

https://dspace.org/

Comparing changes in gene expression across three myeloid cell lines during monocyte-to-macrophage differentiation

Cosser, Duncan Alexander

2022
https://hdl.handle.net/10539/36950

Downloaded from a DSpace repository. https://dspace.org/



Duncan Alexander Cosser
578000

Comparing changes in gene expression across three myeloid cell lines during monocyte-to-macrophage differentiation

Dissertation
Submitted to the Faculty of Science, University of the Witwatersrand
in fulfilment of the requirements for the degree of:

Master of Science

Supervisor: Dr Nikki Gentle

July 2022 – Johannesburg

Declaration

I, Duncan Alexander Cosser (578000), am a student registered for the degree of Master of Science (M.Sc.) by research in the academic year 2021. I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where explicitly indicated otherwise and acknowledged.
- I have not submitted this work before for any other degree or examination at this or any other University.
- The information used in the Dissertation has not been obtained by me while employed by, or working under the aegis of, any person or organization other than the University.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.



Signature: Duncan Cosser **Date**: 08 July 2022

Acknowledgements

I would like to thank my supervisor, Nikki Gentle, for being constantly encouraging, cynical, and direct. Without her, this dissertation would have been 2^{-3} times as good.

My thanks as always go to my family (especially Michael Cosser, Kirsten Cosser, and Steph Dawson-Cosser), to my dear Gemma Hart, and to the fantastic women of 44 Judith road. This dissertation is dedicated to my mum, the late and great Elaine Cosser, who would have loved to read it.

Finally, my thanks to all the members of the Functional Genomics and Immunogenetics (FunGenIgs) laboratory at WITS, headed by Drs Gentle and Meyer, who have done something wonderful in creating such a helpful and productive working environment (even through a global pandemic).

This work was supported by the National Research Foundation (NRF) through a grantholder-linked bursary (grant number 122000) under Dr Nikki Gentle.

This work was supported by WITS University's postgraduate merit award (PMA) in 2020 and 2021.

Research outputs

Oral presentations

- Comparing transcription in model myeloid cell lines during macrophage differentiation (2021). *Molecular Biosciences Research Thrust Annual Research Day. University of the Witwatersrand, Johannesburg. (Winner: Oral presentations)*
- Do model myeloid cell lines express genes equally? (2021). *Cross Faculty Postgraduate Symposium. University of the Witwatersrand, Johannesburg. (2nd place: Oral presentations)*
- Are model myeloid cell lines transcriptionally comparable? (2021). *SASBiSC/SAGS 2021 Virtual Student Symposium. Stellenbosch University, Stellenbosch.*

Poster presentations

- PMA induces both common and distinct genes and pathways across heterogeneous promonocytic cell lines. Cosser, DA., and Gentle, N. (2021). *ESHG European Human Genetics Conference 2021.*

Contents

1 Introduction **9**
- 1.1 Innate and adaptive immunity . . . 9
  - 1.1.1 Innate immune behaviours . . . 9
  - 1.1.2 Adaptive immune behaviours . . . 11
- 1.2 Monocytes . . . 12
  - 1.2.1 Monocyte behaviour . . . 13
  - 1.2.2 Monocyte classification . . . 14
- 1.3 Macrophages . . . 15
  - 1.3.1 Macrophage behaviour . . . 15
  - 1.3.2 Macrophage classification . . . 16
- 1.4 Cellular differentiation . . . 18
- 1.5 The monocyte-to-macrophage transition . . . 19
- 1.6 Cell lines . . . 20
- 1.7 Aims and objectives . . . 22

2 Methodology **22**
- 2.1 Development of an RNA-seq analysis pipeline . . . 22
  - 2.1.1 Modules . . . 23
  - 2.1.2 Logic flow . . . 24
- 2.2 Study design . . . 26
  - 2.2.1 Sample time points . . . 26
- 2.3 Sequence quality control using `FastQC` . . . 27
- 2.4 Transcript quantification with `Salmon` . . . 27
  - 2.4.1 The `Salmon` algorithm . . . 29
- 2.5 Transcript annotation and summarization to gene level with `tximeta` 31
- 2.6 Differential gene expression analysis with `DESeq2` . . . 32
  - 2.6.1 Library normalization . . . 32
  - 2.6.2 `DESeq2`'s approach to differential expression analysis . . . 32
  - 2.6.3 Variance dependence on the mean: heteroskedasticity . . . 33
  - 2.6.4 Independent filtering and shrinkage . . . 33
- 2.7 Functional enrichment analysis with `clusterProfiler` . . . 35
- 2.8 Pathway over-representation visualization with `pathView` . . . 36

3 Results **38**
- 3.1 Pipeline . . . 38
- 3.2 Sequencing quality . . . 38
- 3.3 Transcript abundances . . . 38
- 3.4 Differential gene expression analysis . . . 39
- 3.5 Analysing functional enrichment . . . 40
- 3.6 Analysing pathway over-representation . . . 42
- 3.7 Enriched pathways in commonly differentially expressed genes . . . 42
- 3.8 Enriched pathways in uniquely differentially expressed genes . . . 48

4 Discussion **53**
- 4.1 Core expression signature . . . 53
- 4.2 Unique expression signatures . . . 55
- 4.3 Study limitations . . . 57

5 Conclusion **59**

Bibliography **61**

6 Supplemental **78**
6.1 Pipeline results . . . 78
6.2 Sequencing quality results from `FastQC` . . . 81
6.3 `DESeq2` quality control data . . . 84
6.4 Functional enrichment analysis: `KEGG` pathway . . . 92
6.5 Commonly differentially expressed genes regulated in different directions . . . 93
6.6 Configuration files . . . 95

Bibliography - Supplemental materials **97**

List of Figures

1 Overview of DAMPs and PAMPs. . . . 10
2 Overview of MHC class I and II. . . . 13
3 Overview of macrophage activation. . . . 17
4 Overview of human haematopoiesis. . . . 19
5 Diagram showing the layout of the `rna-pipe` analysis pipeline. . . . 25
6 Timepoints chosen for study design. . . . 28
7 Common and unique differentially expressed genes across all cell lines. 40
8 Lysosome pathway – common DEGs . . . 44
9 Phagosome pathway – common DEGs . . . 46
10 Cell cycle – common DEGs . . . 47
11 Ribosome pathway – HL60 DEGs . . . 49
12 Oxidative phosphorylation pathway – HL60 DEGs . . . 50
13 Cytokine-cytokine receptor interaction pathway – THP-1 DEGs . . . 52
S1 Pipeline output folders. . . . 78
S2 Condensed sample output log from `rna-pipe`. . . . 81
S3 Representative FastQC results. . . . 83
S4 Rejected SRR3214278 `FastQC` results. . . . 84
S5 Unexpected HL60 clustering results. . . . 86
S6 Unexpected U937 counts and clustering results. . . . 88
S7 `DESeq2` results for all cell lines. . . . 90
S8 Functional enrichment analysis of common and unique DEGs. . . . 92
S9 Sample configuration file for `rna-pipe`. . . . 96

List of Tables

1 Data sources for the three cell lines analysed. . . . 26
2 Comparison of the sequencing platforms used to generate the experimental datasets . . . 32
3 Mapping rates for all samples. . . . 39
4 Functionally enriched pathways common to all cell lines . . . 41
5 Functionally enriched pathways unique to HL60 cells . . . 42

6 Functionally enriched pathways unique to THP-1 cells 42
S1 Example list of differentially expressed genes. 91
S2 Genes in lysosome pathway expressed in different directions. 93
S3 Genes in phagosome pathway expressed in different directions. 94
S4 Sample table for `DESeq2` study design. 96

Abbreviations

AML – Acute myeloid leukaemia

APC – Antigen presenting cell

ATP – Adenosine triphosphate

CC – Chemokine family with two adjacent cysteines

CCL – Chemokine C-C motif ligand

CCR – Chemokine C-C motif receptor

CD – Cluster of differentiation

CXC – Chemokine family with N-terminal cysteines separated by a single amino acid

cDNA – Complementary DNA

CPM – Counts per million

CPU – Central processing unit

DAMP – Damage-associated molecular pattern

DEA – Differential expression analysis

DEG – Differentially expressed gene

DMEM – Dulbecco's Modified Eagle Medium

DNA – Deoxyribonucleic acid

ECM – Extracellular matrix

EGF – Epidermal growth factor

EMR – EGF-like module containing mucin-like hormone receptor

ENA – European Nucleotide Archive

FBS – Fetal bovine serum

FDR – False discovery rate

FEA – Functional enrichment analysis

GO – Gene ontology

HIV – Human immunodeficiency virus

HSC – Haematopoietic stem cell

ICAM – Intercellular adhesion molecule

IFN – Interferon

KEGG – Kyoto Encyclopedia of Genes and Genomes

LDL – Low-density lipoprotein

MCM – Minichromosome maintenance protein complex

MHC – Major histocompatibility complex

MMP – Matrix metalloproteinase

mRNA – Messenger RNA

NADH – Nicotinamide adenine dinucleotide

PAMP – Pathogen-associated molecular pattern

PBMC – Peripheral blood mononuclear cell

PCA – Principal components analysis

PCR – Polymerase chain reaction

PMA – Phorbol 12-myristate 13-acetate

PRR – Pattern recognition receptor

RNA – Ribonucleic acid

RNA-Seq – Ribonucleic acid sequencing

ROS – Reactive oxygen species

rRNA – Ribosomal RNA

RPMI – Roswell Park Memorial Institute

TCR – T cell receptor

TEM – Transendothelial migration

TLR – Toll-like receptor

TNF – Tumour necrosis factor

TPA – 12-O-Tetradecanoylphorbol-13-acetate

TPM – Transcripts per million

tRNA – Transfer ribonucleic acid

VST – Variance stabilizing transformation

WSL – Windows Subsystem for Linux

Abstract

Monocytes and macrophages exhibit wide heterogeneity in phenotype and function. Despite this variety, cells from model cell lines are often used in experiments without accounting for these differences. We therefore sought to characterize the gene expression profiles of three model myeloid cell lines as they differentiate from a monocyte-like state to a macrophage-like state in order to identify core pathways across cell lines central to the differentiation process, as well as pathways uniquely involved in differentiation in specific cell lines.

Using an RNA-seq analysis pipeline we developed, we examined gene expression in three myeloid cell lines – HL60, U937, and THP-1 – during phorbol 12-myristate 13-acetate (PMA)-induced macrophage differentiation. Using gene expression data from three experiments which induced macrophage differentiation in cells from these cell lines using PMA, we constructed gene expression profiles for the cells before and after PMA treatment. We performed differential gene expression analysis to find the genes that are differentially expressed in differentiation, followed by pathway over-representation analysis to identify key pathways involved in this process.

We identified 1789 genes which were differentially expressed in all cell lines, as well as 2131 genes marked as differentially expressed only in HL60 cells, 1711 in U937 cells, and 2159 genes uniquely differentially expressed in THP-1 cells. Through subsequent investigation of specific gene clusters, we found that there were conserved pathways involved in the monocyte-to-macrophage transition in all cell lines, particularly those involved in phagocytosis and adherence. All cell lines also down-regulated maintenance pathways involved in DNA replication and cell cycle control, a behaviour associated with differentiation. Additionally, we found that there were differences in differentiation between the cell lines: the ribosome and oxidative phosphorylation pathways were strongly down-regulated during the transition in HL60 cells, and multiple different components in the cytokine-cytokine receptor interaction pathway were up- and down-regulated in THP-1 cells.

A core gene expression profile is associated with PMA-induced differentiation of myeloid cell lines to a macrophage state. However, differentiation of HL60 and THP-1 cells also induces distinct changes in metabolism and cellular signalling. Our research recommends discretion in cross-comparison between experimental results incorporating gene expression studies of these myeloid cells.

1 Introduction

1.1 Innate and adaptive immunity

The human immune system is composed of two main sections – the innate system and the adaptive system. The innate system serves as the primary and general defence system against pathogens, deterring and destroying them wherever they intrude. Its main functions are to prevent the entry of pathogens into the body, inactivate and destroy foreign substances, recruit professional immune cells to sites of infection, and to activate the adaptive immune system. The adaptive immune system mounts specific and targeted defence against pathogens through recognition of their constituents and the development of unique anti-pathogenic cells and proteins. The main behaviours exhibited by these two systems are discussed in the following sections.

1.1.1 Innate immune behaviours

The first line of defence presented by the innate immune system is provided by mechanical and chemical barriers against the entry and proliferation of pathogens found in epithelial cells (Aderem *et al.* 1999). Mechanical and chemical barriers are present all over the body wherever there is contact with the outside world: in the skin, the respiratory system, urogenital system, and across the whole gastrointestinal system. The mechanical barrier of the skin requires that pathogens make their way through several layers of cells before being able to enter the body (Madison 2003), with mucosal secretions and desquamation (continuous renewal of the upper skin layers) making microbial infection of healthy skin almost impossible (Schröder *et al.* 2006). Should the mechanical and chemical barriers be breached, the innate immune system activates the inflammatory response.

The activation of the inflammatory response starts with the recognition of tissue injury or infection: the former being detected through damage to the body's own cells, the latter through the recognition of foreign bodies. The detection of foreign bodies is primarily handled through the recognition of Pathogen-Associated Molecular Patterns (PAMPs) by pattern recognition receptors (PRRs) present on various leukocytes (Akira *et al.* 2006) including macrophages. The molecular patterns recognized by the PRRs are evolutionarily conserved structural and genetic motifs common to many pathogens. The set of PRRs includes lipopolysaccharides (LPS), the main component of the outer membrane of Gram-negative bacteria, which is recognized by Toll-like receptors (Yang *et al.* 1998) and other carbohydrates in the bacterial cell wall. Another sugar present in fungal cell walls, mannose, is recognized by its own dedicated receptor (CD206) (Stahl *et al.* 1998). PRRs can also bind the internal components of pathogenic cells, like the unmethylated CpG motifs in bacterial DNA (Pisetsky 1996; Krieg 2000), which are recognized by TLR-9 (Takeshita *et al.* 2001).

In addition to being able to detect molecular signatures of pathogens, immune cells are sensitive to molecular signals of damage. These signals convey that there has been some destruction or damage to cells of the body, rather than that a pathogen has been detected (Matzinger 1994). These stimuli, known as Damage-Associated Molecular Patterns (DAMPs) are sensed by various receptors in immune cells (Pic-

cinini *et al.* 2010), most notably the Toll-like receptors (TLRs) (Wheeler *et al.* 2009; Asea *et al.* 2002; Vabulas *et al.* 2002). Once these molecular signatures of damage or pathogen are detected by PRRs in leukocytes (particularly tissue-resident macrophages) and epithelial cells (Svanborg *et al.* 1999), the cells are triggered to produce cytokine and chemokine signalling molecules to alert the rest of the immune system to the danger (Janeway *et al.* 2002). An overview of DAMPs and PAMPs with some examples of their forms is given below (Figure 1).

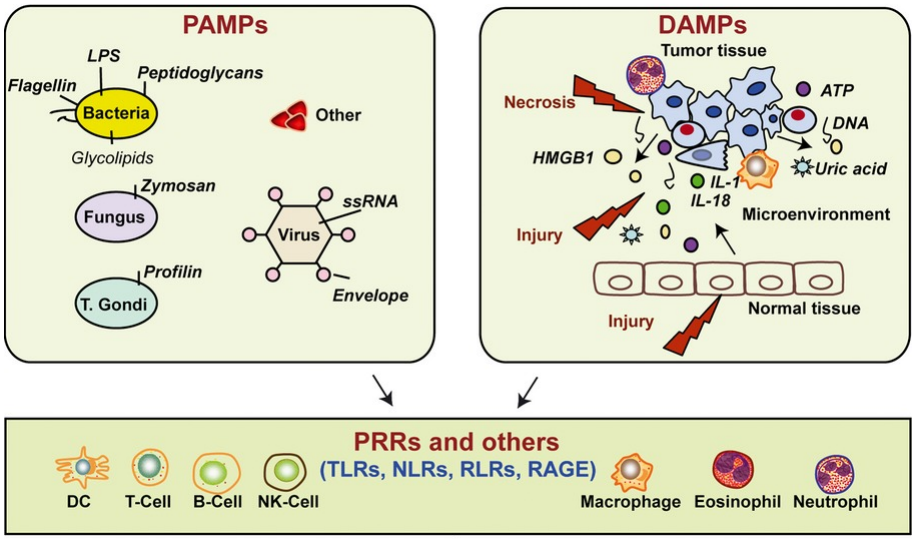


Figure 1: *Diagram showing an overview of Damage-Associated Molecular Patterns (DAMPs) and Pathogen-Associated Molecular Patterns (PAMPs).* PAMPs are associated with pathogens, whereas DAMPs are associated with injury, necrosis, and tumours. Both systems are detected by PRRs in leukocytes. Figure adapted from Tang *et al.* (2012). LPS, lipopolysaccharide. ssRNA, single-stranded RNA. IL, interleukin. HMGB1, high mobility group box protein 1. DC, dendritic cell. PRR, pattern recognition receptor. TLR, toll-like receptor. NLR, NOD-like receptor. RLR, RIG-1-like receptor. RAGE, receptor for AGE (advanced glycation end-product).

Cytokines and chemokines are protein molecules that act as short- and long-range signals, respectively, to recruit leukocytes with the appropriate receptors to the site of infection or damage (Waksman *et al.* 1976). Inflammation is mediated by myriad chemokines, many of which are in turn controlled by tumour necrosis factor-α (TNF-α). This signalling molecule acts as a main regulator of the inflammatory process, and recognition of the protein by its receptors (TNFR1 and TNFR2) controls many down-stream signalling cascades (Parameswaran *et al.* 2010). In initiating the inflammatory response, cytokines and chemokines induce changes in the membranes of epithelial cells that promote adhesion to various leukocytes (Pober *et al.* 1990). This change in adhesion allows circulating leukocytes from the blood to settle near the site of infection, leading to the effective recruitment of more immune cells to bolster the defence of the affected tissue (Auffray *et al.* 2007).

The second major action in inflammation is vasodilation – the process in which blood vessels dilate around the affected tissue and produce the characteristic swelling, redness, pain, and heat associated with inflammation. This vasodilation brings blood-borne antimicrobial proteins (such as antibodies, complement proteins, and

acute-phase proteins) and leukocytes to the site of infection, and transports antigen presenting cells to the nearest lymph nodes to aid in the initiation of the adaptive immune response (Bone 1991).

The next major part of the innate immune response is the complement system. Complement refers to a set of plasma proteins which circulate in the blood in an inactive form and respond to the presence of pathogens through interaction with their cellular membrane (Sjöberg *et al.* 2009). Pathogens which end up coated with complement proteins are more easily recognized by phagocytic cells and thereby destroyed. Although complement is active in innate immunity, it can also function in the adaptive system where it is activated by circulating antibodies (Merle *et al.* 2015).

Besides the effects of the barrier and complement systems, most functions of the innate immune system are carried out by white blood cells. The main leukocytes involved in the innate immune system are mast cells, phagocytes (dendritic cells, macrophages, and neutrophils), granulocytes (basophils and eosinophils), and natural killer (NK) cells (Galli *et al.* 2011). Together these cells are responsible for phagocytosing pathogens, inducing inflammation (especially through the release of histamine, cytokines, and chemokines), destroying pathogens (with antimicrobial agents released from granules and other vesicles such as reactive oxygen species and hydrogen peroxide), and acquiring antigens for presentation to other immune cells, subsequently activating the adaptive immune system (Janeway *et al.* 2002).

1.1.2 Adaptive immune behaviours

The adaptive immune system is organized around the recognition of, and specific response to, pathogens detected within the body. The adaptive system is arranged into two main sections, the humoral and cellular responses. The humoral immune response controls pathogens in extracellular spaces (the blood and lymph network historically referred to as the *humours*), chiefly through the circulation of plasma B cells and their secreted antibodies (Medzhitov 2007). The cellular branch of the adaptive immune system controls intracellular pathogens (such as viruses and bacteria within cells) and cellular misregulation through roaming T cells. Both B and T cells are initially activated through the sensing of a peptide molecule, through the process of antigen presentation (Parker 1993).

Antigen presentation in human cells falls under two main types: amateur and professional. Constant communication with the immune system is one of the myriad requirements of normal nucleated body cells, a task which they complete through the presentation of peptide fragments on their membrane surfaces. All nucleated cells are thus classified as amateur antigen presenting cells (APCs), as opposed to the professional APCs which present antigen as one of their main functions in the body (Gaudino *et al.* 2019). The purpose of communication between immune cells and regular body cells is to maintain the health of regular cells on two fronts: misregulation and infection. The protection against both misregulation and infection is managed using a system of molecular recognition (not unlike the antibody / antigen recognition process) mediated by the major histocompatibility complex (MHC) cell membrane proteins. MHC presents sections of proteins (peptides) to the outside of the cell where they can be recognized by other cells, thus communicating some

information about the cell's internal functioning to patrolling immune cells (Neefjes *et al.* 2011).

MHC molecules appear in humans in two main types, MHC class I (MHCI) and MHC class II (MHCII), which are found on the membranes of different cells in the body. MHC class I molecules are present on the surfaces of all nucleated body cells (amateur APCs), thus conferring on them the ability to present antigens (in most cases normal peptides) to the immune system. As part of healthy metabolism, endogenous peptides are produced through the recycling of unnecessary cellular components and proteins. Through the process of antigen processing, segments of these recycled proteins are complexed with MHC class I proteins and sent to the cell membrane for presentation (Braciale *et al.* 1987).

Upon chance contact between a healthy cell and a roaming helper T cell, the peptide fragment:MHC class I complexes are detected by T cell receptors (TCRs), recognized as normal products, and ignored. In an abnormal cell, the antigen processing process will lead to the display of non-self peptides. Should a $CD8^+$ cytotoxic T lymphocyte (CTL, also known as a killer T cell) come across a diseased cell presenting misregulated peptides on its surface in an MHC Class I complex, the CTL either induces apoptosis or releases granzymes and perforins to disrupt the membrane and destroy the cell (de la Roche *et al.* 2016). This MHC Class I antigen presentation system also protects against intracellular pathogens which have infected normal body cells – the antigen processing and protein recycling system will inevitably present foreign pathogen-derived peptides on the cell's surface, signalling to any passing leukocytes that this cell is infected.

MHC class II molecules are only present on the surfaces of professional APCs such as dendritic cells (Guermonprez *et al.* 2002), macrophages (Harding *et al.* 1992), and B cells (Katikaneni *et al.* 2019). Rather than attaching to endogenous peptides from within the cell, MHC class II complexes with exogenously derived peptide fragments produced by the degradation of extracellular material. In both phagocytosis and receptor-mediated endocytosis, engulfed pathogens are destroyed in endosomes and lysosomes. The exogenous antigen processing pathway then combines fragments of pathogenic proteins with MHC class II molecules and presents them on the cell surface, alerting other immune cells to the detection of a pathogen. Should a $CD4^+$ helper T cell with a matching TCR come into contact with this professional APC and recognize the pathogenic peptide:MHC class II complex on its membrane, the helper T cell will become activated (de la Roche *et al.* 2016). Activated helper T cells begin replicating and split into two populations: one is sent to the site of infection for cytokine production and macrophage activation, and the other initiates the process of antibody production. A graphical summary of the differences between MHC class I and II molecule function is presented below (Figure 2).

1.2 Monocytes

Monocytes are a group of leukocytes (initially termed mononuclear phagocytes) derived from the bone marrow – they circulate through the blood and act in inflammation and pathogen defence (Ginhoux *et al.* 2014). These cells are present in birds (Kelly *et al.* 2000), bony fish (Herbomel *et al.* 1999), and mammals (Virolainen 1968), with similar cells present in insects (Williams 2007). These travelling leuko-

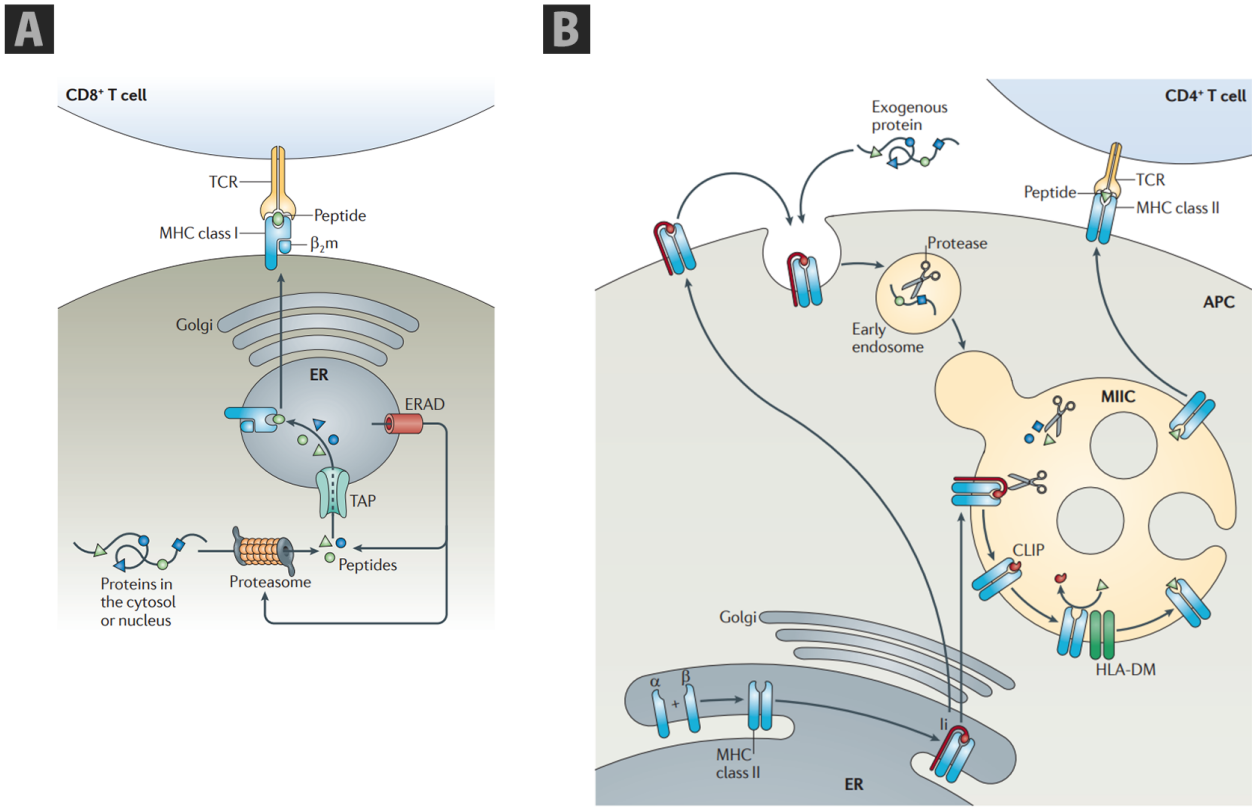


Figure 2: *Diagram showing an overview of MHC class I and II.* **A**: MHC class I molecules present antigenic peptides to T cells via their T cell receptors (TCR). Peptides are generated from intracellular digestion of the cell's own proteins and are trafficked through the endoplasmic reticulum (ER) and Golgi apparatus. **B**: MHC class II molecules are assembled in the ER and fuse with the plasma membrane or move directly to the MHC class II compartment (MIIC). Endocytosis brings exogenous proteins into the MIIC where they are degraded into peptide fragments by proteases and complexed with MHC class II molecules for presentation to T cells via their TCR. Images taken from Neefjes *et al.* (2011). β2m, β2-microglobulin. ER, endoplasmic reticulum. ERAD, ER-associated protein degradation. TCR, T cell receptor. APC, antigen-presenting cell. HLA-DM, human leukocyte antigen DM. CLIP, class II-associated invariant chain peptide.

cytes are released from the bone marrow in an immature state (Volkman 1968), and develop into maturity as they circulate in the blood (Furth *et al.* 1968).

1.2.1 Monocyte behaviour

Monocytes contain various digestive enzymes involved in phagocytosis including oxidases (Rebuck 1952), peroxidases, and phosphatases (Rabinovitch *et al.* 1949). These chemicals are stored in the granules – cytosolic vesicle structures present in these and other leukocytes (Ehrlich 1904) and known to be involved in the digestion of phagocytosed material (Cohn *et al.* 1960). Although not as effective in phagocytosis as macrophages, monocytes can scavenge their surroundings for toxins and other small components (Auffray *et al.* 2009; Geissmann *et al.* 2010).

The mature monocyte cells patrol the endothelium by flowing through the vessels in the blood, or otherwise moving along the vessel walls in a "rolling" motion (Auffray *et al.* 2007). When macrophages or other tissue resident white blood cells are exposed to pathogenic bacteria from an infection they release chemokines which attract other leukocytes to the site of danger (Roth *et al.* 1995). These molecules, specifically the C-C class and C-XXX-C class chemokines, induce adhesion between monocytes and the endothelium, and begin the process which attracts monocytes to cross the endothelial wall (Moser *et al.* 1998).

The process of monocytes exiting the bloodstream and entering infected tissues is known as trans-endothelial migration (TEM). Chemokines and histamine are released by leukocytes at the site of infection (Roth *et al.* 1995), triggering nearby endothelial cells to produce selectins and present them on their surfaces, allowing passing leukocytes to roll along their surface (Auffray *et al.* 2007). This adhesion is then increased by the binding of intercellular adhesion molecules (ICAMs) on the endothelium to integrins on the monocyte's membrane such as the α_L:β_2 integrin (also known as LFA-1), α_M:β_2 integrin (also known as Mac-1), and α_5:β_1 integrin (also known as VLA-4) – these bind ICAMs and fibronectin (Issekutz *et al.* 1999; Gerhardt *et al.* 2015).

After stopping its movement, the monocyte can squeeze through the gap between endothelial cells. Once through the endothelial wall, the monocyte releases matrix metalloproteinase enzymes (MMPs), which break through the tough extracellular matrix protein, allowing it entry into the sub-endothelial tissue (Brundula *et al.* 2002). Inside the affected tissue monocytes follow a concentration gradient of chemokines like chemokine C-C motif ligand (CCL)-2 using their chemokine C-C motif receptor (CCR)-2 until they reach the site of infection (Geissmann *et al.* 2003).

1.2.2 Monocyte classification

Monocytes were originally referred to in tandem with their macrophage descendents as *mononuclear phagocytes* (Furth *et al.* 1968). Though initially believed to be a uniform population, a difference within monocyte subsets was first recognized in unique propensities for enzymatic degradation (Akiyama *et al.* 1983). One monocyte subset (those described first) had the peroxidase enzyme in large amounts in their cytoplasmic vesicles, but the cells of the subset discovered later contained the enzyme in very low amounts. This difference led to the labels of "regular" monocytes and "intermediate" monocytes – the transitionary naming convention growing from the idea that these cells which produced only small amounts of peroxidase were in the process of becoming mature "regular" cells with full phagocytic ability (Akiyama *et al.* 1983).

Later, these two subsets were distinguished more formally by their presentation of the CD14 and CD16 cell surface receptors (Passlick *et al.* 1989; Ziegler-Heitbrock 2007). These demarcations persisted, but the naming changed, with "regular" being replaced by "classical" in later research. The generally accepted convention of separating human monocytes continues to use the CD14 and CD16 markers, but has evolved to include three primary subsets: $CD14^{++}$ $CD16^{-}$ "classical", $CD14^{++}$ $CD16^{+}$ "intermediate" and $CD14^{+}$ $CD16^{++}$ "non-classical" (with the + denoting an expression level of ~10-fold above control and ++ denoting ~100-fold above

the isotype control) (Ziegler-Heitbrock *et al.* 2010). This classification system also takes into account the presence of the CCR2 cell-surface marker – a chemokine C-C motif receptor (CCR) involved in attracting monocytes to sites of infection, and CX_3CR1, a chemokine C-XXX-C motif receptor which controls resident monocyte levels (Geissmann *et al.* 2003).

1.3 Macrophages

Macrophages are large leukocytes found in various tissues in the body and in sites of infection where they engulf harmful pathogens. Some of these professional phagocytes are present in the embryo and then in healthy tissues, but more can be produced through the differentiation of monocytes in times of infection. The investigation of macrophages and of phagocytosis by Metchnikoff in the 19th century was the starting point for modern immunology (Tauber 2003; Gordon 2016). These phagocytic cells are found in almost all adult tissues (as tissue-specific resident macrophages) and originate from haematopoietic stem cells (HSCs) and their progenitors in the bone marrow. Macrophage levels in the body can be maintained through the differentiation of monocytes (Geissmann *et al.* 2010; Fogg *et al.* 2006; van Furth *et al.* 1972), and from generation in the tissues themselves (Hashimoto *et al.* 2013; Yona *et al.* 2013). They are present in vertebrate embryos from a short time after conception (Bertrand *et al.* 2005).

1.3.1 Macrophage behaviour

Macrophages are a diverse group of immune cells involved in innate and acquired immunity, inflammation, and phagocytosis (Taylor *et al.* 2005; Ovchinnikov 2008). Macrophages are involved in various biological processes in the body and are active in all stages of life – starting in development, where they are involved in clearing of apoptotic cells during tissue remodelling (Hopkinson-Woolley *et al.* 1994). In adults of other species that can regenerate limbs, macrophages also drive the regeneration process (Godwin *et al.* 2013). Beyond maintenance of homeostasis, macrophages are critical to wound healing where they perform their regular phagocytic functions (chiefly destroying pathogens by phagocytosis) and attract other immune cells to sites of damage and infection through the release of growth factors and cytokines (Barrientos *et al.* 2008). In the next stages of wound healing, macrophages then encourage blood vessel construction (angiogenesis) (Greenhalgh 1998).

