEN, I., SOUKKA, T., PETTERSSON, K. & WARIS, M. 2017. Lateral flow immunoassay with upconverting nanoparticle-based detection for indirect measurement of interferon response by the level of MxA. *J Med Virol,* 89**,** 598-605.

KABELITZ, D., SERRANO, R., KOUAKANOU, L., PETERS, C. & KALYAN, S. 2020. Cancer immunotherapy with gammadelta T cells: many paths ahead of us. *Cell Mol Immunol,* 17**,** 925-939.

KANEHISA, M. & GOTO, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res,* 28**,** 27-30.

KARBSTEIN, K. 2013. Quality control mechanisms during ribosome maturation. *Trends Cell Biol,* 23**,** 242-50.

KAWASHIMA, Y., NAGAI, H., KONNO, R., ISHIKAWA, M., NAKAJIMA, D., SATO, H., NAKAMURA, R., FURUYASHIKI, T. & OHARA, O. 2022. Single-Shot 10K Proteome Approach: Over 10,000 Protein Identifications by Data-Independent Acquisition-Based Single-Shot Proteomics with Ion Mobility Spectrometry. *J Proteome Res,* 21**,** 1418-1427.

KELLEHER, N. L. 2004. Top-down proteomics. *Anal Chem,* 76**,** 197A-203A.

KHAN, S., SHAFIEI, M. S., LONGORIA, C., SCHOGGINS, J., SAVANI, R. C. & ZAKI, H. 2021. SARS-CoV-2 spike protein induces inflammation via TLR2-dependent activation of the NF-kappaB pathway. *bioRxiv*.

KILPINEN, H., GONCALVES, A., LEHA, A., AFZAL, V., ALASOO, K., ASHFORD, S., BALA, S., BENSADDEK, D., CASALE, F. P., CULLEY, O. J., DANECEK, P., FAULCONBRIDGE, A., HARRISON, P. W., KATHURIA, A., MCCARTHY, D., MCCARTHY, S. A., MELECKYTE, R., MEMARI, Y., MOENS, N., SOARES, F., MANN, A., STREETER, I., AGU, C. A., ALDERTON, A., NELSON, R., HARPER, S., PATEL, M., WHITE, A., PATEL, S. R., CLARKE, L., HALAI, R., KIRTON, C. M., KOLB-KOKOCINSKI, A., BEALES, P., BIRNEY, E., DANOVI, D., LAMOND, A. I., OUWEHAND, W. H., VALLIER, L., WATT, F. M., DURBIN, R., STEGLE, O. & GAFFNEY, D. J. 2017. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature,* 546**,** 370-375.

KIM, D. H., SARBASSOV, D. D., ALI, S. M., KING, J. E., LATEK, R. R., ERDJUMENT-BROMAGE, H., TEMPST, P. & SABATINI, D. M. 2002. mTOR interacts with raptor to form a nutrient-sensitive complex that signals to the cell growth machinery. *Cell,* 110**,** 163-75.

KIM, S., MISCHERIKOW, N., BANDEIRA, N., NAVARRO, J. D., WICH, L., MOHAMMED, S., HECK, A. J. & PEVZNER, P. A. 2010. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Mol Cell Proteomics,* 9**,** 2840-52.

KIMBREL, E. A. & LANZA, R. 2015. Current status of pluripotent stem cells: moving the first therapies to the clinic. *Nat Rev Drug Discov,* 14**,** 681-92.

KIRKWOOD, K. J., AHMAD, Y., LARANCE, M. & LAMOND, A. I. 2013. Characterization of native protein complexes and protein isoform variation using size-fractionation-based quantitative proteomics. *Mol Cell Proteomics,* 12**,** 3851-73.

KLAEGER, S., HEINZLMEIR, S., WILHELM, M., POLZER, H., VICK, B., KOENIG, P. A., REINECKE, M., RUPRECHT, B., PETZOLDT, S., MENG, C., ZECHA, J., REITER, K., QIAO, H., HELM, D., KOCH, H., SCHOOF, M., CANEVARI, G., CASALE, E., DEPAOLINI, S. R., FEUCHTINGER, A., WU, Z., SCHMIDT, T., RUECKERT, L., BECKER, W., HUENGES, J., GARZ, A. K., GOHLKE, B. O., ZOLG, D. P., KAYSER, G., VOODER, T., PREISSNER, R., HAHNE, H., TONISSON, N., KRAMER, K., GOTZE, K., BASSERMANN, F., SCHLEGL, J., EHRLICH, H. C., AICHE, S., WALCH, A., GREIF, P. A., SCHNEIDER, S., FELDER, E. R., RULAND, J., MEDARD, G., JEREMIAS, I., SPIEKERMANN, K. & KUSTER, B. 2017. The target landscape of clinical kinase drugs. *Science,* 358.

KNIGHT, S. R., HO, A., PIUS, R., BUCHAN, I., CARSON, G., DRAKE, T. M., DUNNING, J., FAIRFIELD, C. J., GAMBLE, C., GREEN, C. A., GUPTA, R., HALPIN, S., HARDWICK, H. E., HOLDEN, K. A., HORBY, P. W., JACKSON, C., MCLEAN, K. A., MERSON, L., NGUYEN-VAN-TAM, J. S., NORMAN, L., NOURSADEGHI, M., OLLIARO, P. L., PRITCHARD, M. G., RUSSELL, C. D., SHAW, C. A., SHEIKH, A., SOLOMON, T., SUDLOW, C., SWANN, O. V., TURTLE, L. C., OPENSHAW, P. J., BAILLIE, J. K., SEMPLE, M. G., DOCHERTY, A. B., HARRISON, E. M. & INVESTIGATORS, I. C. 2020. Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score. *BMJ,* 370**,** m3339.

KONDO, M. 2010. Lymphoid and myeloid lineage commitment in multipotent hematopoietic progenitors. *Immunol Rev,* 238**,** 37-46.

KONG, A. T., LEPREVOST, F. V., AVTONOMOV, D. M., MELLACHERUVU, D. & NESVIZHSKII, A. I. 2017. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods,* 14**,** 513-520.

KRUG, K., JAEHNIG, E. J., SATPATHY, S., BLUMENBERG, L., KARPOVA, A., ANURAG, M., MILES, G., MERTINS, P., GEFFEN, Y., TANG, L. C., HEIMAN, D. I., CAO, S., MARUVKA, Y. E., LEI, J. T., HUANG, C., KOTHADIA, R. B., COLAPRICO, A., BIRGER, C., WANG, J., DOU, Y., WEN, B., SHI, Z., LIAO, Y., WIZNEROWICZ, M., WYCZALKOWSKI, M. A., CHEN, X. S., KENNEDY, J. J., PAULOVICH, A. G., THIAGARAJAN, M., KINSINGER, C. R., HILTKE, T., BOJA, E. S., MESRI, M., ROBLES, A. I., RODRIGUEZ, H., WESTBROOK, T. F., DING, L., GETZ, G., CLAUSER, K. R., FENYO, D., RUGGLES, K. V., ZHANG, B., MANI, D. R., CARR, S. A., ELLIS, M. J., GILLETTE, M. A. & CLINICAL PROTEOMIC TUMOR ANALYSIS, C. 2020. Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. *Cell,* 183**,** 1436-1456 e31.

KURI-CERVANTES, L., PAMPENA, M. B., MENG, W., ROSENFELD, A. M., ITTNER, C. A. G., WEISMAN, A. R., AGYEKUM, R. S., MATHEW, D., BAXTER, A. E., VELLA, L. A., KUTHURU, O., APOSTOLIDIS, S. A., BERSHAW, L., DOUGHERTY, J., GREENPLATE, A. R., PATTEKAR, A., KIM, J., HAN, N., GOUMA, S., WEIRICK, M. E., AREVALO, C. P., BOLTON, M. J., GOODWIN, E. C., ANDERSON, E. M., HENSLEY, S. E., JONES, T. K., MANGALMURTI, N. S., LUNING PRAK, E. T., WHERRY, E. J., MEYER, N. J. & BETTS, M. R. 2020. Comprehensive mapping of immune perturbations associated with severe COVID-19. *Sci Immunol,* 5.

LAING, A. G., LORENC, A., DEL MOLINO DEL BARRIO, I., DAS, A., FISH, M., MONIN, L., MUNOZ-RUIZ, M., MCKENZIE, D. R., HAYDAY, T. S., FRANCOS-QUIJORNA, I., KAMDAR, S., JOSEPH, M., DAVIES, D., DAVIS, R., JENNINGS, A., ZLATAREVA, I., VANTOUROUT, P., WU, Y., SOFRA, V., CANO, F., GRECO, M., THEODORIDIS, E., FREEDMAN, J. D., GEE, S., CHAN, J. N. E., RYAN, S., BUGALLO-BLANCO, E., PETERSON, P., KISAND, K., HALJASMAGI, L., CHADLI, L., MOINGEON, P., MARTINEZ, L., MERRICK, B., BISNAUTHSING, K., BROOKS, K., IBRAHIM, M. A. A., MASON, J., LOPEZ GOMEZ, F., BABALOLA, K., ABDUL-JAWAD, S., CASON, J., MANT, C., SEOW, J., GRAHAM, C., DOORES, K. J., DI ROSA, F., EDGEWORTH, J., SHANKAR-HARI, M. & HAYDAY, A. C. 2020. A dynamic COVID-19 immune signature includes associations with poor prognosis. *Nat Med,* 26**,** 1623-1635.

LARANCE, M., KIRKWOOD, K. J., TINTI, M., BRENES MURILLO, A., FERGUSON, M. A. & LAMOND, A. I. 2016. Global Membrane Protein Interactome Analysis using In vivo Crosslinking and Mass Spectrometry-based Protein Correlation Profiling. *Mol Cell Proteomics,* 15**,** 2476-90.

LARANCE, M. & LAMOND, A. I. 2015. Multidimensional proteomics for cell biology. *Nat Rev Mol Cell Biol,* 16**,** 269-80.

LARANCE, M., POURKARIMI, E., WANG, B., BRENES MURILLO, A., KENT, R., LAMOND, A. I. & GARTNER, A. 2015. Global Proteomics Analysis of the Response to Starvation in C. elegans. *Mol Cell Proteomics,* 14**,** 1989-2001.

LAWRENCE, S. M., CORRIDEN, R. & NIZET, V. 2018. The Ontogeny of a Neutrophil: Mechanisms of Granulopoiesis and Homeostasis. *Microbiol Mol Biol Rev,* 82.

LAZAR, C., GATTO, L., FERRO, M., BRULEY, C. & BURGER, T. 2016. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J Proteome Res,* 15**,** 1116-25.

LEE, G., PAPAPETROU, E. P., KIM, H., CHAMBERS, S. M., TOMISHIMA, M. J., FASANO, C. A., GANAT, Y. M., MENON, J., SHIMIZU, F., VIALE, A., TABAR, V., SADELAIN, M. & STUDER, L. 2009. Modelling pathogenesis and treatment of familial dysautonomia using patient-specific iPSCs. *Nature,* 461**,** 402-6.

LEE, J. S. & SHIN, E. C. 2020. The type I interferon response in COVID-19: implications for treatment. *Nat Rev Immunol,* 20**,** 585-586.

LEHRER, R. I., SZKLAREK, D., BARTON, A., GANZ, T., HAMANN, K. J. & GLEICH, G. J. 1989. Antibacterial properties of eosinophil major basic protein and eosinophil cationic protein. *J Immunol,* 142**,** 4428-34.

