Bioinformatic analysis of host cell gene expression and chromatin accessibility in response to *Chlamydia trachomatis* infection

Regan J. Hayward

University of Technology Sydney

Faculty of Science,

School of Life Sciences, The i3 Institute

Supervised by:

Associate Professor Garry Myers

Associate Professor Wilhelmina Huston

A thesis submitted in fulfilment of the requirements for the degree:
Doctor of Philosophy

May 2020

Certificate of original authorship

I, Regan Hayward, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Science, School of Life Sciences at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by an Australian Government Research Training Program.

Signature:

Production Note:
Signature removed prior to publication.

Date: 10th May 2020

Acknowledgements

The decision to return to university was based on a desire to challenge myself, but also to have fun and enjoy the experience along the way. This journey has spanned almost 8 years, and is now nearly complete. There have been plenty of high and lows over that time, but overall, I'm glad I made this decision and that I've been able to follow it through.

I've been fortunate to have the support from my amazing wife Giselle, who has listened, given advice, read through countless drafts and probably never wants to hear the words PhD or *Chlamydia* again! You have been amazing throughout this journey – thank you.

I'd also like to thank my family, who have supported in a range of different ways, particularly during these final months when they have had to travel internationally to allow me the time to focus on writing, especially since time has become a limited commodity with full time work.

I would like to thank my principal supervisor Garry Myers for the opportunity to work on this project and providing continuous support, guidance and mentorship over the course of my candidature. I'd also like to thank my co-supervisor Willa Huston, for her support and being available for many formal and informal chats. I would also like to thank all group members past and present for their support.

I would also like to say thank you to William King, Mark Thomas and Rami Mazraani who have been excellent colleagues and friends, who have made the course of my candidature all the more enjoyable.

Lastly, I'd like to thank the University of Technology Sydney and the i3 Institute for supporting me financially with a Doctoral Scholarship and the ability to travel to conferences locally and internationally.

Publications arising

Chapter 4 (Published)

Regan J. Hayward, James W. Marsh, Michael S. Humphrys, Wilhelmina M. Huston, Garry S. A. Myers. 2019: Early transcriptional landscapes of *Chlamydia trachomatis*-infected epithelial cells at single cell resolution: *Front. Cell. Infect. Microbiol*; vol 9: doi=10.3389/fcimb.2019.00392
Preprint: https://doi.org/10.1101/724641

Chapter 3 (Under peer-review and available as a preprint)

Regan J. Hayward, James W. Marsh, Michael S. Humphrys, Wilhelmina M. Huston, Garry S. A. Myers. 2019: Chromatin accessibility dynamics of *Chlamydia*-infected epithelial cells *bioRxiv*: https://doi.org/10.1101/681999
Preprint: https://doi.org/10.1101/681999

James W. Marsh, Regan Hayward, Amol Shetty, Anup Mahurkar, Michael S. Humphrys and Garry S. A. Myers. 2019. 'Dual RNA-Seq of *Chlamydia* and host cells', in Amanda Claire Brown (Ed.) *Chlamydia trachomatis*: Methods and Protocols. Humana Press. DOI: 10.1007/978-1-4939-9694-0

James W. Marsh, Regan J. Hayward, Amol C. Shetty, Anup Mahurkar, Michael S. Humphrys, Garry S. A. Myers; 2017, Bioinformatic analysis of bacteria and host cell dual RNA-sequencing experiments, *Briefings in Bioinformatics*, https://doi.org/10.1093/bib/bbx043
Preprint: https://doi.org/10.1101/098715

Table of contents

Certificate of original authorship ..... **i**

Acknowledgements ..... **ii**

Publications arising ..... **iii**

List of tables ..... **vi**

List of figures ..... **vii**

Abbreviations ..... **viii**

Abstract ..... **xii**

Chapter 1 Literature review **1**

1.1. *Chlamydia trachomatis* 1

1.2. Chlamydial biology 12

1.3. Next generation sequencing and bioinformatic analyses 18

1.4. Thesis summary 59

1.5. References 62

Chapter 2 Research methodology **93**

2.1. Laboratory-based methods and materials 94

2.2. Bioinformatic methods 96

2.3. References 102

Chapter 3 Chromatin accessibility dynamics of *Chlamydia*-infected epithelial cells **105**

3.1. Abstract 106

3.2. Introduction 107

3.3. Methods 109

3.4. Results and Discussion 114

3.5. Conclusions 140

3.6. Supplementary figures 143

3.7. Supplementary files 151

3.8. References 152

Chapter 4 Early transcriptional landscapes of *Chlamydia trachomatis*-infected epithelial cells at single-cell resolution **165**

4.1. Abstract 166

4.2. Introduction 167

4.3. Results 169

4.4. Methods 181

4.5. Discussion 186
4.6. Supplementary figures 191
4.7. References 199
Chapter 5 Comparative analysis using different MOIs from *Chlamydia*-infected epithelial cells 207
5.1. Introduction 208
5.2. Methods 211
5.3. Results 215
5.4. Discussion 233
5.5. Conclusion 236
5.6. Supplementary files 237
5.7. References 238
Chapter 6 General discussion and future directions 245
6.1. General discussion 246
6.2. Future directions 262
6.3. References 265

List of tables

Table 3.1: Summary of mapped reads, separated by time and condition 114
Table 3.2: Motifs and enriched transcription factors 135

List of figures

Figure 1.1:	Chlamydial species, primary hosts and *Chlamydia trachomatis* biovars	4
Figure 1.2:	Developmental cycle of *Chlamydia trachomatis* within a mucosal epithelial cell	13
Figure 1.3:	Bioinformatic analysis of genome sequencing data	23
Figure 1.4:	Gene expression-based research focused on chlamydial infection	28
Figure 1.5:	Bioinformatic analysis of RNA-seq data	32
Figure 1.6:	Bioinformatic analysis of single cell RNA-seq data	41
Figure 1.7:	Waddington's developmental landscape and epigenetic interactions	44
Figure 1.8:	Ways in which epigenetic changes can alter gene expression	49
Figure 1.9:	Bioinformatic analysis of chromatin accessibility data	57
Figure 3.1:	Identifying significant peaks and creating consensus peaksets	115
Figure 3.2:	Changes in chromatin accessibility throughout the chlamydial developmental cycle	117
Figure 3.3:	Annotation of significant peaks	119
Figure 3.4:	Differential chromatin accessibility within promoter regions	121
Figure 3.5:	Differential chromatin accessibility within enhancer regions	126
Figure 3.6:	Conserved host cell response to infection	129
Figure 3.7:	Enrichment of time-specific differential chromatin regions	131
Figure 4.1:	Experimental design and analysis	170
Figure 4.2:	Cell cycle classification	173
Figure 4.3:	Pseudotime analysis	176
Figure 4.4:	Differentially expressed genes and enriched pathways	180
Figure 5.1:	Experimental process and design	212
Figure 5.2:	Human and chlamydial mapped reads	217
Figure 5.3:	Chlamydial-based expression differences between depletion methods	219
Figure 5.4:	Host-based expression differences of protein coding and non-protein coding genes between depletion methods	222
Figure 5.5:	Top 25 expressed host and chlamydial genes across MOIs	226
Figure 5.6:	Comparison of differentially expressed host genes across MOIs	229
Figure 5.7:	Comparison of differentially expressed chlamydial genes across MOIs	232
Figure 6.1:	Annotation of significant peaks	254
Figure 6.2:	Overlapping time points	258
Figure 6.3:	Single cell sequencing methods	260

Abbreviations

3’UTR	3-Prime untranslated region
AB	Aberrant body
aRB	Aberrant reticulate body
ATAC-seq	Assaying for transposase-accessible chromatin sequencing
ATCC	American type culture collection
AutoML	Automated machine learning
BAM	Binary alignment map
CDC	Centre for disease control and prevention
ChIA-PET	Chromatin interaction analysis by paired-end tag sequencing
ChIP–seq	Chromatin immunoprecipitation followed by sequencing
COREs	Clusters of open regulatory elements
CpG sites	Cytosine and guanine appearing consecutively on the same strand
CRISPR	Clustered regularly interspaced short palindromic repeats
DE	Differentially expressed
DMEM	Dulbecco's modified eagle medium
DMR	Differentially methylated regions
EB	Elementary body
EGFR	Epidermal growth factor receptor
EMT	Epithelial-mesenchymal transition
ENCODE	Encyclopaedia of DNA elements
FACS	fluorescence-activated cell sorting
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements
FBS	Fetal bovine serum

FDR	False discovery rate
FRAEM	Fluorescence-reported allelic exchange mutagenesis
FRiP	Fraction of reads in peaks
GEO	Gene expression omnibus
GO	Gene ontology
HDACs	Histone deacetylases
HEp-2	Human epithelial type 2 cells
Hi-C	Method to examine chromatin interactions from a 3D landscape
HPI	Hours post infection
HtrA	High temperature requirement protein A
HVG	Highly variable genes
IF	Intermediate filament
IFC	Integrated fluidic circuit
IGS	Institute for genome sciences
IGV	Integrative genomics viewer
KEGG	Kyoto encyclopaedia of genes and genomes
KLF	Krüppel-like-factors
KNN	K-nearest neighbour
LGV	Lymphogranuloma venereum
lincRNA	Long intergenic non-coding RNA
miRNA	MicroRNA
miscRNA	Miscellaneous RNA
MNase-seq	Micrococcal nuclease sequencing
MOMP	Major outer membrane protein
MOI	Multiplicity of infection