As professional phagocytes, macrophages are critical at sites of tissue damage and infection where they remove dead cells and cellular debris, and eventually present antigens to T-cells. The beginning of their defensive abilities in sites of infection is their recognition of the signature of pathogens (via their generic PAMPs) by PRRs on their surfaces (Akira *et al.* 2006). This molecular recognition initiates phagocytosis, the macrophage engulfing the pathogen recognized by its cell surface receptors. Once in the cytoplasm, the engulfed pathogen is held in a phagosome – the vesicle with membrane-bound proteins which eventually attract it to a lysosome vesicle (Desjardins 1995) containing catabolic enzymes to degrade the pathogen within. After the enzymatic destruction of the pathogen is completed, the exogenous antigen presentation pathway is activated, which extracts peptides from the degraded pathogen to present as antigens to the immune system (Kinchen *et al.* 2008).

Phagocytosis completed, the macrophage combines antigenic peptide fragments from the degraded pathogen with its MHC class II molecules and presents them on its surface, ready to alert passing helper T cells to the presence of the pathogen (Braciale *et al.* 1987). Should this recognition happen, the helper T cell will initiate the activation of the adaptive immune system and its antibody production (Guermonprez *et al.* 2002).

Prior to the full activation of the adaptive immune response, macrophages will induce the next wave of the innate immune response by signalling to other leukocytes and endothelial cells that there has been an infection (Roth *et al.* 1995). The macrophage releases cytokine and chemokine molecules to signal this intrusion, particularly TNF-α (Beutler 1999). This release of signalling molecules, when combined with the release of histamine by mast cells, induces inflammation in the area. These signalling molecules also initiate the leukocyte trans-endothelial migration process, attracting monocytes to the afflicted location. The monocytes which enter the tissue will later differentiate into macrophages themselves, and some will differentiate into dendritic cells also, thereby replenishing the supply of leukocytes which die in the area.

Macrophages can play host to intracellular parasitic bacteria (Armstrong *et al.* 1971), and the immune system has developed a strategy to destroy these parasites known as macrophage activation. When T helper cells recognize the parasites' presented antigenic peptides in complex with MHC molecules on the macrophage's surfaces, they activate the macrophage to attack the parasites (Bekker *et al.* 2001). This activation takes place using membrane proteins and various secreted cytokines like interferon gamma (IFN-γ) (Janeway *et al.* 2001). A diagram summarizing the main routes to macrophage activation is presented below (Figure 3).

1.3.2 Macrophage classification

As a cell type, macrophages are identified *in vitro* by their expression of cell surface markers (Khazen *et al.* 2005), chiefly CD14, CD68, and EMR1, with the latter being the most specific (Martinez-Pomares *et al.* 1996; Taylor *et al.* 2005). Within this set, macrophages are further classified into two groups based on their propensity to repair or destroy objects they encounter and by their secreted products (Mantovani *et al.* 2002), with M1 macrophages generally producing reactive oxygen species and nitrous oxide to destroy pathogens and induce inflammation (Gordon *et al.* 2005), and M2 macrophages calming the immune response and stimulating tissue repair (Mills 2012).

M1 macrophages are classified by their production of various interleukin cytokines (Gordon *et al.* 2005), particularly IL-12 (which induces the development of T_H1 helper T cells – Hsieh *et al.* (1993)) and their marked absence of secretion of IL-10 (which downregulates inflammation – O'Garra *et al.* (2007)). These inflammatory macrophages are activated by IFN-γ (Adams 1989) and LPS in bacterial cell walls (Guha *et al.* 2001).

Alternatively activated (M2) macrophages are characterized *in vitro* by their IL-12^{low} IL-23^{low} IL-10^{high} TGF-β^{high} phenotype, and generally have high levels of scavenger, mannose, and galactose-type receptors (Martinez *et al.* 2008). These M2 macrophages are further subdivided into several subsets based on their activation stimulus. M2a macrophages are characterized by their expression of the interleukin

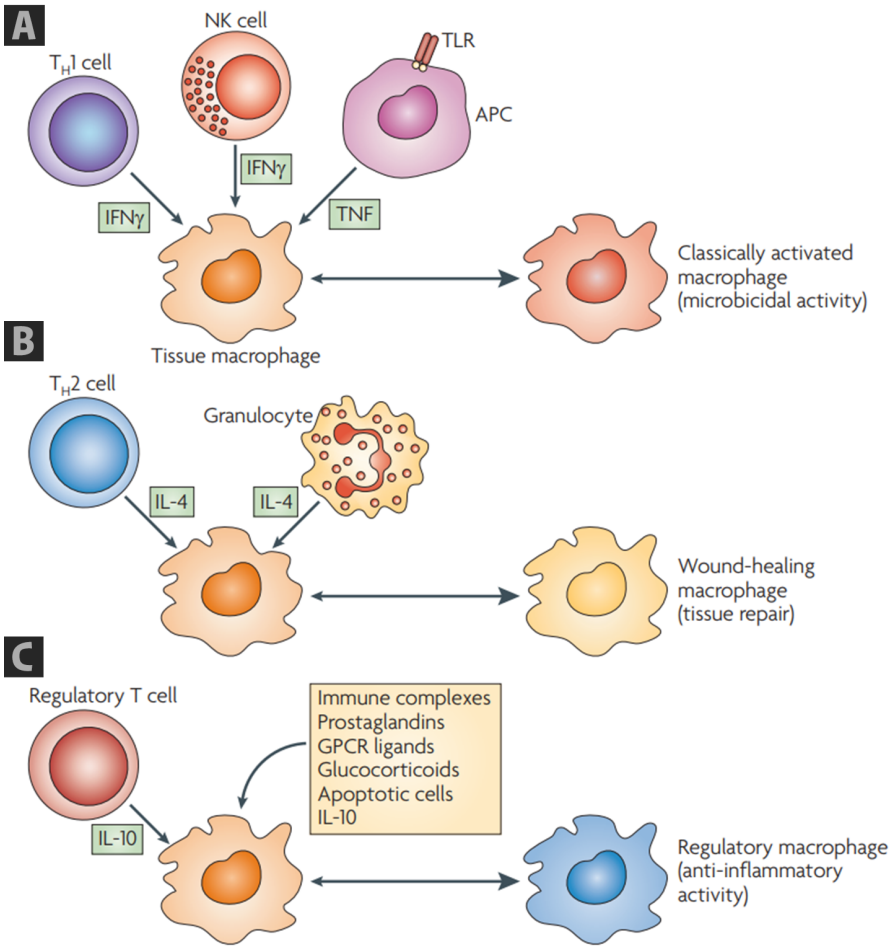


Figure 3: *Diagram showing an overview of macrophage activation.* This representation of macrophages as microbicidal, wound-healing, and anti-inflammatory stems from earlier delineations of monocytes into similar categories. **A**: Classically activated (M1) macrophages are activated by IFN-γ released by T helper 1 cells, and natural killer cells, or by TNF released by antigen presenting cells (APC). **B,C**: Alternatively activated (M2) macrophages are activated by IL-4 from T helper 2 cells and granulocytes, or by IL-10 from regulatory T-cells. Image adapted from Mosser *et al.* (2008). IFN-γ, interferon gamma. TNF, tumour necrosis factor. IL, interleukin. TLR, toll-like receptor. APC, antigen-presenting cell. T_H, T helper.

receptors IL-4 and IL-13 (Gordon 2003; Mantovani *et al.* 2004). They are activated through the alternative pathway by cytokines produced by T helper 2 cells (Gordon *et al.* 2010) – a subset of the helper T cell population involved in B cell regulation and adaptive immune response mediation (Paul *et al.* 2010). M2b macrophages are characterized by their expression of CCL1, and are activated by TLR agonists and decoy IL-1 receptor (Mosser *et al.* 2008). M2c macrophages are characterized by their expression of IL-10 and TGF-β (Rőszer 2015), and are also activated by IL-10 (Ambarus *et al.* 2012). M2c cells also express the MerTK protein (Gautier *et al.* 2012), which is involved in the phagocytosis of apoptotic cells (Zizzo *et al.* 2012), uniquely distinguishing these cells from those in the other subsets.

1.4 Cellular differentiation

Differentiation is the process of cells replicating along a trajectory from multipotency to specialization. In the case of myeloid cells, for example, differentiation is the process of pluripotent myeloid progenitors like haematopoietic stem cells (HSCs, primitive precursor cells that can give rise to the entire lineage of blood cells, both myeloid and lymphoid – Lee *et al.* (2020)) replicating and giving rise to common myeloid progenitors, which then eventually divide to produce monocytes and dendritic cells (Geissmann *et al.* 2010). An overview of the whole process of haematopoiesis in humans is presented below (Figure 4).

In classifying the products of differentiation, cells have traditionally been placed into groups of cell type and cell state (Trapnell 2015). Cell types were historically delineated according to functional similarity: cells which performed the same functions were classified within groups (Bernstein 2021). Within the group boundary demarcated by functional similarity, however, there is often further variation – in cases where this is encountered within cells of the same *type*, and especially when changes are observable over time, cell *state* is used to distinguish between the subsets.

The cell type must be encoded in the cells' genomic arrangement (both spatial and chemical), its transcriptomic profile, its metabolic state, its progress through the cell cycle, and in the arrangement of its organelles and cytoplasmic components (Trapnell 2015). Taken as a whole, this set of temporal information represents everything about the cell which is not encoded by its genome alone, and thus this information must be the source of its unique behaviour among the other cells in the body, and its distinction from them. With increasing technical sensitivity we can categorize cells into types and also into states – higher resolution snapshots of cellular behaviour at a more granular level than cell type (for example, the cell type of T lymphocyte can now be further subdivided into the cell states of naive or activated). With modern techniques (particularly single-cell sequencing technologies which can track cellular behaviours on the individual scale – Stuart *et al.* (2019)) cell types must be described as the combination configurations of the cell's machinery at many different levels.

A powerful method of investigating cellular dynamics exists in the technology of RNA sequencing (RNA-seq). This method of studying the gene expression profiles of cells has unlocked an understanding of differentiation at a level between that of the genome and the proteome. In all indirect RNA-seq experiments (the vast majority of them, and the method used in all the experiments discussed in this work), the following steps are performed: reverse transcription of RNA into cDNA, PCR amplification of the cDNA products, and sequencing of the PCR products (Weber 2015). With the sequences in hand, researchers must align the short reads to the transcriptome (or genome, in some cases) in order to identify the biological origin of each fragment of RNA. In this analysis, RNA-seq and its related technologies provide the perfect toolset for investigating the monocyte-to-macrophage differentiation process.

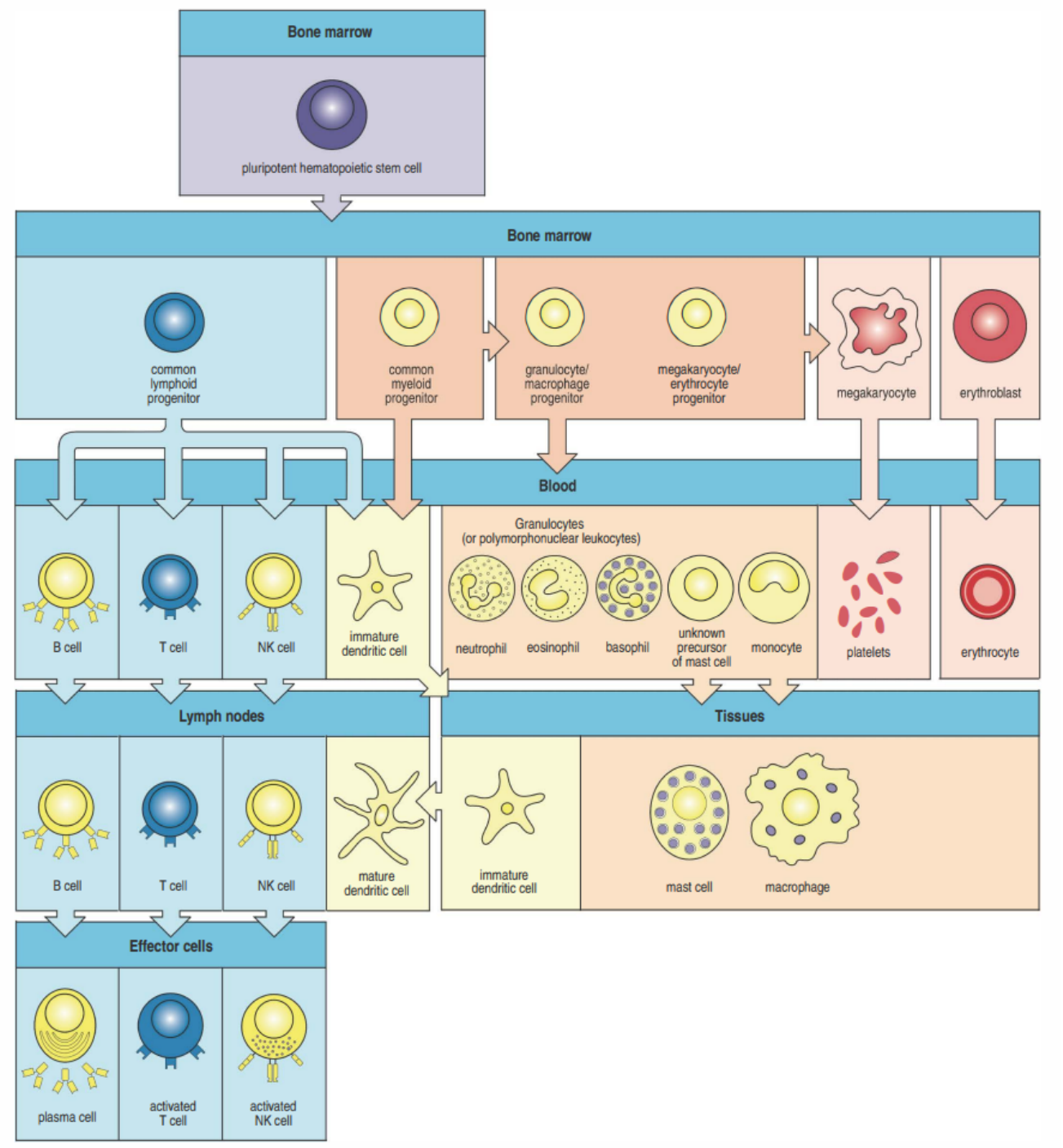


Figure 4: *Diagram showing an overview of human haematopoiesis.* Note the pluripotent haematopoietic stem cell (HSC) at the top of the hierarchy which gives rise to the granulocyte-macrophage progenitor and eventually the monocyte and monocyte-derived-macrophage. Image taken from *Janeway's immunobiology 8th edition* (Murphy, K *et al.* 2012).

1.5 The monocyte-to-macrophage transition

The monocyte-to-macrophage transition is often studied *in vivo* in the context of wounds, in which circulating monocytes exit vasculature, accumulating around the wound site before differentiating (Crane *et al.* 2014). The transition from mono-

cyte to macrophage involves some universal morphological changes (macrophages increasing in size, acquiring more granules in the cytoplasm, and producing more lysosomes), and some changes that are unique to the tissue environment in which the transition occurs (Douglas *et al.* 2010). In some experiments, the phenotype of the macrophage created by the transition seems to depend on the properties of the monocyte from which it descends (Brancato *et al.* 2011). This suggests that the monocyte-to-macrophage transition produces macrophages that retain some properties of their parent monocytes, and which show traits associated with both classical and alternative macrophage activation (Daley *et al.* 2010).

In vivo, the monocyte-to-macrophage transition is driven by the secretion of various regulatory molecules. Among this group of chemicals are granulocyte-macrophage colony-stimulating factor (GM-CSF), which induces proliferation and development of blood monocytes as well as boosting the monocyte-to-macrophage transition (Burgess *et al.* 1980). Specifically, GM-CSF has been shown to regulate genes controlling the cell cycle (Martinez *et al.* 2006). GM-CSF also acts on various other leukocytes and can induce their proliferation and antigen presentation activity. (Fischer *et al.* 1988) A related factor, CSF-1, also induces proliferation and development of blood monocytes and is involved in their differentiation into macrophages (Bonifer *et al.* 2008).

The classical and intermediate monocytes (the two subsets with $CCR2^{high}$ expression, both deemed pro-inflammatory) are both believed to differentiate into M1 (pro-inflammatory) macrophages, while non-classical and anti-inflammatory $CCR2^{low}$ monocytes are believed to differentiate into anti-inflammatory M2 macrophages (Auffray *et al.* 2007).

Existing studies of the gene expression profiles of monocytes transitioning into the macrophage state have shown that these cells change their metabolic activity quite profoundly while undergoing the transition. Regulation of various metabolic pathways in general and lipid metabolism in particular have been noted during the transition, especially through the up-regulation of cyclooxygenases (which can vary based on eventual macrophage polarization) (Martinez *et al.* 2006). Prior work has pointed to epigenetic control as the major regulator of the aforementioned changes in gene expression. Specifically, DNA methylation changes at transcription factor binding sites have been implicated in controlling the gene regulation involved in the transition (Dekkers *et al.* 2019). The breadth of the epigenetic control involved is interesting, with one study finding ~8000 regions under active epigenetic regulation during the macrophage transition and subsequent training – these regions covering ~12% of known human transcription factors (Saeed *et al.* 2014). Studies in mice suggest that distinct populations of macrophages can arise in response to particular injury (Rizzo *et al.* 2020). This propensity, combined with the presence of various tissue-specific macrophage populations, suggests that macrophage differentiation is dependent on factors unique to the tissue micro-environment in which the stimulated monocytes are found at the time of induction.

1.6 Cell lines

In this work, three cell lines widely used as model systems for studying the process of monocyte-to-macrophage differentiation were studied (Chanput *et al.* 2014). The

HL60 cell line was first cultured in 1979 of cells isolated from a 36-year-old female person with acute promyelocytic leukaemia. Subsequent evaluation concluded that HL-60 cells are from a case of acute myeloid leukaemia (AML) FAB-M2 (also referred to as AML with maturation). The **U937** cell line was first cultured in 1976 of cells isolated from a histiocytic lymphoma (cancerous tissue macrophage or dendrite) of a 37-year-old male person, and show promonocytic morphology (Sundström *et al.* 1976). The **THP-1** cell line was first cultured in 1980 of cells isolated from peripheral blood of a 1-year-old male person with AML, and show largely promonocytic morphology (Tsuchiya *et al.* 1980). Due to their ease of use, long-running popularity, and broad differentiation potential, these cell lines are used routinely in basic immunological research (Chanput *et al.* 2014).

These cell lines are known to be able to differentiate into various leukocytic states: HL60 cells can differentiate into granulocytes, monocytes, macrophages, or eosinophils with different stimuli (Gallagher *et al.* 1979). The U937 cells are of tissue origin and show mature promonocytic morphology (Sundström *et al.* 1976), and they can differentiate into a monocytic and then monocyte-derived-macrophage state when exposed to TPA – 12-O-Tetradecanoylphorbol-13-acetate (Koeffler 1983), or into a less differentiated macrophage phenotype upon exposure to 1,25-dihydroxyvitamin D3 (VD3) (Schwende *et al.* 1996). Cells of the THP-1 cell line also show promonocytic morphology, but were cultured from the blood and thus start at a less mature stage of development (Tsuchiya *et al.* 1980).

HL60 cells were initially cultured in order to investigate the cancer from which they came (Gallagher *et al.* 1979). These cells resemble myeloblasts and promyelocytes, and secrete growth factor for continuous proliferation (Brennan *et al.* 1981). The cells have some Fc-γ receptors (which are up-regulated after exposure to phorbol 12-myristate 13-acetate [PMA]) (Guyre *et al.* 1983) and some complement protein 3 subunit b (C3b) receptors (Chiao *et al.* 1982). The cells contain myeloperoxidase and specific esterase but no alkaline phosphatase, unlike *in vivo* promyelocytes (Yam *et al.* 1971). HL60 cells produce low amounts of reactive oxygen species (ROS) like H_2O_2 and $O_{2\text{-}}$ when undifferentiated, but when exposed to PMA they produce it in abundance (Gately *et al.* 1983).

U937 cells were also extracted to investigate the cancer with which they were associated, though unlike the HL60 cells these were harvested from the pleural fluid of a patient with histiocytic lymphoma (Sundström *et al.* 1976). The enzymatic profile of the cells in their undifferentiated state suggests a promonocytic lineage (Yam *et al.* 1971). These cells are able to differentiate into macrophage-like cells (which is measured by their expression of the CD14 marker) through stimulation with PMA (García *et al.* 1999) or vitamin D3 (VD3) (Liu *et al.* 2005). As in HL60 cells, these cells increase their esterase production when exposed to PMA (Nilsson *et al.* 1981). Along with the esterase production, U937 cells release lysozyme into their culture media (Ralph *et al.* 1976), a sign of their phagocytic trajectory. When induced with vitamin D3, the cells express TNF-α (Prehn *et al.* 1992) and TLR4 (Yang *et al.* 2001).

THP-1 cells were cultured from the blood of a boy with acute monocytic leukaemia and are less mature than those of the other two lines investigated here (Tsuchiya *et al.* 1980). These cells can also be induced to differentiate into macrophage-like

cells using PMA (Daigneault *et al.* 2010) or VD3 (Auwerx 1991). When induced with VD3, they express TLR4 and TLR2 in high amounts (Yang *et al.* 2001). The induction of differentiation using VD3 produces macrophage-like cells that show less phagocytic activity than their peripheral blood mononuclear cell (PBMC)-derived biological analogues (Auwerx 1991). THP-1 cells differentiated in this way also produce much lower levels of IL-1β than primary monocytes, although induction using PMA rather than VD3 improves the output slightly (Daigneault *et al.* 2010). When stimulated with LPS, THP-1 monocytes produce less TNF-α, IL-6, IL-8 and IL-10 (all inflammation-related cytokines) than PBMC monocytes (Schildberger *et al.* 2013). THP-1 monocytes' expression of TNF-α, IL-1β, IL-6, CD14 and CD68 when exposed to oxidized LDL (low-density lipoprotein, a proinflammatory protein in the blood – Poznyak *et al.* (2021)) is much weaker than in PBMC monocytes (Bruckmeier *et al.* 2012).

1.7 Aims and objectives

In this work, the differentiation from monocytes (partially differentiated but still specialized and functional cells) into macrophages (terminally differentiated, specialized, and functional cells) was investigated.

The main aim of this work was to identify the genes and gene regulation pathways involved in the monocyte-to-macrophage transition in three different commonly used myeloid cell lines, and to investigate any differences in the regulation of these pathways between the cell lines.

Using an RNA sequencing based approach to probe the monocyte-to-macrophage transition in detail, this project had three main objectives:

1. To develop a pipeline for reliably processing gene expression data
2. To use this pipeline to identify key pathways in the monocyte-to-macrophage differentiation process
3. To compare these pathways between cells from commonly used myeloid cell lines

2 Methodology

2.1 Development of an RNA-seq analysis pipeline

In this work, being able to reliably analyse RNA-seq data across different cell lines was of paramount importance. An RNA analysis pipeline was designed to process bulk RNA-Seq data reproducibly across different experiments. A software analysis pipeline was thus chosen over a set of scripts in order to formalize the process of analysis across different experiments. The main considerations in the design of the software analysis pipeline included:

- *Modularity.* The pipeline should be usable as a whole or in parts, with no penalties for either approach (e.g., running sequence quality control should be as easy as running the whole analysis pipeline).

- *Trackable configuration.* Running the pipeline should be a deterministic process that uses configuration that can be saved, backed up, and referred to later for re-use (thus prioritizing reproducibility).
- *Configuration templating.* If the majority of an analysis experiment is standard and only some sections change (e.g., when one has multiple datasets running `DESeq2` and wants to experiment with different p-value thresholds for differentially expressed gene [DEG] significance), the configuration for *those sections only* should be changeable without having to modify the entire configuration for the rest of the analysis.
- *Logging.* All commands, software versions, and parameters should be logged clearly, and the configuration of an experiment should be interpretable and reproducible from the logs alone.
- *Versioning.* All individual components of the pipeline should be individually versioned and tracked in a source control system, so that changes in function can be traced back to specific versions of the components. This should include the versions of the third party analysis modules such as `FastQC` and `Salmon`.
- *Clarity and verbosity.* Running an experiment through the pipeline should be interactive and easy to read. The pipeline should be expansive in its explanations of what it is doing.

Though many languages could have been used to produce the pipeline, `Bash shell script` and `R` were chosen for their familiarity, popularity, and interoperability. `Shell script` runs well on unix-like systems that have `Bash` or `zsh` (like `Ubuntu`, `MacOS`, and `Windows` with `WSL`), and `R` is fully cross-compatible with all major operating systems. Being popular languages (both `Bash` and `R` are highly ranked in a large developer survey by Stack Overflow (2021)) support and documentation are readily accessible. The easy availability of these languages make the pipeline readily accessible to wet-lab experimental scientists undertaking computational analysis of their results, as well as to more experienced bioinformaticians.

2.1.1 Modules

Analyses in this pipeline were carried out by several distinct programs, each controlled by a *module* – a script in charge of communicating with the analysis program and relaying its results to the main pipeline. The modules available in the pipeline were chosen according to their practicality: these analysis modules are well-supported, and form part of standard RNA-seq analysis pipelines. They include sequencing read quality control with `FastQC` (Andrews 2010), transcript quantification with `Salmon` (Patro *et al.* 2017), transcript annotation and summarization with `tximeta` (Love *et al.* 2019), differential gene expression analysis with `DESeq2` (Love *et al.* 2014), functional enrichment analysis with `clusterProfiler` (Yu *et al.* 2012), and pathway over-representation analysis with `pathView` (Luo *et al.* 2017). The modular design of the pipeline allows for the future addition of new analysis packages without changing the performance of the whole system.

2.1.2 Logic flow

The pipeline is run from a single entry-point, a script called `rna-pipe.sh` (Figure 5). This is a script that parses the configuration files supplied to determine which analysis modules should be run. The main `rna-pipe` script initiates logging, configures the input and output directories (which are set in the configuration file), checks for the presence of the programs upon which it depends (like `R` or `Salmon`), and then activates the required modules one at a time (usually starting with `fastqc-pipe` for sequence quality control).

Once activated, each module starts logging, creates its own output subdirectory, and then converts its configuration into a set of arguments to supply to its analysis package (e.g., the locations of the sequence files for `Salmon`'s quantification). The module then calls its analysis package with all the requisite arguments and logs its output, before finally handing back control to the main script. Once the main script is back in control, it continues processing all the required modules until they have completed their individual analyses, at which point it finishes writing its logs and quits. This triggers a final message to be displayed, indicating that the analysis is complete, and that the results are ready for inspection.

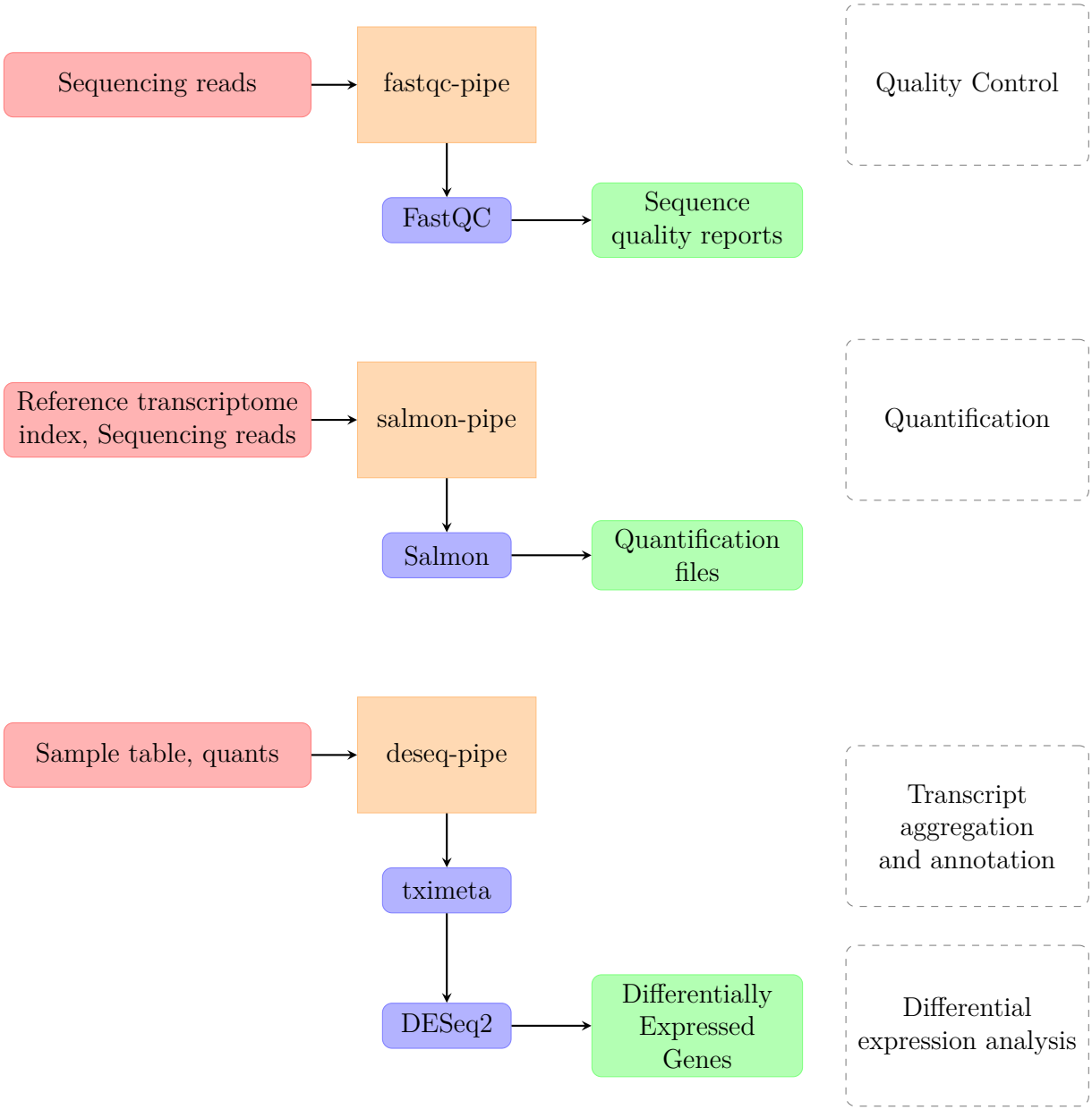


Figure 5: *Diagram showing the layout of the `rna-pipe` analysis pipeline orchestrated by `rna-pipe.sh`.* In the **Quality Control** section, `FastQC` takes the raw sequencing reads and produces sequencing quality reports. In the **Quantification** section, `Salmon` takes the raw sequencing reads and the reference transcriptome index and produces quantification files of transcript abundance. In the **Transcript aggregation and annotation** section, `tximeta` takes the raw quantification files and produces summarized, annotated quantification files. In the **Differential expression analysis** section, `DESeq2` takes the annotated quantification files and the experimental sample information and produces a list of differentially expressed genes. Red represents input data, orange represents pipeline modules, blue represents analysis software programs, and green represents output data.

2.2 Study design

As this work aimed to investigate gene expression in model myeloid cell lines during macrophage differentiation, suitable datasets containing the necessary expression data needed to be sourced from existing experiments in order to perform the requisite analyses. When identifying appropriate datasets, we considered the following inclusion criteria: the experiments used model myeloid cell lines with the potential to differentiate into monocytes, induced their differentiation into monocyte-derived macrophages, and performed RNA-seq multiple times over the experiment (at least once at the beginning of the process and once at the end). In addition, we decided to limit the datasets to those that used the same stimulus for differentiation induction, common examples of which include *phorbol 12-myristate 13-acetate* (PMA, also written as *12-O-Tetradecanoylphorbol-13-acetate* or TPA – Rovera *et al.* 1979) and 1,25-dihydroxyvitamin D3 (vitamin D3 – Murao *et al.* 1983).

This study combined RNA-Seq data from three separate datasets that included three myeloid cell lines. The data were sourced from the *European Nucleotide Archive* (ENA, at the European Molecular Biology Laboratory's European Bioinformatics Institute – *EMBL-EBI* 2021) as follows:

Table 1: Data sources for the three cell lines analysed.

Cell line	Time points	Replicates	ENA Accession number	Reference
HL60	T_0 - T_{+96h}	3	PRJNA314802	(Ramirez *et al.* 2017)
U937	T_0 - T_{+96h}	3	PRJNA381911	(Rossetti *et al.* 2017)
THP-1	T_0 - T_{+72h}	3	PRJNA449980	(Liu *et al.* 2018)

2.2.1 Sample time points

The time points at which measurements were taken differed across the experiments, with the HL60 dataset including three replicate samples for each timepoint of 0h (pre-induction), 48h, 96h, and 120h post-induction. The U937 dataset contained three replicate samples for 0h, 48h, and 96h post-induction. The THP-1 dataset contained three replicate samples for 0h and 72h post-induction. Although we had aimed to use data from experiments with identical timing schemes (or at least with data collected at the same time points), good quality RNA-Seq datasets with perfectly matched timing schemes were difficult to find, and these three datasets represent the most compatible grouping we could compile.

HL60 and U937 datasets contained samples obtained at the same time points (0h, 48h, and 96h post-induction), and the THP-1 dataset gave a reasonable approximation of terminal differentiation with its 72h post-induction data. Although the best approximation of primary cell monocyte-derived-macrophage state is obtained when myeloid cell lines like these are treated with PMA for three days (72h) and then left to rest in culture medium for a further five days (Daigneault *et al.* 2010), a compromise was reached for this analysis which aimed to make the most reasonable biological comparison between these cell lines. Of the HL60 and U937 cell lines' total available data (daily readings from 0-120h in HL60 and 0-96h in U937), the 96h data were chosen as the most logical midpoint at which to measure the "differentiation

induced" state. This 96h timepoint in HL60 and U937 was deemed closer to the 72h dataset from the THP-1 cell line than to that THP-1 experiment's 192h data (recorded after the 72h of induction followed by five days' rest). As such, this 72h dataset was chosen for THP-1 comparison, leaving the final time points as shown in Figure 6. In summary, we chose the final timepoint in the dataset for each cell line as representing the terminal stage of differentiation.