LEITNER, A., REISCHL, R., WALZTHOENI, T., HERZOG, F., BOHN, S., FORSTER, F. & AEBERSOLD, R. 2012. Expanding the chemical cross-linking toolbox by the use of multiple proteases and enrichment by size exclusion chromatography. *Mol Cell Proteomics,* 11**,** M111 014126.

LI, J., CAI, Z., BOMGARDEN, R. D., PIKE, I., KUHN, K., ROGERS, J. C., ROBERTS, T. M., GYGI, S. P. & PAULO, J. A. 2021. TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing. *J Proteome Res,* 20**,** 2964-2972.

LI, J., VAN VRANKEN, J. G., PONTANO VAITES, L., SCHWEPPE, D. K., HUTTLIN, E. L., ETIENNE, C., NANDHIKONDA, P., VINER, R., ROBITAILLE, A. M., THOMPSON, A. H., KUHN, K., PIKE, I., BOMGARDEN, R. D., ROGERS, J. C., GYGI, S. P. & PAULO, J. A. 2020. TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat Methods,* 17**,** 399-404.

LI, Y. F., ARNOLD, R. J., LI, Y., RADIVOJAC, P., SHENG, Q. & TANG, H. 2009. A bayesian approach to protein inference problem in shotgun proteomics. *J Comput Biol,* 16**,** 1183-93.

LIAO, Y., SMYTH, G. K. & SHI, W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics,* 30**,** 923-30.

LIM, M. Y., PAULO, J. A. & GYGI, S. P. 2019. Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model. *J Proteome Res,* 18**,** 4020-4026.

LIMPERT, E., STAHEL, W. & ABBT, M. 2001. Log-normal Distributions across the Sciences: Keys and Clues. *BioSciences,* 51**,** 341-352.

LIN, L. I. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics,* 45**,** 255-68.

LIU, G. H., QU, J., SUZUKI, K., NIVET, E., LI, M., MONTSERRAT, N., YI, F., XU, X., RUIZ, S., ZHANG, W., WAGNER, U., KIM, A., REN, B., LI, Y., GOEBL, A., KIM, J., SOLIGALLA, R. D., DUBOVA, I., THOMPSON, J., YATES, J., 3RD, ESTEBAN, C. R., SANCHO-MARTINEZ, I. & IZPISUA BELMONTE, J. C. 2012. Progressive degeneration of human neural stem cells caused by pathogenic LRRK2. *Nature,* 491**,** 603-7.

LIU, P., BEER, L. A., KY, B., BARNHART, K. T. & SPEICHER, D. W. 2017. Quantitative Comparisons of Large Numbers of Human Plasma Samples Using TMT10plex Labeling. *Methods Mol Biol,* 1619**,** 319-337.

LIU, Y., BEYER, A. & AEBERSOLD, R. 2016. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell,* 165**,** 535-50.

LIVERNOIS, A. M., GRAVES, J. A. & WATERS, P. D. 2012. The origin and evolution of vertebrate sex chromosomes and dosage compensation. *Heredity (Edinb),* 108**,** 50-8.

LONG, M. B., HOWDEN, A. J. M., KEIR, H., ROLLINGS, C. M., GIAM, Y. H., PEMBRIDGE, T., ABO-LEYAH, H., LLOYD, A., SOLLBERGER, G., HULL, R., GILMOUR, A., NEW, B. J. M., CASSIDY, D., SHOEMARK, A., RICHARDSON, H., LAMOND, A. I., CANTRELL, D. A., CHALMERS, J. D. & BRENES, A. J. 2022. Neutrophil proteomics identifies temporal changes and hallmarks of delayed recovery in COVID19. medRxiv.

LOVEN, J., ORLANDO, D. A., SIGOVA, A. A., LIN, C. Y., RAHL, P. B., BURGE, C. B., LEVENS, D. L., LEE, T. I. & YOUNG, R. A. 2012. Revisiting global gene expression analysis. *Cell,* 151**,** 476-82.

LUCAS, C., WONG, P., KLEIN, J., CASTRO, T. B. R., SILVA, J., SUNDARAM, M., ELLINGSON, M. K., MAO, T., OH, J. E., ISRAELOW, B., TAKAHASHI, T., TOKUYAMA, M., LU, P., VENKATARAMAN, A., PARK, A., MOHANTY, S., WANG, H., WYLLIE, A. L., VOGELS, C. B. F., EARNEST, R., LAPIDUS, S., OTT, I. M., MOORE, A. J., MUENKER, M. C., FOURNIER, J. B., CAMPBELL, M., ODIO, C. D., CASANOVAS-MASSANA, A., YALE, I. T., HERBST, R., SHAW, A. C., MEDZHITOV, R., SCHULZ, W. L., GRUBAUGH, N. D., DELA CRUZ, C., FARHADIAN, S., KO, A. I., OMER, S. B. & IWASAKI, A. 2020. Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature,* 584**,** 463-469.

LUDWIG, T. E., BERGENDAHL, V., LEVENSTEIN, M. E., YU, J., PROBASCO, M. D. & THOMSON, J. A. 2006. Feeder-independent culture of human embryonic stem cells. *Nat Methods,* 3**,** 637-46.

LUNDBERG, E., FAGERBERG, L., KLEVEBRING, D., MATIC, I., GEIGER, T., COX, J., ALGENAS, C., LUNDEBERG, J., MANN, M. & UHLEN, M. 2010. Defining the transcriptome and proteome in three functionally different human cell lines. *Mol Syst Biol,* 6**,** 450.

LY, T., AHMAD, Y., SHLIEN, A., SOROKA, D., MILLS, A., EMANUELE, M. J., STRATTON, M. R. & LAMOND, A. I. 2014. A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells. *Elife,* 3**,** e01630.

LY, T., ENDO, A., BRENES, A., GIERLINSKI, M., AFZAL, V., PAWELLEK, A. & LAMOND, A. I. 2018. Proteome-wide analysis of protein abundance and turnover remodelling during oncogenic transformation of human breast epithelial cells. *Wellcome Open Res,* 3**,** 51.

LY, T., ENDO, A. & LAMOND, A. I. 2015. Proteomic analysis of the response to cell cycle arrests in human myeloid leukemia cells. *Elife,* 4.

LY, T., WHIGHAM, A., CLARKE, R., BRENES-MURILLO, A. J., ESTES, B., MADHESSIAN, D., LUNDBERG, E., WADSWORTH, P. & LAMOND, A. I. 2017. Proteomic analysis of cell cycle progression in asynchronous cultures, including mitotic subphases, using PRIMMUS. *Elife,* 6.

LYST, M. J. & BIRD, A. 2015. Rett syndrome: a complex disorder with simple roots. *Nat Rev Genet,* 16**,** 261-75.

MACKAY, L. K., RAHIMPOUR, A., MA, J. Z., COLLINS, N., STOCK, A. T., HAFON, M. L., VEGA-RAMOS, J., LAUZURICA, P., MUELLER, S. N., STEFANOVIC, T., TSCHARKE, D. C., HEATH, W. R., INOUYE, M., CARBONE, F. R. & GEBHARDT, T. 2013. The developmental pathway for CD103(+)CD8+ tissue-resident memory T cells of skin. *Nat Immunol,* 14**,** 1294-301.

MACLEAN, B., TOMAZELA, D. M., SHULMAN, N., CHAMBERS, M., FINNEY, G. L., FREWEN, B., KERN, R., TABB, D. L., LIEBLER, D. C. & MACCOSS, M. J. 2010. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics,* 26**,** 966-8.

MAES, E., HADIWIKARTA, W. W., MERTENS, I., BAGGERMAN, G., HOOYBERGHS, J. & VALKENBORG, D. 2016. CONSTANd : A Normalization Method for Isobaric Labeled Spectra by Constrained Optimization. *Mol Cell Proteomics,* 15**,** 2779-90.

MAHERALI, N., AHFELDT, T., RIGAMONTI, A., UTIKAL, J., COWAN, C. & HOCHEDLINGER, K. 2008. A high-efficiency system for the generation and study of human induced pluripotent stem cells. *Cell Stem Cell,* 3**,** 340-5.

MALUMBRES, M. 2014. Cyclin-dependent kinases. *Genome Biol,* 15**,** 122.

MANN, E. R., MENON, M., KNIGHT, S. B., KONKEL, J. E., JAGGER, C., SHAW, T. N., KRISHNAN, S., RATTRAY, M., USTIANOWSKI, A., BAKERLY, N. D., DARK, P., LORD, G., SIMPSON, A., FELTON, T., HO, L. P., TRC, N. R., FELDMANN, M., CIRCO, GRAINGER, J. R. & HUSSELL, T. 2020. Longitudinal immune profiling reveals key myeloid signatures associated with COVID-19. *Sci Immunol,* 5.

MANN, M. 2019. The ever expanding scope of electrospray mass spectrometry-a 30 year journey. *Nat Commun,* 10**,** 3744.

MANNING, G., WHYTE, D. B., MARTINEZ, R., HUNTER, T. & SUDARSANAM, S. 2002. The protein kinase complement of the human genome. *Science,* 298**,** 1912-34.

MANZ, M. G. & BOETTCHER, S. 2014. Emergency granulopoiesis. *Nat Rev Immunol,* 14**,** 302-14.

MARAHRENS, Y., PANNING, B., DAUSMAN, J., STRAUSS, W. & JAENISCH, R. 1997. Xist-deficient mice are defective in dosage compensation but not spermatogenesis. *Genes Dev,* 11**,** 156-66.

MARCHINGO, J. M. & CANTRELL, D. A. 2019. Proteome of naive and TCR activated wild-type, Myc-deficient and Slc7a5-deficient T cells. *Proteome of naive and TCR activated wild-type, Myc-deficient and Slc7a5-deficient T cells.* https://www.ebi.ac.uk/pride/archive/projects/PXD016105.

MARCHINGO, J. M. & CANTRELL, D. A. 2020. OT1 T cell activation time course. *OT1 T cell activation time course.* https://www.ebi.ac.uk/pride/archive/projects/PXD016443.

MARCHINGO, J. M. & CANTRELL, D. A. 2022. Protein synthesis, degradation, and energy metabolism in T cell immunity. *Cell Mol Immunol,* 19**,** 303-315.

MARCHINGO, J. M., SINCLAIR, L. V., HOWDEN, A. J. & CANTRELL, D. A. 2020. Quantitative analysis of how Myc controls T cell proteomes and metabolic pathways during T cell activation. *Elife,* 9.

MARDAKHEH, F. K., SAILEM, H. Z., KUMPER, S., TAPE, C. J., MCCULLY, R. R., PAUL, A., ANJOMANI-VIRMOUNI, S., JORGENSEN, C., POULOGIANNIS, G., MARSHALL, C. J. & BAKAL, C. 2016. Proteomics profiling of interactome dynamics by colocalisation analysis (COLA). *Mol Biosyst,* 13**,** 92-105.

MARGUERAT, S., SCHMIDT, A., CODLIN, S., CHEN, W., AEBERSOLD, R. & BAHLER, J. 2012. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell,* 151**,** 671-83.

MARTIN-PEREZ, M. & VILLEN, J. 2017. Determinants and Regulation of Protein Turnover in Yeast. *Cell Syst,* 5**,** 283-294 e5.

MARTINEZ-VAL, A., BEKKER-JENSEN, D. B., STEIGERWALD, S., KOENIG, C., OSTERGAARD, O., MEHTA, A., TRAN, T., SIKORSKI, K., TORRES-VEGA, E., KWASNIEWICZ, E., BRYNJOLFSDOTTIR, S. H., FRANKEL, L. B., KJOBSTED, R., KROGH, N., LUNDBY, A., BEKKER-JENSEN, S., LUND-JOHANSEN, F. & OLSEN, J. V. 2021. Spatial-proteomics reveals phospho-signaling dynamics at subcellular resolution. *Nat Commun,* 12**,** 7113.