MSM	Men who have sex with men
MT	Metallothionein
MT rRNA	Mitochondrial rRNA
MTOC	Microtubule-organizing centre
NAAT	Nucleic acid amplification test
NCBI	National Centre for Biotechnology Information
ncRNA	Non-coding RNA
NF-*κB*	Nuclear factor-κB
NGS	Next generation sequencing
NUE	Nuclear effector
PA	Phosphatidic acid
PBS	Phosphate-buffered saline
PCA	Principal component analysis
PID	Pelvic inflammatory disease
PolyA	Polyadenylated
QC	Quality control
RB	Reticulate body
RCA	Rolling circle amplification
ROS	Reactive oxygen species
Ribo-SPIA	RNA-based single-primer isothermal amplification
RLE	Relative log expression
ROS	Reactive oxygen species
RPKM	Reads per kilobase of transcript, per million mapped reads
rRNA	Ribosomal RNA
SC3	Single-cell consensus clustering

scBS-seq	Single cell bisulfite sequencing
scRNA-seq	Single cell RNA sequencing
snoRNA	Small nucleolar RNA
SPG	Succinic phosphate glycine buffer
sRNA	Small non-coding RNA
S.D.	Standard deviation
STI	Sexually transmitted infection
T3SS	Type III secretion system
TF	Transcription factor
TFI	Tubal factor infertility
TMM	Trimmed mean of M-values
TNF	Tumour necrosis factor
TPM	Transcripts per kilobase million
tRNA	Transfer RNA
TSS	Transcription start site
TU	Transcription unit
TTS	Transcription termination site
UMI	Unique molecular identifiers
UV	Ultraviolet
VIM	Vimentin
WHO	World health organisation

Abstract

Chlamydia are Gram-negative, obligate intracellular bacterial pathogens responsible for a wide range of human and animal diseases. In humans, *Chlamydia trachomatis* is the most prevalent bacterial sexually transmitted infection (STI) worldwide and is the leading cause of trachoma (infectious blindness) in disadvantaged populations. If left untreated, infections can lead to more complex disease outcomes including infertility, ectopic pregnancy, epididymitis, prostatitis, and pelvic inflammatory disease. Due to widespread rates of infection and disease around the world and the associated economic costs, chlamydial infections remain a serious public health concern. All chlamydial species are defined by their unique intracellular developmental cycle. However, this has been a significant barrier restricting traditional molecular microbial investigation, such as transformation. As a result, we still do not have a comprehensive understanding of chlamydial gene function, particularly secreted effector proteins that modulate many host cell interactions. In the absence of a reliable and efficient transformation system, next generation sequencing (NGS) approaches enable the recovery of genome-wide expression patterns from a chlamydial or host point of view to aid in uncovering these functions and interactions.

To help with further characterisation and identification of these host-chlamydial interactions, this work applied three novel NGS approaches using *in vitro* models of infection with *C. trachomatis*. Chapter 3 examines chromatin accessibility dynamics across the developmental cycle (1, 12, 24 and 48 hours) to identify epigenomic changes to host cells; Chapter 4 utilises single cell RNA-sequencing (scRNA-seq) from host cells to examine early developmental time points (3, 6 and 12 hours); and Chapter 5 simultaneously examines host and chlamydial expression (dual RNA-seq) from two time points (1 and 24 hours), with an experimental design aimed to examine different depletion techniques and to optimise the ratio of EBs per cell for infection models.

Examination of the host cell epigenome identified both conserved and distinct temporal changes genome-wide. Differentially accessible chromatin regions were associated with immune responses, re-direction of host cell nutrients, intracellular signalling, cell-cell adhesion, extracellular matrix, metabolism and apoptosis. Temporally enriched transcription factors identified a novel family of Krüppel-like-factors (KLFs) which are ubiquitously expressed in reproductive tissues and associated with a variety of uterine pathologies.

Analyses from scRNA-seq highlight infection-specific host cell biology, including two distinct clusters separating 3 hour cells from 6 and 12 hours. Pseudotime analysis identified a possible infection-specific cellular trajectory for *Chlamydia*-infected cells, and differential expression identified temporally expressed genes involved with cell cycle regulation, innate immune responses, cytoskeletal components, lipid biosynthesis and cellular stress.

Dual RNA-seq analysis showed that combining depletion methods (polyA and rRNA) increases the capture rate of chlamydial transcripts, but negatively impacts host-cell expression. Different MOIs (0.1, 1 and 10) highlighted that an MOI of 10 captures significantly more transcripts and is more beneficial for capturing chlamydial transcripts.

Overall, this work highlights the complex nature of chlamydial infections, uncovering novel biological functions and regulatory activities. These results and analyses also provide further considerations and improvements for future *in vitro* experiments, but also enable the application of these genome-scale techniques to the investigation of complex disease models *in vivo* and in human tissues *ex vivo*.

Chapter 1

Literature review

1.1. Chlamydia trachomatis

1.1.1. Introduction

Chlamydia trachomatis is an obligate intracellular, human-specific bacterial pathogen that causes trachoma and urogenital infections, including lymphogranuloma venereum (LGV). Trachoma is predominantly found in countries with populations that have limited access to basic public health infrastructure, and can cause irreversible blindness if left untreated (Burton and Mabey, 2009). Australia is the only high resource country with endemic trachoma, where it is predominantly found in remote Aboriginal communities (Lange et al., 2017). Sexually transmitted *C. trachomatis* infections are the most common bacterial sexually transmitted infection (STI) worldwide with 131 million new cases occurring annually according to the World Health Organisation (WHO) (WHO, 2016b). Disease outcomes disproportionately impact women, and in many cases are largely asymptomatic (~70%) (Menon et al., 2015). If left untreated, ascending infection can cause pelvic inflammatory disease (PID), ectopic pregnancy and in some instances infertility (Menon et al., 2015). Although less common, urogenital infections in men can spread to the testicles and cause epididymitis (Ostaszewska et al., 2000). LGV infections are more invasive and virulent, and predominantly localised to men who have sex with men (MSM) (White, 2009). If left untreated, bowel obstruction in severe cases can lead to loss of life (Mabey and Peeling, 2002). Although chlamydial infections are treatable with antibiotics, symptoms are not present in many cases and can lead to more complex disease outcomes. Due to the widespread rates of infection and disease around the world and the associated economic costs, chlamydial infections remain a serious public health concern.

1.1.2. History

Ancient civilizations as far back as 8,000 B.C have described disease manifestations similar to what is now called trachoma (Mohammadpour et al., 2016). The word trachoma originates from Greek (τράχωμα) which translates to 'roughness', and has been known by various other names throughout history including 'opthalmia' and 'aspiritudo' (Al-Rifai, 1988). There is speculation where trachoma originated from, with origins including the Middle East, specifically Egypt, or from Mongolia where invading nomads helped spread the disease westwards (Al-Rifai, 1988; Taborisky, 1952). The Ebers papyrus is one of the oldest preserved medical documents originating from ancient Egypt and dating back to 1550 B.C. It provides a detailed account of the disease, highlighting the persistent burden trachoma had on Egyptian society (The-Papyrus-Ebers, 1932). More recently in the 19th century, trachoma became heavily prevalent in Europe when soldiers were returning from the Napoleonic Wars, often from military camps with sub-standard hygiene levels. On their return, infected soldiers accelerated the spread across Europe (Larner, 2004). Improvements in hygiene and living standards helped reduce the epidemic, and by the early 20th century, the disease was essentially under control (Feibel, 2011). The discovery in 1938 that sulphonamide antibiotics can successfully treat trachoma helped to almost eliminate the disease from many countries including Europe and North America (Thygeson, 1939). However, in many developing countries with large populations living without access to reliable water or latrines, trachoma is still prevalent (Burton and Mabey, 2009). Surprisingly, Australia is the only high resource country that has not eradicated the disease, where remote indigenous communities with inadequate sanitation are still heavily burdened (Lange et al., 2017).

The identification of the underlying infectious entities within cells (inclusions) was discovered in 1907 by Halberstaedter and von Prowazek from conjunctival scrapings in orang-utans (Halberstaedter, 1907). Exhibiting virus-like characteristics, it wasn't until 1966

that the infection was correctly classified as bacterial in origin (Moulder, 1966). Shortly after this time, urogenital infections were identified and established as a sexually transmitted infection, which could lead to more complex disease outcomes (Dawson and Schachter, 1978; Schachter, 1977). Although the origins of the sexually transmitted strains of *Chlamydia* are unclear, cultural movements during the 1960-1970's with slogans such as "make love, not war", led to more casual attitudes towards sex and helped spread STIs (including *Chlamydia*) considerably. Genital strains have likely existed for a long time but have not been documented as well as trachoma, thus their natural history remains elusive (Szreter, 2019).

1.1.3. Classification/taxonomy

The genus *Chlamydia* belongs to the order *Chlamydiales* and the family *Chlamydiaceae*. To date, there are currently 13 taxonomically classified species and three *Candidatus* species recognised within the genus (Bommana and Polkinghorne, 2019). New species are discovered regularly, and have been more frequent with the increased use of genomic sequencing approaches.

Evolutionary studies have estimated that the order *Chlamydiales* diverged from a common ancestor approximately 700 million years ago, whereas chlamydial species have been co-evolving within their hosts over the last several million years (Horn et al., 2004). Despite the high similarity of chlamydial genomes (~99%), considerable variability exists within particular areas to distinguish each of the species. These include the commonly used 16S and 23S rRNA genes (Meijer et al., 1997), over 700 orthologous genes (Nunes and Gomes, 2014), the *omp*A gene (Brunelle and Sensabaugh, 2006), chlamydial-specific protein signatures (Griffiths and Gupta, 2002; Griffiths et al., 2006), and the plasticity zone, which is a highly variable region close to the origin of replication (Read et al., 2000).