2.3 Sequence quality control using `FastQC`

Quality control analysis was performed on the raw RNA-seq reads using `FastQC v0.11.9` (Andrews 2010), a popular DNA and RNA sequencing quality control tool that assesses the quality of sequence data based on various heuristics. The program produces the following quality control metrics:

- Average read quality on a per-base level (illuminating any read quality degradation over the length of a read or at the very beginning)
- Distribution of average sequence quality throughout the dataset (giving a global picture of the quality of the reads in aggregate)
- The average base identity at each base location (revealing any unexpected patterns, such as a repeated motif with certain bases at the same location or a prevalence of GC bases at the start of the read)
- Distribution of GC content throughout the dataset (giving a global picture of the mean GC content of the reads in aggregate)
- Global sequence length distribution (for platforms in which reads are of variable length)
- Sequence duplication (giving a visual indication of the proportion of the dataset which is made up of unique vs duplicated reads)
- Overrepresented sequences (which identifies adapters or other known fragments to be later filtered out of the dataset)

2.4 Transcript quantification with `Salmon`

Transcript quantification is a prerequisite to downstream analysis, and many tools exist to perform it (Külahoglu *et al.* 2014; Griffith *et al.* 2015). In this work, the aims did not require transcript alignment *per se*, only the quantified transcript abundances (as the individual sequencing read locations are not important in differential expression analysis, which simply uses the estimated abundances of each transcript). Traditional alignment-based RNA quantification strategies like those deployed in `Cufflinks` (Trapnell *et al.* 2010) and `RSEM` (Li *et al.* 2011) produce a rich alignment of reads to the reference genome, and use the per-base locations of the reads to infer transcript abundance (there is an option of incorporating transcriptome data where that is available). This process provides more information than pseudoalignment with its accurate positioning of reads onto the reference genome, but that information comes at the cost of taking much longer to perform and requiring much more memory to complete – alignment of a typical human RNA-seq dataset consisting of 40 million double-stranded reads in `STAR`, for example, requires

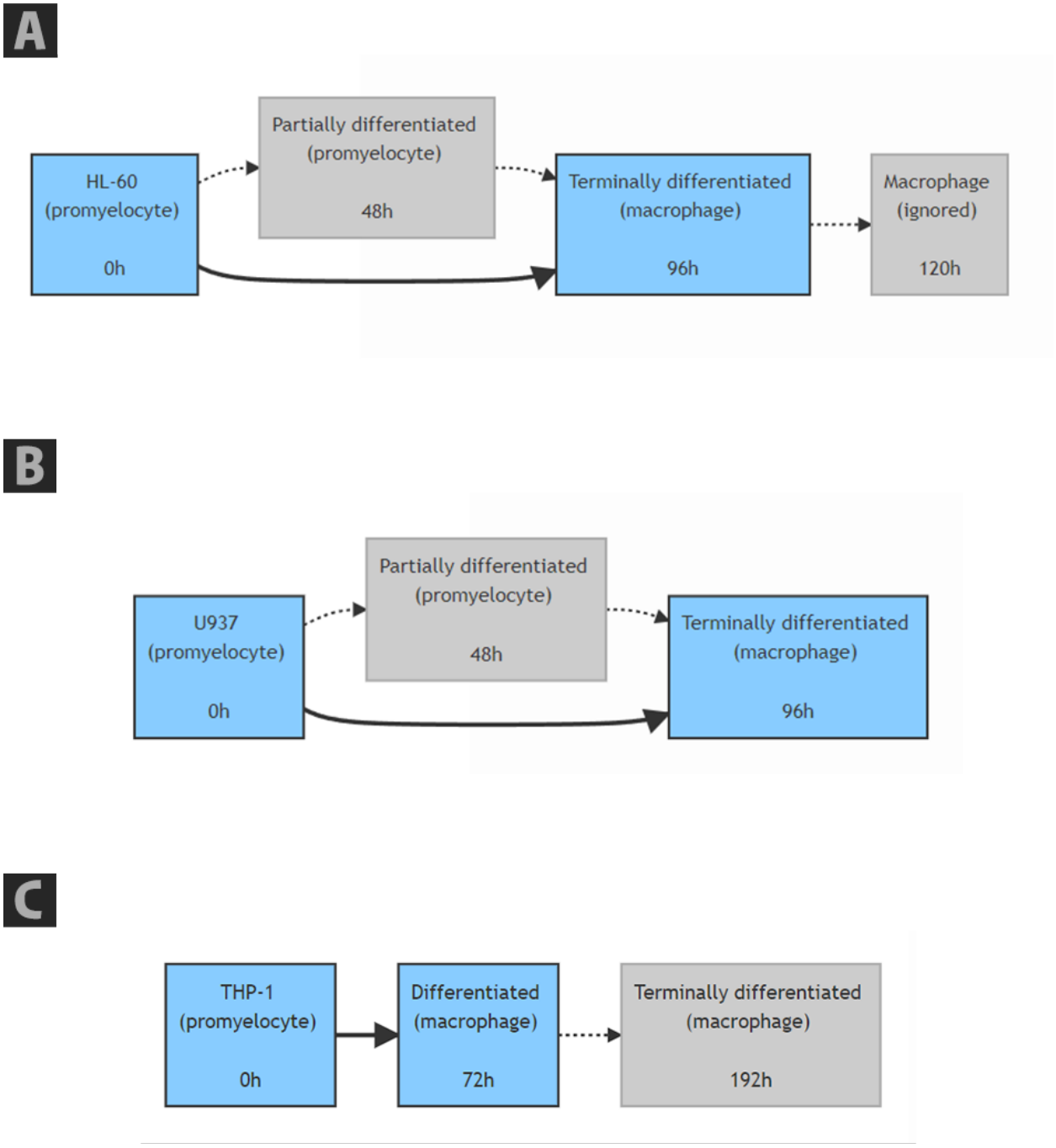


Figure 6: *Diagram showing time points chosen for study design.* Blue shaded boxes indicate time points chosen for analysis, while grey shaded boxes show time points from the original datasets which were ignored. Bold arrows indicate the time points compared in final differential expression analysis, while dashed arrows show the chronological progression of the experiment. **A**: HL60 cell line, **B**: U937 cell line, **C**: THP-1 cell line.

over 26Gb of memory (Dobin *et al.* 2013). In recent benchmarking studies, `Salmon` has been shown to be slightly more accurate and sensitive in quantification than

kallisto (Sarantopoulou *et al.* 2021), a similar modern pseudoalignment-based quantification tool (Bray *et al.* 2016). Salmon v1.2.1 was thus used for transcript abundance estimation due to its rapid output, low memory requirements, and high accuracy.

The quantification process requires an index file, which is built from the reference genome and reference transcriptome. The refgenie website provides pre-built versions of this file and was used to source the index for these experiments to save the time required to produce the index manually (the process is deterministic and would produce equivalent results) (*Refgenie Server - Hg38 Salmon Index* 2021). The refgenie index was built using the hg38 reference genome (Schneider *et al.* 2016; *GRCh38.P13 Genome Assembly - NCBI* 2019). The partial selective alignment index was used instead of the larger full index due to computational memory constraints. The use of this partial index only has a minor impact on the accuracy of the pseudoalignment process. The full index produced using the entire genome simply provides a more comprehensive set of decoy transcripts to avoid accidental mapping to unannotated regions (Srivastava *et al.* 2020).

2.4.1 The Salmon algorithm

Salmon breaks up its computational task of estimating gene abundance in its sample data by looking at individual fragments in the dataset – short stretches of cDNA around 35-200 bases long – that have been read directly by the sequencing machine. The overall aim of Salmon's analysis is to estimate the nucleotide fraction for each fragment in the experimental dataset (nucleotide fraction here representing the proportion of all bases in the entire sample that come from a particular fragment – effectively the fragment abundance percentage). Knowing the nucleotide fraction of a fragment allows Salmon to estimate the source of the fragment in the transcriptome. The algorithm goes through each fragment in turn and creates an estimation model for it that incorporates some basic information about it (the transcript sequence, length, the number of copies read) and a bias term, which it then modulates until the likelihood of a particular transcript being the true source of the fragment is maximized. This process is repeated until the estimation model converges to a point and the nucleotide fractions for each transcript are deemed to be estimated accurately. A more detailed outline of the process is explained below:

The program creates a bias model intended to deal with the non-uniformity of sampling in the sequencing experiment. The model includes terms for fragment length, fragment starting position, fragment orientation, and fragment alignment (accounting for indels and matches). This model thus represents all the features that do not depend on the nucleotide fraction, thus encoding sample-specific biases. The next step in the process is to run an inference procedure that fine-tunes the above parameters to maximize the likelihood of the model. The inference procedure is done in three parts:

1. The algorithm reads through the fragments in batches and estimates the nucleotide fraction for each one, using the bias model and the global nucleotide fraction parameters. The program then updates the global bias model after each batch is complete by adding a weighted version of this local bias model to the global one, thus improving the accuracy of the bias model with every

batch processed. Batches are run in parallel, exploiting multiple processing cores on modern machines.

2. Next, the program generates equivalence classes: groups of similar fragments that could conceivably have come from the same transcript. These equivalence classes encode a much reduced representation of the experimental data that preserves all the necessary features for investigation but hugely decreases the computational power required for further analysis (the number of equivalence classes are often two orders of magnitude smaller than the number of transcripts; Kingsford 2016). In addition, the number of classes only grows with transcriptome complexity, not with transcriptome size (i.e. more fragments do not necessarily mean more equivalence classes – bounding the computational power required for analysis). The equivalence classes are then assigned affinities – probabilities that the members of the class did indeed originate from the transcript in question – made by aggregating the bias terms across all the reads in the class.

3. The final part of the inference procedure is offline (done in a separate pass to the initial data streaming) and involves updating the number of reads assigned to each transcript by multiplying their abundance in the equivalence class with the affinity value. This is repeated until the maximum likelihood model converges.

Finally, an alignment is performed to find the putative transcript locations for all equivalence classes. Central to the speed of these pseudoalignment methods is the lack of any actual base-wise alignment information. The pseudoalignment process only produces links between equivalence classes and transcripts from which they could have been generated (so-called compatible transcripts). In the newer versions of `Salmon`, a selective alignment method is used. This technique uses an alignment scoring system to avoid this decoy problem and other errors to which lightweight alignment was susceptible, and improves mapping sensitivity and specificity (Srivastava *et al.* 2020).

The quantification process requires an index file, which is built from the reference genome and reference transcriptome. The index generation process involves the identification and extraction of decoy regions from the genome using `MashMap` (Jain *et al.* 2018). These decoy regions are unannotated genomic loci that have some sequence similarity to valid transcripts in the transcriptome – enough similarity that fragments arising from these regions could partially match with valid transcripts in the transcriptome, producing spurious mappings (Srivastava *et al.* 2020).

`Salmon` also provides an extrinsic quality control heuristic in its mapping rates. This measure, produced for each sample, reveals what percentage of the input fragments could be confidently mapped to the transcriptome – with low mapping rates implying that only a small fraction of the reads could be mapped. Low mapping rates suggest either low quality data, contamination with many RNA fragments from rRNA or tRNA (the loci of which would not appear in the transcriptome file and thus would not be mapped or counted), or contamination with genetic material from bacteria and other non-target organisms. Though samples with low mapping rates were not removed outright at this step, they were eventually discarded as they proved to be outliers in downstream PCA analysis.

Raw sequence reads were imported into `Salmon` as `fastq` files, for all the time points, across all the cell lines, along with the index file described above. The commands performed by the pipeline module for `Salmon` were as follows (actual file names omitted for brevity):

```
salmon quant
    -i salmon_partial_sa_index
    --libType A
    --threads 4
    --validateMappings
    -r sequence_reads.fastq.gz
    -o sequence_quants
```

The analysis module was then given instructions to use as many available CPU threads as possible, and employ the `automatic` library type argument (in which `Salmon` detects the strandedness and arrangement of the input data without requiring further user input). `Salmon` then performed the analysis as outlined above, with the selective alignment algorithm for its read mapping. Once pseudoalignment was completed, `Salmon` produced quantification files for each sequencing read in its native `.sf` format. The resulting `quant.sf` files contained a list of all known transcripts (with Ensembl format transcript identifiers like ENST00000631435.1), their lengths and effective lengths (as fragment length normalization is estimated in the model above), and the abundance in each sample (both as the raw number of reads and the calculated transcripts per million or TPM).

2.5 Transcript annotation and summarization to gene level with **`tximeta`**

With the RNA abundances calculated, the next processing step involved summarization of the transcripts to the gene level and annotation of the transcripts with biological metadata. Given that it is possible for multiple transcripts to originate from one gene, a decision was needed at this point on how to process the sequence reads further: they could be left as raw transcripts, or condensed to their source genes. We decided to summarize the expression data to the gene level.

`tximeta v1.8.3` (Love *et al.* 2019) automatically obtained transcript metadata for each transcript in a quantification file using a connection to the online *Ensembl* database (*Homo sapiens transcriptome release 97* – Schneider *et al.* 2016). This step produced a file containing useful annotations for the transcript (including accession numbers, chromosomal locations, and associated gene identifiers). Next, `tximeta` obtained a mapping table from Ensembl which it used to summarize all the transcripts to their source genes – combining the abundances of multiple transcripts into one count at the gene level. This transcript-to-gene mapping process produced a list of all the genes detected in each sequencing run, along with their corresponding abundances, ready for differential expression analysis.

2.6 Differential gene expression analysis with DESeq2

2.6.1 Library normalization

The first step in DESeq2's differential expression analysis is library normalization. Here, a library refers to the set of all genes found to be expressed in a sample. The normalization process involves two main steps: sequence depth normalization and library composition normalization.

As library sizes are not the same, samples containing a higher number of sequencing reads need to be adjusted so that their counts are comparable with those from other libraries that have been sequenced at a lower depth. If, for example, one was comparing data from a modern sequencing platform with excellent read depth to older data with shallower read depth, all the genes could appear as differentially expressed simply due to the different sequencing depths in the experiment.

Table 2: Comparison of the sequencing platforms used to generate the experimental datasets

Cell line	Sequencing platform	Read length (base pairs)	Reads generated (million)
HL60	Illumina NextSeq 500 single-end	86	14-20 (average 17.5)
U937	Illumina NextSeq 500 paired-end	101	40-54 (average 48.2)
THP-1	Illumina HiSeq 2500 paired-end	101	39-68 (average 51.9)

2.6.2 DESeq2's approach to differential expression analysis

The DESeq2 algorithm begins with the generation of scaling factors for each gene, and proceeds as follows (Love *et al.* 2014):

1. Gene counts are natural log transformed – their logs are taken with base e to minimize the effect of outliers when comparing between samples, as log encoding shrinks large differences between data points.
2. The transformed count values are averaged across all samples to get a mean expression level for the gene across all samples. This is a geometric mean, as it is taken across log values.
3. The dataset is filtered to remove genes which have no counts in any of the samples. This step normalizes the libraries under the same conditions from different replicates. These filtered genes are not removed from the dataset, but their scaling factors are not calculated (thus in the final normalization step, samples with zero counts in any gene are not affected).
4. The geometric mean count between samples for each gene is subtracted from the count for that gene in every sample. Due to these values being log transformations, this subtraction of one log value from another gives the ratio between each sample's expression and the average expression across all samples for that gene – effectively this particular sample's gene expression compared to the average expression level in all the samples.
5. The ratios generated in the last step are then arranged in order and a median is calculated in each sample, giving an impression of the average gene expression

level in the sample. The median is used instead of mean here because the mean is easily swayed by outlier genes with massive expression.

6. The median ratio is taken as the scaling factor, so each sample has their scaling factor transformed back into base `10` from base `e`, giving a normal decimal number for each sample's scaling factor.

Once scaling factors have been produced, every gene read count for each sample is divided by that sample's scaling factor. This step scales each sample's read counts according to the normalization steps applied above, leaving gene counts that are insensitive to library composition differences.

2.6.3 Variance dependence on the mean: heteroskedasticity

The next step in `DESeq2` analysis is to deal with an intrinsic problem in data generated from RNA sequencing experiments: heteroskedasticity. `DESeq2` uses a negative binomial distribution to model the distribution of sampled RNA-seq counts (Robinson *et al.* 2007), and in this distribution the variance increases as the mean increases (this property is the opposite of what we find in normally distributed data which is homoskedastic). In real RNA-seq data, the variance scales more than linearly with the mean, so that the data is *over-dispersed* (Love *et al.* 2014), with expression values from abundant genes showing much more variance than from genes expressed at a low level. This relationship between mean count and variance in the negative binomial distribution is captured by a dispersion parameter that represents the true inter-replicate variation in expression – rather than the measured variation that is influenced by the mean only.

This extreme variance requires that counts be adjusted in order to be statistically trustworthy, and in order to effect this adjustment correctly, `DESeq2` applies a procedure called "shrinkage" (otherwise known as the variance stabilizing transformation, or VST). In this process, `DESeq2` tabulates the variance in all genes to get a global picture of variance throughout the different samples, and then shares this data between genes to infer what the expected variance for a gene in this sample should be, given its expression level. By estimating the variance using this data from all genes, `DESeq2` can come up with a metric for how dispersed each individual gene in the sample is, and try to correct for this extra variation. This correction process or "shrinkage" adjusts the counts of each datapoint detected as over-dispersed towards an amount in line with the calculated mean variance.

`DESeq2` thus reduces some of the extra variance of over-dispersion and allows gene counts to be compared more accurately. As the mean-dependent variation increases with gene count, lowly-expressed genes are not corrected much by this step (which is why they are often removed by the false discovery rate [FDR] correction process). Instead, many genes with high counts are corrected towards the mean (either having their count values increased or decreased, depending on the side of the mean line they fall).

2.6.4 Independent filtering and shrinkage

To avoid a high FDR with so many samples (FDR referring to the proportion of false positive results – genes incorrectly identified as differentially expressed when

they are not), DESeq2 performs a process called independent filtering. Due to the systematic uncertainty inherent in any technical experiment, gene expression levels below a certain threshold cannot be distinguished from experimental noise, and thus genes with low expression levels only contribute more statistical tests to the analysis and not much differential expression information, worsening the multiple hypothesis testing problem. In order to improve statistical power, these low-information genes need to be discarded.

In some earlier differential expression analysis packages such as edgeR (Robinson *et al.* 2010), an explicit numerical threshold such as "at least one count per million reads (CPM)" was enforced: genes producing fewer reads than this threshold were removed. Though this filtering was effective in improving statistical power by reducing the FDR, the selection of a cut-off value of "at least one CPM in at least one sample" is arbitrary. In DESeq2, the cut-off threshold is generated dynamically from the dataset by estimating the number of genes that would be filtered out at a range of potential CPM cut-off values. The CPM threshold at which the peak number of DEGs is retained then becomes the actual CPM cut-off, and is applied to each gene's average expression value throughout the samples. This reduces the number of false positives (the FDR), and increases sensitivity to true DEGs. After this independent filtering step, DESeq2 performs regular multiple testing correction using the Benjamini-Hochberg method (Benjamini *et al.* 1995).

After independent filtering, the final differential expression testing procedure can begin. The testing process involves calculating a test statistic for each gene (the shrunken fold-change is divided by its standard error to produce a z-statistic), which is then compared against a normal distribution. Each gene's fold-change is then marked as differentially expressed or not, and a p-value is generated for the comparison, indicating the trustworthiness of the "differentially expressed" designation for the gene.

The next step in DESeq2's analysis is to account for dispersion (variability within replicates representing the same condition) as discussed above, a process which produces a plot of dispersion which shows the strength of the shrinkage applied to the data (Love *et al.* 2014). After all the normalization has been finished, DESeq2 performs the actual hypothesis testing on the data (using the Wald test) to determine, for each gene, whether or not it is significantly differentially expressed between the two conditions – here between the two time points. The null hypothesis used in each test (for each gene) is that there is no difference between the expression levels of the gene across the two conditions – in this analysis, before and after differentiation. Accepting the alternative hypothesis in this case implies that there is a significant difference in the expression of this gene between time points.

Given the arrangement of our study datasets, we had a pair of time points for each analysis – T_0 and T_{+n} (with n being 72h or 96h, depending on the dataset). For each analysis, the earlier timepoint in the pair was used as the reference sample, and the later timepoint as the comparison (for example in the HL60 analysis, T_0 was chosen as the reference and T_{96} the comparison). This was chosen so that DEGs with a positive change in expression were up-regulated, which made basic analysis more intuitive (as just looking at the data one could see that *negative* values were *down*-regulated). Changes in gene expression are expressed as log (base 2) fold change

according to standard practice. The base 2 is chosen for comparisons rather than the raw amount of change to make comparisons numerically simpler: 0 in log notation means the same expression level (because $2^0 = 1$) and 1 shows twice the expression (two-fold or $2^1 = 2$ times the expression), while -1 shows half the expression (as $2^{-1} = \frac{1}{2}$).

As part of `DESeq2`'s command arguments, we supply an argument called study design which controls how `DESeq2` searches for DEGs. The terms in the study design parameter are just the names of columns in the original sample table dataset (Supplemental Table S4).

In the analysis pipeline we devised, `DESeq2` first performed library normalization and produced plots showing the raw and normalized counts for each run, along with a side-by-side plot comparing the transcript counts before and after normalization (this was also useful in identifying low-quality replicates, where replicates with unexpected counts were easily identified). After performing variance stabilizing transformation (VST), `DESeq2` then performed principal components analysis (PCA) on the data and produced plots showing clustering of the different datasets when separated according to their top two principal components.

`DESeq2` produced a list of differentially expressed genes for each analysis run. This list, with one row per differentially expressed gene, included the mean expression level in the baseline condition (the data extracted from the cell line at T_0), the $\log_2$ fold change in expression in the test condition (the data from the cell line after differentiation), and raw and adjusted p-values for the Wald test. An example of the resulting list is available in Supplemental Table S1.

2.7 Functional enrichment analysis with `clusterProfiler`

With a list of differentially expressed genes identified, focus can turn towards investigating biological function. The technique of functional enrichment analysis (FEA) allows lists of genes to be tagged with information about the functions they've been associated with through experimental evidence. Databases exist to combine and arrange gene-to-function relationships like these, and are crucial to the FEA process. The `clusterProfiler` package (Yu *et al.* 2012) performs this analysis against various online biological databases including The Gene Ontology Consortium (GO) (Ashburner *et al.* 2000) and Kyoto encyclopedia of genes and genomes (KEGG) (Kanehisa *et al.* 2000; Kanehisa 2019). The KEGG system contains a set of 16 databases of biological information, including *KEGG Pathway*, a curated list of ~100 pathways representing molecular interactions for known human cellular processes and systems. The GO system is arranged into different sections (like the databases in KEGG) including cellular component, biological process, and molecular function. This enormous system has over 7.9 million annotations across the entire database.

Though the two FEA databases are similar in general application, the KEGG Pathway and GO systems differ in their main approach. GO aims to be an encyclopedic resource for thousands of organisms, while KEGG focuses largely on human pathways and attends them in more detail. This decision has positive and negative ramifications: it makes KEGG Pathway the more detailed and specific of the two

approaches, providing useful, high-quality information on specific pathways, but at the cost of bias - the manual curation of the pathways is not agnostic to their contents, and KEGG does not have nearly as many individual datapoints as GO, even just for *Homo sapiens*.

In this application, KEGG Pathway seemed to provide information that we thought more relevant to the environments in which these cell lines are used - generally settings where human health and immunology are studied. The KEGG Pathway database contains dedicated sections for *Human Diseases* and *Organismal Systems* - groups of multiple pathways which describe processes much more specific to the questions at hand than the general (but vast) *Biological Process* section in GO.

The aim of this step in the analysis pipeline was to take the DEGs found in each dataset and find common DEGs across all cell lines and DEGs unique to each cell line, and then to see if these DEGs were overrepresented in any well known human cellular pathways. To that end, the individual lists of DEGs obtained for the different cell lines were combined in `R v4.0.5` using `RStudio v1.4.1103` and filtered to extract only those genes that were common to all datasets (these genes appeared as differentially expressed in all the three cell lines). After extracting these commonly differentially expressed genes, we investigated DEGs that were unique to each cell line (e.g., these genes were commonly expressed in HL60 but not in U937 or THP-1). Using `RStudio` all three cell lines' lists of differentially expressed genes were filtered to remove genes in common with any others, leaving only those DEGs unique to each cell line. This analysis step produced four key datasets which were used in subsequent analysis: genes uniquely differentially expressed in HL60, U937, and THP-1, and genes differentially expressed in the same way regardless of cell line.

The combined dataset with common DEGs was annotated with *entrez* format gene identifiers for input into downstream analysis tools using `BioMart v2.46.3` (Durinck *et al.* 2005; Durinck *et al.* 2009), as the original gene identifiers from `tximeta` were in *Ensembl* gene format. As the mapping between genes across different databases can be inexact (some genes appear in multiple copies in one database but singularly in another), the dataset was deduplicated by removing genes with non-unique identifiers in their *entrez* format. The gene identifiers in the dataset were then sent to `clusterProfiler v3.99.2` (Yu *et al.* 2012) using its `enrichKEGG` function to perform functional enrichment analysis on the genes against all known human pathways in the Kyoto Encyclopedia of Genes and Genomes pathway database (KEGG Pathway).

For this functional enrichment analysis step, the `target organism` parameter supplied to the KEGG database was set to *Homo sapiens*, restricting the search to human pathways, and a p-value cut-off of `0.05` was supplied. This module returned a list of pathways (specifically their KEGG pathway identifiers) in which the supplied genes were statistically overrepresented compared to chance.

2.8 Pathway over-representation visualization with `pathView`

At the final step in the analysis pipeline, `pathView v1.30.1` (Luo *et al.* 2017) was used to annotate KEGG pathway diagrams with the change in gene expression data obtained by `DESeq2`. The `pathView` module was supplied with KEGG path-

way identifiers produced in the last step alongside the relative levels of differential expression for each gene in the pathway (expressed as $\log_2$ fold change). The module fetched KEGG pathway diagrams of the overrepresented pathways (Kanehisa *et al.* 2000; Kanehisa 2019). These diagrams were annotated by `pathView` to include the relative up- and down-regulation of each relevant DEG in the sample, visually indicating the pathway's regulation in each experiment.

3 Results

3.1 Pipeline

The pipeline ran successfully with minimal intervention and produced well organized folders and files for all the expected outputs. As per the aims outlined in the Methodology section Development of an RNA-seq analysis pipeline, the pipeline is usable as a whole and in parts – each module produces its own self-contained output folder, and these output folders are produced in the same way whether the study uses a single analysis module or all of them at once. This system also arranges the results into folders to keep them easily accessible. An image of the resulting folder structure is available in Supplemental Figure S1.

As envisioned in the aims section, log files were produced at every step of the analysis, and all parameters and software versions were recorded therein for reproducibility. A sample of the log produced by a run of the `deseq-pipe` module is available in Supplemental Figure S2.

As is visible in the log file, the pipeline produces an extensive log every time it is run, with output information on every process being undertaken. Lines 30-32 show the version of the module used in this run, the date and time of the run, and the latest change string which was written into the module configuration file. This sample log was produced during a run of the `rna-pipe` pipeline with only the `deseq-pipe` module selected, thus its outputs are only for that module (see lines 35-39 for note of this). The next section of the log (lines 58-69) output the parameters chosen for this run (allowing the run to be reproduced exactly when combined with the version numbers in the previous sections). The remainder of the log file (lines 71-140) are output from the `DESeq2` analysis itself (including some useful logs of which commands were run and what their outputs were). As seen in Figure S1 subfigure **D**, however, the bulk of the output data is written to individual files for further analysis (the log information is just a reminder for posterity).

3.2 Sequencing quality

The results from the RNA-seq quality control analysis suggested that the data was of relatively good and consistent quality, with the most relevant metric of actual sequence quality being acceptable for all reads. As expected with technical experiments, there were some errors in the analysis, and the quality control step illuminated these and motivated us to discard runs where appropriate. The full sequencing quality explanation is available in Supplemental section Sequencing quality results from `FastQC`.

3.3 Transcript abundances

The results from `Salmon`'s quantification showed the mapping rates to be quite low in HL60 and U937 samples (an average mapping rate of 43.12% in HL60 and 62.43% in U937), indicating that nearly half of the fragments read by the sequencing machines could not be mapped to a compatible location on the transcriptome. The HL60 dataset was sequenced on a machine which produced 86bp fragments, a length

Table 3: Mapping rates for all samples.

Cell line	Sample	Mapping rate %
HL60	SRR3214253	32.48
	SRR3214255	32.64
	SRR3214274	54.32
	SRR3214276	54.98
	HL60 average	43.60
U937	SRR5429556	56.04
	SRR5429557	55.65
	SRR5429562	57.84
	SRR5429563	59.50
	U937 average	57.26
THP-1	SRR6997095	90.88
	SRR6997096	90.22
	SRR6997118	91.68
	SRR6997119	91.91
	SRR6997121	93.57
	SRR6997123	93.24
	THP-1 average	91.92

well within the range known to be usable in differential expression analysis experiments of 50-100+ base pairs (Chhangawala *et al.* 2015). The U937 did include some reads with lengths of 35-50bp, below this threshold, however. Though the overall mapping rates were not high for the HL60 and U937 datasets, we decided to continue the analysis while keeping the mapping quality in mind during the interpretation of the results.

3.4 Differential gene expression analysis

To confirm that the analysis work had been successful so far, we compiled a list of DEGs common to all cell lines undergoing differentiation to test for functional enrichment. This list contained 1789 genes which appeared in the list of DEGs from all of the differential expression analysis experiments. Secondly, a list of DEGs unique to each individual cell line were extracted in order to examine the differences in differential expression between the cell lines. The HL60 set contained 2131 DEGs which were not differentially expressed in any other cell line (referred to hereafter as uniquely differentially expressed genes), the U937 set contained 1711 uniquely differentially expressed genes, and the THP-1 set contained 2159 uniquely differentially expressed genes.

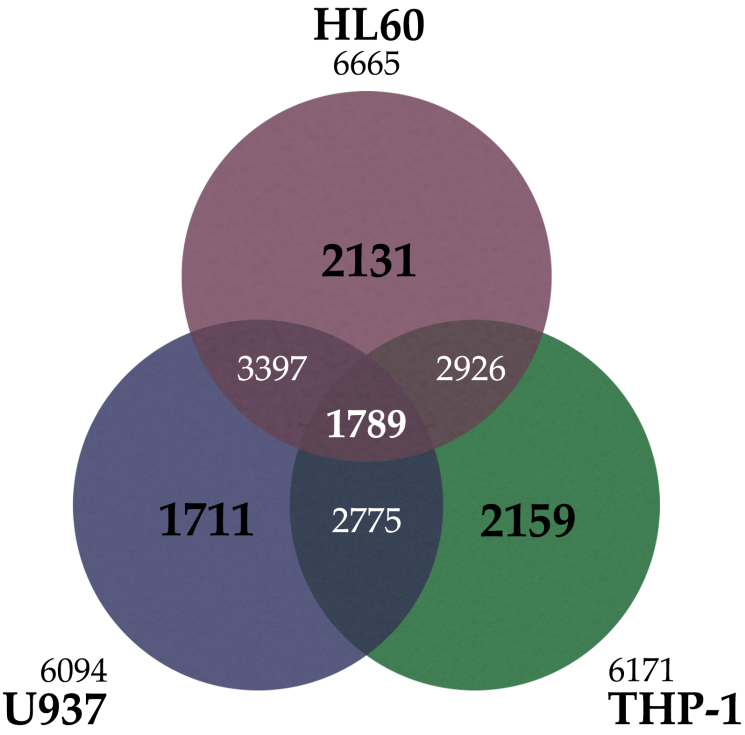


Figure 7: **Common and unique differentially expressed genes across all cell lines.** Numbers outside the circles represent the total number of differentially expressed genes in a cell line. Numbers inside the circles represent the number of genes uniquely differentially expressed in that cell line only. Numbers between two circles represent the number of genes commonly differentially expressed in those two cell lines only. The number in the centre of the graph represents the number of genes commonly differentially expressed across all cell lines.

Differential expression analysis revealed several thousand genes differentially expressed in the myeloid cell lines during their differentiation process. Of those DEGs, 1789 were found to be common across all lines – they were detected as differentially expressed independently in all three datasets. The cell lines also revealed several thousand DEGs which were not in this common set, and were not common to any other line – suggesting unique differential expression in only one dataset. Further analysis of the quality control metrics from differential expression analysis are available in Supplemental section `DESeq2` quality control data.

3.5 Analysing functional enrichment

We first obtained lists of overrepresented KEGG pathways (Kanehisa *et al.* 2000) applicable to all common DEGs (Supplemental Figure S8 **A**), then proceeded to identify overrepresented pathways from each cell line in turn. The HL60 and THP-1 cell lines revealed several pathways in which the supplied DEGs were statistically overrepresented compared to chance – Supplemental Figure S8 subfigures **B** and **C**. The U937 dataset, however, did not show any evidence of unique functional enrichment. The list of pathways functionally enriched in all cell lines is presented in Table 4.