MASOPUST, D., CHOO, D., VEZYS, V., WHERRY, E. J., DURAISWAMY, J., AKONDY, R., WANG, J., CASEY, K. A., BARBER, D. L., KAWAMURA, K. S., FRASER, K. A., WEBBY, R. J., BRINKMANN, V., BUTCHER, E. C., NEWELL, K. A. & AHMED, R. 2010. Dynamic T cell migration program provides resident memory within intestinal epithelium. *J Exp Med,* 207**,** 553-64.

MATHEW, D., GILES, J. R., BAXTER, A. E., OLDRIDGE, D. A., GREENPLATE, A. R., WU, J. E., ALANIO, C., KURI-CERVANTES, L., PAMPENA, M. B., D'ANDREA, K., MANNE, S., CHEN, Z., HUANG, Y. J., REILLY, J. P., WEISMAN, A. R., ITTNER, C. A. G., KUTHURU, O., DOUGHERTY, J., NZINGHA, K., HAN, N., KIM, J., PATTEKAR, A., GOODWIN, E. C., ANDERSON, E. M., WEIRICK, M. E., GOUMA, S., AREVALO, C. P., BOLTON, M. J., CHEN, F., LACEY, S. F., RAMAGE, H., CHERRY, S., HENSLEY, S. E., APOSTOLIDIS, S. A., HUANG, A. C., VELLA, L. A., UNIT, U. P. C. P., BETTS, M. R., MEYER, N. J. & WHERRY, E. J. 2020. Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. *Science,* 369.

MCALISTER, G. C., NUSINOW, D. P., JEDRYCHOWSKI, M. P., WUHR, M., HUTTLIN, E. L., ERICKSON, B. K., RAD, R., HAAS, W. & GYGI, S. P. 2014. MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal Chem,* 86**,** 7150-8.

MCGETTRICK, A. F. & O'NEILL, L. A. J. 2020. The Role of HIF in Immunity and Inflammation. *Cell Metab,* 32**,** 524-536.

MEIER, F., GEYER, P. E., VIRREIRA WINTER, S., COX, J. & MANN, M. 2018. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat Methods*.

MEIER, F., KOHLER, N. D., BRUNNER, A. D., WANKA, J. H., VOYTIK, E., STRAUSS, M. T., THEIS, F. J. & MANN, M. 2021. Deep learning the collisional cross sections of the peptide universe from a million experimental values. *Nat Commun,* 12**,** 1185.

MEIZLISH, M. L., PINE, A. B., BISHAI, J. D., GOSHUA, G., NADELMANN, E. R., SIMONOV, M., CHANG, C. H., ZHANG, H., SHALLOW, M., BAHEL, P., OWUSU, K., YAMAMOTO, Y., ARORA, T., ATRI, D. S., PATEL, A., GBYLI, R., KWAN, J., WON, C. H., DELA CRUZ, C., PRICE, C., KOFF, J., KING, B. A., RINDER, H. M., WILSON, F. P., HWA, J., HALENE, S., DAMSKY, W., VAN DIJK, D., LEE, A. I. & CHUN, H. J. 2021. A neutrophil activation signature predicts critical illness and mortality in COVID-19. *Blood Adv,* 5**,** 1164-1177.

MEKHOUBAD, S., BOCK, C., DE BOER, A. S., KISKINIS, E., MEISSNER, A. & EGGAN, K. 2012. Erosion of dosage compensation impacts human iPSC disease modeling. *Cell Stem Cell,* 10**,** 595-609.

MELLACHERUVU, D., WRIGHT, Z., COUZENS, A. L., LAMBERT, J. P., ST-DENIS, N. A., LI, T., MITEVA, Y. V., HAURI, S., SARDIU, M. E., LOW, T. Y., HALIM, V. A., BAGSHAW, R. D., HUBNER, N. C., AL-HAKIM, A., BOUCHARD, A., FAUBERT, D., FERMIN, D., DUNHAM, W. H., GOUDREAULT, M., LIN, Z. Y., BADILLO, B. G., PAWSON, T., DUROCHER, D., COULOMBE, B., AEBERSOLD, R., SUPERTI-FURGA, G., COLINGE, J., HECK, A. J., CHOI, H., GSTAIGER, M., MOHAMMED, S., CRISTEA, I. M., BENNETT, K. L., WASHBURN, M. P., RAUGHT, B., EWING, R. M., GINGRAS, A. C. &

NESVIZHSKII, A. I. 2013. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat Methods,* 10**,** 730-6.

MELMS, J. C., BIERMANN, J., HUANG, H., WANG, Y., NAIR, A., TAGORE, S., KATSYV, I., RENDEIRO, A. F., AMIN, A. D., SCHAPIRO, D., FRANGIEH, C. J., LUOMA, A. M., FILLIOL, A., FANG, Y., RAVICHANDRAN, H., CLAUSI, M. G., ALBA, G. A., ROGAVA, M., CHEN, S. W., HO, P., MONTORO, D. T., KORNBERG, A. E., HAN, A. S., BAKHOUM, M. F., ANANDASABAPATHY, N., SUAREZ-FARINAS, M., BAKHOUM, S. F., BRAM, Y., BORCZUK, A., GUO, X. V., LEFKOWITCH, J. H., MARBOE, C., LAGANA, S. M., DEL PORTILLO, A., TSAI, E. J., ZORN, E., MARKOWITZ, G. S., SCHWABE, R. F., SCHWARTZ, R. E., ELEMENTO, O., SAQI, A., HIBSHOOSH, H., QUE, J. & IZAR, B. 2021. A molecular single-cell lung atlas of lethal COVID-19. *Nature,* 595**,** 114-119.

MERRICK, W. C. 2015. eIF4F: a retrospective. *J Biol Chem,* 290**,** 24091-9.

MESSNER, C. B., DEMICHEV, V., WENDISCH, D., MICHALICK, L., WHITE, M., FREIWALD, A., TEXTORIS-TAUBE, K., VERNARDIS, S. I., EGGER, A. S., KREIDL, M., LUDWIG, D., KILIAN, C., AGOSTINI, F., ZELEZNIAK, A., THIBEAULT, C., PFEIFFER, M., HIPPENSTIEL, S., HOCKE, A., VON KALLE, C., CAMPBELL, A., HAYWARD, C., PORTEOUS, D. J., MARIONI, R. E., LANGENBERG, C., LILLEY, K. S., KUEBLER, W. M., MULLEDER, M., DROSTEN, C., SUTTORP, N., WITZENRATH, M., KURTH, F., SANDER, L. E. & RALSER, M. 2020. Ultra-High-Throughput Clinical Proteomics Reveals Classifiers of COVID-19 Infection. *Cell Syst,* 11**,** 11-24 e4.

MI, H., HUANG, X., MURUGANUJAN, A., TANG, H., MILLS, C., KANG, D. & THOMAS, P. D. 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res,* 45**,** D183-D189.

MICHALSKI, A., COX, J. & MANN, M. 2011. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J Proteome Res,* 10**,** 1785-93.

MIRAUTA, B. A., SEATON, D. D., BENSADDEK, D., BRENES, A., BONDER, M. J., KILPINEN, H., HIPSCI, C., AGU, C. A., ALDERTON, A., DANECEK, P., DENTON, R., DURBIN, R., GAFFNEY, D. J., GONCALVES, A., HALAI, R., HARPER, S., KIRTON, C. M., KOLB-KOKOCINSKI, A., LEHA, A., MCCARTHY, S. A., MEMARI, Y., PATEL, M., BIRNEY, E., CASALE, F. P., CLARKE, L., HARRISON, P. W., KILPINEN, H., STREETER, I., DENOVI, D., STEGLE, O., LAMOND, A. I., MELECKYTE, R., MOENS, N., WATT, F. M., OUWEHAND, W. H., BEALES, P., STEGLE, O. & LAMOND, A. I. 2020. Population-scale proteome variation in human induced pluripotent stem cells. *Elife,* 9.

MISHRA, A., HOGAN, S. P., LEE, J. J., FOSTER, P. S. & ROTHENBERG, M. E. 1999. Fundamental signals that regulate eosinophil homing to the gastrointestinal tract. *J Clin Invest,* 103**,** 1719-27.

MORRISSEY, S. M., GELLER, A. E., HU, X., TIERI, D., DING, C., KLAES, C. K., COOKE, E. A., WOESTE, M. R., MARTIN, Z. C., CHEN, O., BUSH, S. E., ZHANG, H. G., CAVALLAZZI, R., CLIFFORD, S. P., CHEN, J., GHARE, S., BARVE, S. S., CAI, L., KONG, M., ROUCHKA, E. C., MCLEISH, K. R., URIARTE, S. M., WATSON, C. T., HUANG, J. & YAN, J. 2021. A specific low-density neutrophil population correlates with hypercoagulation and disease severity in hospitalized COVID-19 patients. *JCI Insight,* 6.

MULVEY, C. M., BRECKELS, L. M., GELADAKI, A., BRITOVSEK, N. K., NIGHTINGALE, D. J. H., CHRISTOFOROU, A., ELZEK, M., DEERY, M. J., GATTO, L. & LILLEY, K. S. 2017. Using hyperLOPIT to perform high-resolution mapping of the spatial proteome. *Nat Protoc,* 12**,** 1110-1135.

MUNCHBACH, M., QUADRONI, M., MIOTTO, G. & JAMES, P. 2000. Quantitation and facilitated de novo sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety. *Anal Chem,* 72**,** 4047-57.

MUNTEL, J., GANDHI, T., VERBEKE, L., BERNHARDT, O. M., TREIBER, T., BRUDERER, R. & REITER, L. 2019a. Surpassing 10 000 identified and quantified proteins in a single run by optimizing current LC-MS instrumentation and data analysis strategy. *Mol Omics,* 15**,** 348-360.

MUNTEL, J., KIRKPATRICK, J., BRUDERER, R., HUANG, T., VITEK, O., ORI, A. & REITER, L. 2019b. Comparison of Protein Quantification in a Complex Background by DIA and TMT Workflows with Fixed Instrument Time. *J Proteome Res,* 18**,** 1340-1351.

MYCKO, M. P., FERRERO, I., WILSON, A., JIANG, W., BIANCHI, T., TRUMPP, A. & MACDONALD, H. R. 2009. Selective requirement for c-Myc at an early stage of V(alpha)14i NKT cell development. *J Immunol,* 182**,** 4641-8.

MYERS, S. A., KLAEGER, S., SATPATHY, S., VINER, R., CHOI, J., ROGERS, J., CLAUSER, K., UDESHI, N. D. & CARR, S. A. 2018. Evaluation of Advanced Precursor Determination for Tandem Mass Tag (TMT)-Based Quantitative Proteomics across Instrument Platforms. *J Proteome Res*.

NAG, N., LIN, K. Y., EDMONDS, K. A., YU, J., NADKARNI, D., MARINTCHEVA, B. & MARINTCHEV, A. 2016. eIF1A/eIF5B interaction network and its functions in translation initiation complex assembly and remodeling. *Nucleic Acids Res,* 44**,** 7441-56.

NAKAGAWA, M., KOYANAGI, M., TANABE, K., TAKAHASHI, K., ICHISAKA, T., AOI, T., OKITA, K., MOCHIDUKI, Y., TAKIZAWA, N. & YAMANAKA, S. 2008. Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nat Biotechnol,* 26**,** 101-6.

NESVIZHSKII, A. I. & AEBERSOLD, R. 2005. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics,* 4**,** 1419-40.