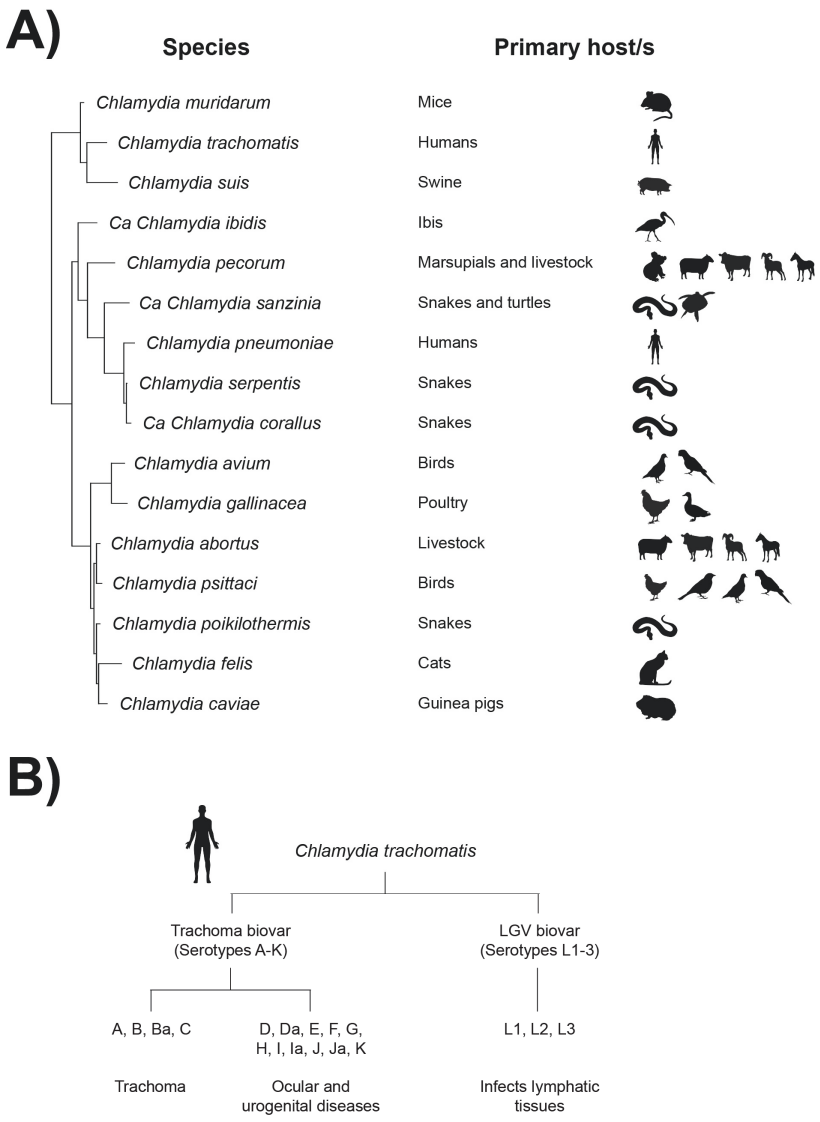


Figure 1.1: Chlamydial species, primary hosts and *Chlamydia trachomatis* biovars

A) There are currently 16 identified chlamydial species comprised of 13 taxonomically classified and three *Candidatus* species. The bacterium has evolved to infect a wide range of hosts, with multiple species being able to infect humans. **B)** *Chlamydia trachomatis* primarily infects human tissues that are characterised into different biovars and serovars exhibiting distinct disease outcomes. Adapted from (Phillips et al., 2019).

Chlamydial species can infect a wide range of hosts (**Figure 1.1A**). Within humans, *C. trachomatis* and *C. pneumoniae* are the primary pathogens, while other chlamydial species such as *C. abortus* and *C. pecorum*, have been able to co-evolve within and infect multiple hosts (Bachmann et al., 2014; Thomson et al., 2005). In addition, *C. abortus*, *C. psittaci* and *C. suis* can also cause zoonotic infections in humans upon exposure to infected animals or tissues (De Puysseleyr et al., 2017; Kieckens et al., 2018; Rohde et al., 2010).

Everett et al (Everett et al., 1999) proposed to split the genus *Chlamydia* into two (*Chlamydia* and *Chlamydophila*) on the basis of 16S sequence data. However, after much debate, the proposal was dismissed, and a single genus remains. The underlying reasons for the reversion included that the proposal was based on limited sequence data, and that it ignored the highly conserved and unique biology shared by all chlamydial species (Sachse et al., 2015; Stephens et al., 2009).

C. trachomatis is further divided into 18 different serovars, based on serological responses to the major outer membrane protein (MOMP). These 18 serovars are broadly categorised into two different biovars on different phenotypic characteristics. The trachoma biovars (A-C) predominantly infect ocular epithelial cells, while the D-K serotypes cause both ocular and urogenital infections. The lymphogranuloma venereum (LGV) biovars have serotypes of L1-3 and infect lymphatic tissues (Stephens et al., 1982; Wang et al., 1985) (**Figure 1.1B**). There are limitations however to using MOMP for characterisation, as serotypes are not reflective of disease presentation or virulence (Byrne, 2010). Ideally identification would be based on taxonomic clustering from whole genomes (Seth-Smith et al., 2009), but this isn't currently feasible due to the costs involved and the high number of clinical samples examined from hospitals and screening programs. Interestingly, serovar B strains are able to infect both ocular and genital epithelial cells, but have only been found in rare cases (Caldwell et al., 2003). Furthermore, a recent study identified trachoma-based isolates that were phylogenetically

placed within the urogenital clade, highlighting that tissue tropism can be dependent on only a small number of genes (Andersson et al., 2016).

1.1.4. Disease manifestations

1.1.4.1. Ocular infections

C. trachomatis serotypes A-C cause ocular infections in humans, commonly referred to as trachoma. Early symptoms are discharge from an infected eye and can resemble conjunctivitis (Prost and Négrel, 1989). The bacteria can be easily spread if the discharge comes into contact with clothes, towels, hands, eye seeking flies, or even from coughing or sneezing (CDC, 2019b; WHO, 2016a). Additionally, the bacteria can be passed from a mother infected with a *C. trachomatis* genital infection directly to a new-born infant, termed neonatal conjunctivitis (Darville, 2005).

If caught early, infection can be eliminated by oral antibiotics (WHO, 2016b), but if left untreated, can eventually lead to irreversible blindness. Loss of sight is caused by repeated or long-term infections that ultimately induce scarring inside the eyelid. Scar contraction causes the eyelashes to turn in (entropion), abrading the corneal surface and causing large amounts of pain (Burton, 2009). As of early 2019, WHO estimates an estimated 1.9 million people are visually impaired by trachoma, and 142 million people are at the risk of blindness (WHO, 2019).

1.1.4.2. Urogenital infections

Serotypes D-K are primarily linked to STI. *Chlamydia* is the most common bacterial STI worldwide, with latest data from the Centre for Disease Control and Prevention (CDC) reporting 1.7 million cases (a 22% increase since 2013) - the highest infection rates are in

woman, and in people below the age of 24 years (CDC, 2017). In Australia, *Chlamydia* is the most commonly reported bacterial STI, with 110,775 identified cases in 2017; people younger than 30 have the greatest risk of infection (Kirby-Institute, 2018). These numbers however are likely to be underestimated, due to high rates of undiagnosed and/or asymptomatic infections (CDC, 2019a). Symptoms vary between men and woman, but asymptomatic infections are common and if left untreated, can lead to more complicated infections and diseases (Menon et al., 2015). However, if caught early, oral antibiotics can be used to treat the infection (WHO, 2016b).

In woman, an infection can cause inflammation of the cervix (cervicitis); symptoms include vaginal discharge, abnormal bleeding, and pain or a burning sensation when urinating (CDC, 2019a). If the infection is left to spread into the upper genital tract, outcomes include PID, salpingitis, scarring and occlusion, which may result in an ectopic pregnancy or even result in future infertility (Haggerty et al., 2010; Oakeshott et al., 2010). The infection can also be passed on to unborn infants from an untreated mother, and lead to neonatal conjunctivitis and pneumonia (CDC, 2019a). In men, an infection can cause inflammation of the urethra (urethritis); symptoms include discharge from the penis and pain or a burning sensation when urinating, or pain and swelling in the testicles (CDC, 2019a). If the infection is left to spread into the testicles, it can lead to epididymitis, which in rare cases can lead to sterility (Bebear and de Barbeyrac, 2009).

In rare cases, infection can lead to Reiter's syndrome, also known as reactive arthritis. This autoimmune disease is associated with a genetic predisposition and appears to more common among men than women (Rahman et al., 1992). Inflammation of joints (commonly knees, ankles and feet) develops in response to infection from other parts of the body, particularly the urogenital or gastrointestinal tract (Amor, 1983). Antibiotic treatments exist, however in some patients the arthritis can become chronic (Gaston, 2000).

1.1.4.3. Lymphogranuloma Venereum (LGV)

Serotypes L1-3 induce LGV; these serotypes are more invasive and more virulent compared to serotypes A-K (White, 2009). LGV is generally an uncommon STI, with infection rates that are generally between MSM, and in much fewer instances, in woman (CDC, 2015b). Although generally found in countries such as India and South America, in the last ten years LGV infections have begun to be recognised in Europe, the United Kingdom, North America and Australia (Kapoor, 2008; Stark et al., 2007). Infection occurs during sexual activity, where the bacterium predominantly infects macrophages and monocytes, allowing passage through the epithelial surface to regional lymph nodes, and can cause systemic disease (Mabey and Peeling, 2002). If diagnosed early, oral antibiotics can be used to treat the infection (WHO, 2016b). However, if untreated, disfiguration from ulceration, enlargement of external genitalia, and possible lymphatic obstructions may result (Stoner and Cohen, 2015).

LGV occurs in three stages. Stage 1 consists of a primary lesion, which is generally an unnoticed genital papule, pustule or ulcer that is painless and rapidly heals. Stage 2 occurs 2-6 weeks after the primary lesion and consists of a painful inguinal lymphadenopathy. Stage 3 which can occur a number of years after the original infection, is identified by proctocolitis, and is more common in MSM than woman (Meyer, 2016; Stoner and Cohen, 2015).