Table 4: Functionally enriched pathways common to all cell lines

ID	Description	Gene Ratio	Bg Ratio	p-value	p.adjust	qvalue	Count
hsa03030	DNA replication	27/827	36/8113	4.36E-20	1.39E-17	1.10E-17	27
hsa04142	Lysosome	40/827	132/8113	9.97E-11	1.59E-08	1.26E-08	40
hsa03430	Mismatch repair	13/827	23/8113	5.07E-08	5.41E-06	4.27E-06	13
hsa04380	Osteoclast differentiation	33/827	128/8113	3.30E-07	2.64E-05	2.08E-05	33
hsa05152	Tuberculosis	40/827	180/8113	1.34E-06	8.58E-05	6.77E-05	40
hsa04110	Cell cycle	31/827	126/8113	2.22E-06	1.18E-04	9.35E-05	31
hsa03440	Homologous recombination	15/827	41/8113	5.85E-06	2.68E-04	2.11E-04	15
hsa03420	Nucleotide excision repair	16/827	47/8113	8.44E-06	3.38E-04	2.66E-04	16
hsa03410	Base excision repair	13/827	33/8113	9.60E-06	3.42E-04	2.70E-04	13
hsa05135	Yersinia infection	31/827	137/8113	1.41E-05	4.51E-04	3.56E-04	31
hsa04137	Mitophagy - animal	20/827	72/8113	2.12E-05	6.17E-04	4.87E-04	20
hsa04218	Cellular senescence	33/827	156/8113	3.30E-05	8.80E-04	6.95E-04	33
hsa05323	Rheumatoid arthritis	23/827	93/8113	4.10E-05	1.01E-03	7.96E-04	23
hsa04145	Phagosome	32/827	152/8113	4.78E-05	1.09E-03	8.62E-04	32
hsa05221	Acute myeloid leukaemia	18/827	67/8113	8.79E-05	1.88E-03	1.48E-03	18
hsa04216	Ferroptosis	13/827	41/8113	1.36E-04	2.72E-03	2.15E-03	13
hsa05120	Epithelial cell sig. *H. pylori* infection	18/827	70/8113	1.62E-04	3.05E-03	2.41E-03	18
hsa04010	MAPK signalling pathway	50/827	294/8113	1.74E-04	3.10E-03	2.45E-03	50
hsa05140	Leishmaniasis	19/827	77/8113	1.94E-04	3.27E-03	2.58E-03	19
hsa04979	Cholesterol metabolism	14/827	50/8113	3.31E-04	5.29E-03	4.18E-03	14
hsa05134	Legionellosis	15/827	57/8113	4.28E-04	6.53E-03	5.15E-03	15
hsa00240	Pyrimidine metabolism	15/827	58/8113	5.25E-04	7.63E-03	6.02E-03	15
hsa04668	TNF signalling pathway	23/827	112/8113	7.75E-04	1.06E-02	8.35E-03	23
hsa05222	Small cell lung cancer	20/827	92/8113	7.93E-04	1.06E-02	8.35E-03	20
hsa05161	Hepatitis B	30/827	162/8113	8.40E-04	1.08E-02	8.49E-03	30
hsa04657	IL-17 signalling pathway	20/827	94/8113	1.05E-03	1.30E-02	1.02E-02	20
hsa05203	Viral carcinogenesis	35/827	204/8113	1.36E-03	1.57E-02	1.24E-02	35
hsa05212	Pancreatic cancer	17/827	76/8113	1.37E-03	1.57E-02	1.24E-02	17
hsa05418	Fluid shear stress and atherosclerosis	26/827	139/8113	1.57E-03	1.73E-02	1.36E-02	26
hsa04621	NOD-like receptor signalling pathway	32/827	184/8113	1.71E-03	1.82E-02	1.44E-02	32
hsa04613	Neutrophil extracellular trap formation	32/827	190/8113	2.89E-03	2.97E-02	2.34E-02	32
hsa05146	Amoebiasis	20/827	102/8113	2.97E-03	2.97E-02	2.34E-02	20
hsa05417	Lipid and atherosclerosis	35/827	215/8113	3.38E-03	3.22E-02	2.54E-02	35
hsa05202	Transcriptional misregulation in cancer	32/827	192/8113	3.42E-03	3.22E-02	2.54E-02	32
hsa05131	Shigellosis	39/827	247/8113	3.54E-03	3.24E-02	2.56E-02	39
hsa05144	Malaria	12/827	50/8113	3.66E-03	3.25E-02	2.57E-02	12
hsa04130	SNARE interactions in vesicular transport	9/827	33/8113	4.61E-03	3.99E-02	3.15E-02	9

The `clusterProfiler` program identified 37 functionally enriched pathways which were differentially expressed in all cell lines undergoing differentiation at a p-value of 0.05. The Gene Ratio shows how many of the total number of DEGs were represented in that pathway: in Table 4, of the 1789 common DEGs supplied, only 827 were possible matches to pathways – so each numerator shows how many of those 827 genes mapped to the pathway in question. The Bg Ratio (background ratio) shows how many genes (including non DEGs) in the sample were represented in the pathway (of the tens of thousands of genes in the human genome, only 8113 are annotated in KEGG pathways, thus this number is the denominator).

Using `clusterProfiler` we identified 17 pathways as functionally enriched in the set of 2131 genes that were uniquely differentially expressed in HL60 cells. This list of pathways is presented in Table 5.

We found 13 pathways that were functionally enriched in the set of 2159 genes that were uniquely differentially expressed in THP-1 cells. This list of pathways uniquely differentially expressed in THP-1 is presented in Table 6. No pathways were found to be functionally enriched in the set of 1711 genes that were uniquely differentially expressed in U937 cells.

Table 5: Functionally enriched pathways unique to HL60 cells

ID	Description	Gene Ratio	Bg Ratio	p-value	p.adjust	qvalue	Count
hsa03010	Ribosome	79/876	158/8113	1.08E-35	3.51E-33	3.25E-33	79
hsa05171	Coronavirus disease – COVID-19	81/876	232/8113	2.15E-23	3.48E-21	3.22E-21	81
hsa05016	Huntington disease	95/876	306/8113	5.27E-23	5.69E-21	5.27E-21	95
hsa05012	Parkinson disease	85/876	266/8113	1.43E-21	1.16E-19	1.07E-19	85
hsa00190	Oxidative phosphorylation	56/876	134/8113	9.58E-21	6.21E-19	5.75E-19	56
hsa05020	Prion disease	82/876	273/8113	6.01E-19	3.24E-17	3.00E-17	82
hsa04714	Thermogenesis	73/876	232/8113	3.24E-18	1.50E-16	1.39E-16	73
hsa05415	Diabetic cardiomyopathy	67/876	203/8113	5.13E-18	2.08E-16	1.92E-16	67
hsa05014	Amyotrophic lateral sclerosis	96/876	364/8113	1.02E-17	3.66E-16	3.39E-16	96
hsa05022	Pathways of neurodegen. – multiple diseases	111/876	476/8113	4.41E-16	1.43E-14	1.32E-14	111
hsa05010	Alzheimer disease	93/876	384/8113	1.19E-14	3.50E-13	3.24E-13	93
hsa05208	Chemical carcinogenesis – ROS	64/876	223/8113	5.79E-14	1.56E-12	1.45E-12	64
hsa04932	Non-alcoholic fatty liver disease	46/876	155/8113	6.38E-11	1.59E-09	1.47E-09	46
hsa03040	Spliceosome	35/876	147/8113	4.31E-06	9.96E-05	9.23E-05	35
hsa04723	Retrograde endocannabinoid signalling	35/876	148/8113	5.07E-06	1.09E-04	1.01E-04	35
hsa03060	Protein export	11/876	23/8113	8.68E-06	1.76E-04	1.63E-04	11
hsa05017	Spinocerebellar ataxia	31/876	143/8113	1.03E-04	1.97E-03	1.82E-03	31

Table 6: Functionally enriched pathways unique to THP-1 cells

ID	Description	Gene Ratio	Bg Ratio	p-value	p.adjust	qvalue	Count
hsa04360	Axon guidance	41/841	182/8113	1.05E-06	3.39E-04	3.06E-04	41
hsa05332	Graft-versus-host disease	15/841	42/8113	1.01E-05	1.40E-03	1.26E-03	15
hsa05168	Herpes simplex virus 1 infection	81/841	495/8113	1.48E-05	1.40E-03	1.26E-03	81
hsa04640	Hematopoietic cell lineage	25/841	99/8113	1.74E-05	1.40E-03	1.26E-03	25
hsa04020	Calcium signalling pathway	46/841	240/8113	2.42E-05	1.56E-03	1.41E-03	46
hsa04060	Cytokine-cytokine receptor interaction	52/841	295/8113	7.74E-05	4.15E-03	3.75E-03	52
hsa05323	Rheumatoid arthritis	21/841	93/8113	4.33E-04	1.99E-02	1.80E-02	21
hsa04390	Hippo signalling pathway	30/841	157/8113	6.48E-04	2.61E-02	2.35E-02	30
hsa04310	Wnt signalling pathway	31/841	167/8113	8.81E-04	3.15E-02	2.84E-02	31
hsa04940	Type I diabetes mellitus	12/841	43/8113	1.05E-03	3.37E-02	3.04E-02	12
hsa04514	Cell adhesion molecules	28/841	149/8113	1.26E-03	3.70E-02	3.34E-02	28
hsa05146	Amoebiasis	21/841	102/8113	1.54E-03	3.90E-02	3.52E-02	21
hsa05217	Basal cell carcinoma	15/841	63/8113	1.57E-03	3.90E-02	3.52E-02	15

3.6 Analysing pathway over-representation

In order to better visualize the results from the functional enrichment analysis we produced diagrams of the pathways in which the data had overrepresented DEGs and annotated them to show relative level of under- and over-expression.

Though there were many pathways in which the DEGs were overrepresented for each grouping (common, unique to HL60, unique to THP-1), we selected pathways for further visualization and analysis based on their relevance to our biological question, vis-à-vis the key pathways involved in the monocyte-to-macrophage transition. In addition to these common pathways we searched through the pathways overrepresented in the unique groupings for behaviours likely to distinguish the behaviour of these cell lines in practice.

3.7 Enriched pathways in commonly differentially expressed genes

Of the 37 pathways in which the commonly differentially expressed genes were overrepresented (Table 4 in the supplemental materials), the Lysosome and Phagosome pathways were selected as being of interest to our study because they make up

one of the defining characteristics of differentiated macrophages. These pathways are relevant to the monocyte-to-macrophage transition and are discussed further in subsection 4.1: Core expression signature. These pathways of interest were then investigated to compare the direction of change of expression for each gene across the different cell lines. This analysis revealed a few minor discrepancies wherein some genes which we picked up as significantly differentially expressed in all cell lines were expressed in different directions (a list of these results is available in Supplemental subsection 6.5: Commonly differentially expressed genes regulated in different directions).

After performing over-representation analysis with all common DEGs as a whole, we split the commonly DEGs into up- and down-regulated groups and performed functional enrichment analysis in the same way. As a result, we received a list of pathways included only with genes that were up- or down-regulated over differentiation. Although the results from these tests with only single-direction net changes in expression are less dependable than those from the experiments including both up-and down-regulated genes together (as we are artificially splitting our true list of common DEGs), we found some of the results interesting. Although the majority of the enriched pathways were the same in the combined versus split groups, some pathways only appeared in the up-regulated group, such as the Leukocyte transendothelial migration pathway.

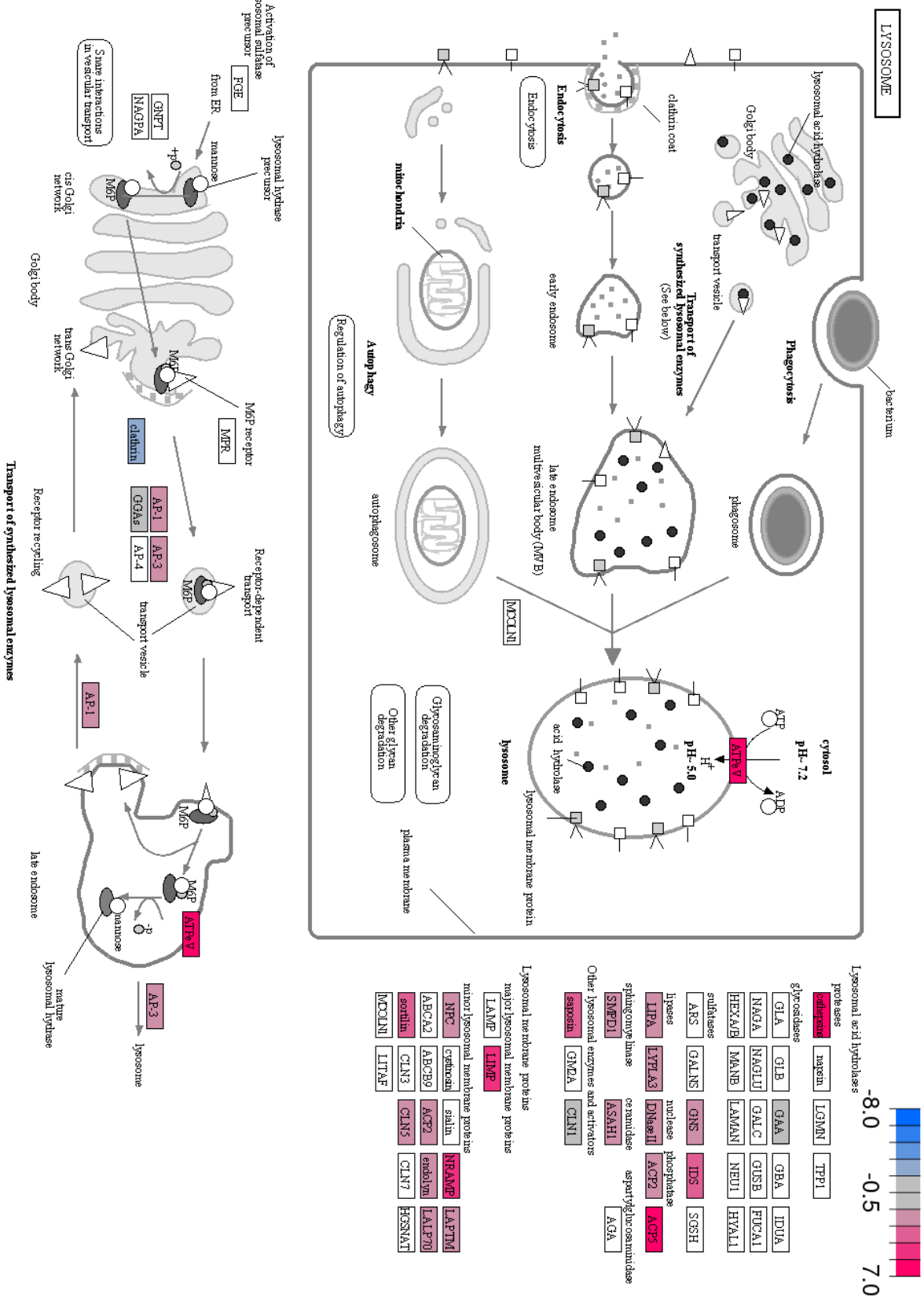


Figure 8: **KEGG Lysosome pathway – commonly differentially expressed genes**. All cell lines' differentially expressed genes in the lysosome pathway and their average relative levels of up- and down-regulation during differentiation. Expression levels are written as $\log_2$ fold change. Note the up-regulation of genes encoding hydrolase enzymes and lysosomal membrane proteins. Image adapted from Kanehisa *et al.* (2000).

All cell lines differentially expressed several genes in the lysosome pathway – with all but one of the genes being up-regulated over the course of differentiation. The lipase enzymes lysosomal acid lipase (LIPA) and lysophospholipase LYPLA3 and saposin (which all degrade lipids and membrane phospholipids), phosphatases ACP2 and ACP5, sphingomyelinase SMPD1, and ASAH1 cermidase were all up-regulated. Cathepsin protease was up-regulated, as was the production of LIMP major lysosomal membrane protein and various minor lysosomal membrane proteins which make up the lysosome membrane.

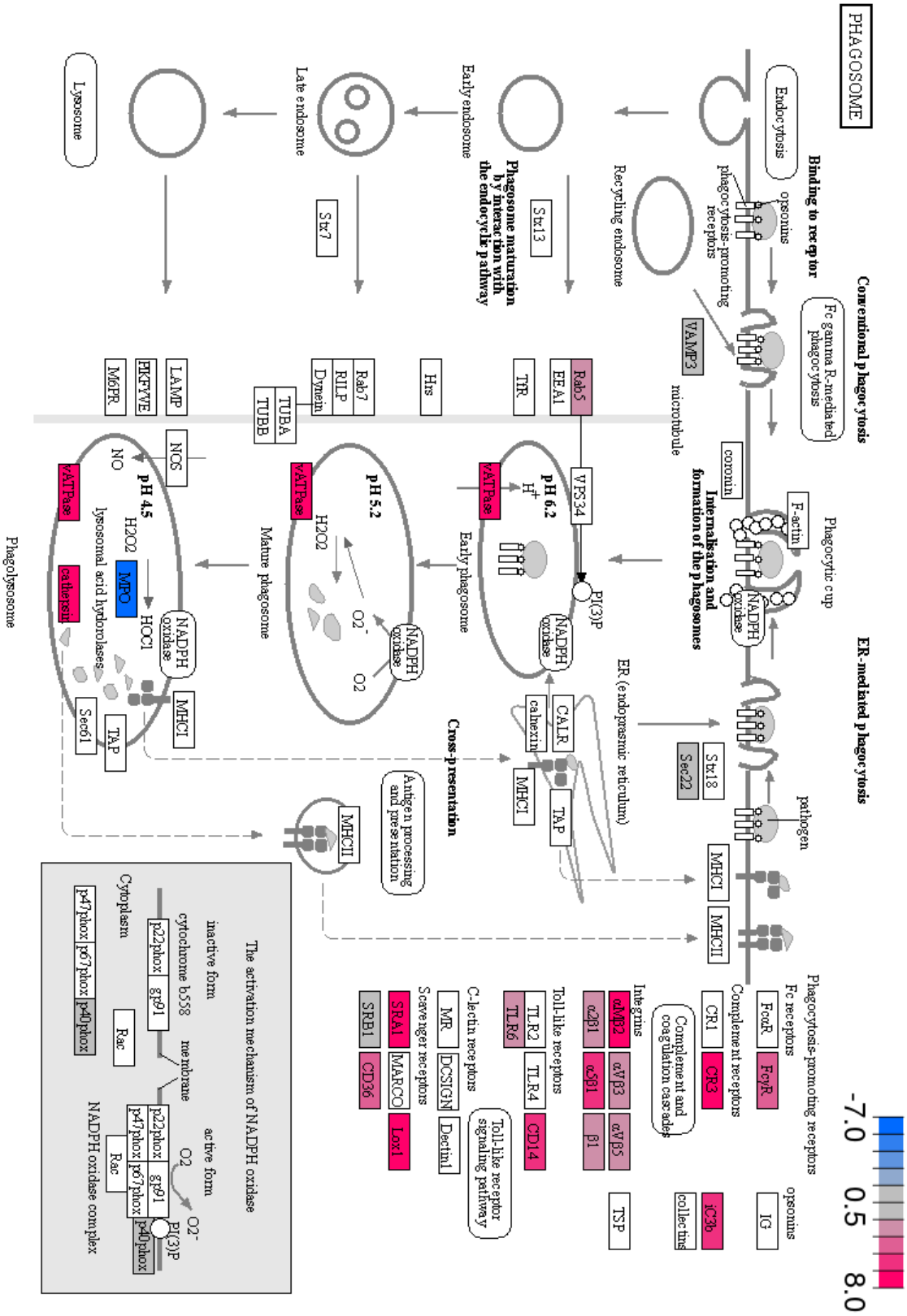


Figure 9: **KEGG Phagosome pathway – commonly differentially expressed genes**. All cell lines' differentially expressed genes in the phagosome pathway and their average relative levels of up- and down-regulation during differentiation. Expression levels are written as $\log_2$ fold change. Note the up-regulation of genes encoding integrins, scavenger receptors, and phagolysosome components. Image adapted from Kanehisa *et al.* (2000).

Production of various genes in the phagosome pathway was up-regulated in all cell lines. Phagocytosis-promoting receptors Fcγ, complement receptor CR3, and markers of macrophage phenotype CD14 and TLR6 toll-like receptors were all up-regulated. The production of various integrin genes was increased, including the α_M:β_2, α_V:β_3 and α_V:β_5, α_2:β_1, α_5:β_1 and the β_1 subunit alone. Various scavenger receptors were also up-regulated, specifically the Lox1 scavenger receptor and the class A1 scavenger receptor SRA1.

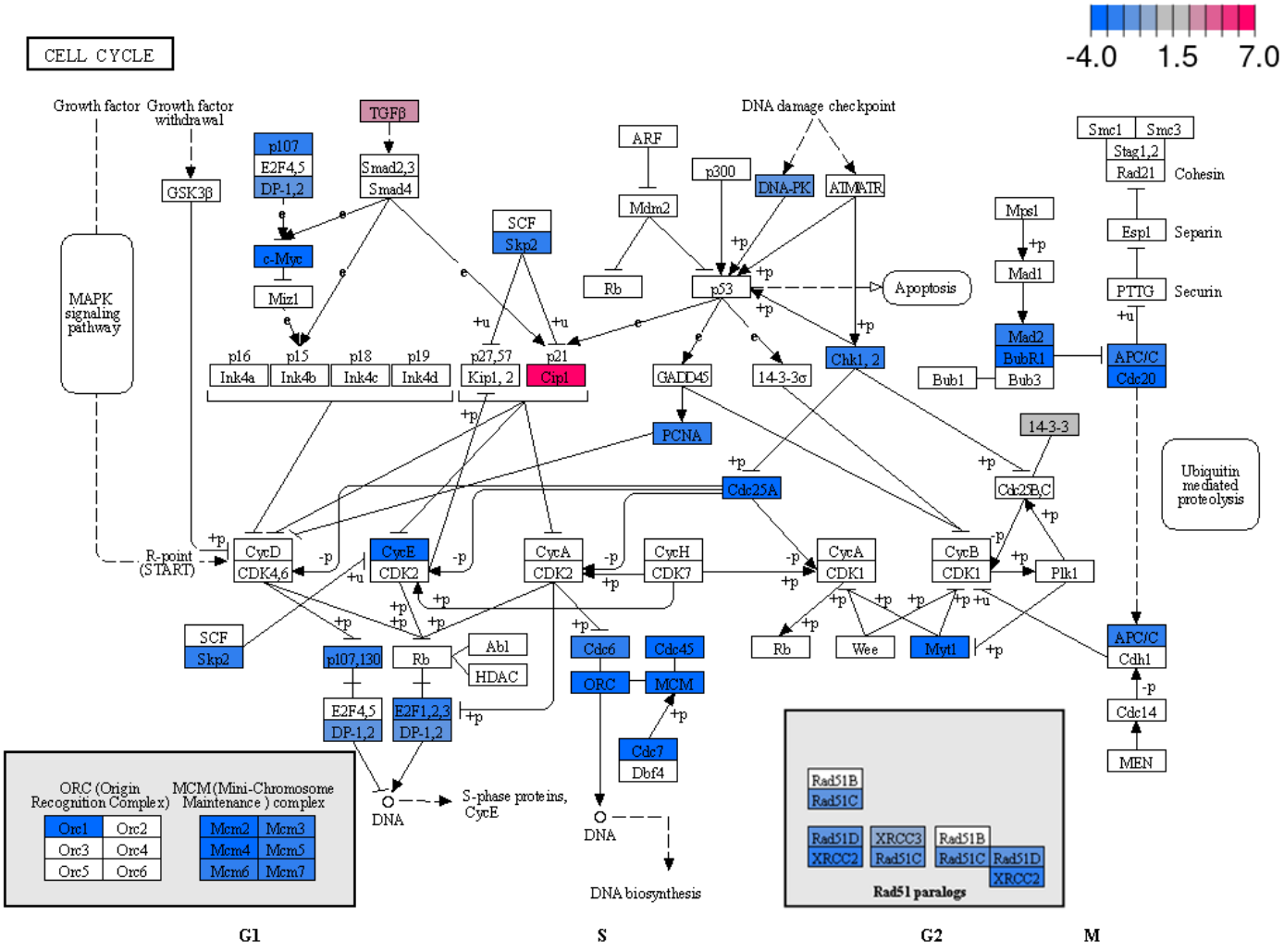


Figure 10: **KEGG Cell cycle pathway – commonly differentially expressed genes**. All cell lines' differentially expressed genes in the cell cycle pathway and their average relative levels of up- and down-regulation during differentiation. Expression levels are written as $\log_2$ fold change. Note the down-regulation of Rad51 paralogs, MCM genes, and c-Myc. Image adapted from Kanehisa *et al.* (2000).

Production of the c-Myc gene was down-regulated in all cell lines, along with several cell division cycle genes (CDC). All members of the minichromosome maintenance protein complex (MCM) DNA replication control genes were down-regulated. Finally, all cell lines also showed down-regulation of several Rad51 paralogs – controllers of double-stranded DNA break repair.

3.8 Enriched pathways in uniquely differentially expressed genes

After ascertaining that there were several enriched pathways in commonly differentially expressed genes, we searched for the existence of enriched pathways within the sets of DEGs unique to the HL60, U937, and THP-1 cell lines (as mentioned in subsection 3.5). The searches revealed 17 enriched pathways in the HL60 DEG set and 13 enriched pathways in the THP-1 DEG set (Table 5 and Table 6). Interestingly, no KEGG pathways were functionally enriched in the set of U937 DEGs. In the HL60 set, the Ribosome and Oxidative phosphorylation pathways were selected as being of interest to our study, and in the THP-1 set we decided to investigate the Cytokine-cytokine receptor interaction pathway.

After discovering these uniquely regulated pathways we decided to produce visualizations of the pathways we found of interest in each cell line (Figure 11, Figure 12, and Figure 13).

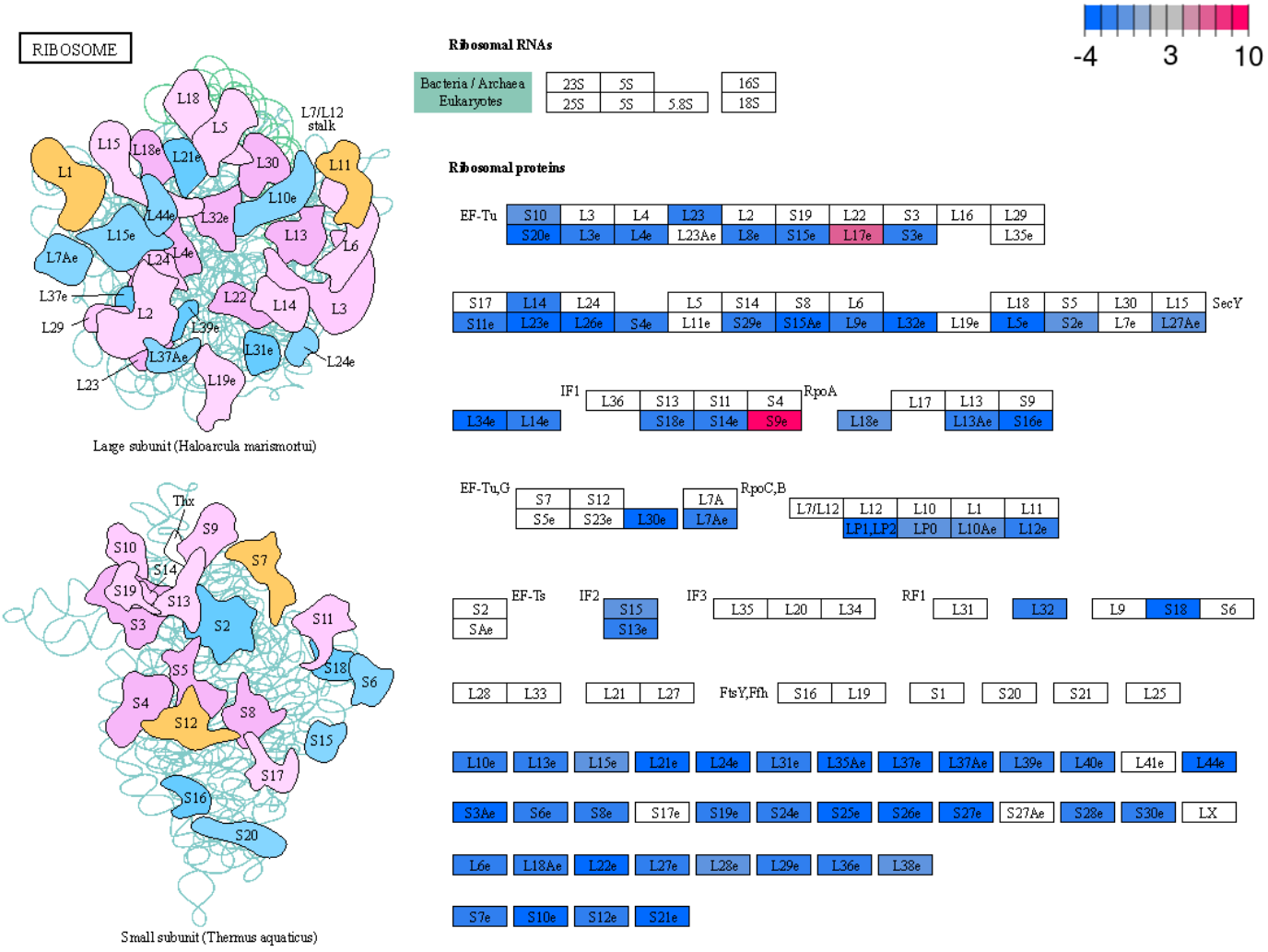


Figure 11: **KEGG Ribosome pathway – differentially expressed genes in HL60**. HL60 cells' differentially expressed genes in the ribosome pathway and their relative levels of up- and down-regulation during differentiation. Expression levels are written as $\log_2$ fold change. Note the resounding down-regulation of genes encoding ribosomal proteins. Image adapted from Kanehisa *et al.* (2000).

The HL60 cells differentially expressed several genes in the ribosome KEGG pathway – almost universally down-regulating them. Over 60 such ribosomal genes showed diminished gene expression by the end of the experimental procedure, with genes for both halves of the ribosome (large and small subunit proteins) being affected. Interestingly, the S9e small subunit protein and the L17e large subunit protein showed increased expression.

Figure 12: **KEGG Oxidative phosphorylation pathway – differentially expressed genes in HL60**. HL60 cells' differentially expressed genes in the oxidative phosphorylation pathway and their relative levels of up- and down-regulation during differentiation. Expression levels are written as $\log_2$ fold change. Note the profound down-regulation of NADH dehydrogenase components, cytochrome c reductase and oxidases, and F-type ATPases. (Kanehisa *et al.* 2000)

HL60 cells also showed decreased expression of many genes involved in oxidative phosphorylation. Large enzyme complexes such as the NADH-ubiquinone-1 reductase complex (7.1.1.2 in the diagram), the quinol—cytochrome-c reductase complex (7.1.1.8 in the diagram), cytochrome-c oxidase (7.1.1.9 in the diagram), and H^+-transporting two-sector ATPase (marked 7.1.2.2) were down-regulated significantly. Many of the components of these complexes, and some stand-alone proteins involved in the electron transport chain, showed various levels of down-regulation.