NESVIZHSKII, A. I., KELLER, A., KOLKER, E. & AEBERSOLD, R. 2003. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem,* 75**,** 4646-58.

NETT, I. R., MARTIN, D. M., MIRANDA-SAAVEDRA, D., LAMONT, D., BARBER, J. D., MEHLERT, A. & FERGUSON, M. A. 2009. The phosphoproteome of bloodstream form Trypanosoma brucei, causative agent of African sleeping sickness. *Mol Cell Proteomics,* 8**,** 1527-38.

NICOLAI, L., LEUNIG, A., BRAMBS, S., KAISER, R., WEINBERGER, T., WEIGAND, M., MUENCHHOFF, M., HELLMUTH, J. C., LEDDEROSE, S., SCHULZ, H., SCHERER, C., RUDELIUS, M., ZOLLER, M., HÖCHTER, D., KEPPLER, O., TEUPSER, D., ZWISSLE, B., VON BERGWELT-BAILDON, M., KÄÄB, S., MASSBERG, S., PEKAYVAZ, K. & STARK, K. 2020. Immunothrombotic Dysregulation in COVID-19 Pneumonia Is Associated With Respiratory Failure and Coagulopathy. *Circulation,* 142**,** 1176-1189.

NUSINOW, D. P., SZPYT, J., GHANDI, M., ROSE, C. M., MCDONALD, E. R., 3RD, KALOCSAY, M., JANE-VALBUENA, J., GELFAND, E., SCHWEPPE, D. K., JEDRYCHOWSKI, M., GOLJI, J., PORTER, D. A., REJTAR, T., WANG, Y. K., KRYUKOV, G. V., STEGMEIER, F., ERICKSON, B. K., GARRAWAY, L. A., SELLERS, W. R. & GYGI, S. P. 2020. Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell,* 180**,** 387-402 e16.

O'CONNELL, J. D., PAULO, J. A., O'BRIEN, J. J. & GYGI, S. P. 2018. Proteome-Wide Evaluation of Two Common Protein Quantification Methods. *J Proteome Res,* 17**,** 1934-1942.

OCHOA, D., JARNUCZAK, A. F., VIEITEZ, C., GEHRE, M., SOUCHERAY, M., MATEUS, A., KLEEFELDT, A. A., HILL, A., GARCIA-ALONSO, L., STEIN, F., KROGAN, N. J., SAVITSKI, M. M., SWANEY, D. L., VIZCAINO, J. A., NOH, K. M. & BELTRAO, P. 2020. The functional landscape of the human phosphoproteome. *Nat Biotechnol,* 38**,** 365-373.

OKITA, K., NAKAGAWA, M., HYENJONG, H., ICHISAKA, T. & YAMANAKA, S. 2008. Generation of mouse induced pluripotent stem cells without viral vectors. *Science,* 322**,** 949-53.

ONG, S. E., BLAGOEV, B., KRATCHMAROVA, I., KRISTENSEN, D. B., STEEN, H., PANDEY, A. & MANN, M. 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics,* 1**,** 376-86.

ONO, M., BOLLAND, S., TEMPST, P. & RAVETCH, J. V. 1996. Role of the inositol phosphatase SHIP in negative regulation of the immune system by the receptor Fc(gamma)RIIB. *Nature,* 383**,** 263-6.

ONO, M., OKADA, H., BOLLAND, S., YANAGI, S., KUROSAKI, T. & RAVETCH, J. V. 1997. Deletion of SHIP or SHP-1 reveals two distinct pathways for inhibitory signaling. *Cell,* 90**,** 293-301.

OSNATO, A., BROWN, S., KRUEGER, C., ANDREWS, S., COLLIER, A. J., NAKANOH, S., QUIROGA LONDONO, M., WESLEY, B. T., MURARO, D., BRUMM, A. S., NIAKAN, K. K., VALLIER, L., ORTMANN, D. & RUGG-GUNN, P. J. 2021. TGFbeta signalling is required to maintain pluripotency of human naive pluripotent stem cells. *Elife,* 10.

OVELAND, E., MUTH, T., RAPP, E., MARTENS, L., BERVEN, F. S. & BARSNES, H. 2015. Viewing the proteome: how to visualize proteomics data? *Proteomics,* 15**,** 1341-55.

OW, S. Y., SALIM, M., NOIREL, J., EVANS, C., REHMAN, I. & WRIGHT, P. C. 2009. iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly". *J Proteome Res,* 8**,** 5347-55.

PALLETT, L. J., DAVIES, J., COLBECK, E. J., ROBERTSON, F., HANSI, N., EASOM, N. J. W., BURTON, A. R., STEGMANN, K. A., SCHURICH, A., SWADLING, L., GILL, U. S., MALE, V., LUONG, T., GANDER, A., DAVIDSON, B. R., KENNEDY, P. T. F. & MAINI, M. K. 2017. IL-2(high) tissue-resident T cells in the human liver: Sentinels for hepatotropic infection. *J Exp Med,* 214**,** 1567-1580.

PALUDAN, S. R., PRADEU, T., MASTERS, S. L. & MOGENSEN, T. H. 2021. Constitutive immune mechanisms: mediators of host defence and immune regulation. *Nat Rev Immunol,* 21**,** 137-150.

PANCHAUD, A., SCHERL, A., SHAFFER, S. A., VON HALLER, P. D., KULASEKARA, H. D., MILLER, S. I. & GOODLETT, D. R. 2009. Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Anal Chem,* 81**,** 6481-8.

PANDA, R., CASTANHEIRA, F. V., SCHLECHTE, J. M., SUREWAARD, B. G., SHIM, H. B., ZUCOLOTO, A. Z., SLAVIKOVA, Z., YIPP, B. G., KUBES, P. & MCDONALD, B. 2022. A functionally distinct neutrophil landscape in severe COVID-19 reveals opportunities for adjunctive therapies. *JCI Insight,* 7.

PANDEY, A. & MANN, M. 2000. Proteomics to study genes and genomes. *Nature,* 405**,** 837-46.

PAPA, S., CHOY, P. M. & BUBICI, C. 2019. The ERK and JNK pathways in the regulation of metabolic reprogramming. *Oncogene,* 38**,** 2223-2240.

PAPPIREDDI, N., MARTIN, L. & WUHR, M. 2019. A Review on Quantitative Multiplexed Proteomics. *Chembiochem,* 20**,** 1210-1224.

PATRO, R., DUGGAL, G., LOVE, M. I., IRIZARRY, R. A. & KINGSFORD, C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods,* 14**,** 417-419.

PAULO, J. A., O'CONNELL, J. D. & GYGI, S. P. 2016. A Triple Knockout (TKO) Proteomics Standard for Diagnosing Ion Interference in Isobaric Labeling Experiments. *J Am Soc Mass Spectrom,* 27**,** 1620-5.

PAVON, E. J., GARCIA-RODRIGUEZ, S., ZUMAQUERO, E., PERANDRES-LOPEZ, R., ROSAL-VELA, A., LARIO, A., LONGOBARDO, V., CARRASCAL, M., ABIAN, J., CALLEJAS-RUBIO, J. L., ORTEGO-CENTENO, N., ZUBIAUR, M. & SANCHO, J. 2012. Increased expression and phosphorylation of the two S100A9 isoforms in mononuclear cells from patients with systemic lupus erythematosus: a proteomic signature for circulating low-density granulocytes. *J Proteomics,* 75**,** 1778-91.

PEARCE, E. L. & PEARCE, E. J. 2013. Metabolic pathways in immune cell activation and quiescence. *Immunity,* 38**,** 633-43.

PENA, C., HURT, E. & PANSE, V. G. 2017. Eukaryotic ribosome assembly, transport and quality control. *Nat Struct Mol Biol,* 24**,** 689-699.

PENNY, G. D., KAY, G. F., SHEARDOWN, S. A., RASTAN, S. & BROCKDORFF, N. 1996. Requirement for Xist in X chromosome inactivation. *Nature,* 379**,** 131-7.

PEREZ-RIVEROL, Y., BAI, J., BANDLA, C., GARCIA-SEISDEDOS, D., HEWAPATHIRANA, S., KAMATCHINATHAN, S., KUNDU, D. J., PRAKASH, A., FRERICKS-ZIPPER, A., EISENACHER, M., WALZER, M., WANG, S., BRAZMA, A. & VIZCAINO, J. A. 2022. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res,* 50**,** D543-D552.

PEREZ-RIVEROL, Y., CSORDAS, A., BAI, J., BERNAL-LLINARES, M., HEWAPATHIRANA, S., KUNDU, D. J., INUGANTI, A., GRISS, J., MAYER, G., EISENACHER, M., PEREZ, E., USZKOREIT, J., PFEUFFER, J., SACHSENBERG, T., YILMAZ, S., TIWARY, S., COX, J., AUDAIN, E., WALZER, M., JARNUCZAK, A. F., TERNENT, T., BRAZMA, A. & VIZCAINO, J. A. 2018. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res*.

PERKINS, D. N., PAPPIN, D. J., CREASY, D. M. & COTTRELL, J. S. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis,* 20**,** 3551-67.

PHIFER-RIXEY, M. & NACHMAN, M. W. 2015. Insights into mammalian biology from the wild house mouse Mus musculus. *Elife,* 4.

PING, L., DUONG, D. M., YIN, L., GEARING, M., LAH, J. J., LEVEY, A. I. & SEYFRIED, N. T. 2018. Global quantitative analysis of the human brain proteome in Alzheimer's and Parkinson's Disease. *Sci Data,* 5**,** 180036.

PIRCHER, H., BURKI, K., LANG, R., HENGARTNER, H. & ZINKERNAGEL, R. M. 1989. Tolerance induction in double specific T-cell receptor transgenic mice varies with antigen. *Nature,* 342**,** 559-61.

PLUBELL, D. L., WILMARTH, P. A., ZHAO, Y., FENTON, A. M., MINNIER, J., REDDY, A. P., KLIMEK, J., YANG, X., DAVID, L. L. & PAMIR, N. 2017. Extended Multiplexing of Tandem Mass Tags (TMT) Labeling Reveals Age and High Fat Diet Specific Proteome Changes in Mouse Epididymal Adipose Tissue. *Mol Cell Proteomics,* 16**,** 873-890.

PLUMB, R. S., JOHNSON, K. A., RAINVILLE, P., SMITH, B. W., WILSON, I. D., CASTRO-PEREZ, J. M. & NICHOLSON, J. K. 2006. UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Commun Mass Spectrom,* 20**,** 1989-94.

POULOS, R. C., HAINS, P. G., SHAH, R., LUCAS, N., XAVIER, D., MANDA, S. S., ANEES, A., KOH, J. M. S., MAHBOOB, S., WITTMAN, M., WILLIAMS, S. G., SYKES, E. K., HECKER, M., DAUSMANN, M., WOUTERS, M. A., ASHMAN, K., YANG, J., WILD, P. J., DEFAZIO, A., BALLEINE, R. L., TULLY, B., AEBERSOLD, R., SPEED, T. P., LIU, Y., REDDEL, R. R., ROBINSON, P. J. & ZHONG, Q. 2020. Strategies to enable large-scale proteomics for reproducible research. *Nat Commun,* 11**,** 3793.

PURVINE, S., EPPEL, J. T., YI, E. C. & GOODLETT, D. R. 2003. Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics,* 3**,** 847-50.