1.1.5. Detection and treatment

1.1.5.1. Detection

The first step in detecting a *C. trachomatis* infection requires obtaining a sample with a swab from the infected area, by either an invasive or non-invasive method (Haugland et al., 2010). Non-invasive methods such as first-void urine tests, can be collected privately, whereas invasive methods (which are becoming less common), such as cervical, urethral or vaginal swabs are generally carried out at a medical establishment (Chernesky, 2005; WHO, 2016b). Since invasive methods extract specimens directly from the environment, infections can be detected through culturing or by antigen or nucleic acid tests (Carlson et al., 2008). Non-invasive methods however require a more sensitive Nucleic Acid Amplification Test (NAAT) due to the bacteria being difficult to culture outside of their preferred cellular environment (Chernesky, 2005). Today, NAATs are the preferred method as they are highly sensitive and specific, and samples can be obtained non-invasively (CDC, 2014; Meyer, 2016).

1.1.5.2. Treatment

A diagnosed *Chlamydia* infection can be treated with a course of oral antibiotics. Treatment options are currently azithromycin, which is usually prescribed in a single large dose, or a seven day course of doxycycline (CDC, 2015a; WHO, 2016b). Other antibiotics, such as levofloxacin, erythromycin and ofloxacin may also be employed with varying doses depending on the chlamydial serovar, severity of the infection, if the patient is pregnant, or existing conditions and medications cannot be taken together (CDC, 2015a; WHO, 2016b).

1.1.6. Chlamydial vaccine

Although antibiotics can be used for treatment, three out of four infections are asymptomatic and as a result are often untreated (Newman et al., 2015). Untreated or repeat infections are the main driving force behind *Chlamydia*-associated morbidity and have negatively contributed to being able to control infection rates (Davies et al., 2016). The best way to control the epidemic would be through a preventative vaccine that could provide protective immunity against *C. trachomatis*, alleviating the current burden many people face. However, developing an effective human-based vaccine has remained elusive even after 220 reported vaccine trials that have examined a range of different chlamydial species and different hosts (Phillips et al., 2019). The main challenge has been from our incomplete knowledge of the underlying infection-based mechanisms, thus making successful vaccine targets difficult (Davies et al., 2016). A further challenge is that the majority (85%) of vaccine trials designed for humans have been performed in different hosts. These substitute hosts have been used as they are generally easier to obtain, administer, monitor and examine in detail (Phillips et al., 2019). The most targeted chlamydial species has *been C. muridarum* reflecting the predominance of mouse models for vaccine development (77 trials), followed by *C. trachomatis* (67 trials), but with a range of different hosts including mice, non-human primates, pigs, guinea pigs and rabbits (Phillips et al., 2019). The advantage of using mice and guinea pigs as surrogate models is that their disease pathology reflects diseases in humans (Darville and Hiltke, 2010; Miyairi et al., 2010). However, most of the current vaccine-based knowledge has come from these surrogate models, and the main difficulty has been replicating results in other host species (Phillips et al., 2019).

Across the 220 vaccine trials which have spanned over 70 years, a range of methods and techniques have been tried but ultimately not succeeded. These include different delivery sites (systemically and/or mucosally), different antigens (live attenuated or inactivated bacteria,

polymorphic membrane proteins, the major outer membrane protein, heat shock proteins and the chlamydial protease-like activity factor), all with and without different adjuvants (de la Maza et al., 2017; Hafner et al., 2014; Poston et al., 2017). To date, all of this research and knowledge has helped produce two commercially available animal-based chlamydial vaccines targeting *C. felis* in cats and *C. abortus* in sheep (Phillips et al., 2019). In humans, a recent breakthrough phase 1 trial was shown to be safe and tolerated, and produce antibodies against *C. trachomatis* (Davies et al., 2016). This is the first in-human trial, and although still in its early stages, the signs are encouraging.

1.2. Chlamydial biology

1.2.1. Elementary bodies and reticulate bodies

All members of the Order *Chlamydiales* are obligate intracellular Gram-negative bacteria that share a unique biphasic developmental cycle. In an uninterrupted cycle, the bacterium differentiates between two forms: an extracellular form called the elementary body (EB) and an intracellular form called the reticulate body (RB) (Matsumoto, 2019). Each form of the bacteria is morphologically and functionally distinct. EBs are able to survive in the harsh extracellular environment, they are infectious, have a tightly packed nucleoid, are morphologically similar with a diameter 0.3 μm, and have limited metabolic activity (Grieshaber et al., 2018). RBs are intracellular and more fragile, their nucleoid is less compacted, are larger and vary more in size with an average diameter of 1 μm (AbdelRahman and Belland, 2005; Omsland et al., 2014). RBs are metabolically active but do not have the capability to create all the proteins required throughout the developmental cycle. To account for these deficiencies, nutrients from the host's cell cytoplasm, are obtained, which are primarily used for replication, but also allow a wider range of activity that includes general protein synthesis, additional nutrient transport mechanisms and defence (Bastidas et al., 2013; Saka et al., 2011).

1.2.2. The development cycle

C. trachomatis infection begins when an EB attaches and enters a suitable host cell through endocytosis (Elwell et al., 2016). Once inside the cell, the EB will remain within this membrane bound vacuole, which has been termed an inclusion. Within the first 2 hours of internalisation, the bacteria will transform from an EB into an RB. This involves the reduction of the disulphide-linked outer membrane, decondensation of the nucleoid, and the bacteria

increasing in size (AbdelRahman and Belland, 2005; Cocchiaro and Valdivia, 2009). Once transformed, the inclusion provides the EB an intracellular niche from which bacterial protein synthesis can begin. In the early stages (~6-8 hours), the RBs begin to replicate through a budding process (Abdelrahman et al., 2016), and if the conditions within the host cell are favourable, will continue to replicate through the mid-cycle stages (~8-24 hours) (Brunham and Rey-Ladino, 2005). In the late-cycle stages (~24-72 hours), RBs will asynchronously transition into EBs, and through either extrusion or host cell lysis, are released; providing new infectious bodies that can begin the cycle again (Hybiske and Stephens, 2007).

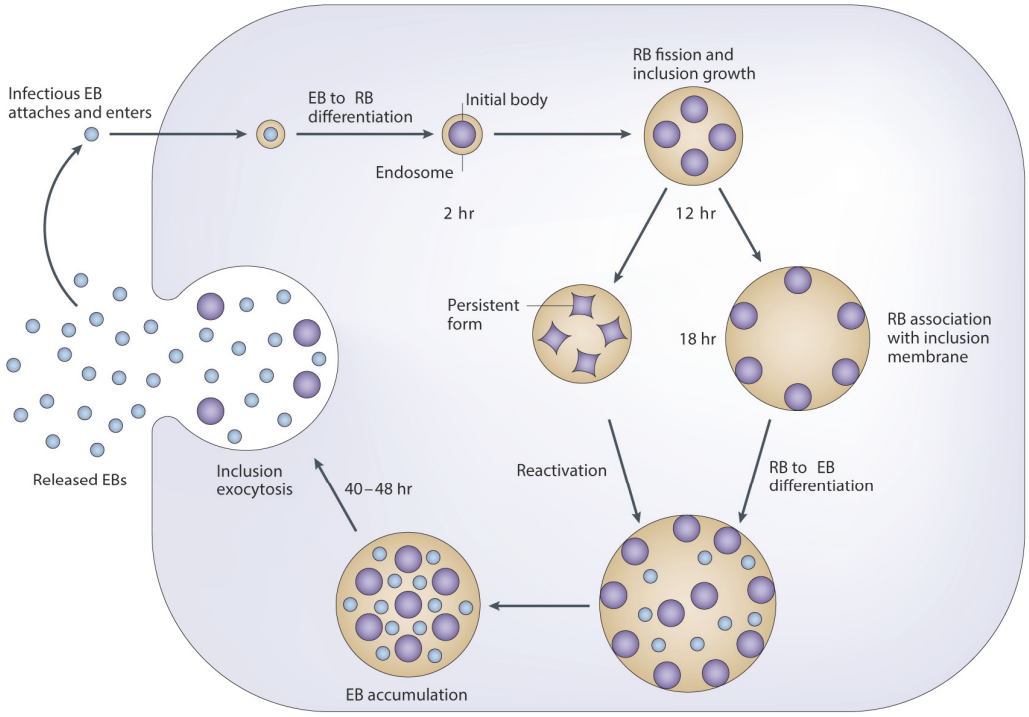


Figure 1.2: Developmental cycle of *Chlamydia trachomatis* within a mucosal epithelial cell

Infection begins when an elementary body (EB) attaches and enters the cell, creating an inclusion. In the early stages the EB differentiates into a reticulate body (RB). The RB will continue to replicate into the mid-cycle stages. The inclusion during this time can take over

most of the size of cell if the conditions are favourable. In the late cycle stages, RBs asynchronously transition back into EBs, and are released through cell lysis or exocytosis. In the presence of non-favourable conditions, RBs can transform into a non-replicating aberrant form (shown here as a persistent form). If the conditions become favourable again, the bacterium will revert back into the RB form and continue in the same late-cycle stages. Reproduced from (Brunham and Rey-Ladino, 2005).

1.2.3. Persistence and recurrence

If conditions are not favourable during the early to mid-cycle stages, the bacterium may transform into a non-replicating persistent form. This is either called an aberrant body (AB) (Brunham and Rey-Ladino, 2005), or an aberrant RB (aRB) (Bavoil, 2014). Non-favourable conditions can be induced from environmental factors, various stressors that include nutrient deficiency (Mpiga and Ravaoarinoro, 2006), the presence of host-factors such as interferon-γ (Kazar et al., 1971), in addition to antibiotics and cytokines which interfere with cell wall synthesis (Mpiga and Ravaoarinoro, 2006). The aberrant form provides a useful way of remaining dormant and avoiding the host immune system. Within this 'persistent' state, *Chlamydia* continues to transcribe genes but slows down DNA replication; it also stops dividing, becoming viable, but non-cultivable (Muramatsu et al., 2016; Ouellette et al., 2006). If the conditions do become favourable again, the aberrant form can transform back in to RBs and follow the same late stage cycle as previously discussed (Brunham and Rey-Ladino, 2005).