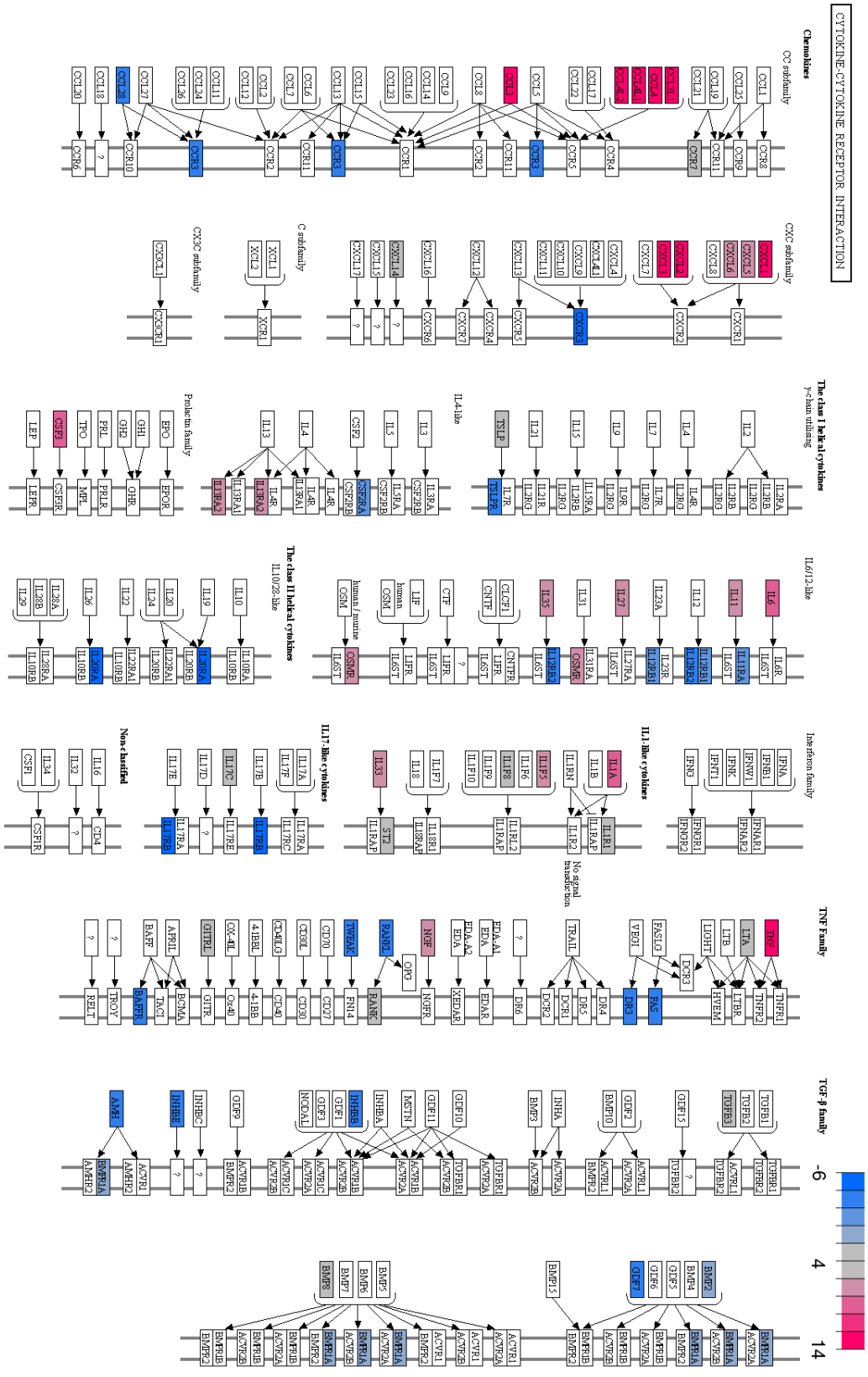


Figure 13: **KEGG Cytokine-cytokine receptor interaction pathway – differentially expressed genes in THP-1**. THP-1 cells' differentially expressed genes in the Cytokine-cytokine receptor interaction pathway and their relative levels of up- and down-regulation during differentiation. Expression levels are written as $\log_2$ fold change. Note up- and down-regulated chemokines (CC and CXC subfamilies) and various IL-like cytokines. (Kanehisa *et al.* 2000)

THP-1 cells showed up- and down-regulation of various genes in the massive Cytokine-cytokine receptor interaction pathway. Within the CC subfamily of chemokines, five ligands for the CCR5 receptor (namely CCL3, CCL3L1, CCL4, CCL4L1, and CCL4L2) were upregulated by several hundred times (though the receptor itself was not differentially regulated). The expression of the CCR3 receptor was heavily down-regulated, as was one of its ligands: CCL28. Several ligands in the CXC chemokine subfamily were upregulated: CXCL1, 5, and 6 (all ligands for CXCR1 and CXCR2), as well as CXCL2 and 3, both ligands for CXCR2 only. The expression of the CXCR3 receptor was down-regulated strongly, though none of its ligands were differentially regulated.

In the interleukin 6- and 12-like chemokine family four ligands were upregulated by several fold (IL6, IL11, IL27, and IL35) – but their receptors were down-regulated by similar amounts (IL11RA, IL12RB1 and IL12RB2). Of note, both subunits of the IL12 receptor, IL12RB1 and IL12RB2, were down-regulated. Some other interleukin receptors were down-regulated more severely, notably IL20RA and IL17RB.

In the TNF family, TNF-α and NGF ligands were upregulated, while ligands RANKL and TWEAK as well as receptors FAS, DR3, and BAFFR were down-regulated.

Finally, in the TGF-β family, the INHBB, INHBE, and AMH ligands were all down-regulated.

4 Discussion

The results of our analysis suggested that the cells in the cell lines we investigated differentiate as expected, showing signs of phagocytosis and exiting the cell cycle to prepare for differentiation, and presenting markers of the macrophage state. The pathways enriched in the differentially expressed genes suggested also that the HL60 cells in particular shift their metabolism away from aerobic energy generation, and down-regulate their overall protein production. The chemokine signalling systems of THP-1 cells in particular indicated that these cells were progressing towards an M1 (pro-inflammatory) macrophage phenotype.

4.1 Core expression signature

Of the 1789 genes identified as differentially expressed in all the cell lines undergoing differentiation, only a small subset were involved in the 37 enriched pathways discovered through pathway analysis. Of interest was the finding that all cell lines universally down-regulated genes involved in cell division. Genes controlled by c-Myc are involved in cell proliferation and growth (Schmidt 1999), and these genes are known to be down-regulated during the differentiation of some cells (Shalhoub *et al.* 1989). In our findings, all cell lines showed down-regulation of the minichromosome maintenance protein complex (MCM), which primes DNA for replication and controls entry into the S-phase of the cell cycle (Tye 1999) as well as universal down-regulation of c-Myc. All cell lines also showed down-regulation of several Rad51 paralogs – proteins in the Rad51 family which control homologous recombination of DNA during double strand break repair (Shinohara *et al.* 1992), suggesting

DNA replication halting and withdrawal from the cell cycle. The down-regulation of these genes implies a reduction in cell growth – behaviour consistent with differentiation (Shalhoub *et al.* 1989).

In addition to the regulation of genes involved in cell cycle control, all cell lines showed differential expression of a variety of genes involved in the behaviour of adherence and trans-endothelial migration. Monocytes circulate through the blood in the natural state and only change this behaviour when differentiating into macrophages – at which point they need to start adhering to endothelial cells to start crossing the blood vessel wall and enter a site of infection. This process of transendothelial migration (TEM) is observed in many leukocytes (Schimmel *et al.* 2017), and is up-regulated in the cells of all lines we studied – with many genes in the pathways expressed at far above the pre-induction reference level. The matrix metalloproteinases or MMPs effect the breakdown of extracellular matrix proteins to allow leukocyte infiltration of sub-endothelial tissue (Brundula *et al.* 2002), and their expression was up-regulated across all the cell lines. Notably, this process is associated more with monocytes than with macrophages, as the monocytes are the cell type which need to exit circulation and migrate across the endothelial wall. This reinforces the notion that these cells are in the process of differentiating, and some of their number may still have phenotypes associated with monocytes, not yet with monocyte-derived-macrophages.

Another key process in the negotiation between leukocytes and endothelial cells is the regulation of the cells' extracellular matrix (ECM) interactions. Distinct from the action of MMPs, the interactions of ICAMs and fibronectin on the endothelial cells with various integrins on the leukocyte are crucial in mediating TEM (Issekutz *et al.* 1999; Schimmel *et al.* 2017) – and the genes coding for these proteins and their receptors are up-regulated in all the cell lines we observed. Integrins LFA-1 Mac-1 and VLA-4 were all up-regulated by several fold, suggesting these cells' preparation for TEM (Meerschaert *et al.* 1995).

The third major behaviour we found to be commonly differentially expressed across all cell lines was that of phagocytosis. We came to this conclusion by noting that a multitude of genes in two pathways involved in the behaviour were highly up-regulated – the Lysosome and Phagosome KEGG pathways. Within the phagosome pathway, the production of phagocytosis-promoting receptors Fcγ, various integrins, toll-like receptor TLR6 and the TLR4 co-receptor CD14, and various scavenger receptors were all up-regulated. These receptors are known to be required to some degree or another for phagocytosis to take place (Aderem *et al.* 1999).

This change in expression suggests a shift towards production of much of the cellular machinery required for the phagosome, and thus preparation for phagocytosis. Fcγ receptors are present on the membranes of various leukocytes and directly mediate their ability to perform phagocytosis (Anderson *et al.* 1990). These integrins, particularly α_M:β_2, are crucial in macrophages' ability to phagocytose pathogens (Jawhara *et al.* 2017).

CD14 is used as a monocyte classification marker (Ziegler-Heitbrock 2007) and is the co-receptor for LPS in bacterial cell walls (Zanoni *et al.* 2009), crucial in inducing pathogen recognition before phagocytosis (Zanoni *et al.* 2011). TLR6 is a key leukocyte receptor and PRR involved in pathogen recognition – specifically the

recognition of peptidoglycan from bacteria and mycoplasma which have been captured in the phagosome (Ozinsky *et al.* 2000). The scavenger receptors CD36 (which recognizes oxidized phospholipids and lipoproteins (Silverstein *et al.* 2009) and mediates some forms of phagocytosis – Smith *et al.* 2003), and SRA1 (which mediates phagocytosis in other leukocytes, particularly dendritic cells – Amiel *et al.* 2007) were all up-regulated by several fold. The Lox1 scavenger receptor (which scavenges oxidized low-density lipoprotein – Yoshimoto *et al.* 2011) is notably known to be expressed in monocyte-derived macrophages and differentiated THP-1 cells, and not in immature THP-1 cells or monocytes (Yoshida *et al.* 1998).

Within the lysosome pathway, we observed that the expression of various genes encoding proteins involved in lysosome production and functioning were up-regulated: saposins are glycoproteins responsible for membrane digestion and lipid degradation in lysosomes (Darmoise *et al.* 2010), and cathepsin proteases are endopeptidases involved in lysosomal protein breakdown and antigen processing in the endosome (Brix 2013). Sulfatases like glucosamine (N-acetyl)-6-sulfatase (GNS) and iduronate 2-sulfatase (IDS), which catalyse the breakdown of heparin, heparan sulphate, and keratan sulphate (GNS – Freeman *et al.* 1987), and of heparan sulphate and dermatan sulphate (IDS – Bielicki *et al.* 1990) were also up-regulated. These changes also point to lysosome construction as a new target of these cells' metabolism – reinforcing the idea that the cells are preparing for phagocytosis.

4.2 Unique expression signatures

The unique expression signatures detected in HL60 and THP-1 were identified using functional enrichment analysis with `clusterProfiler`. This technique identifies enriched pathways based on the statistical over-representation of the supplied genes in the annotated list of pathways in a database. Some of the differentially expressed genes in the list supplied to `clusterProfiler` are doubtless of little interest to the present study, but the pathway over-representation analysis requires that all genes be submitted. Because of this arrangement, most of the genes in a particular pathway need to be differentially expressed in our dataset in order for that pathway to show up as functionally enriched and for us to take an interest in it. This means that the noise pattern in the datasets needs to match up – in order for pathways to be flagged as functionally enriched, genes of no interest need to be differentially expressed alongside those of relevance for this study. As a result of this study design, the fact that no pathways were functionally enriched in U937 cells does not imply that no relevant genes were differentially expressed in those lines: the functional enrichment analysis must be interpreted as an exploratory technique.

A confounding factor in the analysis must also be noted here: that the experiments which generated the data we used were performed using different culturing techniques at different laboratories. The HL60 were cultured in Dulbecco's Modified Eagle Medium [DMEM] (Ramirez *et al.* 2017), and the THP-1 cells were cultured in Roswell Park Memorial Institute Medium [RPMI]-1640 (Rossetti *et al.* 2017). The experiment which used the U937 cells did not report directly on their culturing media, though the listed supplier for their cells (the American Type Culture Collection) mentions the RPMI-1640 medium as the standard culture medium for growing the U937 / ATCC CRL-1593.2 cells (ATCC 2022), which would make it the

same medium as for the THP-1 cultures. The amount of fetal bovine serum (FBS) in the growth media differ slightly, with HL60 receiving 20% FBS in their media, and THP-1 receiving 10% FBS. The antibiotics used in the growth media are (where listed) all Penicillin, with THP-1 cells also receiving Streptomycin. These differences are slight and do not suggest a substantial change in conditions, thus we would not expect the culture environment to significantly affect the comparisons of differential gene expression. Finally, the amount of PMA used to induce differentiation also differed between the experiments, with HL60 cells receiving 10µM PMA, U937 cells receiving 80nM PMA, and THP-1 cells being exposed to ~162nM PMA (reported as 100 ng/ml).

The DEGs unique to HL60 cells were functionally enriched in two pathways, both indicating changes in metabolism. The annotation produced for the Ribosome KEGG pathway shows enormous down-regulation of ribosomal proteins in both the large and small subunits – over 60 ribosomal proteins are down-regulated to some degree. Regulation of these genes is mostly monolithic – the multiple subunit genes behave as a single regulatory unit, being up- or down-regulated together (Petibon *et al.* 2021). The ribosomes' prime function is in maintaining steady-state levels of cellular proteins (Lafontaine *et al.* 2001), and the expression of their genes is mostly controlled by stress and nutrient access – with stress and starvation inducing down-regulation of ribosomal proteins (Warner 1999). Ribosomal genes are known to be tightly controlled during differentiation, and rRNA expression and ribosome synthesis have been found to be down-regulated during lineage progression (Ali *et al.* 2008; Neben *et al.* 2017). The fact that these changes in ribosomal genes were so pronounced in the HL60 line and not the others is likely due to nutrient availability and cell stress changes, as the pattern would be expected in all cell lines and was likely more pronounced in this line because of variation in the culturing technique.

The second pathway to be down-regulated in HL60 cells is the Oxidative phosphorylation KEGG pathway, where many components of the Type I NADH dehydrogenase complex were found to be down-regulated (NADH-ubiquinone oxidoreductase chain proteins (ND)1-6, iron-sulphur proteins, and alpha and beta subcomplexes all show down-regulation). The Type I NADH dehydrogenase complex is responsible for proton transport in oxidative phosphorylation (Esposti 1998), and thus its down-regulation suggests movement in these cells away from oxidative metabolism. This change has been recognized in M1 macrophages as a metabolic switch from oxidative phosphorylation to glycolysis (Rodríguez-Prados *et al.* 2010). Alternatively activated M2 macrophages perform the metabolic rearrangement in the other direction: they have been shown to have lower glycolytic rates and higher levels of oxidative metabolism after induction (O'Neill *et al.* 2013). The corollary of this process has also been tested, and it showed that the treatment of HL60 cells with NADH dehydrogenase inhibitors can induce CD14 marker up-regulation (Mills *et al.* 1999).

Given that leukocytes use chemokines and cytokines to communicate as part of their functions in the innate and adaptive immune systems it is not surprising that DEGs observed in monocyte-to-macrophage differentiation would appear as functionally enriched in the Cytokine-cytokine receptor interaction pathway.

In the THP-1 dataset, over 25 chemokine and cytokine receptors were differentially

regulated (in both directions). Within the CC chemokine subfamily, the CCL3, CCL3L1, CCL4, CCL4L1, and CCL4L2 ligands for the CCR5 receptor were strongly up-regulated. Almost all the known ligands for the CXCR1 and CXCR2 receptors (the two heterodimeric halves of the Interleukin 8 receptor) were up-regulated significantly, reiterating that these cells are inducing neutrophil chemotaxis (a process dependent on the CXCR1 and CXCR2 receptors – Bergin *et al.* 2010).

CCL3 and CCL4 (and their variants) are the heterodimeric components of the macrophage inflammatory protein (MIP)-1 (Sherry *et al.* 1988). MIP-1 is a heterodimeric CC subfamily chemokine protein comprised of MIP-1α and MIP-1β subunits (Fahey *et al.* 1992). The chemokine is produced by macrophages and monocytes after they are stimulated with bacterial endotoxin (O'Grady *et al.* 1999) or IL-1β (Menten *et al.* 2002). As a proinflammatory chemokine, MIP-1 has been shown to attract neutrophils to sites of infection and to induce leukocyte transendothelial migration (Wolpe *et al.* 1988; Menten *et al.* 2002). Interestingly, the molecule acts as an autocrine regulator in macrophages, with the two subunits of which it is comprised showing opposite behaviours. MIP-1α stimulates macrophage activation (as evidenced by the TNF-α, IL-1 alpha, and IL-6 release), but MIP-1β does not have this effect, and instead diminishes TNF-α release and its resulting macrophage activation (Fahey *et al.* 1992).

The expression of the CCR3 receptor was strongly down-regulated, as was the expression of one of its ligands: CCL28. CCR3 is a receptor for eotaxins (of which CCL28 is an example – John *et al.* (2005)), a chemoattractant for eosinophils (Struyf *et al.* 2001). The up-regulation of these CC family chemokines points towards a macrophage-like phenotype – specifically an M1 phenotype – due to the release of proinflammatory signals and recruitment of antimicrobial eosinophils. The M2 phenotype is also hinted at, however, by the profound down-regulation of CXCR3, a receptor known to induce M2 polarization when knocked out (Oghumu *et al.* 2014). TNF-α up-regulation reinforces this suggestion, as it has been shown to induce differentiation in macrophages in an autocrine manner (Witsell *et al.* 1992), and is required for long-term macrophage survival (Lombardo *et al.* 2007). Interestingly, the THP-1 cells did not show differential expression in CCR2 and CCL2, suggesting that they are not preparing for the induction of or response to monocyte chemotaxis (Geissmann *et al.* 2003), or that they were already in that mode before induction and that the expression of those genes was constant over the process.

In addition to the increase in TNF-α expression, THP-1 cells showed a marked decrease in Fas receptor expression, the receptor responsible for Fas-induced apoptosis in the extrinsic apoptosis pathway (Wajant 2002). This result, combined with the down-regulation of Receptor activator of nuclear factor ϰ-B ligand (RANKL, another TNF-α-related apoptotic regulator – Mizukami *et al.* (2002)) suggests that the differentiating THP-1 cells are markedly down-regulating apoptosis. Differentiating macrophages are known to resist cell death by apoptosis (Rijal *et al.* 2018), so this result is in line with the general shift towards the macrophage state.

4.3 Study limitations

The data acquired for this study were incomplete in that they lacked a measurement of THP-1 cells at the 96h timepoint. In choosing the time points for our analysis,

we thus had to compromise and use the THP-1 data from 72h post-induction, which was not ideal. Having access to datasets with matched time points would have improved the reliability of our results.

The mapping rates of genes in two of the datasets (HL60 and U937) were lower than expected. RNA-seq sequence mapping rates have increased over time from their range in early experiments of 17-52% (Li *et al.* 2010) to the now standard 70-90% read mapping against the human genome (Conesa *et al.* 2016). Using a reference transcriptome as part of the mapping process, we expected to reach at least 80% read mapping in our experiments (Zhao 2014). This low mapping rate may have contributed to lower sensitivity in the differential gene expression analysis.

Differential expression analysis was successful but would have benefited from more replicates. Though the authors of `DESeq2` mention that small sample sizes are technically workable, they also note that two is the bare minimum number of samples required for any analysis at all, and recommend more wherever possible (Love *et al.* 2014). More recent meta-analysis has suggested that although `DESeq2` is best in its class at producing good results from datasets with few samples, at least six samples are the recommended minimum number required for good sensitivity in differential expression analysis with RNA-seq data (Schurch *et al.* 2016). The accuracy of this analysis could thus be improved in future by incorporating the information from some of the other time points present in the original datasets (in addition to increasing the number of experimental datasets included for each cell line).

Our study design was lacking in detailed analysis of the levels of gene expression between the cell lines. Our analysis only took the relative up- and down-regulation of differentially expressed genes into account and took only the direction of change, rather than its magnitude, into account. Using effect size (e.g., the amount of change in some gene within a cell line) would be possible only with many more replicates and multiple datasets for each line.

5 Conclusion

Our analysis suggests that HL60, U937, and THP-1 cells differentiate similarly and show comparable patterns of gene expression, with some minor differences in key areas. A core gene expression signature is associated with PMA-induced differentiation of myeloid cell lines to a macrophage state. The gene expression signature shows up-regulation of phagocytosis generally, and involves increased expression of FC-γ, integrins, and TLRs (regulators of pattern recognition) and of saposins, proteases, sulfatases, and nucleases (which are all components of the lysosome). Additionally, the trans-endothelial migration pathway was found to be up-regulated (especially in the production of MMPs, integrins, and signalling proteins that interact with endothelial cells). All cells also showed down-regulation of genes responsible for cell proliferation and growth (especially MCM proteins and Rad51 paralogs), a finding that signifies the exiting of the cell cycle associated with differentiation. These changes suggest that the induction of differentiation of these cell lines was successful and that the cells were in the process of differentiating as expected.

Differentiation of HL60 and THP-1 cells induces distinct changes in metabolism and cellular signalling. HL60 cells show down-regulation of protein production (emphasized by their marked slowing of ribosomal gene expression) and a shift towards anaerobic metabolism (specifically in their down-regulation of NADH dehydrogenase and enzymes controlling oxidative phosphorylation – behaviours associated with M1 macrophage polarization). THP-1 cells show a shift towards M1 macrophage phenotype by their increased expression of various pro-inflammatory chemokines, and also show a down-regulation of apoptosis-inducing genes.

We recommend that laboratory analysis that makes use of these cell lines is undertaken with an awareness of the aforementioned differences in expression profiles. Should analysis focus on general macrophage-like behaviours such as phagocytosis (and its associated production of lysosomes), researchers should be free to use any cells from the three lines we examined. Experiments that focus on cell metabolism in general and anaerobic metabolism in particular should be undertaken with care when using HL60 cells, and experiments which probe chemokine induced inflammation and apoptosis should be performed with care when cells of the THP-1 lineage are used. Based on these findings we expect THP-1 cells to show increased microbicidal activity due to their signs of M1 type macrophage polarization as well as some resistance to apoptosis. We also expect HL60 cells to show reduced metabolic activity and protein production after progressing through PMA-induced differentiation due to their down-regulation of oxidative phosphorylation and ribosomal gene expression pathways, respectively.

Future work

Further analysis of the gene expression differences between HL60, U937, and THP-1 cells undergoing the monocyte-to-macrophage transition should be undertaken. Specifically, we recommend an expanded experiment which incorporates multiple datasets for each cell line. With several experiments that induce differentiation in the cells using the same stimulus and over the same time period, we could compare gene expression changes in terms of magnitude of change of expression as well as direction only, as we have here. This would provide a quantitative estimate for gene

expression levels in each cell line, and thus allow for exploratory comparisons of gene expression levels across the cell lines.

Future work could also include an analysis of the transcription factors involved in regulating the observed changes in gene expression during differentiation. Additionally, we would be interested in investigating the changes in behaviour (if any) in the situation where these cell lines are induced to differentiate with various other stimuli – for example a comparison of the differentially expressed genes in U937 cells being induced with PMA versus with vitamin D3.

Finally, wet-lab experiments should be performed to confirm the changes in gene expression we observed here in live cells, and further experiments should be devised to investigate the eventual protein expression changes induced by these shifting expression patterns.

Bibliography

Adams, D. O. (1989). "Molecular Interactions in Macrophage Activation". *Immunology Today* 10.2. DOI: 10.1016/0167-5699(89)90298-3. pmid: 2665771.

Aderem, A. and D. M. Underhill (1999). "Mechanisms of Phagocytosis in Macrophages". *Annual Review of Immunology* 17.1. DOI: 10.1146/annurev.immunol.17.1.593.

Akira, S., S. Uematsu, and O. Takeuchi (2006). "Pathogen Recognition and Innate Immunity". *Cell* 124.4. DOI: 10.1016/j.cell.2006.02.015. pmid: 16497588.

Akiyama, Y., P. J. Miller, G. B. Thurman, R. H. Neubauer, C. Oliver, T. Favilla, J. A. Beman, R. K. Oldham, and H. C. Stevenson (1983). "Characterization of a Human Blood Monocyte Subset with Low Peroxidase Activity." *Journal of Clinical Investigation* 72.3. DOI: 10.1172/JCI111034.

Ali, S. A., S. K. Zaidi, C. S. Dacwag, N. Salma, D. W. Young, A. R. Shakoori, M. A. Montecino, J. B. Lian, A. J. van Wijnen, A. N. Imbalzano, G. S. Stein, and J. L. Stein (2008). "Phenotypic Transcription Factors Epigenetically Mediate Cell Growth Control". *Proceedings of the National Academy of Sciences of the United States of America* 105.18. DOI: 10.1073/pnas.0800970105. pmid: 18445650.

Ambarus, C. A., K. C. M. Santegoets, L. van Bon, M. H. Wenink, P. P. Tak, T. R. D. J. Radstake, and D. L. P. Baeten (2012). "Soluble Immune Complexes Shift the TLR-Induced Cytokine Production of Distinct Polarized Human Macrophage Subsets towards IL-10". *PLoS ONE* 7.4. DOI: 10.1371/journal.pone.0035994. pmid: 22563430.

Amiel, E., S. Nicholson-Dykstra, J. J. Walters, H. Higgs, and B. Berwin (2007). "Scavenger Receptor-A Functions in Phagocytosis of *E. Coli* by Bone Marrow Dendritic Cells". *Experimental cell research* 313.7. DOI: 10.1016/j.yexcr.2007.02.011. pmid: 17362929.

Anderson, C. L., L. Shen, D. M. Eicher, M. D. Wewers, and J. K. Gill (1990). "Phagocytosis Mediated by Three Distinct Fc Gamma Receptor Classes on Human Leukocytes". *The Journal of Experimental Medicine* 171.4. DOI: 10.1084/jem.171.4.1333. pmid: 2139103.

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Version 0.11.9. Cambridge, United Kingdom: Babraham Bioinformatics, Babraham Institute. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Armstrong, J. A. and P. D. Hart (1971). "Response of Cultured Macrophages to Mycobacterium Tuberculosis, with Observations on Fusion of Lysosomes with Phagosomes". *Journal of Experimental Medicine* 134.3. DOI: 10.1084/jem.134.3.713.

Asea, A., M. Rehli, E. Kabingu, J. A. Boch, O. Bare, P. E. Auron, M. A. Stevenson, and S. K. Calderwood (2002). "Novel Signal Transduction Pathway Utilized by Extracellular HSP70: Role of Toll-like Receptor (TLR) 2 and TLR4". *The Journal of Biological Chemistry* 277.17. DOI: 10.1074/jbc.M200497200. pmid: 11836257.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000). "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium". *Nature Genetics* 25.1. DOI: 10.1038/75556. pmid: 10802651.

ATCC (2022). *U-937 | ATCC*. CRL-1593.2 handling information. URL: https://www.atcc.org/products/crl-1593.2 (visited on 07/05/2022).

Auffray, C., D. Fogg, M. Garfa, G. Elain, O. Join-Lambert, S. Kayal, S. Sarnacki, A. Cumano, G. Lauvau, and F. Geissmann (2007). "Monitoring of Blood Vessels and Tissues by a Population of Monocytes with Patrolling Behavior". *Science* 317.5838. DOI: 10.1126/science.1142883.

Auffray, C., M. H. Sieweke, and F. Geissmann (2009). "Blood Monocytes: Development, Heterogeneity, and Relationship with Dendritic Cells". *Annual Review of Immunology* 27. DOI: 10.1146/annurev.immunol.021908.132557. pmid: 19132917.

Auwerx, J. (1991). "The Human Leukemia Cell Line, THP-1: A Multifacetted Model for the Study of Monocyte-Macrophage Differentiation". *Experientia* 47.1. DOI: 10.1007/BF02041244. pmid: 1999239.

Barrientos, S., O. Stojadinovic, M. S. Golinko, H. Brem, and M. Tomic-Canic (2008). "Growth Factors and Cytokines in Wound Healing". *Wound Repair and Regeneration: Official Publication of the Wound Healing Society [and] the European Tissue Repair Society* 16.5. DOI: 10.1111/j.1524-475X.2008.00410.x. pmid: 19128254.

Bekker, L. G., S. Freeman, P. J. Murray, B. Ryffel, and G. Kaplan (2001). "TNF-alpha Controls Intracellular Mycobacterial Growth by Both Inducible Nitric Oxide Synthase-Dependent and Inducible Nitric Oxide Synthase-Independent Pathways". *Journal of Immunology (Baltimore, Md.: 1950)* 166.11. DOI: 10.4049/jimmunol.166.11.6728. pmid: 11359829.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

Bergin, D. A., E. P. Reeves, P. Meleady, M. Henry, O. J. McElvaney, T. P. Carroll, C. Condron, S. H. Chotirmall, M. Clynes, S. J. O'Neill, and N. G. McElvaney (2010). "α-1 Antitrypsin Regulates Human Neutrophil Chemotaxis Induced by Soluble Immune Complexes and IL-8". *The Journal of Clinical Investigation* 120.12. DOI: 10.1172/JCI41196. pmid: 21060150.

Bernstein, M. (2021). *On Cell Types and Cell States*. Matthew N Bernstein. URL: https://mbernste.github.io/posts/cell_types_cell_states/ (visited on 03/15/2022).

Bertrand, J. Y., A. Jalil, M. Klaine, S. Jung, A. Cumano, and I. Godin (2005). "Three Pathways to Mature Macrophages in the Early Mouse Yolk Sac". *Blood* 106.9. DOI: 10.1182/blood-2005-02-0461. pmid: 16020514.

Beutler, B. A. (1999). "The Role of Tumor Necrosis Factor in Health and Disease". *The Journal of Rheumatology. Supplement* 57. pmid: 10328138.

Bielicki, J., C. Freeman, P. R. Clements, and J. J. Hopwood (1990). "Human Liver Iduronate-2-Sulphatase. Purification, Characterization and Catalytic Properties". *The Biochemical Journal* 271.1. DOI: 10.1042/bj2710075. pmid: 2222422.

Bone, R. C. (1991). "The Pathogenesis of Sepsis". *Annals of Internal Medicine* 115.6. DOI: 10.7326/0003-4819-115-6-457.

Bonifer, C. and D. A. Hume (2008). "The Transcriptional Regulation of the Colony-Stimulating Factor 1 Receptor (Csf1r) Gene during Hematopoiesis". *Front Biosci* 13.13.

Braciale, T. J., L. A. Morrison, M. T. Sweetser, J. Sambrook, M. J. Gething, and V. L. Braciale (1987). "Antigen Presentation Pathways to Class I and Class II

MHC-restricted T Lymphocytes". *Immunological reviews* 98. DOI: 10.1111/j.1600-065x.1987.tb00521.x. pmid: 2443444.

Brancato, S. K. and J. E. Albina (2011). "Wound Macrophages as Key Regulators of Repair: Origin, Phenotype, and Function". *The American Journal of Pathology* 178.1. DOI: 10.1016/j.ajpath.2010.08.003. pmid: 21224038.

Bray, N. L., H. Pimentel, P. Melsted, and L. Pachter (2016). "Near-Optimal Probabilistic RNA-seq Quantification". *Nature Biotechnology* 34.5. DOI: 10.1038/nbt.3519.

Brennan, J. K., C. N. Abboud, J. F. DiPersio, G. H. Barlow, and M. A. Lichtman (1981). "Autostimulation of Growth by Human Myelogenous Leukemia Cells (HL-60)". *Blood* 58.4. DOI: 10.1182/blood.V58.4.803.803.

Brix, K. (2013). *Lysosomal Proteases: Revival of the Sleeping Beauty*. Landes Bioscience. URL: https://www.ncbi.nlm.nih.gov/books/NBK6348/ (visited on 03/19/2022).

Bruckmeier, M., A. Kuehnl, M. Culmes, J. Pelisek, and H.-H. Eckstein (2012). "Impact of oxLDL and LPS on C-type Natriuretic Peptide System Is Different between THP-1 Cells and Human Peripheral Blood Monocytic Cells". *Cellular Physiology and Biochemistry: International Journal of Experimental Cellular Physiology, Biochemistry, and Pharmacology* 30.1. DOI: 10.1159/000339044. pmid: 22759967.

Brundula, V., N. B. Rewcastle, L. M. Metz, C. C. Bernard, and V. W. Yong (2002). "Targeting Leukocyte MMPs and Transmigration". *Brain* 125.6. DOI: 10.1093/brain/awf133.

Burgess, A. W. and D. Metcalf (1980). "The Nature and Action of Granulocyte-Macrophage Colony Stimulating Factors". *Blood* 56.6. pmid: 7002232.

Chanput, W., J. J. Mes, and H. J. Wichers (2014). "THP-1 Cell Line: An in Vitro Cell Model for Immune Modulation Approach". *International Immunopharmacology* 23.1. DOI: 10.1016/j.intimp.2014.08.002.

Chhangawala, S., G. Rudy, C. E. Mason, and J. A. Rosenfeld (2015). "The Impact of Read Length on Quantification of Differentially Expressed Genes and Splice Junction Detection". *Genome Biology* 16.1. DOI: 10.1186/s13059-015-0697-y.

Chiao, J. W., J. Bressler, C. Pinksy, Y. Hirshaut, H. Oettgen, and B. Clarkson (1982). "Induction of Differentiation of HL-60 Leukemic-Cells by Post-Endotoxin Serum". In *Proceedings of the American Association for Cancer Research*. Vol. 23. American association for cancer research, pp. 226–226. ISBN: 0197-016X.

Cohn, Z. A. and J. G. Hirsch (1960). "The Isolation and Properties of the Specific Cytoplasmic Granules of Rabbit Polymorphonuclear Leucocytes". *The Journal of Experimental Medicine* 112. DOI: 10.1084/jem.112.6.983. pmid: 13694490.

Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi (2016). "A Survey of Best Practices for RNA-seq Data Analysis". *Genome Biology* 17.1. DOI: 10.1186/s13059-016-0881-8.

Crane, M. J., J. M. Daley, O. van Houtte, S. K. Brancato, W. L. Henry, and J. E. Albina (2014). "The Monocyte to Macrophage Transition in the Murine Sterile Wound". *PLoS ONE* 9.1. DOI: 10.1371/journal.pone.0086660. pmid: 24466192.

Daigneault, M., J. A. Preston, H. M. Marriott, M. K. B. Whyte, and D. H. Dockrell (2010). "The Identification of Markers of Macrophage Differentiation in PMA-Stimulated THP-1 Cells and Monocyte-Derived Macrophages". *PLOS ONE* 5.1. DOI: 10.1371/journal.pone.0008668.

Daley, J. M., S. K. Brancato, A. A. Thomay, J. S. Reichner, and J. E. Albina (2010). "The Phenotype of Murine Wound Macrophages". *Journal of Leukocyte Biology* 87.1. DOI: 10.1189/jlb.0409236. pmid: 20052800.

Darmoise, A., P. Maschmeyer, and F. Winau (2010). "The Immunological Functions of Saposins". *Advances in immunology* 105. DOI: 10.1016/S0065-2776(10)05002-9. pmid: 20510729.