QUAIL, D. F., AMULIC, B., AZIZ, M., BARNES, B. J., ERUSLANOV, E., FRIDLENDER, Z. G., GOODRIDGE, H. S., GRANOT, Z., HIDALGO, A., HUTTENLOCHER, A., KAPLAN, M. J., MALANCHI, I., MERGHOUB, T., MEYLAN, E., MITTAL, V., PITTET, M. J., RUBIO-PONCE, A., UDALOVA, I. A., VAN DEN BERG, T. K., WAGNER, D. D., WANG, P., ZYCHLINSKY, A., DE VISSER, K. E., EGEBLAD, M. & KUBES, P. 2022. Neutrophil phenotypes and functions in cancer: A consensus statement. *J Exp Med,* 219.

REID, G. E. & MCLUCKEY, S. A. 2002. 'Top down' protein characterization via tandem mass spectrometry. *J Mass Spectrom,* 37**,** 663-75.

RENDEIRO, A. F., RAVICHANDRAN, H., BRAM, Y., CHANDAR, V., KIM, J., MEYDAN, C., PARK, J., FOOX, J., HETHER, T., WARREN, S., KIM, Y., REEVES, J., SALVATORE, S., MASON, C. E., SWANSON, E. C., BORCZUK, A. C., ELEMENTO, O. & SCHWARTZ, R. E. 2021. The spatial landscape of lung pathology during COVID-19 progression. *Nature,* 593**,** 564-569.

REUSCH, N., DE DOMENICO, E., BONAGURO, L., SCHULTE-SCHREPPING, J., BASSLER, K., SCHULTZE, J. L. & ASCHENBRENNER, A. C. 2021. Neutrophils in COVID-19. *Front Immunol,* 12**,** 652470.

REYES, L., M, A. S.-G., MORRISON, T., HOWDEN, A. J. M., WATTS, E. R., ARIENTI, S., SADIKU, P., COELHO, P., MIRCHANDANI, A. S., ZHANG, A., HOPE, D., CLARK, S. K., SINGLETON, J., JOHNSTON, S., GRECIAN, R., POON, A., MCNAMARA, S., HARPER, I., FOURMAN, M. H., BRENES, A. J., PATHAK, S., LLOYD, A., BLANCO, G. R., VON KRIEGSHEIM, A., GHESQUIERE, B., VERMAELEN, W., COLOGNA, C. T., DHALIWAL, K., HIRANI, N., DOCKRELL, D. H., WHYTE, M. K. B., GRIFFITH, D., CANTRELL, D. A. & WALMSLEY, S. R. 2021. -------A type I IFN, prothrombotic hyperinflammatory neutrophil signature is distinct for COVID-19 ARDS. *Wellcome Open Res,* 6**,** 38.

RICHTER, J. D. & SONENBERG, N. 2005. Regulation of cap-dependent translation by eIF4E inhibitory proteins. *Nature,* 433**,** 477-80.

RIECKMANN, J. C., GEIGER, R., HORNBURG, D., WOLF, T., KVELER, K., JARROSSAY, D., SALLUSTO, F., SHEN-ORR, S. S., LANZAVECCHIA, A., MANN, M. & MEISSNER, F. 2017. Social network architecture of human immune cells unveiled by quantitative proteomics. *Nat Immunol,* 18**,** 583-593.

RIFFELMACHER, T., CLARKE, A., RICHTER, F. C., STRANKS, A., PANDEY, S., DANIELLI, S., HUBLITZ, P., YU, Z., JOHNSON, E., SCHWERD, T., MCCULLAGH, J., UHLIG, H., JACOBSEN, S. E. W. & SIMON, A. K. 2017. Autophagy-Dependent Generation of Free Fatty Acids Is Critical for Normal Neutrophil Differentiation. *Immunity,* 47**,** 466-480 e5.

RITCHIE, M. E., PHIPSON, B., WU, D., HU, Y., LAW, C. W., SHI, W. & SMYTH, G. K. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res,* 43**,** e47.

RITORTO, M. S., COOK, K., TYAGI, K., PEDRIOLI, P. G. & TROST, M. 2013. Hydrophilic strong anion exchange (hSAX) chromatography for highly orthogonal peptide separation of complex proteomes. *J Proteome Res,* 12**,** 2449-57.

ROBINSON, M. D., MCCARTHY, D. J. & SMYTH, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26**,** 139-40.

ROLFS, Z., FREY, B. L., SHI, X., KAWAI, Y., SMITH, L. M. & WELHAM, N. V. 2021. An atlas of protein turnover rates in mouse tissues. *Nat Commun,* 12**,** 6778.

ROLLINGS, C. M., SINCLAIR, L. V., BRADY, H. J. M., CANTRELL, D. A. & ROSS, S. H. 2018. Interleukin-2 shapes the cytotoxic T cell proteome and immune environment-sensing programs. *Sci Signal,* 11.

RORVIG, S., HONORE, C., LARSSON, L. I., OHLSSON, S., PEDERSEN, C. C., JACOBSEN, L. C., COWLAND, J. B., GARRED, P. & BORREGAARD, N. 2009. Ficolin-1 is present in a highly mobilizable subset of human neutrophil granules and associates with the cell surface after stimulation with fMLP. *J Leukoc Biol,* 86**,** 1439-49.

RORVIG, S., OSTERGAARD, O., HEEGAARD, N. H. & BORREGAARD, N. 2013. Proteome profiling of human neutrophil granule subsets, secretory vesicles, and cell membrane: correlation with transcriptome profiling of neutrophil precursors. *J Leukoc Biol,* 94**,** 711-21.

ROSS, P. L., HUANG, Y. N., MARCHESE, J. N., WILLIAMSON, B., PARKER, K., HATTAN, S., KHAINOVSKI, N., PILLAI, S., DEY, S., DANIELS, S., PURKAYASTHA, S., JUHASZ, P., MARTIN, S., BARTLET-JONES, M., HE, F., JACOBSON, A. & PAPPIN, D. J. 2004. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics,* 3**,** 1154-69.

ROSS, S. H. & CANTRELL, D. A. 2017. Interleukin-2 shapes the cytotoxic T cell proteome and immune-environment sensing. *Interleukin-2 shapes the cytotoxic T cell proteome and immune-environment sensing.* https://www.ebi.ac.uk/pride/archive/projects/PXD008112.

ROSS, S. H. & CANTRELL, D. A. 2018. Signaling and Function of Interleukin-2 in T Lymphocytes. *Annu Rev Immunol,* 36**,** 411-433.

ROSS, S. H. & CANTRELL, D. A. 2021. The impact of hypoxia on the cytotoxic T lymphocyte proteome. *The impact of hypoxia on the cytotoxic T lymphocyte proteome.* https://www.ebi.ac.uk/pride/archive/projects/PXD026223.

ROSS, S. H., ROLLINGS, C. M. & CANTRELL, D. A. 2021. Quantitative Analyses Reveal How Hypoxia Reconfigures the Proteome of Primary Cytotoxic T Lymphocytes. *Front Immunol,* 12**,** 712402.

ROST, H. L., MALMSTROM, L. & AEBERSOLD, R. 2015. Reproducible quantitative proteotype data matrices for systems biology. *Mol Biol Cell,* 26**,** 3926-31.

ROST, H. L., ROSENBERGER, G., NAVARRO, P., GILLET, L., MILADINOVIC, S. M., SCHUBERT, O. T., WOLSKI, W., COLLINS, B. C., MALMSTROM, J., MALMSTROM, L. & AEBERSOLD, R. 2014. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol,* 32**,** 219-23.

ROUX, P. P., SHAHBAZIAN, D., VU, H., HOLZ, M. K., COHEN, M. S., TAUNTON, J., SONENBERG, N. & BLENIS, J. 2007. RAS/ERK signaling promotes site-specific ribosomal protein S6 phosphorylation via RSK and stimulates cap-dependent translation. *J Biol Chem,* 282**,** 14056-64.

SADIKU, P., WILLSON, J. A., RYAN, E. M., SAMMUT, D., COELHO, P., WATTS, E. R., GRECIAN, R., YOUNG, J. M., BEWLEY, M., ARIENTI, S., MIRCHANDANI, A. S., SANCHEZ GARCIA, M. A., MORRISON, T., ZHANG, A., REYES, L., GRIESSLER, T., JHEETA, P., PATERSON, G. G., GRAHAM, C. J., THOMSON, J. P., BAILLIE, K., THOMPSON, A. A. R., MORGAN, J. M., ACOSTA-SANCHEZ, A., DARDE, V. M., DURAN, J., GUINOVART, J. J., RODRIGUEZ-BLANCO, G., VON KRIEGSHEIM, A., MEEHAN, R. R., MAZZONE, M., DOCKRELL, D. H., GHESQUIERE, B., CARMELIET, P., WHYTE, M. K. B. & WALMSLEY, S. R. 2021. Neutrophils Fuel Effective Immune Responses through Gluconeogenesis and Glycogenesis. *Cell Metab,* 33**,** 411-423 e4.

SADOWSKI, P. G., DUNKLEY, T. P., SHADFORTH, I. P., DUPREE, P., BESSANT, C., GRIFFIN, J. L. & LILLEY, K. S. 2006. Quantitative proteomic approach to study subcellular localization of membrane proteins. *Nat Protoc,* 1**,** 1778-89.

SAGIROGLU, F. & SINANC, D. 2013. Big data: A review. *2013 International Conference on Collaboration Technologies and Systems*.

SAGIV, J. Y., MICHAELI, J., ASSI, S., MISHALIAN, I., KISOS, H., LEVY, L., DAMTI, P., LUMBROSO, D., POLYANSKY, L., SIONOV, R. V., ARIEL, A., HOVAV, A. H., HENKE, E., FRIDLENDER, Z. G. & GRANOT, Z. 2015. Phenotypic diversity and plasticity in circulating neutrophil subpopulations in cancer. *Cell Rep,* 10**,** 562-73.

SAITO, Y. D., JENSEN, A. R., SALGIA, R. & POSADAS, E. M. 2010. Fyn: a novel molecular target in cancer. *Cancer,* 116**,** 1629-37.

SAKATA, Y., NAGAO, K., HOKI, Y., SASAKI, H., OBUSE, C. & SADO, T. 2017. Defects in dosage compensation impact global gene regulation in the mouse trophoblast. *Development,* 144**,** 2784-2797.

SALTZMAN, A. B., LENG, M., BHATT, B., SINGH, P., CHAN, D. W., DOBROLECKI, L., CHANDRASEKARAN, H., CHOI, J. M., JAIN, A., JUNG, S. Y., LEWIS, M. T., ELLIS, M. J. & MALOVANNAYA, A. 2018. gpGrouper: A Peptide Grouping Algorithm for Gene-Centric Inference and Quantitation of Bottom-Up Proteomics Data. *Mol Cell Proteomics,* 17**,** 2270-2283.

SCHULTE-SCHREPPING, J., REUSCH, N., PACLIK, D., BASSLE, K., SCHLICKEISER, S., ZHANG, B., KRÄMER, B., KRAMMER, T., BRUMHARD, S., BONAGURO, L., DE DOMENICO, E., WENDISCH, D., GRASSHOFF, M., KAPELLOS, T. S., BECKSTETTE, M., PECHT, T., SAGLAM, A., DIETRICH, O., MEI, H. E., SCHULZ, A. R., CONRAD, C., KUNKEL, D., VAFADARNEJAD, E., XU, C. J., HORNE, A., HERBERT, M., DREWS, A., THIBEAULT, C., PFEIFFER, M., HIPPENSTIEL, S., HOCKE, A., MÜLLER-REDETZKY, H., HEIM, K. M., MACHLEIDT, F., UHRIG, A., BOSQUILLON DE JARCY, L., JÜRGENS, L., STEGEMANN, M., GLÖSENKAMP, C. R., VOLK, H. D., GOFFINET, C., LANDTHALER, M., WYLER, E., GEORG, P., SCHNEIDER, M., DANG-HEINE, C., NEUWINGER, N., KAPPERT, K., TAUBER, R., CORMAN, V., RAABE, J., KAISER, K. M., VINH, M. T., RIEKE, G., MEISEL, C., ULAS, T., BECKER, M., GEFFERS, R., WITZENRATH, M., DROSTEN, C., SUTTORP, N., VON KALLE, C., KURTH, F., HÄNDLER, K., SCHULTZE, J. L., ASCHENBRENNER, A. C., LI, Y., NATTERMANN, J., SAWITZKI, B., SALIBA, A. E. & SANDER, L. E. 2020. Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. *Cell,* 182**,** 1419-1440.e23.