Persistent states and resulting aberrant forms have been widely identified and examined from *in vitro* studies, whereas *in vivo* studies have been limited in number and contained small

sample sizes (Panzetta et al., 2018). Although it is likely to occur within human infections, there is still limited evidence for persistence *in vivo*.

1.2.4. Infection mechanisms

All chlamydial species encode a Type III secretion system to release proteins that interact with host cells. Secreted proteins are initially used by EBs to help modify the cytoskeleton of host cells allowing uptake, entry, and separation from degradative pathways (Elwell et al., 2016; Nans et al., 2015). Once the inclusion is formed, synthesised proteins are released and either inserted into the inclusion membrane (Incs), or released into the host cell cytosol (effector proteins). Different effectors and Inc proteins are expressed throughout the course of infection and at different developmental stages that assist growth and development (Kleba and Stephens, 2008; Valdivia, 2008). An example during the early stages of infection are Incs that interact with host cell dynein, utilising motor proteins that move the inclusion along microtubule filaments towards the microtubule-organizing centre (MTOC) (Grieshaber et al., 2003). Once in this peri-nuclear location, effector proteins are in closer proximity to the Golgi apparatus and can redirect exocytic vesicles containing sphingomyelin and cholesterol, which are essential for growth (Hackstadt et al., 1996; Scidmore et al., 1996). Additionally, the inclusion engages with other host organelles, including the endoplasmic reticulum (Derré, 2015), mitochondria (Liang et al., 2018), the cytoskeleton (Molloy, 2014), and the nucleus (Hobolt-Pedersen et al., 2009). Interactions provide the acquisition of amino acids, iron, lipids, energy metabolites, dampening pro-apoptotic signals (Betts-Hampikian and Fields, 2010; Elwell et al., 2016), and avoiding host cell defences (Bastidas et al., 2013; Redgrove and McLaughlin, 2014).

1.2.5. Genomic manipulation

An optimal way to comprehensively understand the chlamydial bacterium and its requirements for pathogenesis and growth is to attempt to understand the functionality and interactions of all the encoded proteins. In traditional bacterial models such as *Escherichia coli*, methods that include transposon mutagenesis, are used to inactivate genes to understand their importance in pathways and biological processes, such as virulence (Shuman and Silhavy, 2003). However, in *Chlamydia*, traditional molecular methods have not yet been possible due to the chlamydial intracellular developmental cycle, as the host cell, the inclusion, and bacterial membranes all act as barriers against genetic manipulation (Hooppaw and Fisher, 2016). With these restrictions, chemical mutagenesis methods have been used, allowing forward and reverse genetic screening of mutant libraries in combination with whole genome sequencing (Kari et al., 2011; Kokes et al., 2015; Nguyen and Valdivia, 2012). Other methods include site-specific group II intron insertion to inactivate genes (Johnson and Fisher, 2013). Transformation systems in *Chlamydia* have also been hindered by these intracellular barriers. Ideally, genetically transformed mutants are grown and monitored based on predicted or unknown genetic changes. As a result, recombinant DNA vectors are constructed to propagate in two different hosts (shuttle vectors), and are only a recent development (Hooppaw and Fisher, 2016). Currently, there are various recombinant shuttle vectors that can provide stable transformants, allowing reverse genetic approaches such as homologous recombination (Hooppaw and Fisher, 2016; Wang et al., 2011). A further transformation method utilising a chlamydial suicide vector with Fluorescence-Reported Allelic Exchange Mutagenesis (FRAEM), has been able to generate mutant populations with targeted gene deletions (Mueller et al., 2016). More recently, a landmark study was able to apply transposon mutagenesis to *C. trachomatis*, generating single-insertion mutant clones and discovering a homolog likely involved with lateral gene transfer (LaBrie et al., 2019). Furthermore, a

FACS-based CRISPR screen of host protein coding genes helped to identify key molecules promoting *C. trachomatis* invasion and further our understanding of the underlying host factors (Park et al., 2019). With the discovery of these new methods, combined with the wealth of genome-wide NGS data that has recently accumulated, we may begin unravelling many of the current mysteries surrounding chlamydial pathogenesis.

1.3. Next generation sequencing and bioinformatic analyses

1.3.1. Genome sequencing

1.3.1.1. Chlamydial overview

Owing to the small size of the *Chlamydia* spp. genome (~1Mbp) and the global burden of infection and disease, *C. trachomatis* (serovar D) was one of the first ten bacterial genomes to be completely sequenced following the advent of whole genome sequencing (Stephens et al., 1998). Since then, over 150 chlamydial genomes have been sequenced, encompassing numerous species and many clinical strains (Hooppaw and Fisher, 2016). Chlamydial species often carry a highly conserved plasmid (~7-7.5kbp) that has also been sequenced, with functional associations with virulence and pathogenicity (Zhong, 2017). With constant improvements to sequencing chemistry and technology, the accuracy of base calls is now extremely high, thus providing an accurate method for genome identification, particularly for bacteria (Quainoo et al., 2017). As a result, new species and hosts are constantly being discovered, broadening our understanding of *Chlamydia* and chlamydial-like bacteria (Phillips et al., 2019; Taylor-Brown et al., 2015). The wealth of genetic material that has accumulated has revealed that *Chlamydia* spp. are closely related, sharing a large pool of conserved genes, with differences attributed to the varying hosts and tissues (Sigalova et al., 2019). Genome similarity of strains within the same species such as *C. trachomatis*, is as high as ~99%, with the ~1% allowing different disease manifestations such as trachoma and urogenital diseases (Brunelle et al., 2004; Carlson et al., 2004).

1.3.1.2. Sequencing overview

Genome sequencing can examine either specified genes or whole genomes, and has been applied across all domains of life (Harris et al., 2003). Specified genes require custom primers that isolate the specified gene fragments to be captured, while whole genome sequencing uses universal adaptors to capture as many fragments as possible. When examining specific genes such as the rRNA genes, then likely outcomes are either species identification or will be phylogeny-based (Janda and Abbott, 2007). Although more expensive, whole genome-sequencing provides an increased resolution allowing genome characterisation and comparative analyses (Besser et al., 2018).

1.3.1.3. Bioinformatic analysis

Phase 1 - Quality control: Sequenced reads are presented in Fastq files containing nucleotide base calls and the corresponding accuracy of each base, which are known as Phred quality scores. Phred scores are logarithmically linked to error probabilities, where 10=90% accuracy, 20=99%, 30=99.9% etc. Default cutoffs are generally set at 30, ensuring a high confidence of base calls (Bolger et al., 2014). Initial quality control steps remove reads that have low Phred scores, sequencing adaptors that were added in library preparation, and reads that are below a minimum specified length (Bolger et al., 2014). Further quality control (QC) steps can identify sources of contamination, highlighting when samples may not be usable (Wingett and Andrews, 2018) (**Figure 1.3**). Common QC software includes Trimmomatic (Bolger et al., 2014), Cutadapt (Martin, 2011) and FastQC (Andrews, 2010). These QC steps and tools are not specific to genome sequencing, and can are applied to all NGS datasets. When experiments generate long-reads, the underlying sequences are stored in FAST5 files. Although these short-read tools can be used for QC, more specialised tools are available such as PoreTools (Loman and Quinlan, 2014) and NanoPack (De Coster et al., 2018).

Phase 2 – Genome assembly: During assembly, multiple assemblers are generally run, often with varying parameters to ensure the final assembly is the best representation of the original genome. The choice of software and parameters depend on two main factors: 1) Are short-reads (< 400bp) or long-reads (> 400bp) being assembled. 2) Is there an underlying reference genome that can assist with assembly (reference-based), or not (*De novo*). Ultimately, for the best genome assembly, high quality reads and sufficient sequencing depth is needed (~50x for bacterial genomes (Pightling et al., 2014)), and generally the longer the reads the better. Unfortunately, current sequencing technologies cannot produce all three. Illumina-based reads are high quality, but are often too short at < 400bp (Tan et al., 2019). Longer reads from Oxford Nanopore or Pacific Biosciences can reach upwards of 10kbp, but have a high rate of sequencing errors (5-15%) (Rang et al., 2018). To circumvent these issues, hybrid approaches combining short and long reads are becoming more popular, particularly when reconstructing more complex genomes such as polyploids, or to help determine highly repetitive genomic regions (Dominguez Del Angel et al., 2018).

In instances when no primary genome is available, similar genome references from closely related species are frequently used (Lischer and Shimizu, 2017). Many of the alignment tools used for reference-based assembly were designed specifically for either short or long-reads. Recent adaptations (BWA and BWA-mem (Li, 2013)) or the inclusion of parameter adjustments (Bowtie2 and SOAP2) allow dual use, but often require much longer runtimes (Lindner and Friedel, 2012; Shang et al., 2014). If the reference is highly similar, standard alignment tools will likely be sufficient as they can capture small variants between the reference and the new assembly, such as inserts, deletions and gapped alignments (Langmead and Salzberg, 2012). It should be noted that transposable elements (DNA sequences that can change position within a genome) if present may disrupt alignment algorithms, and specific tools have been developed for their identification such as RepeatMasker (Tarailo-Graovac

and Chen, 2009). Ultimately, reference-based genome assemblies are only as accurate as the underlying reference. When the reference is not similar, and large segments of the aligned genome are missing combined with many unmapped reads, then a *de novo* assembly step is often included to increase resolution.