De la Roche, M., Y. Asano, and G. M. Griffiths (2016). "Origins of the Cytolytic Synapse". *Nature Reviews Immunology* 16.7 (7). DOI: 10.1038/nri.2016.54.

Dekkers, K. F., A. E. Neele, J. W. Jukema, B. T. Heijmans, and M. P. J. de Winther (2019). "Human Monocyte-to-Macrophage Differentiation Involves Highly Localized Gain and Loss of DNA Methylation at Transcription Factor Binding Sites". *Epigenetics & Chromatin* 12.1. DOI: 10.1186/s13072-019-0279-4.

Desjardins, M. (1995). "Biogenesis of Phagolysosomes: The 'kiss and Run' Hypothesis". *Trends in Cell Biology* 5.5. DOI: 10.1016/s0962-8924(00)88989-8. pmid: 14731444.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras (2013). "STAR: Ultrafast Universal RNA-seq Aligner". *Bioinformatics* 29.1. DOI: 10.1093/bioinformatics/bts635.

Douglas, S. D. and F. Tuluc (2010). "Chapter 67. Morphology of Monocytes and Macrophages". *Williams Hematology*. Ed. by M. A. Lichtman, T. J. Kipps, U. Seligsohn, K. Kaushansky, and J. T. Prchal. 8th ed. New York, NY: The McGraw-Hill Companies. URL: accessmedicine.mhmedical.com/content.aspx?aid=6112762 (visited on 04/14/2020).

Durinck, S., Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber (2005). "BioMart and Bioconductor: A Powerful Link between Biological Databases and Microarray Data Analysis". *Bioinformatics (Oxford, England)* 21.16. DOI: 10.1093/bioinformatics/bti525. pmid: 16082012.

Durinck, S., P. T. Spellman, E. Birney, and W. Huber (2009). "Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor Package biomaRt". *Nature Protocols* 4.8. DOI: 10.1038/nprot.2009.97. pmid: 19617889.

Ehrlich, P. (1904). *Gesammelte Arbeiten Zur Immunitätsforschung*. Hirschwald.

EMBL-EBI (2021). The European Bioinformatics Institute: EMBL-EBI. URL: https://www.ebi.ac.uk/ (visited on 12/14/2021).

Esposti, M. (1998). "Inhibitors of NADH–Ubiquinone Reductase: An Overview". *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1364.2. DOI: 10.1016/S0005-2728(98)00029-2.

Fahey, T. J., K. J. Tracey, P. Tekamp-Olson, L. S. Cousens, W. G. Jones, G. T. Shires, A. Cerami, and B. Sherry (1992). "Macrophage Inflammatory Protein 1 Modulates Macrophage Function". *Journal of Immunology (Baltimore, Md.: 1950)* 148.9. pmid: 1573267.

Fischer, H. G., S. Frosch, K. Reske, and A. B. Reske-Kunz (1988). "Granulocyte-Macrophage Colony-Stimulating Factor Activates Macrophages Derived from Bone Marrow Cultures to Synthesis of MHC Class II Molecules and to Augmented Antigen Presentation Function." *The Journal of Immunology* 141.11.

Fogg, D. K., C. Sibon, C. Miled, S. Jung, P. Aucouturier, D. R. Littman, A. Cumano, and F. Geissmann (2006). "A Clonogenic Bone Marrow Progenitor Specific for Macrophages and Dendritic Cells". *Science (New York, N.Y.)* 311.5757. DOI: 10.1126/science.1117729. pmid: 16322423.

Freeman, C. and J. J. Hopwood (1987). "Human Liver N-acetylglucosamine-6-sulphate Sulphatase. Catalytic Properties". *The Biochemical Journal* 246.2. DOI: 10.1042/bj2460355. pmid: 3689315.

Furth, R. van and Z. A. Cohn (1968). "The Origin and Kinetics of Mononuclear Phagocytes". *The Journal of Experimental Medicine* 128.3. DOI: 10.1084/jem.128.3.415. pmid: 5666958.

Gallagher, R., S. Collins, J. Trujillo, K. McCredie, M. Ahearn, S. Tsai, R. Metzgar, G. Aulakh, R. Ting, F. Ruscetti, and R. Gallo (1979). "Characterization of the Continuous, Differentiating Myeloid Cell Line (HL-60) from a Patient with Acute Promyelocytic Leukemia". *Blood* 54.3. pmid: 288488.

Galli, S. J., N. Borregaard, and T. A. Wynn (2011). "Phenotypic and Functional Plasticity of Cells of Innate Immunity: Macrophages, Mast Cells and Neutrophils". *Nature Immunology* 12.11. DOI: 10.1038/ni.2109.

García, A., A. Serrano, E. Abril, P. Jimenez, L. M. Real, J. Cantón, F. Garrido, and F. Ruiz-Cabello (1999). "Differential Effect on U937 Cell Differentiation by Targeting Transcriptional Factors Implicated in Tissue- or Stage-Specific Induced Integrin Expression". *Experimental Hematology* 27.2. DOI: 10.1016/s0301-472x(98)00038-1. pmid: 10029175.

Gately, C. L., S. M. Wahl, and J. J. Oppenheim (1983). "Characterization of Hydrogen Peroxide-Potentiating Factor, a Lymphokine That Increases the Capacity of Human Monocytes and Monocyte-like Cell Lines to Produce Hydrogen Peroxide". *Journal of Immunology (Baltimore, Md.: 1950)* 131.6. pmid: 6438232.

Gaudino, S. J. and P. Kumar (2019). "Cross-Talk Between Antigen Presenting Cells and T Cells Impacts Intestinal Homeostasis, Bacterial Infections, and Tumorigenesis". *Frontiers in Immunology* 10. URL: https://www.frontiersin.org/article/10.3389/fimmu.2019.00360 (visited on 03/16/2022).

Gautier, E. L., T. Shay, J. Miller, M. Greter, C. Jakubzick, S. Ivanov, J. Helft, A. Chow, K. G. Elpek, S. Gordonov, A. R. Mazloom, A. Ma'ayan, W.-J. Chua, T. H. Hansen, S. J. Turley, M. Merad, and G. J. Randolph (2012). "Gene Expression Profiles and Transcriptional Regulatory Pathways Underlying Mouse Tissue Macrophage Identity and Diversity". *Nature immunology* 13.11. DOI: 10.1038/ni.2419. pmid: 23023392.

Geissmann, F., S. Jung, and D. R. Littman (2003). "Blood Monocytes Consist of Two Principal Subsets with Distinct Migratory Properties". *Immunity* 19.1. DOI: 10.1016/s1074-7613(03)00174-2. pmid: 12871640.

Geissmann, F., M. G. Manz, S. Jung, M. H. Sieweke, M. Merad, and K. Ley (2010). "Development of Monocytes, Macrophages and Dendritic Cells". *Science (New York, N.Y.)* 327.5966. DOI: 10.1126/science.1178331. pmid: 20133564.

Gerhardt, T. and K. Ley (2015). "Monocyte Trafficking across the Vessel Wall". *Cardiovascular Research* 107.3. DOI: 10.1093/cvr/cvv147.

Ginhoux, F. and S. Jung (2014). "Monocytes and Macrophages: Developmental Pathways and Tissue Homeostasis". *Nature Reviews Immunology* 14.6. DOI: 10.1038/nri3671.

Godwin, J. W., A. R. Pinto, and N. A. Rosenthal (2013). "Macrophages Are Required for Adult Salamander Limb Regeneration". *Proceedings of the National Academy of Sciences of the United States of America* 110.23. DOI: 10.1073/pnas.1300290110. pmid: 23690624.

Gordon, S. (2003). "Alternative Activation of Macrophages". *Nature Reviews Immunology* 3.1 (1). DOI: 10.1038/nri978.

Gordon, S. (2016). "Phagocytosis: The Legacy of Metchnikoff". *Cell* 166.5. DOI: 10.1016/j.cell.2016.08.017.

Gordon, S. and F. O. Martinez (2010). "Alternative Activation of Macrophages: Mechanism and Functions". *Immunity* 32.5. DOI: 10.1016/j.immuni.2010.05.007. pmid: 20510870.

Gordon, S. and P. R. Taylor (2005). "Monocyte and Macrophage Heterogeneity". *Nature Reviews Immunology* 5.12 (12). DOI: 10.1038/nri1733.

GRCh38.P13 Genome Assembly - NCBI (2019). URL: https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39 (visited on 12/06/2021).

Greenhalgh, D. G. (1998). "The Role of Apoptosis in Wound Healing". *The International Journal of Biochemistry & Cell Biology* 30.9. DOI: 10.1016/s1357-2725(98)00058-2. pmid: 9785465.

Griffith, M., J. R. Walker, N. C. Spies, B. J. Ainscough, and O. L. Griffith (2015). "Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud". *PLoS computational biology* 11.8. DOI: 10.1371/journal.pcbi.1004393. pmid: 26248053.

Guermonprez, P., J. Valladeau, L. Zitvogel, C. Théry, and S. Amigorena (2002). "Antigen Presentation and T Cell Stimulation by Dendritic Cells". *Annual Review of Immunology* 20.1. DOI: 10.1146/annurev.immunol.20.100301.064828.

Guha, M. and N. Mackman (2001). "LPS Induction of Gene Expression in Human Monocytes". *Cellular Signalling* 13.2. DOI: 10.1016/s0898-6568(00)00149-2. pmid: 11257452.

Guyre, P. M., P. M. Morganelli, and R. Miller (1983). "Recombinant Immune Interferon Increases Immunoglobulin G Fc Receptors on Cultured Human Mononuclear Phagocytes". *The Journal of Clinical Investigation* 72.1. DOI: 10.1172/jci110980. pmid: 6192145.

Harding, C. V. and H. J. Geuze (1992). "Class II MHC Molecules Are Present in Macrophage Lysosomes and Phagolysosomes That Function in the Phagocytic Processing of Listeria Monocytogenes for Presentation to T Cells." *Journal of Cell Biology* 119.3. DOI: 10.1083/jcb.119.3.531.

Hashimoto, D. *et al.* (2013). "Tissue-Resident Macrophages Self-Maintain Locally throughout Adult Life with Minimal Contribution from Circulating Monocytes". *Immunity* 38.4. DOI: 10.1016/j.immuni.2013.04.004. pmid: 23601688.

Herbomel, P., B. Thisse, and C. Thisse (1999). "Ontogeny and Behaviour of Early Macrophages in the Zebrafish Embryo". *Development (Cambridge, England)* 126.17. DOI: 10.1242/dev.126.17.3735. pmid: 10433904.

Hopkinson-Woolley, J., D. Hughes, S. Gordon, and P. Martin (1994). "Macrophage Recruitment during Limb Development and Wound Healing in the Embryonic and Foetal Mouse". *Journal of Cell Science* 107 (Pt 5). pmid: 7929625.

Hsieh, C. S., S. E. Macatonia, C. S. Tripp, S. F. Wolf, A. O'Garra, and K. M. Murphy (1993). "Development of TH1 CD4+ T Cells through IL-12 Produced by Listeria-induced Macrophages". *Science (New York, N.Y.)* 260.5107. DOI: 10.1126/science.8097338. pmid: 8097338.

Issekutz, A. C., D. Rowter, and T. A. Springer (1999). "Role of ICAM-1 and ICAM-2 and Alternate CD11/CD18 Ligands in Neutrophil Transendothelial Migration". *Journal of Leukocyte Biology* 65.1. DOI: 10.1002/jlb.65.1.117. pmid: 9886254.

Jain, C., S. Koren, A. Dilthey, A. M. Phillippy, and S. Aluru (2018). "A Fast Adaptive Algorithm for Computing Whole-Genome Homology Maps". *Bioinformatics* 34.17. DOI: 10.1093/bioinformatics/bty597.

Janeway, C. A. and R. Medzhitov (2002). "Innate Immune Recognition". *Annual Review of Immunology* 20.1. DOI: 10.1146/annurev.immunol.20.083001.084359.

Janeway, C. A., P. Travers, M. Walport, and M. J. Shlomchik (2001). "Macrophage Activation by Armed CD4 TH1 Cells". *Immunobiology: The Immune System in Health and Disease. 5th edition.* URL: https://www.ncbi.nlm.nih.gov/books/NBK27153/ (visited on 03/17/2022).

Jawhara, S., E. Pluskota, W. Cao, E. F. Plow, and D. A. Soloviev (2017). "Distinct Effects of Integrins αXβ2 and αMβ2 on Leukocyte Subpopulations during Inflammation and Antimicrobial Responses". *Infection and Immunity* 85.1. DOI: 10.1128/IAI.00644-16. pmid: 27799334.

John, A. E., M. S. Thomas, A. A. Berlin, and N. W. Lukacs (2005). "Temporal Production of CCL28 Corresponds to Eosinophil Accumulation and Airway Hyperreactivity in Allergic Airway Inflammation". *The American Journal of Pathology* 166.2. pmid: 15681819. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1602329/ (visited on 03/22/2022).

Kanehisa, M. and S. Goto (2000). "KEGG: Kyoto Encyclopedia of Genes and Genomes". *Nucleic Acids Research* 28.1. DOI: 10.1093/nar/28.1.27. pmid: 10592173.

Kanehisa, M. (2019). "Toward Understanding the Origin and Evolution of Cellular Organisms". *Protein Science: A Publication of the Protein Society* 28.11. DOI: 10.1002/pro.3715. pmid: 31441146.

Katikaneni, D. S. and L. Jina (2019). "B Cell MHC Class II Signaling: A Story of Life and Death". *Human immunology* 80.1. DOI: 10.1016/j.humimm.2018.04.013. pmid: 29715484.

Kelly, L. M., U. Englmeier, I. Lafon, M. H. Sieweke, and T. Graf (2000). "MafB Is an Inducer of Monocytic Differentiation". *The EMBO journal* 19.9. DOI: 10.1093/emboj/19.9.1987. pmid: 10790365.

Khazen, W., J.-P. M'bika, C. Tomkiewicz, C. Benelli, C. Chany, A. Achour, and C. Forest (2005). "Expression of Macrophage-Selective Markers in Human and Rodent Adipocytes". *FEBS letters* 579.25. DOI: 10.1016/j.febslet.2005.09.032. pmid: 16213494.

Kinchen, J. M. and K. S. Ravichandran (2008). "Phagosome Maturation: Going through the Acid Test". *Nature reviews. Molecular cell biology* 9.10. DOI: 10.1038/nrm2515. pmid: 18813294.

Kingsford, C. (2016). "Accurate, Fast, and Model-Aware Transcript Expression Quantification with Salmon | Simons Institute for the Theory of Computing (Accessed Online at https://www.youtube.com/watch?v=TMLIxwDP7sk)". Symposium. Accurate, Fast, and Model-Aware Transcript Expression Quantification with Salmon (Calvin Lab, Berkeley). URL: https://simons.berkeley.edu/talks/carl-kingsford-03-07-16.

Koeffler, H. P. (1983). "Induction of Differentiation of Human Acute Myelogenous Leukemia Cells: Therapeutic Implications". *Blood* 62.4. pmid: 6192859.

Krieg, A. M. (2000). "The Role of CpG Motifs in Innate Immunity". *Current Opinion in Immunology* 12.1. DOI: 10.1016/s0952-7915(99)00048-5. pmid: 10679406.

Külahoglu, C. and A. Bräutigam (2014). "Quantitative Transcriptome Analysis Using RNA-seq". *Methods in Molecular Biology (Clifton, N.J.)* 1158. DOI: 10.1007/978-1-4939-0700-7_5. pmid: 24792045.

Lafontaine, D. L. and D. Tollervey (2001). "The Function and Synthesis of Ribosomes". *Nature Reviews. Molecular Cell Biology* 2.7. DOI: 10.1038/35080045. pmid: 11433365.

Lee, J. Y. and S.-H. Hong (2020). "Hematopoietic Stem Cells and Their Roles in Tissue Regeneration". *International Journal of Stem Cells* 13.1. DOI: 10.15283/ijsc19127. pmid: 31887851.

Li, B. and C. N. Dewey (2011). "RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome". *BMC Bioinformatics* 12.1. DOI: 10.1186/1471-2105-12-323.

Li, B., V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey (2010). "RNA-Seq Gene Expression Estimation with Read Mapping Uncertainty". *Bioinformatics* 26.4. DOI: 10.1093/bioinformatics/btp692. pmid: 20022975.

Liu, H.-Z., J.-P. Gong, C.-X. Wu, Y. Peng, X.-H. Li, and H.-B. You (2005). "The U937 Cell Line Induced to Express CD14 Protein by 1,25-Dihydroxyvitamin D3 and Be Sensitive to Endotoxin Stimulation". *Hepatobiliary & pancreatic diseases international: HBPD INT* 4.1. pmid: 15730927.

Liu, H., P. A. Lorenzini, F. Zhang, S. Xu, M. S. M. Wong, J. Zheng, and X. Roca (2018). "Alternative Splicing Analysis in Human Monocytes and Macrophages Reveals MBNL1 as Major Regulator". *Nucleic Acids Research* 46.12. DOI: 10.1093/nar/gky401.

Lombardo, E., A. Alvarez-Barrientos, B. Maroto, L. Boscá, and U. G. Knaus (2007). "TLR4-mediated Survival of Macrophages Is MyD88 Dependent and Requires TNF-alpha Autocrine Signalling". *Journal of Immunology (Baltimore, Md.: 1950)* 178.6. DOI: 10.4049/jimmunol.178.6.3731. pmid: 17339471.

Love, M. I., W. Huber, and S. Anders (2014). "Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2". *Genome Biology* 15.12. DOI: 10.1186/s13059-014-0550-8. pmid: 25516281.

Love, M. I., C. Soneson, P. F. Hickey, L. K. Johnson, N. T. Pierce, L. Shepherd, M. Morgan, and R. Patro (2019). "Tximeta: Reference Sequence Checksums for Provenance Identification in RNA-seq". *bioRxiv*. DOI: 10.1101/777888.

Luo, W., G. Pant, Y. K. Bhavnasi, S. G. Blanchard Jr, and C. Brouwer (2017). "Pathview Web: User Friendly Pathway Visualization and Data Integration". *Nucleic Acids Research* 45.W1. DOI: 10.1093/nar/gkx372.

Madison, K. C. (2003). "Barrier Function of the Skin: "La Raison d'Être" of the Epidermis". *Journal of Investigative Dermatology* 121.2. DOI: 10.1046/j.1523-1747.2003.12359.x. pmid: 12880413.

Mantovani, A., S. Sozzani, M. Locati, P. Allavena, and A. Sica (2002). "Macrophage Polarization: Tumor-Associated Macrophages as a Paradigm for Polarized M2 Mononuclear Phagocytes". *Trends in Immunology* 23.11. DOI: 10.1016/S1471-4906(02)02302-5.

Mantovani, A., A. Sica, S. Sozzani, P. Allavena, A. Vecchi, and M. Locati (2004). "The Chemokine System in Diverse Forms of Macrophage Activation and Polarization". *Trends in Immunology* 25.12. DOI: 10.1016/j.it.2004.09.015. pmid: 15530839.

Martinez, F. O., S. Gordon, M. Locati, and A. Mantovani (2006). "Transcriptional Profiling of the Human Monocyte-to-Macrophage Differentiation and Polarization: New Molecules and Patterns of Gene Expression". *Journal of Immunology (Baltimore, Md.: 1950)* 177.10. DOI: 10.4049/jimmunol.177.10.7303. pmid: 17082649.

Martinez, F. O., A. Sica, A. Mantovani, and M. Locati (2008). "Macrophage Activation and Polarization". *Frontiers in Bioscience-Landmark* 13.2 (2). DOI: 10.2741/2692.

Martinez-Pomares, L., N. Platt, A. J. McKnight, R. P. da Silva, and S. Gordon (1996). "Macrophage Membrane Molecules: Markers of Tissue Differentiation and Heterogeneity". *Immunobiology* 195.4-5. DOI: 10.1016/S0171-2985(96)80012-X. pmid: 8933147.

Matzinger, P. (1994). "Tolerance, Danger, and the Extended Family". *Annual Review of Immunology* 12. DOI: 10.1146/annurev.iy.12.040194.005015. pmid: 8011301.

Medzhitov, R. (2007). "Recognition of Microorganisms and Activation of the Immune Response". *Nature* 449.7164 (7164). DOI: 10.1038/nature06246.

Meerschaert, J. and M. B. Furie (1995). "The Adhesion Molecules Used by Monocytes for Migration across Endothelium Include CD11a/CD18, CD11b/CD18, and VLA-4 on Monocytes and ICAM-1, VCAM-1, and Other Ligands on Endothelium". *Journal of Immunology (Baltimore, Md.: 1950)* 154.8. pmid: 7535821.

Menten, P., A. Wuyts, and J. Van Damme (2002). "Macrophage Inflammatory Protein-1". *Cytokine & Growth Factor Reviews* 13.6. DOI: 10.1016/s1359-6101(02)00045-x. pmid: 12401480.

Merle, N. S., R. Noe, L. Halbwachs-Mecarelli, V. Fremeaux-Bacchi, and L. T. Roumenina (2015). "Complement System Part II: Role in Immunity". *Frontiers in Immunology* 6. URL: https://www.frontiersin.org/article/10.3389/fimmu.2015.00257 (visited on 03/09/2022).

Mills, C. (2012). "M1 and M2 Macrophages: Oracles of Health and Disease". *Critical Reviews in Immunology* 32.6. DOI: 10.1615/CritRevImmunol.v32.i6.10.

Mills, K. I., L. J. Woodgate, A. F. Gilkes, V. Walsh, M. C. Sweeney, G. Brown, and A. K. Burnett (1999). "Inhibition of Mitochondrial Function in HL60 Cells Is Associated with an Increased Apoptosis and Expression of CD14". *Biochemical and Biophysical Research Communications* 263.2. DOI: 10.1006/bbrc.1999.1356. pmid: 10491287.

Mizukami, J., G. Takaesu, H. Akatsuka, H. Sakurai, J. Ninomiya-Tsuji, K. Matsumoto, and N. Sakurai (2002). "Receptor Activator of NF-κB Ligand (RANKL) Activates TAK1 Mitogen-Activated Protein Kinase Kinase Kinase through a Signaling Complex Containing RANK, TAB2, and TRAF6". *Molecular and Cellular Biology* 22.4. DOI: 10.1128/MCB.22.4.992-1000.2002. pmid: 11809792.

Moser, B., M. Loetscher, L. Piali, and P. Loetscher (1998). "Lymphocyte Responses to Chemokines". *International Reviews of Immunology* 16.3-4. DOI: 10.3109/08830189809043000.

Mosser, D. M. and J. P. Edwards (2008). "Exploring the Full Spectrum of Macrophage Activation". *Nature reviews. Immunology* 8.12. DOI: 10.1038/nri2448. pmid: 19029990.

Murao, S.-i., M. A. Gemmell, M. F. Callaham, N. L. Anderson, and E. Huberman (1983). "Control of Macrophage Cell Differentiation in Human Promyelocytic HL-60 Leukemia Cells by 1,25-Dihydroxyvitamin D3 and Phorbol-12-myristate-

13-acetate". *Cancer Research* 43.10. pmid: 6576856. URL: https://cancerres.aacrjournals.org/content/43/10/4989 (visited on 12/14/2021).

Murphy, K, C. A. J. Jr, P. Travers, M. Walport, and M. J. Shlomchik (2012). *Immunobiology*. 8th ed. Garland Science. ISBN: 978-0-8153-4243-4.

Neben, C. L., F. D. Lay, X. Mao, C. T. Tuzon, and A. E. Merrill (2017). "Ribosome Biogenesis Is Dynamically Regulated during Osteoblast Differentiation". *Gene* 612. DOI: 10.1016/j.gene.2016.11.010. pmid: 27847259.

Neefjes, J., M. L. M. Jongsma, P. Paul, and O. Bakke (2011). "Towards a Systems Understanding of MHC Class I and MHC Class II Antigen Presentation". *Nature Reviews Immunology* 11.12 (12). DOI: 10.1038/nri3084.

Nilsson, K., K. Forsbeck, M. Gidlund, C. Sundström, T. Tötterman, J. Sällström, and P. Venge (1981). "Surface Characteristics of the U-937 Human Histiocytic Lymphoma Cell Line: Specific Changes During Inducible Morphologic and Functional Differentiation In Vitro". *Modern Trends in Human Leukemia IV*. Ed. by R. Neth, R. C. Gallo, T. Graf, K. Mannweiler, and K. Winkler. Red. by H. Heimpel, D. Huhn, C. Mueller-Eckhardt, and G. Ruhenstroth-Bauer. Vol. 26. Haematology and Blood Transfusion / Hämatologie Und Bluttransfusion. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 215–221. ISBN: 978-3-540-10622-7 978-3-642-67984-1. DOI: 10.1007/978-3-642-67984-1_35.

O'Garra, A. and P. Vieira (2007). "T(H)1 Cells Control Themselves by Producing Interleukin-10". *Nature Reviews. Immunology* 7.6. DOI: 10.1038/nri2097. pmid: 17525751.

O'Grady, N. P., M. Tropea, H. L. Preas, D. Reda, R. W. Vandivier, S. M. Banks, and A. F. Suffredini (1999). "Detection of Macrophage Inflammatory Protein (MIP)-1alpha and MIP-1beta during Experimental Endotoxemia and Human Sepsis". *The Journal of Infectious Diseases* 179.1. DOI: 10.1086/314559. pmid: 9841832.

O'Neill, L. A. J. and D. G. Hardie (2013). "Metabolism of Inflammation Limited by AMPK and Pseudo-Starvation". *Nature* 493.7432. DOI: 10.1038/nature11862. pmid: 23325217.

Oghumu, S., S. Varikuti, C. Terrazas, D. Kotov, M. W. Nasser, C. A. Powell, R. K. Ganju, and A. R. Satoskar (2014). "CXCR3 Deficiency Enhances Tumor Progression by Promoting Macrophage M2 Polarization in a Murine Breast Cancer Model". *Immunology* 143.1. DOI: 10.1111/imm.12293.

Ovchinnikov, D. A. (2008). "Macrophages in the Embryo and beyond: Much More than Just Giant Phagocytes". *genesis* 46.9. DOI: 10.1002/dvg.20417.

Ozinsky, A., D. M. Underhill, J. D. Fontenot, A. M. Hajjar, K. D. Smith, C. B. Wilson, L. Schroeder, and A. Aderem (2000). "The Repertoire for Pattern Recognition of Pathogens by the Innate Immune System Is Defined by Cooperation between Toll-like Receptors". *Proceedings of the National Academy of Sciences of the United States of America* 97.25. DOI: 10.1073/pnas.250476497. pmid: 11095740.

Parameswaran, N. and S. Patial (2010). "Tumor Necrosis Factor-α Signaling in Macrophages". *Critical reviews in eukaryotic gene expression* 20.2. pmid: 21133840. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3066460/ (visited on 03/24/2022).

Parker, D. C. (1993). "T Cell-Dependent B Cell Activation". *Annual Review of Immunology* 11. DOI: 10.1146/annurev.iy.11.040193.001555. pmid: 8476565.

Passlick, B., D. Flieger, and H. Ziegler-Heitbrock (1989). “Identification and Characterization of a Novel Monocyte Subpopulation in Human Peripheral Blood”. *Blood* 74.7. DOI: 10.1182/blood.V74.7.2527.2527.

Patro, R., G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford (2017). “Salmon: Fast and Bias-Aware Quantification of Transcript Expression Using Dual-Phase Inference”. *Nature methods* 14.4. DOI: 10.1038/nmeth.4197. pmid: 28263959.

Paul, W. E. and J. Zhu (2010). “How Are TH2-type Immune Responses Initiated and Amplified?” *Nature reviews. Immunology* 10.4. DOI: 10.1038/nri2735. pmid: 20336151.

Petibon, C., M. Malik Ghulam, M. Catala, and S. Abou Elela (2021). “Regulation of Ribosomal Protein Genes: An Ordered Anarchy”. *WIREs RNA* 12.3. DOI: 10.1002/wrna.1632.

Piccinini, A. M. and K. S. Midwood (2010). “DAMPening Inflammation by Modulating TLR Signalling”. *Mediators of Inflammation* 2010. DOI: 10.1155/2010/672395. pmid: 20706656.

Pisetsky, D. S. (1996). “Immune Activation by Bacterial DNA: A New Genetic Code”. *Immunity* 5.4. DOI: 10.1016/S1074-7613(00)80256-3. pmid: 8885863.

Pober, J. S. and R. S. Cotran (1990). “Cytokines and Endothelial Cell Biology”. *Physiological Reviews* 70.2. DOI: 10.1152/physrev.1990.70.2.427.

Poznyak, A. V., N. G. Nikiforov, A. M. Markin, D. A. Kashirskikh, V. A. Myasoedova, E. V. Gerasimova, and A. N. Orekhov (2021). “Overview of OxLDL and Its Impact on Cardiovascular Health: Focus on Atherosclerosis”. *Frontiers in Pharmacology* 11. URL: https://www.frontiersin.org/article/10.3389/fphar.2020.613780 (visited on 03/21/2022).

Prehn, J. L., D. L. Fagan, S. C. Jordan, and J. S. Adams (1992). “Potentiation of Lipopolysaccharide-Induced Tumor Necrosis Factor-Alpha Expression by 1,25-Dihydroxyvitamin D3”. *Blood* 80.11. pmid: 1450407.

Rabinovitch, M. and D. Andreucci (1949). “A Histochemical Study of "Acid" and "Alkaline" Phosphatase Distribution in Normal Human Bone Marrow Smears”. *Blood* 4.5. DOI: 10.1182/blood.V4.5.580.580.

Ralph, P., M. A. Moore, and K. Nilsson (1976). “Lysozyme Synthesis by Established Human and Murine Histiocytic Lymphoma Cell Lines”. *The Journal of Experimental Medicine* 143.6. DOI: 10.1084/jem.143.6.1528. pmid: 1083890.

Ramirez, R. N., N. C. El-Ali, M. A. Mager, D. Wyman, A. Conesa, and A. Mortazavi (2017). “Dynamic Gene Regulatory Networks of Human Myeloid Differentiation”. *Cell Systems* 4.4. DOI: 10.1016/j.cels.2017.03.005. pmid: 28365152.

Rebuck, J. W. (1952). “Enhanced Nadi Reaction in Lymphocytes in Inflammation”. In *Federation Proceedings*. Vol. 11. 1. FEDERATION AMER SOC EXP BIOL 9650 Rockville Pike, Bethesda, MD 20814-3998, pp. 424–425. ISBN: 0014-9446.

Refgenie Server - Hg38 Salmon Index (2021). URL: http://refgenomes.databio.org/v3/assets/splash/2230c535660fb4774114bfa966a62f823fdb6d21acf138d4/salmon_partial_sa_index?tag=default (visited on 12/06/2021).

Rijal, D., A. Ariana, A. Wight, K. Kim, N. A. Alturki, Z. Aamir, E. S. Ametepe, R. G. Korneluk, C. Tiedje, M. B. Menon, M. Gaestel, S. McComb, and S. Sad (2018). “Differentiated Macrophages Acquire a Pro-Inflammatory and Cell Death–Resistant Phenotype Due to Increasing XIAP and P38-Mediated Inhibition of RipK1”. *Journal of Biological Chemistry* 293.30. DOI: 10.1074/jbc.RA118.003614.

Rizzo, G., E. Vafadarnejad, P. Arampatzi, J.-S. Silvestre, A. Zernecke, A.-E. Saliba, and C. Cochain (2020). *Single-Cell Transcriptomic Profiling Maps Monocyte/Macrophage Transitions after Myocardial Infarction in Mice*. DOI: 10.1101/2020.04.14.040451.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth (2010). "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data". *Bioinformatics* 26.1. DOI: 10.1093/bioinformatics/btp616. pmid: 19910308.

Robinson, M. D. and G. K. Smyth (2007). "Moderated Statistical Tests for Assessing Differences in Tag Abundance". *Bioinformatics* 23.21. DOI: 10.1093/bioinformatics/btm453.

Rodríguez-Prados, J.-C., P. G. Través, J. Cuenca, D. Rico, J. Aragonés, P. Martín-Sanz, M. Cascante, and L. Boscá (2010). "Substrate Fate in Activated Macrophages: A Comparison between Innate, Classic, and Alternative Activation". *Journal of Immunology (Baltimore, Md.: 1950)* 185.1. DOI: 10.4049/jimmunol.0901698. pmid: 20498354.

Rossetti, C., E. Picardi, M. Ye, G. Camilli, A. M. D'Erchia, L. Cucina, F. Locatelli, L. Fianchi, L. Teofili, G. Pesole, A. Gallo, and R. Sorrentino (2017). "RNA Editing Signature during Myeloid Leukemia Cell Differentiation". *Leukemia* 31.12 (12). DOI: 10.1038/leu.2017.134.

Rőszer, T. (2015). "Understanding the Mysterious M2 Macrophage through Activation Markers and Effector Mechanisms". *Mediators of Inflammation* 2015. DOI: 10.1155/2015/816460. pmid: 26089604.

Roth, S. J., M. W. Carr, and T. A. Springer (1995). "C-C Chemokines, but Not the C-X-C Chemokines Interleukin-8 and Interferon-γ Inducible Protein-10, Stimulate Transendothelial Chemotaxis of T Lymphocytes". *European Journal of Immunology* 25.12. DOI: 10.1002/eji.1830251241.

Rovera, G., T. G. O'Brien, and L. Diamond (1979). "Induction of Differentiation in Human Promyelocytic Leukemia Cells by Tumor Promoters". *Science* 204.4395. DOI: 10.1126/science.286421.

Saeed, S. *et al.* (2014). "Epigenetic Programming of Monocyte-to-Macrophage Differentiation and Trained Innate Immunity". *Science* 345.6204. DOI: 10.1126/science.1251086.