SCHWACKE, J. H., HILL, E. G., KRUG, E. L., COMTE-WALTERS, S. & SCHEY, K. L. 2009. iQuantitator: a tool for protein expression inference using iTRAQ. *BMC Bioinformatics,* 10**,** 342.

SCHWANHAUSSER, B., BUSSE, D., LI, N., DITTMAR, G., SCHUCHHARDT, J., WOLF, J., CHEN, W. & SELBACH, M. 2011. Global quantification of mammalian gene expression control. *Nature,* 473**,** 337-42.

SEARLE, B. C., LAWRENCE, R. T., MACCOSS, M. J. & VILLEN, J. 2019. Thesaurus: quantifying phosphopeptide positional isomers. *Nat Methods,* 16**,** 703-706.

SEARLE, B. C., SWEARINGEN, K. E., BARNES, C. A., SCHMIDT, T., GESSULAT, S., KUSTER, B. & WILHELM, M. 2020. Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nat Commun,* 11**,** 1548.

SERANG, O. & NOBLE, W. 2012. A review of statistical methods for protein identification using tandem mass spectrometry. *Stat Interface,* 5**,** 3-20.

SHAATH, H. & ALAJEZ, N. M. 2021. Identification of PBMC-based molecular signature associational with COVID-19 disease severity. *Heliyon,* 7**,** e06866.

SHAATH, H., VISHNUBALAJI, R., ELKORD, E. & ALAJEZ, N. M. 2020. Single-Cell Transcriptome Analysis Highlights a Role for Neutrophils and Inflammatory Macrophages in the Pathogenesis of Severe COVID-19. *Cells,* 9.

SHERMAN, B. T., HAO, M., QIU, J., JIAO, X., BASELER, M. W., LANE, H. C., IMAMICHI, T. & CHANG, W. 2022. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*.

SHULTZ, L. D., RAJAN, T. V. & GREINER, D. L. 1997. Severe defects in immunity and hematopoiesis caused by SHP-1 protein-tyrosine-phosphatase deficiency. *Trends Biotechnol,* 15**,** 302-7.

SHUWA, H. A., SHAW, T. N., KNIGHT, S. B., WEMYSS, K., MCCLURE, F. A., PEARMAIN, L., PRISE, I., JAGGER, C., MORGAN, D. J., KHAN, S., BRAND, O., MANN, E. R., USTIANOWSKI, A., BAKERLY, N. D., DARK, P., BRIGHTLING, C. E., BRIJ, S., CIRCO, FELTON, T., SIMPSON, A., GRAINGER, J. R., HUSSELL, T., KONKEL, J. E. & MENON, M. 2021. Alterations in T and B cell function persist in convalescent COVID-19 patients. *Med (N Y),* 2**,** 720-735 e4.

SILVIN, A., CHAPUIS, N., DUNSMORE, G., GOUBET, A. G., DUBUISSON, A., DEROSA, L., ALMIRE, C., HENON, C., KOSMIDER, O., DROIN, N., RAMEAU, P., CATELAIN, C., ALFARO, A., DUSSIAU, C., FRIEDRICH, C., SOURDEAU, E., MARIN, N., SZWEBEL, T. A., CANTIN, D., MOUTHON, L., BORDERIE, D., DELOGER, M., BREDEL, D., MOURAUD, S., DRUBAY, D., ANDRIEU, M., LHONNEUR, A. S., SAADA, V., STOCLIN, A., WILLEKENS, C., POMMERET, F., GRISCELLI, F., NG, L. G., ZHANG, Z., BOST, P., AMIT, I., BARLESI, F., MARABELLE, A., PENE, F., GACHOT, B., ANDRE, F., ZITVOGEL, L., GINHOUX, F., FONTENAY, M. & SOLARY, E. 2020. Elevated Calprotectin and Abnormal Myeloid Cell Subsets Discriminate Severe from Mild COVID-19. *Cell,* 182**,** 1401-1418 e18.

SINCLAIR, L. V., ROLF, J., EMSLIE, E., SHI, Y. B., TAYLOR, P. M. & CANTRELL, D. A. 2013. Control of amino-acid transport by antigen receptors coordinates the metabolic reprogramming essential for T cell differentiation. *Nat Immunol,* 14**,** 500-8.

SMITH, A. G. 2001. Embryo-derived stem cells: of mice and men. *Annu Rev Cell Dev Biol,* 17**,** 435-62.

SMITH, T. S., ANDREJEVA, A., CHRISTOPHER, J., CROOK, O. M., ELZEK, M. & LILLEY, K. S. 2022. Prior Signal Acquisition Software Versions for Orbitrap Underestimate Low Isobaric Mass Tag Intensities, Without Detriment to Differential Abundance Experiments. *ACS Meas Sci Au,* 2**,** 233-240.

SOEHNLEIN, O., STEFFENS, S., HIDALGO, A. & WEBER, C. 2017. Neutrophils as protagonists and targets in chronic inflammation. *Nat Rev Immunol,* 17**,** 248-261.

SONG, J., JANISZEWSKI, A., DE GEEST, N., VANHEER, L., TALON, I., EL BAKKALI, M., OH, T. & PASQUE, V. 2019. X-Chromosome Dosage Modulates Multiple Molecular and Cellular Properties of Mouse Pluripotent Stem Cells Independently of Global DNA Methylation Levels. *Stem Cell Reports,* 12**,** 333-350.

SPECHT, H., EMMOTT, E., PETELSKI, A. A., HUFFMAN, R. G., PERLMAN, D. H., SERRA, M., KHARCHENKO, P., KOLLER, A. & SLAVOV, N. 2021. Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biol,* 22**,** 50.

SPRENT, J. & SURH, C. D. 2001. Generation and maintenance of memory T cells. *Curr Opin Immunol,* 13**,** 248-54.

STAERK, J., DAWLATY, M. M., GAO, Q., MAETZEL, D., HANNA, J., SOMMER, C. A., MOSTOSLAVSKY, G. & JAENISCH, R. 2010. Reprogramming of human peripheral blood cells to induced pluripotent stem cells. *Cell Stem Cell,* 7**,** 20-4.

STEPHENSON, E., REYNOLDS, G., BOTTING, R. A., CALERO-NIETO, F. J., MORGAN, M. D., TUONG, Z. K., BACH, K., SUNGNAK, W., WORLOCK, K. B., YOSHIDA, M., KUMASAKA, N., KANIA, K., ENGELBERT, J., OLABI, B., SPEGAROVA, J. S., WILSON, N. K., MENDE, N., JARDINE, L., GARDNER, L. C. S., GOH, I., HORSFALL, D., MCGRATH, J., WEBB, S., MATHER, M. W., LINDEBOOM, R. G. H., DANN, E., HUANG, N., POLANSKI, K., PRIGMORE, E., GOTHE, F., SCOTT, J., PAYNE, R. P., BAKER, K. F., HANRATH, A. T., SCHIM VAN DER LOEFF, I. C. D., BARR, A. S., SANCHEZ-GONZALEZ, A., BERGAMASCHI, L., MESCIA, F., BARNES, J. L., KILICH, E., DE WILTON, A., SAIGAL, A., SALEH, A., JANES, S. M., SMITH, C. M., GOPEE, N., WILSON, C., COUPLAND, P., COXHEAD, J. M., KISELEV, V. Y., VAN DONGEN, S., BACARDIT, J., KING, H. W., CAMBRIDGE INSTITUTE OF THERAPEUTIC, I., INFECTIOUS DISEASE-NATIONAL INSTITUTE OF HEALTH RESEARCH, C.-B. C., ROSTRON, A. J., SIMPSON, A. J., HAMBLETON, S., LAURENTI, E., LYONS, P. A., MEYER, K. B., NIKOLIC, M. Z., DUNCAN, C. J. A., SMITH, K. G. C., TEICHMANN, S. A., CLATWORTHY, M. R., MARIONI, J. C., GOTTGENS, B. & HANIFFA, M. 2021. Single-cell multi-omics analysis of the immune response in COVID-19. *Nat Med,* 27**,** 904-916.

SULLIVAN, L., WALL, S. J., CARRINGTON, M. & FERGUSON, M. A. 2013. Proteomic selection of immunodiagnostic antigens for human African trypanosomiasis and generation of a prototype lateral flow immunodiagnostic device. *PLoS Negl Trop Dis,* 7**,** e2087.

SYRETT, C. M., SINDHAVA, V., HODAWADEKAR, S., MYLES, A., LIANG, G., ZHANG, Y., NANDI, S., CANCRO, M., ATCHISON, M. & ANGUERA, M. C. 2017. Loss of Xist RNA from the inactive X during B cell development is restored in a dynamic YY1-dependent two-step process in activated B cells. *PLoS Genet,* 13**,** e1007050.

TABB, D. L., MCDONALD, W. H. & YATES, J. R., 3RD 2002. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res,* 1**,** 21-6.

TADA, M., TAKAHAMA, Y., ABE, K., NAKATSUJI, N. & TADA, T. 2001. Nuclear reprogramming of somatic cells by in vitro hybridization with ES cells. *Curr Biol,* 11**,** 1553-8.

TAKAGI, N. & ABE, K. 1990. Detrimental effects of two active X chromosomes on early mouse development. *Development,* 109**,** 189-201.

TAKAHASHI, K., TANABE, K., OHNUKI, M., NARITA, M., ICHISAKA, T., TOMODA, K. & YAMANAKA, S. 2007. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell,* 131**,** 861-72.

TAKAHASHI, K. & YAMANAKA, S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell,* 126**,** 663-76.

TAN, L., WANG, Q., ZHANG, D., DING, J., HUANG, Q., TANG, Y. Q., WANG, Q. & MIAO, H. 2020. Lymphopenia predicts disease severity of COVID-19: a descriptive and predictive study. *Signal Transduct Target Ther,* 5**,** 33.

TEIJARO, J. R., TURNER, D., PHAM, Q., WHERRY, E. J., LEFRANCOIS, L. & FARBER, D. L. 2011. Cutting edge: Tissue-retentive lung memory CD4 T cells mediate optimal protection to respiratory virus infection. *J Immunol,* 187**,** 5510-4.

THE UNIPROT, C. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res,* 45**,** D158-D169.

THOMAS, H. B., MOOTS, R. J., EDWARDS, S. W. & WRIGHT, H. L. 2015. Whose Gene Is It Anyway? The Effect of Preparation Purity on Neutrophil Transcriptome Studies. *PLoS One,* 10**,** e0138982.

THOMPSON, A., SCHAFER, J., KUHN, K., KIENLE, S., SCHWARZ, J., SCHMIDT, G., NEUMANN, T., JOHNSTONE, R., MOHAMMED, A. K. & HAMON, C. 2003. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem,* 75**,** 1895-904.

THOMSON, J. A., ITSKOVITZ-ELDOR, J., SHAPIRO, S. S., WAKNITZ, M. A., SWIERGIEL, J. J., MARSHALL, V. S. & JONES, J. M. 1998. Embryonic stem cell lines derived from human blastocysts. *Science,* 282**,** 1145-7.