De novo genome assemblies require more complicated algorithms due to the lack of any reference. Many of these algorithms are based on *de Brujin* graphs that construct contigs (continuous sequences) based on small overlapping sequences of a specific length (k), called k-mers. Different software allow different k-mer sizes and different input parameters, making it important that the assembly step is repeated with different software and parameters to find a consensus of contigs that can be merged into scaffolds. Comparisons can be performed with software such as QUAST (Gurevich et al., 2013), which provides metrics for evaluation and to help pick the best assembly. *De novo* software that are bacterial specific include Spades (Bankevich et al., 2012) and SMALT (sanger.ac.uk/resources/software/smalt/), both of which are frequently used in chlamydial genome studies (Hadfield et al., 2017; Harris et al., 2012; Sigar et al., 2014). Hybrid *de novo* assemblies will often generate a *de Brujin* graph assembly using short reads, and then long reads are used to improve assembly by closing gaps, re-ordering contigs and resolving repetitive regions (Wick et al., 2017). Hybrid software tools include LoRDEC (Salmela and Rivals, 2014) and Unicycler, with Unicycler developed specifically for bacterial studies (Wick et al., 2017). Chlamydial studies have not utilised long-reads technology due to the small size of the species genome (~1Mbp), where short reads can be sufficiently assembled.

Phase 3 – Accuracy of assembly: There is no specific tool or metric to determine if the final assembly has been constructed without any errors. However, statistical methods can help to identify potential problems and give an overall evaluation. N50 is a widely used metric to evaluate an assembly, where low numbers represent a high degree of fragmentation and high

numbers indicate less fragmentation and a more contigious assembly. To calculate, all contigs and scaffolds are aligned longest to shortest. Starting from the longest, the lengths are summed until the running total equals one-half of the total length of all the contigs and scaffolds in the assembly (Alhakami et al., 2017). A further consideration to increase the accuracy of short reads assemblies is the underlying sequencing technology used to capture genomic fragments. These include single-end reads, paired-end reads and strand-specific paired-end reads. Generally, the more detail that can be captured from each fragment the better the assembly will be. Single-end runs provide an economical alternative, but lack orientation and information from the opposite end of fragments. Capturing strand-specific paired-end reads results in information from 5' and 3' fragment boundaries and the fragments correct orientation, greatly assisting assembly algorithms (Khan et al., 2018). This is equally beneficial for *de novo* transcriptome assemblies to generate accurate transcriptome maps (Vivancos et al., 2010).

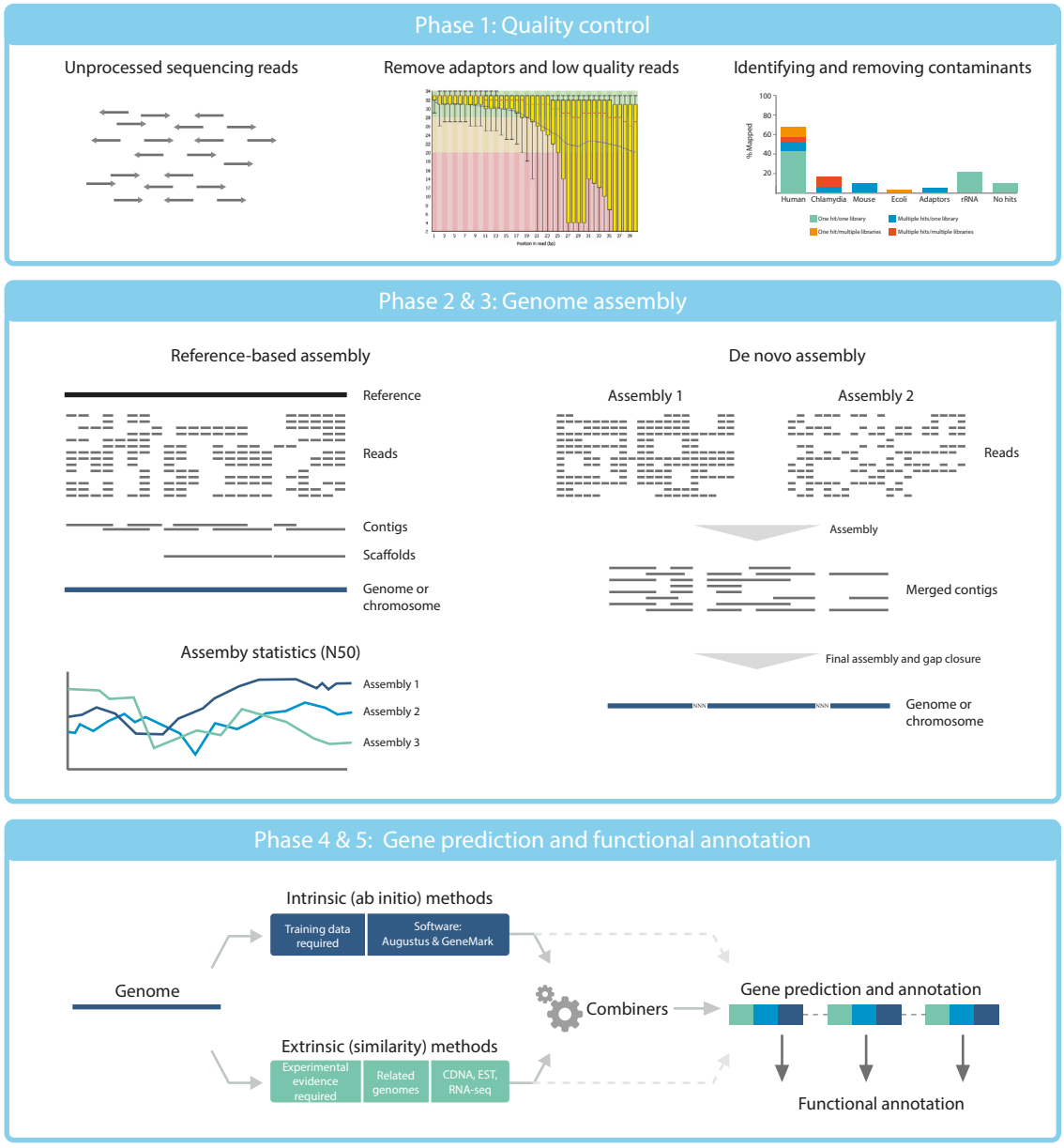


Figure 1.3: Bioinformatic analysis of genome sequencing data

Phase 1) Quality control is performed on unprocessed reads after sequencing. Steps include removing sequencing adaptors and low-quality reads. Additionally, reads can be compared against different genomes to identify contaminants. **Phase 2)** Reads can either be aligned to a reference genome, or *de novo* assembled if no suitable reference exists. Overlapping reads create contigs (contiguous segments), which can be merged to create scaffolds and ultimately genome assemblies. **Phase 3)** Statistical methods are used to identify any

assembly-associated problems or compare different assemblies. N50 is a commonly used metric to evaluate a single assembly or different assemblies for fragmentation. **Phase 4 & 5)** Gene prediction, annotation and associated biological functions can be performed through either intrinsic and/or extrinsic methods. If both methods are used, then combiner software can merge the results and provide a more comprehensive and often more accurate genetic prediction and annotation.

Phase 4 – Gene prediction and annotation: The process of accurately determining the location and structure of genes is well understood, particularly when identifying protein-coding genes which have primarily been the main focus in older studies, and are thus easier to characterise (Dominguez Del Angel et al., 2018). In general, there are three main approaches involved with gene annotation that include intrinsic methods (also called *ab initio*), extrinsic methods and combiners (Yandell and Ence, 2012). Intrinsic methods require either development or training of statistical models which use existing data for gene prediction, such as Augustus (Stanke et al., 2008) and GeneMark (Besemer and Borodovsky, 2005). Disadvantages include that the training data must be accurate and relevant, where advantages include the capability of predicting novel and fast evolving species specific genes (Dominguez Del Angel et al., 2018). Extrinsic methods rely on public data repositories (NCBI non-redundant protein database, RefSeq and UniProt), making them more universally applicable and generally easier to run. Existing sequence data such as RNA-seq can generate transcript maps that can also be utilised in extrinsic methods to increase resolution (Dominguez Del Angel et al., 2018). The job of a combiner such as EuGene (Schiex et al., 2000) and Maker (Campbell et al., 2014), is to integrate information from the intrinsic and/or extrinsic methods to provide accurate gene models, which are ultimately only as accurate as the input data. Each method can be run separately without the use of a combiner if required,

or if suitable associated data is available. Specialist annotation tools exist for bacteria and prokaryotes as standalone tools such as Prokka (Seemann, 2014) and DFAST (Tanizawa et al., 2017), and generally can be run on desktop computers in < 10 minutes due to the small size of their genomes.

Phase 5 – Functional annotation: Depending on experimental outcomes, assigning functional information to genes and proteins will likely be required. Many of these tools such as Uniprot (Apweiler et al., 2004) are available online, where nucleotides from either a single gene or a Fasta file comprising whole genomes can be uploaded and analysed. Further examples include SignalP and TargetP, which can predict signal peptides and the subcellular locations of proteins (Emanuelsson et al., 2007); while tools such as BLAST, HMMER and LAST can perform sequence similarity matches to associate gene names, IDs and closely related species in a high throughput manner (Dominguez Del Angel et al., 2018). To assist in chlamydial annotation and functional analysis, the online chlamydial-specific database ChlamBase (Putman et al., 2019), can be used to characterise evidence-based genes from multiple species and strains that include *C. trachomatis*, *C. muridarum* and *C. pneumoniae*.

1.3.2. Transcriptomics

1.3.2.1. Overview

Transcriptomics provides a genome-scale method to capture RNA (transcripts) within cells, and has provided insights into gene expression from an incredibly wide range of organisms and species (Stark et al., 2019). Early genome-scale approaches used arrays containing DNA oligonucleotide 'probes' that are complementary to a set of target genes. A typical array can contain up to 1 million probes that represent different genes or known variants of these genes (Liu et al., 2010). Fluorescently labelled target RNA is washed over the array, allowing matching transcripts to hybridise to each probe. As more target RNA binds to a probe, the emitted signal intensifies; capturing expression levels within a sample and allowing direct comparisons between conditions (Bumgarner, 2013). The main drawback to using microarrays is that a known gene sequence is needed to generate a probe. Tiling arrays solved this issue by allowing novel genes to be identified (through custom probes), but both methods are still limited by the number of probes that can be placed on a chip (Mockler et al., 2005; Siezen et al., 2010).