Sarantopoulou, D., T. G. Brooks, S. Nayak, A. Mrčela, N. F. Lahens, and G. R. Grant (2021). "Comparative Evaluation of Full-Length Isoform Quantification from RNA-Seq". *BMC Bioinformatics* 22.1. DOI: 10.1186/s12859-021-04198-1.

Schildberger, A., E. Rossmanith, T. Eichhorn, K. Strassl, and V. Weber (2013). "Monocytes, Peripheral Blood Mononuclear Cells, and THP-1 Cells Exhibit Different Cytokine Expression Patterns Following Stimulation with Lipopolysaccharide". *Mediators of Inflammation* 2013. DOI: 10.1155/2013/697972. pmid: 23818743.

Schimmel, L., N. Heemskerk, and J. D. van Buul (2017). "Leukocyte Transendothelial Migration: A Local Affair". *Small GTPases* 8.1. DOI: 10.1080/21541248.2016.1197872. pmid: 27715453.

Schmidt, E. V. (1999). "The Role of C-Myc in Cellular Growth Control". *Oncogene* 18.19 (19). DOI: 10.1038/sj.onc.1202751.

Schneider, V. A. *et al.* (2016). "Evaluation of GRCh38 and de Novo Haploid Genome Assemblies Demonstrates the Enduring Quality of the Reference Assembly". DOI: 10.1101/072116.

Schröder, J. M. and J. Harder (2006). "Antimicrobial Skin Peptides and Proteins". *Cellular and Molecular Life Sciences* 63.4. DOI: 10.1007/s00018-005-5364-0.

Schurch, N. J., P. Schofield, M. Gierliński, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. G. Simpson, T. Owen-Hughes, M. Blaxter, and G. J. Barton (2016). "How Many Biological Replicates Are Needed in an RNA-seq Experiment and Which Differential Expression Tool Should You Use?" *RNA* 22.6. DOI: 10.1261/rna.053959.115. pmid: 27022035.

Schwende, H., E. Fitzke, P. Ambs, and P. Dieter (1996). "Differences in the State of Differentiation of THP-1 Cells Induced by Phorbol Ester and 1,25-Dihydroxyvitamin D3". *Journal of Leukocyte Biology* 59.4. pmid: 8613704.

Shalhoub, V., L. C. Gerstenfeld, D. Collart, J. B. Lian, and G. S. Stein (1989). "Downregulation of Cell Growth and Cell Cycle Regulated Genes during Chick Osteoblast Differentiation with the Reciprocal Expression of Histone Gene Variants". *Biochemistry* 28.13. DOI: 10.1021/bi00439a002. pmid: 2476172.

Sherry, B., P. Tekamp-Olson, C. Gallegos, D. Bauer, G. Davatelis, S. D. Wolpe, F. Masiarz, D. Coit, and A. Cerami (1988). "Resolution of the Two Components of Macrophage Inflammatory Protein 1, and Cloning and Characterization of One of Those Components, Macrophage Inflammatory Protein 1 Beta." *Journal of Experimental Medicine* 168.6. DOI: 10.1084/jem.168.6.2251.

Shinohara, A., H. Ogawa, and T. Ogawa (1992). "Rad51 Protein Involved in Repair and Recombination in S. Cerevisiae Is a RecA-like Protein". *Cell* 69.3. DOI: 10.1016/0092-8674(92)90447-k. pmid: 1581961.

Silverstein, R. L. and M. Febbraio (2009). "CD36, a Scavenger Receptor Involved in Immunity, Metabolism, Angiogenesis, and Behavior". *Science signaling* 2.72. DOI: 10.1126/scisignal.272re3. pmid: 19471024.

Sjöberg, A. P., L. A. Trouw, and A. M. Blom (2009). "Complement Activation and Inhibition: A Delicate Balance". *Trends in Immunology* 30.2. DOI: 10.1016/j.it.2008.11.003. pmid: 19144569.

Smith, T. G., L. Serghides, S. N. Patel, M. Febbraio, R. L. Silverstein, and K. C. Kain (2003). "CD36-Mediated Nonopsonic Phagocytosis of Erythrocytes Infected with Stage I and IIA Gametocytes of Plasmodium Falciparum". *Infection and Immunity* 71.1. DOI: 10.1128/IAI.71.1.393-400.2003. pmid: 12496189.

Srivastava, A., L. Malik, H. Sarkar, M. Zakeri, F. Almodaresi, C. Soneson, M. I. Love, C. Kingsford, and R. Patro (2020). "Alignment and Mapping Methodology Influence Transcript Abundance Estimation". *Genome Biology* 21.1. DOI: 10.1186/s13059-020-02151-8.

Stack Overflow (2021). *Stack Overflow Developer Survey 2021*. Stack Overflow. URL: https://insights.stackoverflow.com/survey/2021/?utm_source=social-share&utm_medium=social&utm_campaign=dev-survey-2021 (visited on 02/04/2022).

Stahl, P. D. and R. A. Ezekowitz (1998). "The Mannose Receptor Is a Pattern Recognition Receptor Involved in Host Defense". *Current Opinion in Immunology* 10.1. DOI: 10.1016/s0952-7915(98)80031-9. pmid: 9523111.

Struyf, S., P. Menten, J.-P. Lenaerts, W. Put, A. D'Haese, E. De Clercq, D. Schols, P. Proost, and J. Van Damme (2001). "Diverging Binding Capacities of Natural LD78β Isoforms of Macrophage Inflammatory Protein-1α to the CC Chemokine Receptors 1, 3 and 5 Affect Their Anti-HIV-1 Activity and Chemotactic Potencies for Neutrophils and Eosinophils". *European Journal of Immunology* 31.7. DOI: 10.1002/1521-4141(200107)31.

Stuart, T. and R. Satija (2019). "Integrative Single-Cell Analysis". *Nature Reviews Genetics* 20.5 (5). DOI: 10.1038/s41576-019-0093-7.

Sundström, C. and K. Nilsson (1976). "Establishment and Characterization of a Human Histiocytic Lymphoma Cell Line (U-937)". *International Journal of Cancer* 17.5. DOI: 10.1002/ijc.2910170504. pmid: 178611.

Svanborg, C., G. Godaly, and M. Hedlund (1999). "Cytokine Responses during Mucosal Infections: Role in Disease Pathogenesis and Host Defence". *Current Opinion in Microbiology* 2.1. DOI: 10.1016/s1369-5274(99)80017-4. pmid: 10047563.

Takeshita, F., C. A. Leifer, I. Gursel, K. J. Ishii, S. Takeshita, M. Gursel, and D. M. Klinman (2001). "Cutting Edge: Role of Toll-Like Receptor 9 in CpG DNA-Induced Activation of Human Cells". *The Journal of Immunology* 167.7. DOI: 10.4049/jimmunol.167.7.3555. pmid: 11564765.

Tang, D., R. Kang, C. B. Coyne, H. J. Zeh, and M. T. Lotze (2012). "PAMPs and DAMPs: Signal 0s That Spur Autophagy and Immunity". *Immunological Reviews* 249.1. DOI: 10.1111/j.1600-065X.2012.01146.x.

Tauber, A. I. (2003). "Metchnikoff and the Phagocytosis Theory". *Nature Reviews Molecular Cell Biology* 4.11 (11). DOI: 10.1038/nrm1244.

Taylor, P. R., L. Martinez-Pomares, M. Stacey, H.-H. Lin, G. D. Brown, and S. Gordon (2005). "Macrophage Receptors and Immune Recognition". *Annual Review of Immunology* 23. DOI: 10.1146/annurev.immunol.23.021704.115816. pmid: 15771589.

Trapnell, C. (2015). "Defining Cell Types and States with Single-Cell Genomics". *Genome Research* 25.10. DOI: 10.1101/gr.190595.115. pmid: 26430159.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter (2010). "Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation". *Nature Biotechnology* 28.5 (5). DOI: 10.1038/nbt.1621.

Tsuchiya, S., M. Yamabe, Y. Yamaguchi, Y. Kobayashi, T. Konno, and K. Tada (1980). "Establishment and Characterization of a Human Acute Monocytic Leukemia Cell Line (THP-1)". *International Journal of Cancer* 26.2. DOI: 10.1002/ijc.2910260208. pmid: 6970727.

Tye, B. K. (1999). "MCM Proteins in DNA Replication". *Annual Review of Biochemistry* 68. DOI: 10.1146/annurev.biochem.68.1.649. pmid: 10872463.

Vabulas, R. M., S. Braedel, N. Hilf, H. Singh-Jasuja, S. Herter, P. Ahmad-Nejad, C. J. Kirschning, C. Da Costa, H.-G. Rammensee, H. Wagner, and H. Schild (2002). "The Endoplasmic Reticulum-Resident Heat Shock Protein Gp96 Activates Dendritic Cells via the Toll-like Receptor 2/4 Pathway". *The Journal of Biological Chemistry* 277.23. DOI: 10.1074/jbc.M200425200. pmid: 11912201.

Van Furth, R., Z. A. Cohn, J. G. Hirsch, J. H. Humphrey, W. G. Spector, and H. L. Langevoort (1972). "The Mononuclear Phagocyte System: A New Classification of Macrophages, Monocytes, and Their Precursor Cells". *Bulletin of the World Health Organization* 46.6. pmid: 4538544.

Virolainen, M. (1968). "Hematopoietic Origin of Macrophages as Studied by Chromosome Markers in Mice". *Journal of Experimental Medicine* 127.5. DOI: 10.1084/jem.127.5.943.

Volkman, A. (1968). "The Function of the Monocyte". *International Society of Blood Transfusion*. Ed. by International Society of Blood Transfusion and L. Hollander.

Vol. 29. S. Karger AG, pp. 86–97. ISBN: 978-3-8055-0129-3 978-3-318-03498-1. DOI: 10.1159/000384596.

Wajant, H. (2002). "The Fas Signaling Pathway: More than a Paradigm". *Science (New York, N.Y.)* 296.5573. DOI: 10.1126/science.1071553. pmid: 12040174.

Waksman, B. H. and Y. Namba (1976). "On Soluble Mediators of Immunologic Regulation". *Cellular Immunology* 21.1. DOI: 10.1016/0008-8749(76)90337-3.

Warner, J. R. (1999). "The Economics of Ribosome Biosynthesis in Yeast". *Trends in Biochemical Sciences* 24.11. DOI: 10.1016/s0968-0004(99)01460-7. pmid: 10542411.

Weber, A. P. M. (2015). "Discovering New Biology through Sequencing of RNA". *Plant Physiology* 169.3. DOI: 10.1104/pp.15.01081. pmid: 26353759.

Wheeler, D. S., M. A. Chase, A. P. Senft, S. E. Poynter, H. R. Wong, and K. Page (2009). "Extracellular Hsp72, an Endogenous DAMP, Is Released by Virally Infected Airway Epithelial Cells and Activates Neutrophils via Toll-like Receptor (TLR)-4". *Respiratory Research* 10.1. DOI: 10.1186/1465-9921-10-31. pmid: 19405961.

Williams, M. J. (2007). "Drosophila Hemopoiesis and Cellular Immunity". *Journal of Immunology (Baltimore, Md.: 1950)* 178.8. DOI: 10.4049/jimmunol.178.8.4711. pmid: 17404248.

Witsell, A. L. and L. B. Schook (1992). "Tumor Necrosis Factor Alpha Is an Autocrine Growth Regulator during Macrophage Differentiation." *Proceedings of the National Academy of Sciences of the United States of America* 89.10. pmid: 1374912. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC49162/ (visited on 03/22/2022).

Wolpe, S. D., G. Davatelis, B. Sherry, B. Beutler, D. G. Hesse, H. T. Nguyen, L. L. Moldawer, C. F. Nathan, S. F. Lowry, and A. Cerami (1988). "Macrophages Secrete a Novel Heparin-Binding Protein with Inflammatory and Neutrophil Chemokinetic Properties." *Journal of Experimental Medicine* 167.2. DOI: 10.1084/jem.167.2.570.

Yam, L. T., C. Y. Li, and W. H. Crosby (1971). "Cytochemical Identification of Monocytes and Granulocytes". *American Journal of Clinical Pathology* 55.3. DOI: 10.1093/ajcp/55.3.283. pmid: 5549896.

Yang, R. B., M. R. Mark, A. Gray, A. Huang, M. H. Xie, M. Zhang, A. Goddard, W. I. Wood, A. L. Gurney, and P. J. Godowski (1998). "Toll-like Receptor-2 Mediates Lipopolysaccharide-Induced Cellular Signalling". *Nature* 395.6699. DOI: 10.1038/26239. pmid: 9751057.

Yang, S., R. Tamai, S. Akashi, O. Takeuchi, S. Akira, S. Sugawara, and H. Takada (2001). "Synergistic Effect of Muramyldipeptide with Lipopolysaccharide or Lipoteichoic Acid To Induce Inflammatory Cytokines in Human Monocytic Cells in Culture". *Infection and Immunity* 69.4. DOI: 10.1128/IAI.69.4.2045-2053.2001. pmid: 11254557.

Yona, S., K.-W. Kim, Y. Wolf, A. Mildner, D. Varol, M. Breker, D. Strauss-Ayali, S. Viukov, M. Guilliams, A. Misharin, D. A. Hume, H. Perlman, B. Malissen, E. Zelzer, and S. Jung (2013). "Fate Mapping Reveals Origins and Dynamics of Monocytes and Tissue Macrophages under Homeostasis". *Immunity* 38.1. DOI: 10.1016/j.immuni.2012.12.001. pmid: 23273845.

Yoshida, H., N. Kondratenko, S. Green, D. Steinberg, and O. Quehenberger (1998). "Identification of the Lectin-like Receptor for Oxidized Low-Density Lipoprotein

in Human Macrophages and Its Potential Role as a Scavenger Receptor." *Biochemical Journal* 334 (Pt 1). DOI: 10.1042/bj3340009. pmid: 9693095.

Yoshimoto, R., Y. Fujita, A. Kakino, S. Iwamoto, T. Takaya, and T. Sawamura (2011). "The Discovery of LOX-1, Its Ligands and Clinical Significance". *Cardiovascular Drugs and Therapy* 25.5. DOI: 10.1007/s10557-011-6324-6. pmid: 21805404.

Yu, G., L.-G. Wang, Y. Han, and Q.-Y. He (2012). "clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters". *OMICS: A Journal of Integrative Biology* 16.5. DOI: 10.1089/omi.2011.0118.

Zanoni, I., R. Ostuni, G. Capuano, M. Collini, M. Caccia, A. E. Ronchi, M. Rocchetti, F. Mingozzi, M. Foti, G. Chirico, B. Costa, A. Zaza, P. Ricciardi-Castagnoli, and F. Granucci (2009). "CD14 Regulates the Dendritic Cell Life Cycle after LPS Exposure through NFAT Activation". *Nature* 460.7252 (7252). DOI: 10.1038/nature08118.

Zanoni, I., R. Ostuni, L. R. Marek, S. Barresi, R. Barbalat, G. M. Barton, F. Granucci, and J. C. Kagan (2011). "CD14 Controls the LPS-Induced Endocytosis of Toll-like Receptor 4". *Cell* 147.4. DOI: 10.1016/j.cell.2011.09.051. pmid: 22078883.

Zhao, S. (2014). "Assessment of the Impact of Using a Reference Transcriptome in Mapping Short RNA-Seq Reads". *PLOS ONE* 9.7. DOI: 10.1371/journal.pone.0101374.

Ziegler-Heitbrock, L. (2007). "The CD14+ CD16+ Blood Monocytes: Their Role in Infection and Inflammation". *Journal of Leukocyte Biology* 81.3. DOI: 10.1189/jlb.0806510. pmid: 17135573.

Ziegler-Heitbrock, L., P. Ancuta, S. Crowe, M. Dalod, V. Grau, D. N. Hart, P. J. M. Leenen, Y.-J. Liu, G. MacPherson, G. J. Randolph, J. Scherberich, J. Schmitz, K. Shortman, S. Sozzani, H. Strobl, M. Zembala, J. M. Austyn, and M. B. Lutz (2010). "Nomenclature of Monocytes and Dendritic Cells in Blood". *Blood* 116.16. DOI: 10.1182/blood-2010-02-258558.

Zizzo, G., B. A. Hilliard, M. Monestier, and P. L. Cohen (2012). "Efficient Clearance of Early Apoptotic Cells by Human Macrophages Requires "M2c" Polarization and MerTK Induction". *Journal of immunology (Baltimore, Md. : 1950)* 189.7. DOI: 10.4049/jimmunol.1200662. pmid: 22942426.

Data availability

The data used to generate the findings in this study are openly available in the ENA under accession numbers PRJNA314802, PRJNA381911, and PRJNA449980.

The software analysis pipeline devised for this project is openly available in the Functional Genomics and Immunogenetics Lab `GitHub` repository, and it is free to use under the MIT licence.

The repository is hosted at https://github.com/fungenigs/rna-pipe

6 Supplemental

6.1 Pipeline results

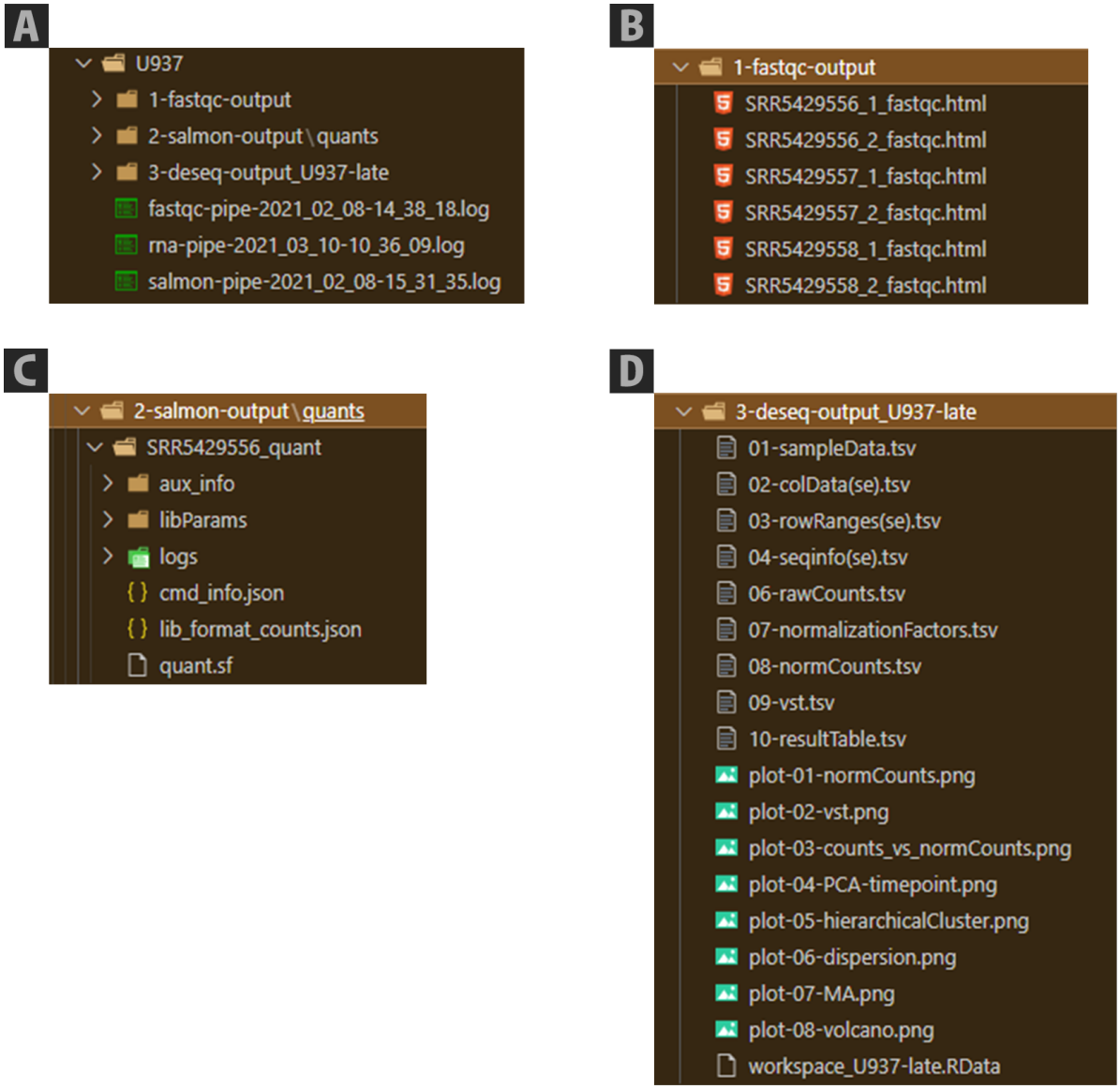


Figure S1: **Pipeline output folders.** **A**: Main output directory. Three subfolders for the three selected modules which were run (`FastQC`, `Salmon`, and `DESeq2`). Note the logs for each module and for the main pipeline (with the date and time of the analysis written into the filename). **B**: `fastqc-pipe` output directory. Contents: `HTML` reports on sequence quality for each input sequence file from `FastQC` which can be opened in the browser. **C**: `salmon-pipe` output directory. Contents: `quant` folders for each input sequence file (or pair, in the case of paired-end data) from `Salmon` – these contain the raw quantification files as well as internal logs and command parameters from `Salmon`. **D**: `deseq-pipe` output directory. Note the name which is supplied from the configuration file (in this case "U937-late") and the contents: various useful outputs from the differential expression analysis process, including eight output graphs and the entire `RStudio` workspace.

Condensed sample output log from `rna-pipe`

```
Running configuration file 'RSP-U937.rnaconf' version '1.5.0'
_____________________________________________________________
           .u    .      u.    u.
         .d88B :@8c   x@88k u@88c.       u
        ="8888f8888r ^"8888""8888"    us888u.
          4888>'88"    8888  888R  .@88 "8888"
          4888> '      8888  888R  9888  9888
          4888>        8888  888R  9888  9888
         .d888L .+     8888  888R  9888  9888
         ^"8888*"     "*88*" 8888" 9888  9888
            "Y"         ""   'Y"   "888*""888"
                                    ^Y"   ^Y'
                .
               @88>
 .d``          %8P    .d``
 @8Ne.   .u      .     @8Ne.   .u        .u
 %8888:u@88N   .@88u   %8888:u@88N    ud8888.
  `888I  888. ''888E`   `888I  888. :888'8888.
   888I  888I   888E     888I  888I d888 '88%"
   888I  888I   888E     888I  888I 8888.+"
 uW888L  888'   888E   uW888L  888' 8888L
'*88888Nu88P    888&  '*88888Nu88P  '8888c. .+
~ '88888F`      R888" ~ '88888F`     "88888%
   888 ^         ""      888 ^         "YP'
   *8E                   *8E
   '8>                   '8>
    "                     "
_____________________________________________________________

rna-pipe version 0.1.0 running at 10/03/2021 10:36:09
**** INFO: The latest changes to this config file were:
Rerunning U937 without LFCT
_____________________________________________________________

=============== rna-pipe > Skipping FASTQC ==============

=============== rna-pipe > Skipping SALMON ==============

=============== rna-pipe > Running DESEQ ================

**** INFO: The configuration file you supplied demands overwriting the default
↪  DESEQ2 config file.
**** INFO: New DESEQ2 config file: located at
↪  './deseq-pipe/rna-seq-paper-project-U937.deseqconf'
**** INFO: Using the following DESEQ2 config file:
↪  rna-seq-paper-project-U937.deseqconf version 1.5.0
**** INFO: The latest changes to this config file were:
Re-running deseq on U937 0h to 96h with no replicate 1, timepoint only, no LFCT
**** WARNING: Using the supplied R executable: /usr/bin/Rscript
**** INFO: Executable now the following: /usr/bin/Rscript
↪  ./deseq-pipe/deseq-pipeline.R
**** INFO: Run name:
_U937-late
============ Creating DESEQ output directory ============

**** OK: Directory created succesfully at '/home/duncan/masters/data/⌋
↪  rna-seq-paper-project/U937//rna-pipe-output//3-deseq-output_U937-late/'
================= Running DESEQ analysis ================

**** INFO: Processing /media/duncan/data/U937/PRJNA381911/ (output from here is
↪  directly from deseq-pipeline.R)


Checking arguments
======================
Input Directory (indir):  /home/duncan/masters/data/rna-seq-paper-project/U937/
Output Directory (outdir):  /home/duncan/masters/data/rna-seq-paper-project/U937//⌋
↪  rna-pipe-output//3-deseq-output_U937-late/
```

```
Sample table (sampleTable):
↪  /media/duncan/data/PRJNA381911/sample_table-late_0to96h_sans_rep-1.tsv
p-value (pvalue):  0.01
Log base 2 fold change threshold (lfct):
Design formula (design):  ~ timepoint
Run name (runname):  _U937-late
======================
Checking arguments completed
======================

Producing sample data.frame
======================
                                                    names timepoint replicate
1 U937 cell line treated with PMA for 96 hrs replicate 3        96         3
2 U937 cell line treated with PMA for 96 hrs replicate 2        96         2
3                           U937 cell line replicate 3          0         3
4                           U937 cell line replicate 2          0         2
                                                                files
1 rna-pipe-output/2-salmon-output/quants/SRR5429556_quant/quant.sf
2 rna-pipe-output/2-salmon-output/quants/SRR5429557_quant/quant.sf
3 rna-pipe-output/2-salmon-output/quants/SRR5429562_quant/quant.sf
4 rna-pipe-output/2-salmon-output/quants/SRR5429563_quant/quant.sf
----------------------
Writing 01-sampleData.tsv
----------------------


Running tximeta
======================
importing quantifications
reading in files with read_tsv
1 2 3 4
found matching transcriptome:
[ Ensembl - Homo sapiens - release 97 ]
loading existing EnsDb created: 2020-08-08 22:03:42
loading existing transcript ranges created: 2020-08-08 22:03:46
loading existing EnsDb created: 2020-08-08 22:03:42
obtaining transcript-to-gene mapping from database
loading existing gene ranges created: 2020-08-08 22:03:53
summarizing abundance
summarizing counts
summarizing length
======================
Running tximeta completed
======================


Running DESeq2
======================
Using design: ~ timepoint
Conditions of interest: timepoint
using counts and average transcript lengths from tximeta


Performing manual filtering
======================
Complete cases according to DESeq2: 16462

----------------------
Using significance level of p-value less than 0.01
----------------------
Significant complete cases according to DESeq2: 6094

----------------------
Not filtering based on log (base 2) fold change threshold
----------------------
======================
Performing manual filtering completed
======================

----------------------
Saving R workspace
```

```
133  ----------------------
134
135
136  DESEQ-PIPE
137  ======================
138  ======================
139  COMPLETED
140  ======================
141
142
143  **** INFO: Completed /media/duncan/data/U937/PRJNA381911/ (output from here is from
     ↪  rna-pipe)
144
145
146  ==================== rna-pipe complete ====================
147
```

Figure S2: **Condensed sample output log from `rna-pipe`.**

6.2 Sequencing quality results from `FastQC`

Though the aim in choosing these datasets was to find very similar experiments for comparing the differentiation process in different cell lines, the sequencing platforms used did vary across the datasets. The data statistics shown in Table 2 indicate that the U937 and THP-1 experiments generated far more reads than the HL60 experiment – though the difference cannot be due to sequencing platform alone, as HL60 and U937 used the same technology.

Almost all samples from all cell lines had overrepresented sequences (the module producing either warnings or failures for this test), high sequence duplication levels, and per-sequence GC content warnings. If this experiment used data from DNA sequencing, these GC content warnings would suggest widespread contamination – in this case they just indicate systematic error as described above. Per-base N content was zero for all the runs, indicating the sequencing platform used was confident in all the bases it called, N being the standard symbol for unresolved bases (Andrews 2020b).

One sample from the HL60 dataset, *SRR3214278*, triggered severe warnings and turned out to have unusable contents: 41.04% of the sequences read in this experiment were *TruSeq* Adapter sequences (artefacts of the sequencing chemistry caused by improper experimental setup), and as such, sequence duplication levels were very high and per-base sequence content warnings were triggered (with 38.95% of the reads matched to a single sequence). Fortunately, the quality metrics for all the other HL60 samples were positive (as per Figure S3), so no other samples were discarded.

The quality control tests on the sequences from the U937 cell lines were generally successful – two failure messages were triggered for per base sequence content (*SRR5429562* and *SRR5429563*), but this can again be attributed to `FastQC`'s idiosyncrasies when processing RNA-seq data. Both failures happened in runs which also raised warnings for sequence duplication and overrepresented sequences, reinforcing the explanation that it was the expected sequence duplication that was responsible for the per base sequence content messages.

As with the other cell lines, all THP-1 runs showed an interesting sequence duplication pattern (Figure S3 subfigure **C**) with peaks in each bin of high duplication (10 – 10 000x duplication) suggesting the presence of a few transcripts with very high abundance – the expected pattern with RNA-seq data (Andrews 2020a). Overall, outputs from `FastQC` were positive and similar across all cell lines – the images chosen for Figure S3 are representative of all the processed data. At the end of this analysis, the data was deemed of sufficient quality to continue, barring the one sample from the HL60 group (the quality control results of which are included in Figure S4).

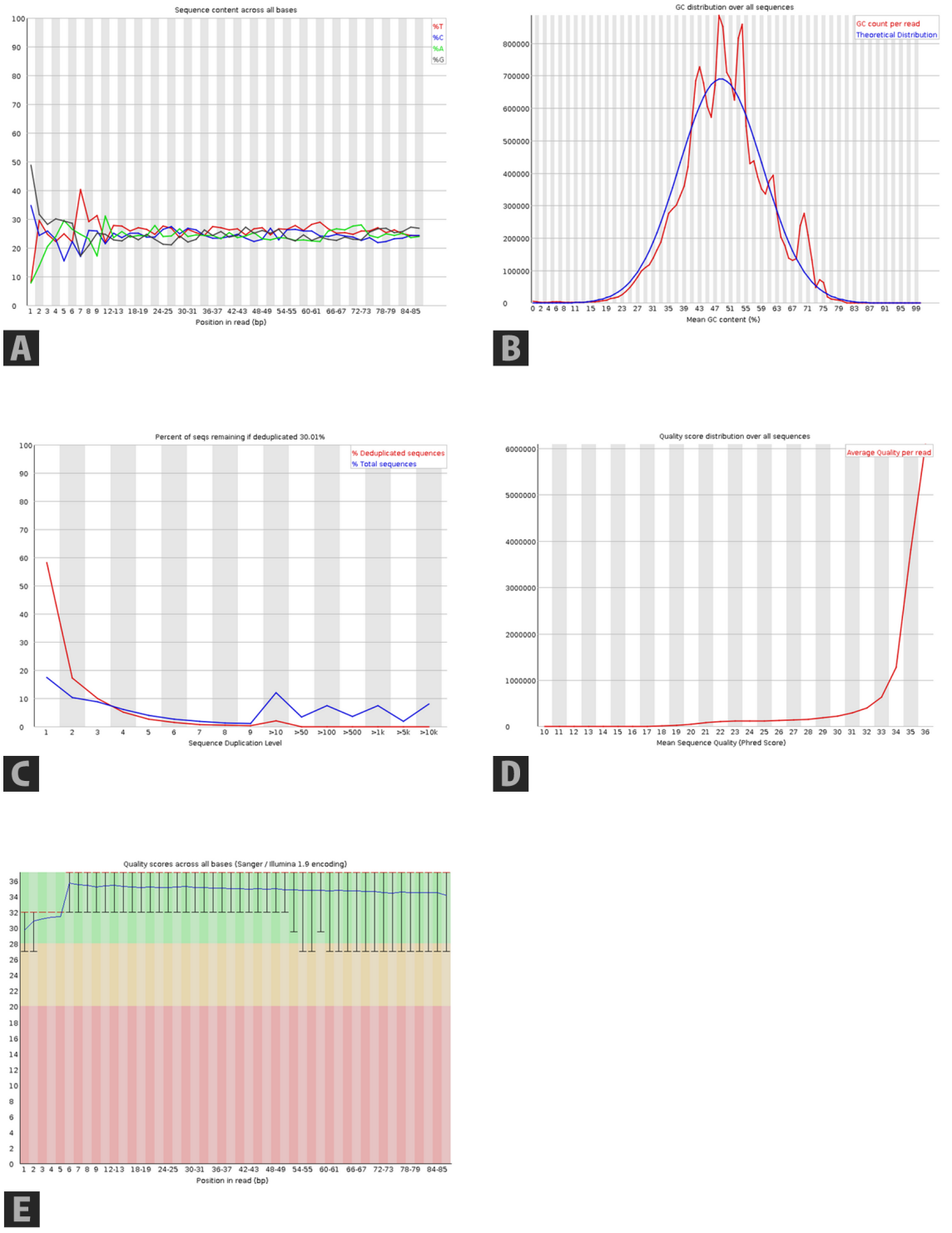
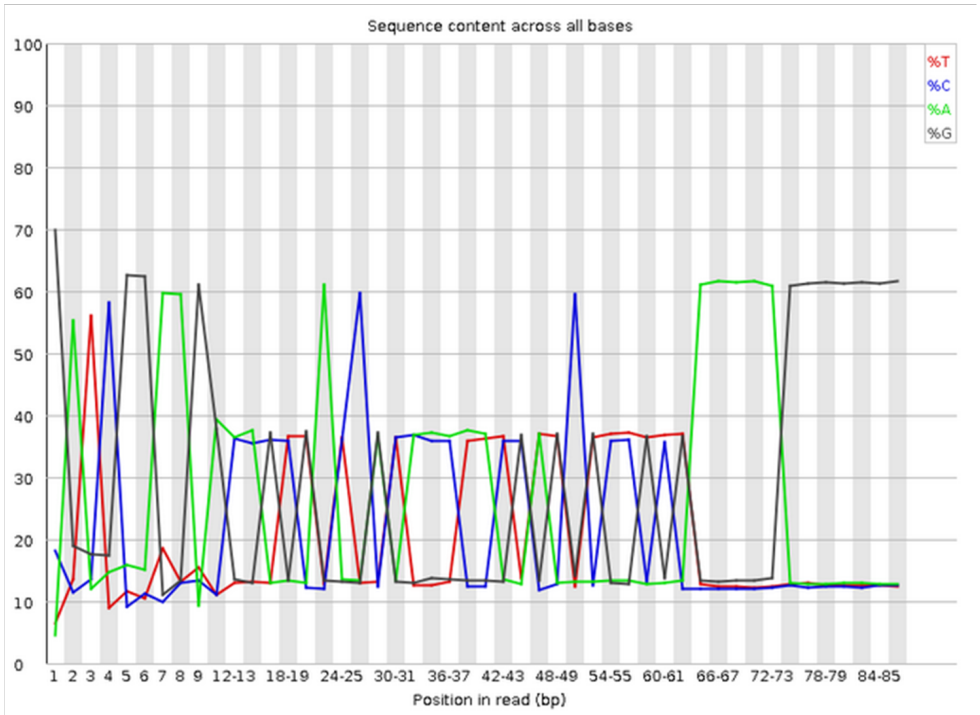


Figure S3: **Representative FastQC results from the HL60 dataset.** **A**: Sequence content across all bases showing expected balanced base content across all positions in the reads, barring the first few positions (which may be influenced by adapter content). **B**: GC content distribution across all sequences which conforms well to theoretical distribution but includes peaks which correspond to highly duplicated sequences. **C**: Sequence duplication levels showing generally low numbers of duplicates but with peaks for very common sequences (10x – 10 000x duplication). **D**: Quality score distribution over all sequences showing very high Phred scores (majority over *Q=30* or 99.9% accuracy). **E**: Quality scores per base showing position-invariant high quality across the sequences.