THUL, P. J., AKESSON, L., WIKING, M., MAHDESSIAN, D., GELADAKI, A., AIT BLAL, H., ALM, T., ASPLUND, A., BJORK, L., BRECKELS, L. M., BACKSTROM, A., DANIELSSON, F., FAGERBERG, L., FALL, J., GATTO, L., GNANN, C., HOBER, S., HJELMARE, M., JOHANSSON, F., LEE, S., LINDSKOG, C., MULDER, J., MULVEY, C. M., NILSSON, P., OKSVOLD, P., ROCKBERG, J., SCHUTTEN, R., SCHWENK, J. M., SIVERTSSON, A., SJOSTEDT, E., SKOGS, M., STADLER, C., SULLIVAN, D. P., TEGEL, H., WINSNES, C., ZHANG, C., ZWAHLEN, M., MARDINOGLU, A., PONTEN, F., VON FEILITZEN, K., LILLEY, K. S., UHLEN, M. & LUNDBERG, E. 2017. A subcellular map of the human proteome. *Science,* 356.

THWAITES, R. S., SANCHEZ SEVILLA URUCHURTU, A., SIGGINS, M. K., LIEW, F., RUSSELL, C. D., MOORE, S. C., FAIRFIELD, C., CARTER, E., ABRAMS, S., SHORT, C. E., THAVENTHIRAN, T., BERGSTROM, E., GARDENER, Z., ASCOUGH, S., CHIU, C., DOCHERTY, A. B., HUNT, D., CROW, Y. J., SOLOMON, T., TAYLOR, G. P., TURTLE, L., HARRISON, E. M., DUNNING, J., SEMPLE, M. G., BAILLIE, J. K., OPENSHAW, P. J. & INVESTIGATORS, I. C. 2021. Inflammatory profiles across the spectrum of disease reveal a distinct role for GM-CSF in severe COVID-19. *Sci Immunol,* 6.

TING, Y. S., EGERTSON, J. D., BOLLINGER, J. G., SEARLE, B. C., PAYNE, S. H., NOBLE, W. S. & MACCOSS, M. J. 2017. PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat Methods,* 14**,** 903-908.

TINTI, M. & FERGUSON, M. A. J. 2022. Visualisation of proteome-wide ordered protein abundances in Trypanosoma brucei. *Wellcome Open Res,* 7**,** 34.

TINTI, M., GUTHER, M. L. S., CROZIER, T. W. M., LAMOND, A. I. & FERGUSON, M. A. J. 2019. Proteome turnover in the bloodstream and procyclic forms of Trypanosoma brucei measured by quantitative proteomics. *Wellcome Open Res,* 4**,** 152.

TIWARY, S., LEVY, R., GUTENBRUNNER, P., SALINAS SOTO, F., PALANIAPPAN, K. K., DEMING, L., BERNDL, M., BRANT, A., CIMERMANCIC, P. & COX, J. 2019. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat Methods,* 16**,** 519-525.

TOBY, T. K., FORNELLI, L., SRZENTIC, K., DEHART, C. J., LEVITSKY, J., FRIEDEWALD, J. & KELLEHER, N. L. 2019. A comprehensive pipeline for translational top-down proteomics from a single blood draw. *Nat Protoc,* 14**,** 119-152.

TODD, R., DONOFF, B. R., CHIANG, T., CHOU, M. Y., ELOVIC, A., GALLAGHER, G. T. & WONG, D. T. 1991. The eosinophil as a cellular source of transforming growth factor alpha in healing cutaneous wounds. *Am J Pathol,* 138**,** 1307-13.

TRINKLE-MULCAHY, L., BOULON, S., LAM, Y. W., URCIA, R., BOISVERT, F. M., VANDERMOERE, F., MORRICE, N. A., SWIFT, S., ROTHBAUER, U., LEONHARDT, H. & LAMOND, A. 2008. Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes. *J Cell Biol,* 183**,** 223-39.

TSOU, C. C., AVTONOMOV, D., LARSEN, B., TUCHOLSKA, M., CHOI, H., GINGRAS, A. C. & NESVIZHSKII, A. I. 2015. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods,* 12**,** 258-64, 7 p following 264.

TUKIAINEN, T., VILLANI, A. C., YEN, A., RIVAS, M. A., MARSHALL, J. L., SATIJA, R., AGUIRRE, M., GAUTHIER, L., FLEHARTY, M., KIRBY, A., CUMMINGS, B. B., CASTEL, S. E., KARCZEWSKI, K. J., AGUET, F., BYRNES, A., CONSORTIUM, G. T., LABORATORY, D. A., COORDINATING CENTER -ANALYSIS WORKING, G., STATISTICAL METHODS GROUPS-ANALYSIS WORKING, G., ENHANCING, G. G., FUND, N. I. H. C., NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, BIOSPECIMEN COLLECTION SOURCE SITE, N., BIOSPECIMEN COLLECTION SOURCE SITE, R., BIOSPECIMEN CORE RESOURCE, V., BRAIN BANK REPOSITORY-UNIVERSITY OF MIAMI BRAIN ENDOWMENT, B., LEIDOS BIOMEDICAL-PROJECT, M., STUDY, E., GENOME BROWSER DATA, I., VISUALIZATION, E. B. I., GENOME BROWSER DATA, I., VISUALIZATION-UCSC GENOMICS INSTITUTE, U. O. C. S. C., LAPPALAINEN, T., REGEV, A., ARDLIE, K. G., HACOHEN, N. & MACARTHUR, D. G. 2017. Landscape of X chromosome inactivation across human tissues. *Nature,* 550**,** 244-248.

TURCK, C. W., FALICK, A. M., KOWALAK, J. A., LANE, W. S., LILLEY, K. S., PHINNEY, B. S., WEINTRAUB, S. T., WITKOWSKA, H. E., YATES, N. A. & ASSOCIATION OF BIOMOLECULA RESOURCE FACILITIES PROTEOMICS RESEARCH, G. 2007. The Association of Biomolecular Resource Facilities Proteomics Research Group 2006 study: relative protein quantitation. *Mol Cell Proteomics,* 6**,** 1291-8.

TYANOVA, S., TEMU, T. & COX, J. 2016a. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc,* 11**,** 2301-2319.

TYANOVA, S., TEMU, T., SINITCYN, P., CARLSON, A., HEIN, M. Y., GEIGER, T., MANN, M. & COX, J. 2016b. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods,* 13**,** 731-40.

UHLEN, M., KARLSSON, M. J., ZHONG, W., TEBANI, A., POU, C., MIKES, J., LAKSHMIKANTH, T., FORSSTROM, B., EDFORS, F., ODEBERG, J., MARDINOGLU, A., ZHANG, C., VON FEILITZEN, K., MULDER, J., SJOSTEDT, E., HOBER, A., OKSVOLD, P., ZWAHLEN, M., PONTEN, F., LINDSKOG, C., SIVERTSSON, A., FAGERBERG, L. & BRODIN, P. 2019. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science,* 366.

URBANIAK, M. D., MARTIN, D. M. & FERGUSON, M. A. 2013. Global quantitative SILAC phosphoproteomics reveals differential phosphorylation is widespread between the procyclic and bloodstream form lifecycle stages of Trypanosoma brucei. *J Proteome Res,* 12**,** 2233-44.

VALLOT, C., PATRAT, C., COLLIER, A. J., HURET, C., CASANOVA, M., LIYAKAT ALI, T. M., TOSOLINI, M., FRYDMAN, N., HEARD, E., RUGG-GUNN, P. J. & ROUGEULLE, C. 2017. XACT Noncoding RNA Competes with XIST in the Control of X Chromosome Activity during Human Early Development. *Cell Stem Cell,* 20**,** 102-111.

VAN DER VOORT, L. F., KOK, A., VISSER, A., OUDEJANS, C. B., CALDANO, M., GILLI, F., BERTOLOTTO, A., POLMAN, C. H. & KILLESTEIN, J. 2009. Interferon-beta bioactivity measurement in multiple sclerosis: feasibility for routine clinical practice. *Mult Scler,* 15**,** 212-8.

VAN REES, D. J., SZILAGYI, K., KUIJPERS, T. W., MATLUNG, H. L. & VAN DEN BERG, T. K. 2016. Immunoreceptors on neutrophils. *Semin Immunol,* 28**,** 94-108.

VANROBAYS, E., GELUGNE, J. P., GLEIZES, P. E. & CAIZERGUES-FERRER, M. 2003. Late cytoplasmic maturation of the small ribosomal subunit requires RIO proteins in Saccharomyces cerevisiae. *Mol Cell Biol,* 23**,** 2083-95.

VENABLE, J. D., DONG, M. Q., WOHLSCHLEGEL, J., DILLIN, A. & YATES, J. R. 2004. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods,* 1**,** 39-45.

VILLEN, J., BEAUSOLEIL, S. A., GERBER, S. A. & GYGI, S. P. 2007. Large-scale phosphorylation analysis of mouse liver. *Proc Natl Acad Sci U S A,* 104**,** 1488-93.

VIRREIRA WINTER, S., MEIER, F., WICHMANN, C., COX, J., MANN, M. & MEISSNER, F. 2018. EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification. *Nat Methods,* 15**,** 527-530.

VIZCAINO, J. A., CSORDAS, A., DEL-TORO, N., DIANES, J. A., GRISS, J., LAVIDAS, I., MAYER, G., PEREZ-RIVEROL, Y., REISINGER, F., TERNENT, T., XU, Q. W., WANG, R. & HERMJAKOB, H. 2016. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res,* 44**,** D447-56.

VIZCAINO, J. A., DEUTSCH, E. W., WANG, R., CSORDAS, A., REISINGER, F., RIOS, D., DIANES, J. A., SUN, Z., FARRAH, T., BANDEIRA, N., BINZ, P. A., XENARIOS, I., EISENACHER, M., MAYER, G., GATTO, L., CAMPOS, A., CHALKLEY, R. J., KRAUS, H. J., ALBAR, J. P., MARTINEZ-BARTOLOME, S., APWEILER, R., OMENN, G. S., MARTENS, L., JONES, A. R. & HERMJAKOB, H. 2014. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol,* 32**,** 223-6.

VOEHRINGER, D., VAN ROOIJEN, N. & LOCKSLEY, R. M. 2007. Eosinophils develop in distinct stages and are recruited to peripheral sites by alternatively activated macrophages. *J Leukoc Biol,* 81**,** 1434-44.

VOLAREVIC, V., MARKOVIC, B. S., GAZDIC, M., VOLAREVIC, A., JOVICIC, N., ARSENIJEVIC, N., ARMSTRONG, L., DJONOV, V., LAKO, M. & STOJKOVIC, M. 2018. Ethical and Safety Issues of Stem Cell-Based Therapy. *Int J Med Sci,* 15**,** 36-45.

WANG, D., HU, B., HU, C., ZHU, F., LIU, X., ZHANG, J., WANG, B., XIANG, H., CHENG, Z., XIONG, Y., ZHAO, Y., LI, Y., WANG, X. & PENG, Z. 2020. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA,* 323**,** 1061-1069.

WANG, J., SYRETT, C. M., KRAMER, M. C., BASU, A., ATCHISON, M. L. & ANGUERA, M. C. 2016. Unusual maintenance of X chromosome inactivation predisposes female lymphocytes for increased expression from the inactive X. *Proc Natl Acad Sci U S A,* 113**,** E2029-38.