RNA-seq was developed in 2009 (Wang et al., 2009) and has since become the method of choice for examining gene expression. The main advantages include high quality single base-pair resolution combined with high throughput sequencing that was not limited by the number of probes. Initial protocols used oligo(dT) primers that only captured transcripts with polyadenylated tails, such as mRNA. Later, randomly designed primers allowed the capture of a greater range of transcripts, making it suitable for organisms where transcripts are not polyadenylated, such as bacteria (Armour et al., 2009). (Wang et al., 2009)Advancements in protocols and sequencing platforms now allow the capture of many types and transcripts from the 5' end (Adiconis et al., 2018), 3' end (Gruber et al., 2016), full length transcripts (Ju et al., 2019), and of late, RNA-sequencing directly within a cell (Depledge et al., 2019).

RNA-seq has been successfully applied to *Chlamydia*-infected cells and has been a highly useful approach to examine gene expression from the host (Alvesalo et al., 2008; Vasileva et al., 2018; Wang et al., 2013), from the chlamydial bacterium (Brinkworth et al., 2018; Carlson et al., 2008; Song et al., 2013), and simultaneously from both organisms (Humphrys et al., 2013). In the absence of an easily applied transformation system, genome-wide expression studies have been useful for enhancing our understanding of infection related mechanisms, pathways and significant gene sets.

1.3.2.2. Chlamydial-specific gene expression sequencing overview

Using the search criteria with a combination of "*Chlamydia*" + "transcriptome", transcriptomics, "RNA-seq", "microarray", "arrays" or "NGS", from Pubmed and bioRxiv, resulted in 77 gene expression-based studies. The collation of these results highlight that that the number of publications is consistently increasing (**Figure 1.4A**), with three main chlamydial species being examined *C. trachomatis* (63%), *C. pneumoniae* (19%), and *C. muridarum* (12%) (**Figure 1.4B**). The proportion of host-based research is nearly 3x higher than chlamydial-based research (**Figure 1.4C**), and in the last 6 years, 10 different sequencing approaches have been utilised to capture expression-based infection changes (**Figure 1.4D**). Although methylation isn't strictly gene expression-based, the studies were included to highlight a new area of bacterial epigenetics and epigenome regulation effecting gene expression.

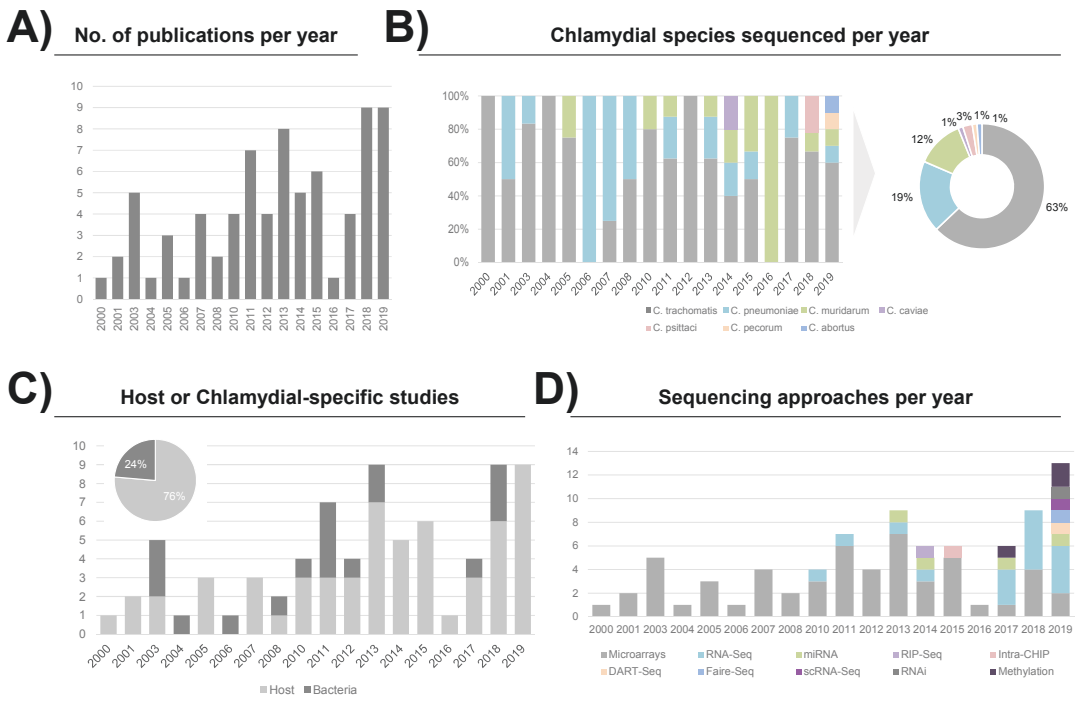


Figure 1.4: Gene expression-based research focused on chlamydial infection

A) The number of chlamydial-based gene expression publications per year. Search criteria from PubMed and bioRxiv included a combination of "*Chlamydia*" + "transcriptome", transcriptomics, "RNA-seq", "microarray", "arrays" or "NGS". **B)** The range of chlamydial species sequenced per year, with the overall percentages summarised in the doughnut chart. **C)** Numbers of studies examining either host or chlamydial-specific responses, with inserted pie chart showing the overall proportions. **D)** The number and type of sequencing approach per year. The FAIRE-seq and single-cell RNA-seq sequencing approaches that appear in 2019 are preprints of **Chapters 3** and **4** from this thesis.

1.3.2.3. Bioinformatic analysis

Phase 1 & 2 – QC and Alignment: Once reads have been quality checked and prepared (see *Phase 1* from the genome sequencing section), they are aligned to genomic coordinates from either an annotated genome or transcriptome (**Figure 1.5**). In more complex organisms where genes are combined of introns and exons (i.e. humans), the aligning software needs to have the capability of determining alternatively spliced transcripts, such as STAR (Dobin et al., 2012). Although isoform-based transcripts have been observed in bacterial systems (Conway et al., 2014), the absence of introns reduces the complexity of the underlying operons, and of the alignment software, where non-splice aware aligners are used, such as Bowtie2 (Langmead and Salzberg, 2012). If longer reads were captured, short-read alignment software can be used, but long-read specific aligners are recommended such as Graphmap2 (Marić et al., 2019) and Minimap2 (Li, 2018).

Phase 3 –Quantification: If reads were aligned to an annotated genome, then additional software is required to quantify read abundances for individual genes. Common software includes HTSeq (Anders et al., 2015) and FeatureCounts (Liao et al., 2014), and can be used irrespective of the underlying organism. These tools will generally disregard many aligned reads to avoid ambiguity. This includes reads that map to multiple genomic locations, and reads that span multiple features. If reads were aligned to transcripts using an alignment-free assembly such as RSEM (Li and Dewey, 2011) or Kallisto (Bray et al., 2016), transcript estimates are produced from the software, removing the need for quantification software. Each tool quantifies differently, with RSEM utilising ambiguous reads using expectation maximisation, while Kallisto directly includes these reads, potentially biasing results. If estimating transcript abundances from long reads, software such as deSALT (Liu et al., 2019) and SQANTI (Tardaguila et al., 2018) are more suited, particularly when dealing with the high error rates that are common. Deciding between a genome-based or transcript-based

assembly ultimately comes down to the underlying experiment. Transcript-based software has demonstrated a strong performance in identifying highly abundant and long transcripts, but appears to be less accurate when quantifying low abundance or short transcripts (Wu et al., 2018). For example, if examining a bacterial experiment that expects to capture low quantities of transcripts from an early time point (i.e. many chlamydial studies), then using a genome-based approach would be more suited as the quantification would be more accurate.

Phase 4 – Filtering and normalisation: The resulting output after quantification is an expression matrix, where rows are genes or transcripts, columns are the samples, and values are read counts or estimated transcript abundances. To improve the accuracy of underlying statistical models which are used for detecting differentially expressed genes, low expressed genes are generally removed during a filtering step (Gentleman R, 2011). Specific parameters depend on the organism and experiment. For example, the human genome contains approximately 50k genes and after filtering may be reduced to 15 or 20k, whereas the chlamydial genome contains ~ 1,000 genes and filters need to be more constrictive.

Normalisation of the expression matrix is more complex, requiring statistical transformations to account for differences in sequencing depth (library size) and technical biases (Risso et al., 2014). Many normalisation methods exist, and are often bundled with differential expression software, including upper quantile and the trimmed mean of M-values (TMM) method (Robinson and Oshlack, 2010). Normalisation methods can be applied to any organism, with multiple methods usually tested and visually inspected using relative log expression (RLE) plots before and after normalisation. Particular methods are chosen based on the underlying statistical assumptions and the fit to the experimental data. A further visualisation tool that can help confirm if the correct normalisation method was used is principal component analysis (PCA) plots. Genes from each sample within the expression matrix are clustered

based on their similarities, where replicates from the same conditions should cluster together; additionally, PCA plots can help to highlight potential outlier samples.