A

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAACTAATCTCGTATGC	5252559	38.94668149824662	TruSeq Adapter, Index 23 (97% over 38bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAACAAATCTCGTATGC	77543	0.5749659401100564	TruSeq Adapter, Index 23 (97% over 37bp)
GAACGGAAGAGCACACGTCTGAACTCCAGTCACCAACTAATCTCGTATGC	49605	0.3678112203443167	TruSeq Adapter, Index 7 (97% over 35bp)
GGGGCATCCATATAGTCACTCCAGGTTTATGGAGGGTTCTTCTACTATTA	25311	0.1876760366522528	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACAAACTAATCTCGTATGC	22945	0.17013261668784085	TruSeq Adapter, Index 5 (97% over 36bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAACAATCTCGTATGCC	22704	0.1683456495655149	TruSeq Adapter, Index 23 (97% over 37bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAACTAATATCGTATGC	22131	0.16409696839915477	TruSeq Adapter, Index 23 (97% over 38bp)
GATCGGAAGAGCACACGTCAGAACTCCAGTCACCAACTAATCTCGTATGC	21132	0.15668958186304002	TruSeq Adapter, Index 7 (97% over 35bp)
GATAGGAAGAGCACACGTCTGAACTCCAGTCACCAACTAATCTCGTATGC	18391	0.1363656113970835	TruSeq Adapter, Index 7 (97% over 35bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCACCTAATCTCGTATGC	16941	0.12561414945777782	TruSeq Adapter, Index 7 (97% over 35bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAACTAAACTCGTATGC	15767	0.11690917268760895	TruSeq Adapter, Index 23 (97% over 38bp)
GATCGGAAGAGCACACGACTGAACTCCAGTCACCAACTAATCTCGTATGC	14816	0.10985769661569188	TruSeq Adapter, Index 7 (97% over 35bp)

B

Figure S4: **Rejected SRR3214278 `FastQC` results from the HL60 dataset. A**: sequence content across all bases showing unexpected base content across many positions in the reads. **B**: table of overrepresented sequences showing huge sequence duplication mainly consisting of TruSeq adapters.

6.3 `DESeq2` quality control data

HL60 (*0h versus 96h*)

In addition to the standard differential expression analysis performed according to `timepoint` only, we performed a run of the `DESeq2` pipeline to test for potential errors in the dataset not picked up by sequence quality and mapping rate checks. In order to effect this, we ran `DESeq2` with `timepoint + replicate` for a single analysis – a scheme in which changes due to time as well as inter-replicate differences are examined. We were not interested in the differential expression analysis produced by this effort, just the resulting PCA plots produced by the pipeline. These plots revealed that replicate 2 of the HL60 dataset was clustering very unexpectedly and thus probably represented a contaminated sample (Figure S5). As a result we decided to remove replicate 2 entirely from the analysis and use replicates 1 and 3.

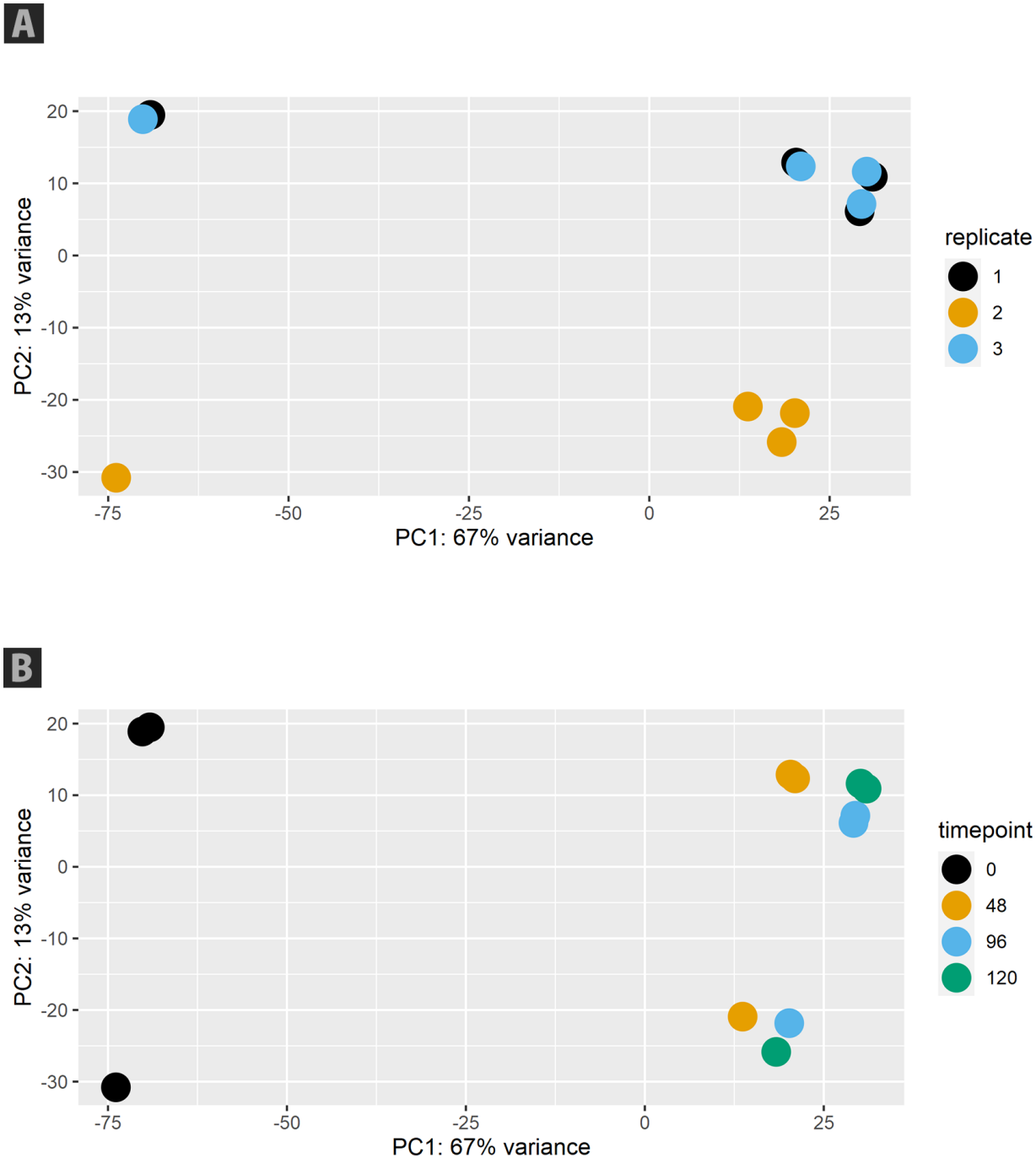


Figure S5: **Unexpected HL60 clustering results.** The two PCA plots produced by `DESeq2` show unexpected clustering of replicate 2 of the dataset. The PCA was performed here on all four time points of the original HL60 dataset. **A**: PCA plot of the datasets when separated by *replicate* show replicate 2 to be a major outlier. The pairs of points represent a single timepoint and the other two replicates (1 and 3) – notice their tight clustering compared with replicate 2. **B**: The same PCA plot represented above but arranged here by *timepoint*. Notice the tight clustering of pairs (from replicates 1 and 3 as above).

Differential expression analysis (DEA) revealed 6665 differentially expressed genes that were statistically significant ($p < 0.01$) between T_0 and T_{+96h} in the HL60 cell line. Log_2 fold change values were between -24.5 and 22.6 across the genes (Figure S7).

U937 (*0h versus 96h*)

One of the output graphs from the `deseq-pipe` module shows the total number of counts (of all genes combined) before and after normalization. This plot of *Counts versus Normalized Counts* (Figure S6 subfigure **A**) showed replicate 1 of the dataset to have far more counts than the other replicates across all time points, suggesting experimental error with this sample. We repeated the testing run of `DESeq2` with *replicate + timepoint* to produce PCA plots for quality control as mentioned above. The resultant PCA plots revealed that *Replicate 1* was clustering unexpectedly (Figure S6 subfigure **B**). As a result we decided to remove *Replicate 1* from the analysis in its entirety.

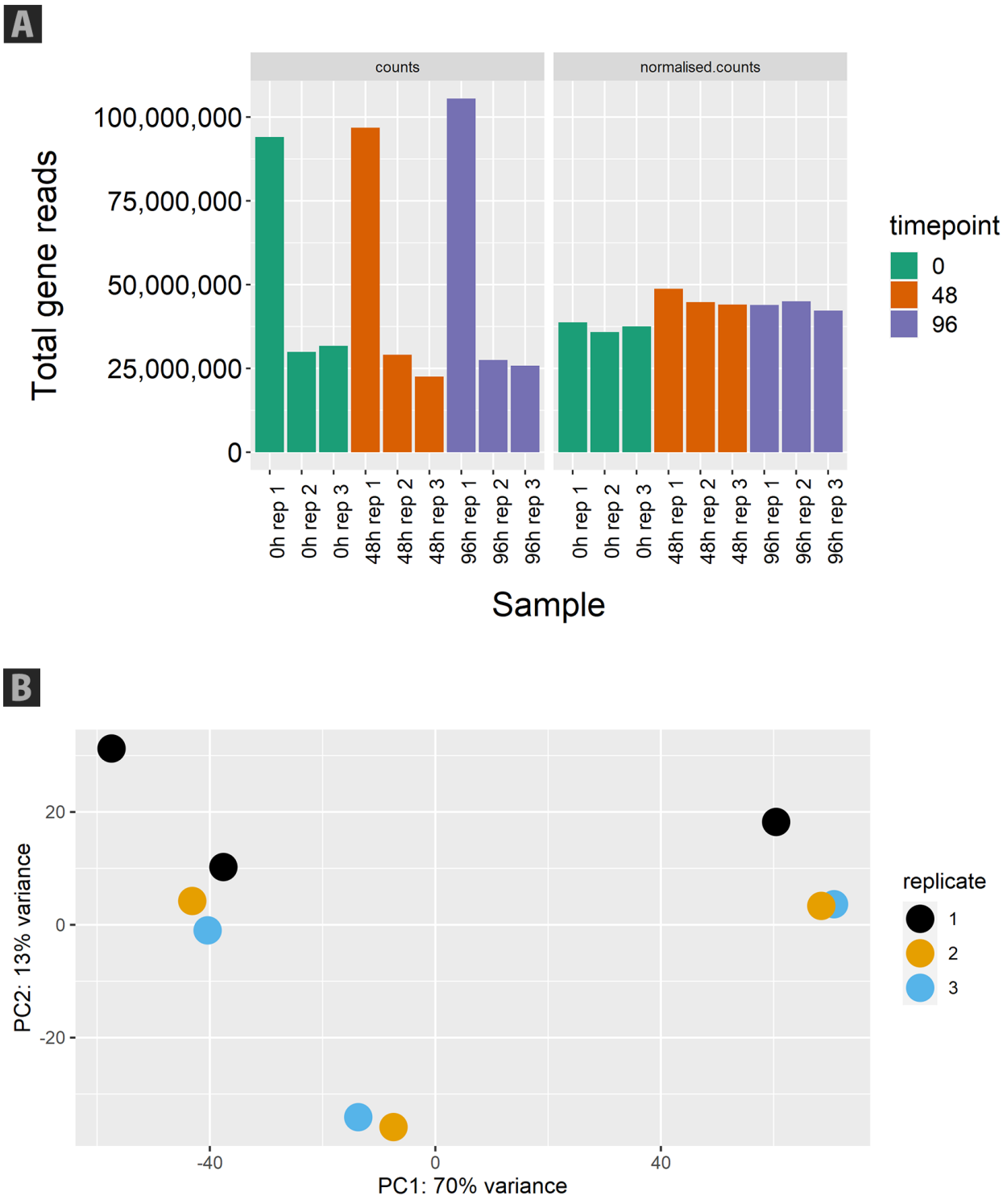


Figure S6: **Unexpected U937 counts and clustering results.** Unexpected counts and clustering of replicate 1 of the U937 dataset. The PCA and counts plotting here were performed on all three time points of the original U937 dataset. **A**: Plot of counts versus normalized counts for each of the time points in the dataset, arranged by replicate. Replicate 1 shows massively more counts than the other two (nearly 4x more), which is unexpected and suggests a problem with the experiment. **B**: PCA plot of the datasets grouped by *replicate* show replicate 1 to be an outlier. Notice the clustering of replicates 2 and 3 compared with replicate 1.

DEA revealed 6094 differentially expressed genes that were statistically significant ($p < 0.01$) between T_0 and T_{+96h} in the U937 cell line. Log_2 fold change values were between -23.6 and 22.7 across the genes (Figure S7).

THP-1 (*0h versus 72h*)

DEA revealed 6171 differentially expressed genes that were statistically significant ($p < 0.01$) between T_0 and T_{+72h} in the THP-1 cell line. Log_2 fold change values were between -12.6 and 21.0 across the genes (Figure S7).

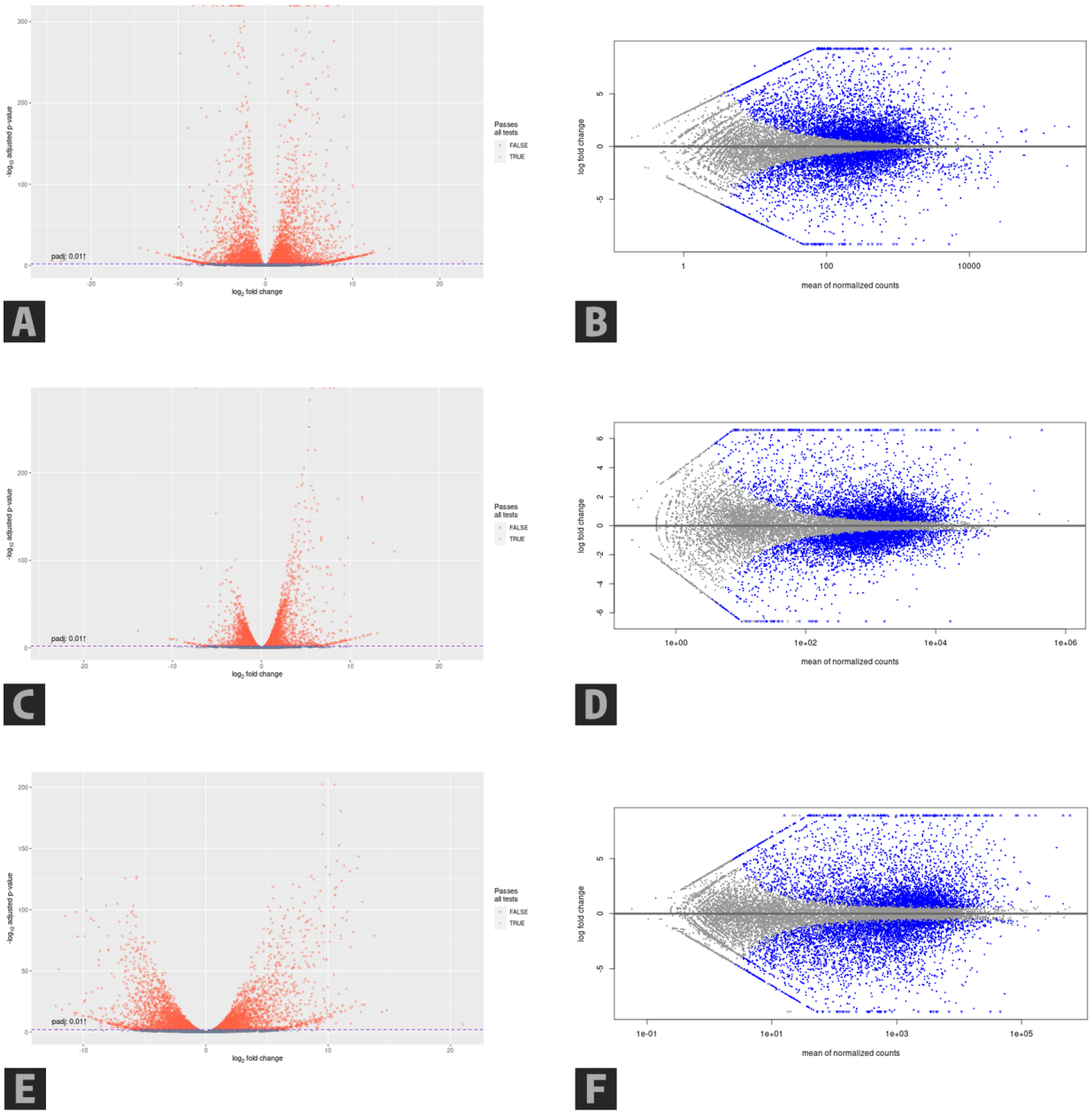


Figure S7: *DESeq2 results for all cell lines.* **Volcano plots** compare $\log_2$ fold change with significance (p-value) for all DEGs, essentially showing how the magnitude of change in expression relates to the statistical significance of the observation. The threshold line indicates an adjusted p-value of 0.01. **MA plots** compare $\log_2$ fold change with mean of normalized counts, showing how the magnitude of change in expression depends on the abundance of a transcript. Darker shading here indicates DEGs that are deemed statistically significant. **A, B**: Volcano and MA plots of HL60 DEGs between T_0 and T_{+96h} **C, D**: Volcano and MA plots of U937 DEGs between T_0 and T_{+96h} **E, F**: Volcano and MA plots of THP-1 DEGs between T_0 and T_{+72h}

Table S1: Example list of differentially expressed genes, showing gene identifiers, mean expression levels, change in expression, and quality control statistics.

gene	baseMean	log2FoldChange	lfcSE	stat	p-value	padj
ENSG00000000003	2.434E+00	-5.008E+00	3.097E+00	-1.617E+00	1.058E-01	1.753E-01
ENSG00000000419	2.687E+02	-7.883E-01	2.478E-01	-3.182E+00	1.463E-03	3.872E-03
ENSG00000000457	1.189E+02	1.022E+00	3.697E-01	2.766E+00	5.683E-03	1.341E-02
ENSG00000000460	8.130E+01	-1.068E+00	4.960E-01	-2.152E+00	3.138E-02	6.217E-02
ENSG00000000938	1.784E+03	2.997E+00	1.223E-01	2.451E+01	1.248E-132	1.158E-130

6.4 Functional enrichment analysis: KEGG pathway

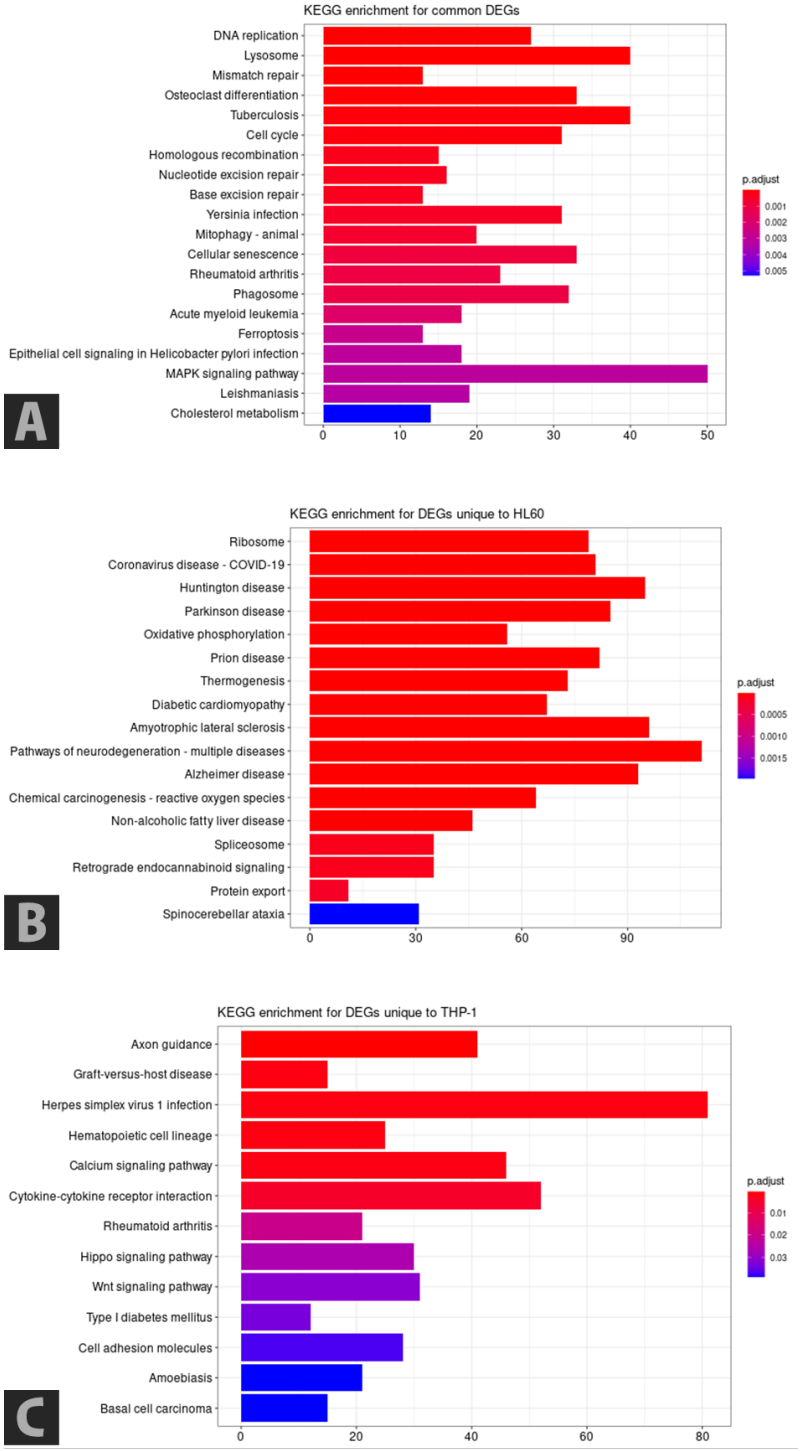


Figure S8: **Functional enrichment analysis of common and unique DEGs.** Terms on the y-axis refer to KEGG pathways in *Homo sapiens*, while the x-axis shows how many DEGs from the dataset appear in that pathway. Shading denotes statistical significance through adjusted p-value. **A**: KEGG enrichment of pathways from DEGs common to all cell lines **B**: KEGG enrichment of pathways from DEGs unique to HL60 **C**: KEGG enrichment of pathways from DEGs unique to THP-1

6.5 Commonly differentially expressed genes regulated in different directions

Table S2: **Commonly differentially expressed genes in the lysosome KEGG pathway showing those which are expressed in different directions between cell lines.** The columns labelled l2fc represent the change in gene expression represented as $\log_2$ fold change for each cell line. Where KEGG pathway labels are different to the standard labels, the "Label in pathview" column shows this difference. Rows in grey show genes with a different direction of expression (chosen based on sign change only, as values are in a log scale so anything negative represents a smaller amount of expression).

Gene ID	l2fc_HL60	l2fc_U937	l2fc_THP1	Gene name	Label in pathview
ENSG00000010404	5.0854	2.3083	3.5037		
ENSG00000018280	1.7173	5.9663	3.0995		
ENSG00000047249	2.4754	2.6462	1.6102		
ENSG00000065000	1.7158	0.7303	0.6800		
ENSG00000070371	-2.8818	-1.7462	-6.4330		
ENSG00000071553	2.7229	2.2570	0.9113		
ENSG00000100448	-4.6498	-7.3063	-9.8948		
ENSG00000101160	1.9824	1.1093	1.7493		
ENSG00000102575	3.7838	10.6377	2.6581		
ENSG00000102805	1.0529	0.6393	1.7407		
ENSG00000103066	2.5691	2.1427	0.9669		
ENSG00000103365	-1.6446	-0.8741	-1.9524		
ENSG00000103811	3.0371	2.1830	3.1043		
ENSG00000104093	-1.2039	1.1904	1.0898	DMXL2 (rabconnectin 3a)	-
ENSG00000104763	2.0767	0.7012	1.2735		
ENSG00000105612	3.0126	1.1211	1.5073		
ENSG00000107798	3.1344	1.8085	1.7439		
ENSG00000110719	2.1557	1.8414	2.6404		
ENSG00000110911	1.2860	1.0416	2.5685		
ENSG00000117410	1.0056	1.2282	1.3311		
ENSG00000119655	0.9621	1.2829	2.1044		
ENSG00000131238	1.0717	0.9317	-1.5169	PPT1 (palmitoyl-protein thioesterase)	CLN1
ENSG00000134243	3.5662	4.0309	1.4196		
ENSG00000134575	2.7924	1.2795	1.3609		
ENSG00000135047	5.7776	8.8946	4.8428		
ENSG00000135404	0.9334	1.2642	1.9583		
ENSG00000135535	3.9485	2.0006	1.1153		
ENSG00000135677	3.0662	1.6130	1.2580		
ENSG00000138760	3.7350	2.5158	2.5594		
ENSG00000141367	2.1824	1.0949	0.9167		
ENSG00000147614	3.6303	5.7538	4.3388		
ENSG00000159720	2.1204	1.1263	0.9494		
ENSG00000162511	1.6223	1.8014	2.3118		
ENSG00000163131	1.2015	1.8230	1.5791		
ENSG00000164733	3.7652	2.9065	4.0322		
ENSG00000166311	4.7654	0.7926	1.8310		
ENSG00000171298	2.8901	0.9083	-0.9989	GAA (lysosomal alpha-glucosidase)	GAA
ENSG00000182287	1.4787	1.8185	2.7802		
ENSG00000197217	1.4577	0.9811	1.1247		
ENSG00000197746	3.2451	2.1969	2.2767		

Table S3: **Commonly differentially expressed genes in the phagosome KEGG pathway showing those which are expressed in different directions between cell lines.** The columns labelled l2fc represent the change in gene expression represented as $\log_2$ fold change for each cell line. Where KEGG pathway labels are different to the standard labels, the "Label in pathview" column shows this difference. Rows in grey show genes with a different direction of expression (chosen based on sign change only, as values are in a log scale so anything negative represents a smaller amount of expression).

Gene ID	l2fc_HL60	l2fc_U937	l2fc_THP1	Gene name	Label in pathview
ENSG00000005381	-7.1432	-2.6591	-8.5133		
ENSG00000038945	5.2010	6.2231	8.3787		
ENSG00000047249	2.4754	2.6462	1.6102		
ENSG00000049245	-1.8174	0.5877	0.7526	VAMP3 (vesicle-associated membrane protein 3)	VAMP3
ENSG00000071553	2.7229	2.2570	0.9113		
ENSG00000072694	4.2340	4.2842	5.0019		
ENSG00000073060	-1.4323	0.5610	-1.3984	SCARB1 (scavenger receptor class B, member 1)	SRB1
ENSG00000100365	-0.9204	-0.5078	-1.5686		
ENSG00000108774	0.9521	0.7141	1.0720		
ENSG00000110719	2.1557	1.8414	2.6404		
ENSG00000117410	1.0056	1.2282	1.3311		
ENSG00000125730	4.3821	7.0674	4.8687		
ENSG00000131100	1.0784	0.6695	1.1439		
ENSG00000135047	5.7776	8.8946	4.8428		
ENSG00000135218	3.7287	3.3531	4.2898		
ENSG00000138448	1.6894	2.6476	3.1269		
ENSG00000143226	1.1387	4.0437	-3.4594	FCGR2A (low affinity FC-γ receptor II-a) [CD32]	FcγR
ENSG00000144566	2.3138	0.7953	1.3207		
ENSG00000147416	1.8406	2.1047	1.1276		
ENSG00000147614	3.6303	5.7538	4.3388		
ENSG00000150093	3.1987	1.5360	2.4496		
ENSG00000150337	-2.2089	0.9782	-2.9354	FCGR1A (high affinity FC-γ receptor I) [CD64]	FcγR
ENSG00000155097	1.5915	1.2489	1.4337		
ENSG00000159720	2.1204	1.1263	0.9494		
ENSG00000160255	3.2332	2.3099	2.8198		
ENSG00000161638	2.7292	2.0818	3.1852		
ENSG00000163131	1.2015	1.8230	1.5791		
ENSG00000169896	4.6850	5.3414	4.2335		
ENSG00000170458	1.7994	7.3247	6.5883		
ENSG00000173391	5.5397	9.3578	6.8193		
ENSG00000174130	2.3465	5.4248	2.2436		
ENSG00000265808	0.8769	1.1331	1.0208		

All of the commonly differentially expressed genes in cell cycle pathway were regulated in the same direction in all cell lines.

6.6 Configuration files

Sample configuration file for `rna-pipe`

```
#====================================
# Configuration file for running the RNA-seq pipeline
#====================================


#====================================

#            configuration

#====================================
# Only change the file version and verbose flags as you go
#====================================
config_UsingConfigFile=1                              # Don't change
config_ConfigFilename=$(basename $BASH_SOURCE)        # This file's name
config_ConfigFileVersion="1.5.0"                      # Update as version progresses
config_Verbose=1                                      # Leave on (1) for more verbose
↪  logs
config_ConfigFileChangeString="Rerunning U937 without LFCT"
#====================================


#====================================

#          module switch

#====================================
# Setting a module below to "1" means the module will be run
#====================================
config_Fastqc=0
config_Salmon=0
config_Deseq=1
#====================================


#====================================

#             general

#====================================
# Location of the directory containing all the raw reads ready for all pipeline
↪  operations
config_InputDirectory=/media/eunoia/SeagateBack/U937/PRJNA381911/
# Location of the directory in which we will create "*-pipe-output" directories to
↪  fill with the results
config_OutputDirectory=/home/eunoia/masters/data/rna-seq-paper-project/U937/
# Path to R executable
config_R_Executable=/usr/bin/Rscript
# Flag for searching for files in subdirectories
config_SubdirectorySearch=1
#====================================


#====================================

#           fastqc-pipe

#====================================
# Should we extract the zipped HTML page which fastqc creates? 1=true, 0=false
config_Fastqc_Extract=1
# How many threads should we let fastqc use?
config_Fastqc_Threads=4
# Should fastqc output be verbose?
config_Fastqc_Verbose=1
#====================================


#====================================

```

```
#             salmon

#====================================
config_Salmon_OverWrite=1                           # If we set this to 1, we will use the supplied salmonconfig
                                                    # file instead of the defaults in salmon's defaults.salmonconf file
                                                    # The config value below won't be used unless overwrite is TRUE
config_Salmon_OverWrite_ConfigFile="rna-seq-paper-project-test.salmonconf"
#====================================



#====================================

#             DESEQ2

#====================================
config_Deseq_OverWrite=1
config_Deseq_OverWrite_ConfigFile="rna-seq-paper-project-U937.deseqconf"
#====================================
```

Figure S9: **Sample configuration file for `rna-pipe`.**

Table S4: **Sample table for `DESeq2`'s differential expression analysis showing the data available for study design.** The options available for `DESeq2`'s *study design* parameter are `Replicate`, `Timepoint`, and `Timepoint + Replicate` (the order of terms doesn't matter).

Sample ID	Sample name	Replicate	Timepoint
SRR5429564	U937 cell line rep 1	1	0
SRR5429563	U937 cell line rep 2	2	0
SRR5429562	U937 cell line rep 3	3	0
SRR5429558	U937 cell line PMA treated @96h rep 1	1	96
SRR5429557	U937 cell line PMA treated @96h rep 2	2	96
SRR5429556	U937 cell line PMA treated @96h rep 3	3	96

Bibliography - Supplemental materials

Andrews, S. (2020a). *Duplicated Sequences*. FastQC Documentation. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/8%20Duplicate%20Sequences.html (visited on 11/10/2021).

Andrews, S. (2020b). *Per Base N Content*. FastQC Documentation. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/6%20Per%20Base%20N%20Content.html (visited on 11/10/2021).