WANG, J., TUCHOLSKA, M., KNIGHT, J. D., LAMBERT, J. P., TATE, S., LARSEN, B., GINGRAS, A. C. & BANDEIRA, N. 2015. MSPLIT-DIA: sensitive peptide identification for data-independent acquisition. *Nat Methods,* 12**,** 1106-8.

WANG, J., VASAIKAR, S., SHI, Z., GREER, M. & ZHANG, B. 2017. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res,* 45**,** W130-W137.

WANG, L., LE MERCIER, I., PUTRA, J., CHEN, W., LIU, J., SCHENK, A. D., NOWAK, E. C., SURIAWINATA, A. A., LI, J. & NOELLE, R. J. 2014. Disruption of the immune-checkpoint VISTA gene imparts a proinflammatory phenotype with predisposition to the development of autoimmunity. *Proc Natl Acad Sci U S A,* 111**,** 14846-51.

WASHBURN, M. P., WOLTERS, D. & YATES, J. R., 3RD 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol,* 19**,** 242-7.

WATERHOUSE, A. M., PROCTER, J. B., MARTIN, D. M., CLAMP, M. & BARTON, G. J. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics,* 25**,** 1189-91.

WATTS, E. R., HOWDEN, A. J., MORRISON, T., SADIKU, P., HUKELMANN, J., VON KRIEGSHEIM, A., GHESQUIERE, B., MURPHY, F., MIRCHANDANI, A. S., HUMPHRIES, D. C., GRECIAN, R., RYAN, E. M., COELHO, P., BLANCO, G. R., PLANT, T. M., DICKINSON, R. S., FINCH, A., VERMAELEN, W., CANTRELL, D. A., WHYTE, M. K. & WALMSLEY, S. R. 2021. Hypoxia drives murine neutrophil protein scavenging to maintain central carbon metabolism. *J Clin Invest,* 131.

WEBB-ROBERTSON, B. J., WIBERG, H. K., MATZKE, M. M., BROWN, J. N., WANG, J., MCDERMOTT, J. E., SMITH, R. D., RODLAND, K. D., METZ, T. O., POUNDS, J. G. & WATERS, K. M. 2015. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J Proteome Res,* 14**,** 1993-2001.

WEINBERGER, L., AYYASH, M., NOVERSHTERN, N. & HANNA, J. H. 2016. Dynamic stem cell states: naive to primed pluripotency in rodents and humans. *Nat Rev Mol Cell Biol,* 17**,** 155-69.

WEISBROD, C. R., ENG, J. K., HOOPMANN, M. R., BAKER, T. & BRUCE, J. E. 2012. Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification. *J Proteome Res,* 11**,** 1621-32.

WELLE, K. A., ZHANG, T., HRYHORENKO, J. R., SHEN, S., QU, J. & GHAEMMAGHAMI, S. 2016. Time-resolved Analysis of Proteome Dynamics by Tandem Mass Tags and Stable Isotope Labeling in Cell Culture (TMT-SILAC) Hyperplexing. *Mol Cell Proteomics,* 15**,** 3551-3563.

WELLER, P. F. & SPENCER, L. A. 2017. Functions of tissue-resident eosinophils. *Nat Rev Immunol,* 17**,** 746-760.

WICKHAM, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.

WILK, A. J., RUSTAGI, A., ZHAO, N. Q., ROQUE, J., MARTINEZ-COLON, G. J., MCKECHNIE, J. L., IVISON, G. T., RANGANATH, T., VERGARA, R., HOLLIS, T., SIMPSON, L. J., GRANT, P., SUBRAMANIAN, A., ROGERS, A. J. & BLISH, C. A. 2020. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat Med,* 26**,** 1070-1076.

WILKINS, M. R., SANCHEZ, J. C., GOOLEY, A. A., APPEL, R. D., HUMPHERY-SMITH, I., HOCHSTRASSER, D. F. & WILLIAMS, K. L. 1996. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnol Genet Eng Rev,* 13**,** 19-50.

WILMUT, I., SCHNIEKE, A. E., MCWHIR, J., KIND, A. J. & CAMPBELL, K. H. 1997. Viable offspring derived from fetal and adult mammalian cells. *Nature,* 385**,** 810-3.

WILSON, J. C., KEALY, D., JAMES, S. R., PLOWMAN, T., NEWLING, K., JAGGER, C., FILBEY, K., MANN, E. R., KONKEL, J. E., MENON, M., KNIGHT, S. B., SIMPSON, A., GROUP,

C. C., PRIHARTADI, A., FORSHAW, G., TODD, N., YATES, D. R. A., GRAINGER, J. R., HUSSELL, T., KAYE, P. M., SIGNORET, N. & LAGOS, D. 2022. Integrated miRNA/cytokine/chemokine profiling reveals severity-associated step changes and principal correlates of fatality in COVID-19. *iScience,* 25**,** 103672.

WISNIEWSKI, J. R., HEIN, M. Y., COX, J. & MANN, M. 2014. A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Mol Cell Proteomics,* 13**,** 3497-506.

WOLF, T., JIN, W., ZOPPI, G., VOGEL, I. A., AKHMEDOV, M., BLECK, C. K. E., BELTRAMINELLI, T., RIECKMANN, J. C., RAMIREZ, N. J., BENEVENTO, M., NOTARBARTOLO, S., BUMANN, D., MEISSNER, F., GRIMBACHER, B., MANN, M., LANZAVECCHIA, A., SALLUSTO, F., KWEE, I. & GEIGER, R. 2020. Dynamics in protein translation sustaining T cell preparedness. *Nat Immunol,* 21**,** 927-937.

WOLTERS, D. A., WASHBURN, M. P. & YATES, J. R., 3RD 2001. An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem,* 73**,** 5683-90.

WU, D., MOLOFSKY, A. B., LIANG, H. E., RICARDO-GONZALEZ, R. R., JOUIHAN, H. A., BANDO, J. K., CHAWLA, A. & LOCKSLEY, R. M. 2011. Eosinophils sustain adipose alternatively activated macrophages associated with glucose homeostasis. *Science,* 332**,** 243-7.

WUTZ, A. 2012. Epigenetic alterations in human pluripotent stem cells: a tale of two cultures. *Cell Stem Cell,* 11**,** 9-15.

XIE, G., DING, F., HAN, L., YIN, D., LU, H. & ZHANG, M. 2021. The role of peripheral blood eosinophil counts in COVID-19 patients. *Allergy,* 76**,** 471-482.

XU, Z., SHI, L., WANG, Y., ZHANG, J., HUANG, L., ZHANG, C., LIU, S., ZHAO, P., LIU, H., ZHU, L., TAI, Y., BAI, C., GAO, T., SONG, J., XIA, P., DONG, J., ZHAO, J. & WANG, F. S. 2020. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *Lancet Respir Med,* 8**,** 420-422.

YANG, C., LI, Z., BHATT, T., DICKLER, M., GIRI, D., SCALTRITI, M., BASELGA, J., ROSEN, N. & CHANDARLAPATY, S. 2017. Acquired CDK6 amplification promotes breast cancer resistance to CDK4/6 inhibitors and loss of ER signaling and dependence. *Oncogene,* 36**,** 2255-2264.

YANG, S. C., TSAI, Y. F., PAN, Y. L. & HWANG, T. L. 2021. Understanding the role of neutrophils in acute respiratory distress syndrome. *Biomed J,* 44**,** 439-446.

YE, Z., BATTH, T. S., RUTHER, P. & OLSEN, J. V. 2022. A deeper look at carrier proteome effects for single-cell proteomics. *Commun Biol,* 5**,** 150.

YOSHIKAWA, H., LARANCE, M., HARNEY, D. J., SUNDARAMOORTHY, R., LY, T., OWEN-HUGHES, T. & LAMOND, A. I. 2018. Efficient analysis of mammalian polysomes in cells and tissues using Ribo Mega-SEC. *Elife,* 7.

YU, J., VODYANIK, M. A., SMUGA-OTTO, K., ANTOSIEWICZ-BOURGET, J., FRANE, J. L., TIAN, S., NIE, J., JONSDOTTIR, G. A., RUOTTI, V., STEWART, R., SLUKVIN, II & THOMSON, J. A. 2007. Induced pluripotent stem cell lines derived from human somatic cells. *Science,* 318**,** 1917-20.

ZECHA, J., MENG, C., ZOLG, D. P., SAMARAS, P., WILHELM, M. & KUSTER, B. 2018. Peptide Level Turnover Measurements Enable the Study of Proteoform Dynamics. *Mol Cell Proteomics,* 17**,** 974-992.

ZERBINO, D. R., ACHUTHAN, P., AKANNI, W., AMODE, M. R., BARRELL, D., BHAI, J., BILLIS, K., CUMMINS, C., GALL, A., GIRON, C. G., GIL, L., GORDON, L., HAGGERTY, L., HASKELL, E., HOURLIER, T., IZUOGU, O. G., JANACEK, S. H., JUETTEMANN, T., TO, J. K., LAIRD, M. R., LAVIDAS, I., LIU, Z., LOVELAND, J. E., MAUREL, T., MCLAREN, W., MOORE, B., MUDGE, J., MURPHY, D. N., NEWMAN, V., NUHN, M., OGEH, D., ONG, C. K., PARKER, A., PATRICIO, M., RIAT, H. S., SCHUILENBURG, H., SHEPPARD, D., SPARROW, H., TAYLOR, K., THORMANN, A., VULLO, A., WALTS, B., ZADISSA, A., FRANKISH, A., HUNT, S. E., KOSTADIMA, M., LANGRIDGE, N., MARTIN, F. J., MUFFATO, M., PERRY, E., RUFFIER, M., STAINES, D. M., TREVANION, S. J., AKEN, B. L., CUNNINGHAM, F., YATES, A. & FLICEK, P. 2018. Ensembl 2018. *Nucleic Acids Res,* 46**,** D754-D761.

ZHANG, B., CHAMBERS, M. C. & TABB, D. L. 2007. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J Proteome Res,* 6**,** 3549-57.

ZHANG, T., WANG, S., LIN, Y., XU, W., YE, D., XIONG, Y., ZHAO, S. & GUAN, K. L. 2012. Acetylation negatively regulates glycogen phosphorylase by recruiting protein phosphatase 1. *Cell Metab,* 15**,** 75-87.

ZHANG, Y., CASTILLO-MORALES, A., JIANG, M., ZHU, Y., HU, L., URRUTIA, A. O., KONG, X. & HURST, L. D. 2013a. Genes that escape X-inactivation in humans have high intraspecific variability in expression, are associated with mental impairment but are not slow evolving. *Mol Biol Evol,* 30**,** 2588-601.

ZHANG, Y., FONSLOW, B. R., SHAN, B., BAEK, M. C. & YATES, J. R., 3RD 2013b. Protein analysis by shotgun/bottom-up proteomics. *Chem Rev,* 113**,** 2343-94.

ZHOU, H., WATTS, J. D. & AEBERSOLD, R. 2001. A systematic approach to the analysis of protein phosphorylation. *Nat Biotechnol,* 19**,** 375-8.

ZHOU, T., BENDA, C., DUNZINGER, S., HUANG, Y., HO, J. C., YANG, J., WANG, Y., ZHANG, Y., ZHUANG, Q., LI, Y., BAO, X., TSE, H. F., GRILLARI, J., GRILLARI-VOGLAUER, R., PEI, D. & ESTEBAN, M. A. 2012. Generation of human induced pluripotent stem cells from urine samples. *Nat Protoc,* 7**,** 2080-9.

ZHOU, W. & FREED, C. R. 2009. Adenoviral gene delivery can reprogram human fibroblasts to induced pluripotent stem cells. *Stem Cells,* 27**,** 2667-74.