Phase 5 & 6 – Differential expression and biological analysis: Depending on the experimental design, differential expression tools can be modelled to examine simple differences between infected and uninfected conditions, or more complex experimental designs containing additive models and blocking effects. Common tools include edgeR (Robinson et al., 2010) and DESeq2 (Love et al., 2014), and typically utilise generalised linear models to allow for these comparisons (Stark et al., 2019). These two tools in particular, are widely used as they provide comparable results and have easy to follow user guides and tutorials. For each comparison, fold changes and p-values are calculated for each gene, identifying their expression differences and significance respectively. Significant gene sets can then be analysed with enrichment software, detecting if subsets of genes are correlated with biological pathways, processes or functions. Common pathway analysis includes the Kyoto encyclopedia of genes and genomes (KEGG) (Kanehisa and Goto, 2000), Panther (Mi et al., 2017) and David (Huang da et al., 2009); while Gene Ontologies (GO) can identify biological processes, molecular functions and predict cellular locations of where transcripts originated (The-Gene-Ontology-Consortium, 2019). The underlying databases that these tools use can often vary significantly, highlighting different biological outcomes (García-Campos et al., 2015; Young et al., 2010). Often multiple sources are used to confirm validity, and if lab resources are available, qPCR can be used for quantitative validation.

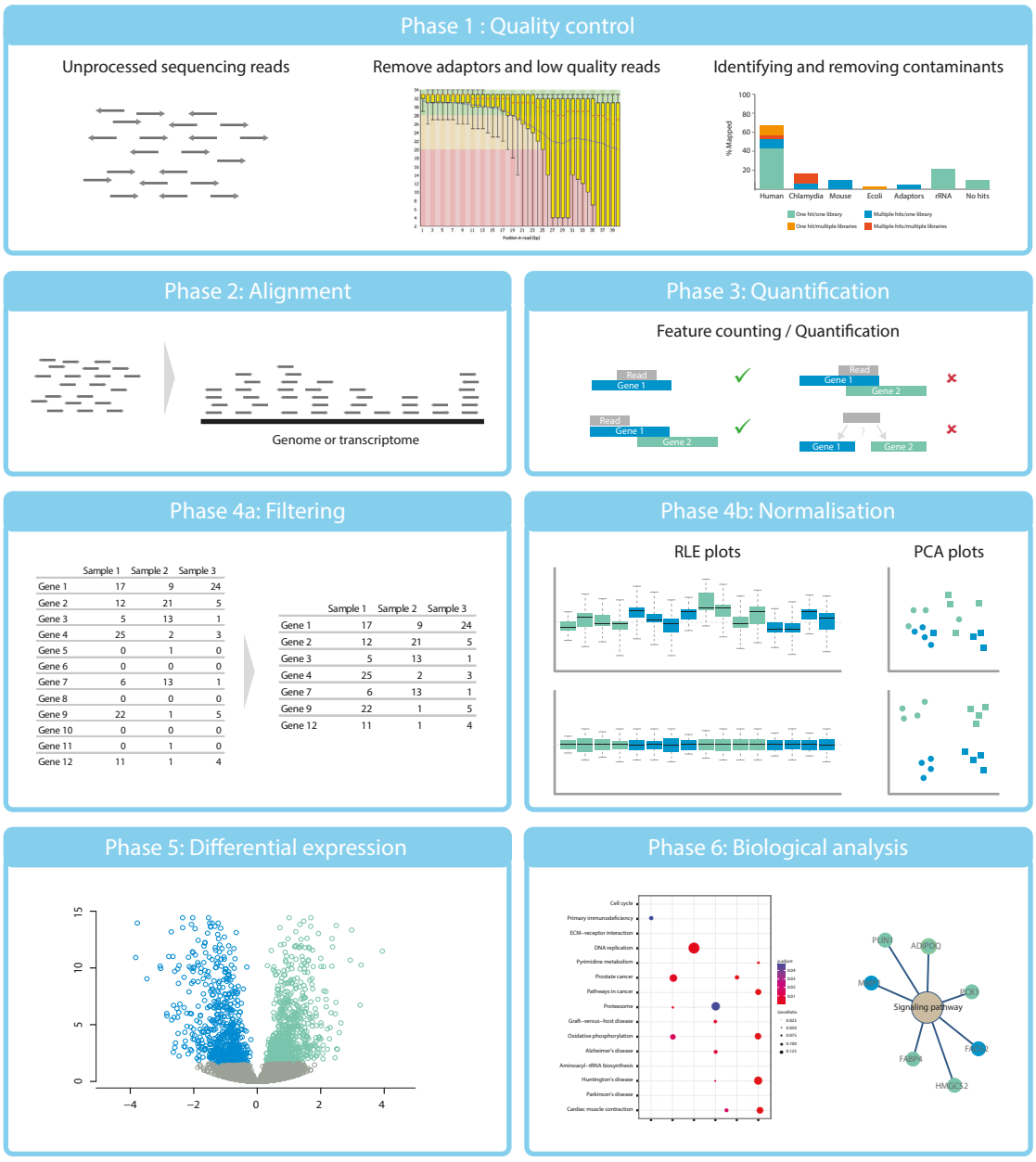


Figure 1.5: Bioinformatic analysis of RNA-seq data

Phase 1) Quality control is performed on unprocessed reads after sequencing. Steps include removing sequencing adaptors and low-quality reads. Additionally, reads can be compared against different genomes to identify contaminants. **Phase 2)** Processed reads can be aligned to either a reference genome or transcriptome. If a genome is used, then the aligner may need to be capable of determining alternatively spliced transcripts. **Phase 3)**

Quantifying aligned reads helps to remove ambiguous reads that map to multiple locations or overlapping features. **Phase 4)** By filtering genes with low expression, library normalisation approaches can be more accurate and help to give more accurate downstream analyses. Visualisation from plots can help determine what the best normalisation methods are, and include relative log expression (RLE) and principal component analysis (PCA). **Phase 5)** Differential expression software allows comparisons to be made between different conditions, such as infected and uninfected samples. Subsets of genes will be highlighted, identifying the significance of each differentially expressed gene and the amount of change that occurred. **Phase 6)** Significant genes are examined for their biological relevance through pathway analysis, enrichment, and interactions with other genes.

1.3.3. Single cell transcriptomics

1.3.3.1. Overview

Until recently, transcriptomic sequencing focused on generating data from large populations of cells or tissues. These datasets represent a summary of all cellular expression profiles in the population, which is essentially an average of the total (Hebenstreit, 2012). Subsets of cells that have a higher or lower expression profile can skew this average (Łabaj et al., 2011), resulting in smaller but potentially equally important subsets and their corresponding expression profiles being obscured. Thus the measured expression profile of a cell population can be misleading (Kolodziejczyk et al., 2015). By sequencing single cells within the originating population, these biases may be removed. The resulting expression data is therefore linked to each cell, and cells with similar expression profiles cluster together. This separation provides different subsets of cells often with varying expression profiles to be further examined, allowing a much deeper understanding of inter-population heterogeneity, cell states and interactions, and gene regulation (Kolodziejczyk et al., 2015; Shapiro et al., 2013).

A major difficulty of working with single cells is the small amounts of available genetic material. For example, a diploid human cell contains approximately 7pg of DNA (Macaulay and Voet, 2014) and less than 1pg of mRNA (Kawasaki, 2004). New methods are continuously been developed, often to increase transcript yields (Svensson et al., 2017), capture specific cell types (Hwang et al., 2018), and increase the speed in which cells can be processed (Svensson et al., 2018). Initial experiments were low-throughput such as the Fluidigm C1 instrument, capturing hundreds of cells, whereas Chromium 10x platforms can generate hundreds of thousands of cells in parallel (Svensson et al., 2018).

The underlying laboratory procedures for scRNA-seq methods can be broken down into three main steps: 1) Efficient isolation of single cells, 2) Isolation of DNA or RNA from each cell, and 3) Accurate amplification of that genetic material, which is then used for sequencing (Kolodziejczyk et al., 2015). Some methods allow the addition of unique molecular identifiers (UMIs), providing a more accurate representation of the transcripts inside each cell by integrating unique barcodes or identifiers into the primers (Islam et al., 2013). Adding a known amount of selectively chosen RNA into the experimental process (spiked-in RNA), can provide a control that is used for calibration during normalisation (Lun et al., 2017). Both UMIs and spiked-in controls help reduce noise in the resulting datasets, which is still a common problem in scRNA-seq (Kolodziejczyk et al., 2015). It should also be noted that these methods predominately capture polyadenylated (PolyA) transcripts, limiting their use to eukaryotic and viral studies.

1.3.3.2. Bacterial-based scRNA-seq

Due to this limitation, many scRNA-seq studies focusing on bacterial infections have predominantly focused on host responses alone (Avraham et al., 2015; Bossel Ben-Moshe et al., 2019; Saliba et al., 2016). Only a limited number of bacterial-specific scRNA-seq studies have been undertaken, examining *Burkholderia thailandensis* (Kang et al., 2011), *Synechocystis* (Wang et al., 2015), malaria parasites (Poran et al., 2017; Reid et al., 2018), and *Salmonella typhimurium* (Avital et al., 2017). These underlying protocols utilise more complex techniques to capture a wider range of transcripts, such as rolling circle amplification (RCA) (Kang et al., 2015), RNA-based single-primer isothermal amplification (Ribo-SPIA) (Wang et al., 2015), random hexamer primers (Avital et al., 2017), and a template switching mechanism (Reid et al., 2018). There are still improvements to be made however, as these methods either have low capture rates, or capture all RNAs including rRNA and tRNA, compromising the overall coverage of mRNA. In addition, no methods currently exist that

can efficiently capture long-reads from single cells, but due to the interest in single cell omics, methods will likely be developed in the not too distant future.

1.3.3.3. Bioinformatic analysis of single cell sequence data

Phase 1 – Quality checking, alignment and quantification: Although many single cells are being sequenced, initial quality control of sequencing reads remains the same (see *Phase 1 - Quality control* from genome sequencing) (**Figure 1.6**). The same applies to alignment and quantification, where the same tools can be used irrespective of whether the data is bulk RNA-seq or scRNA-seq, with methods being repeated relative to the number of cells (see *Phase 2 – Alignment* from transcriptomics). To help organise the large number of corresponding metrics during these initial stages, software such as MultiQC (Ewels et al., 2016) collates and groups this output into single useful reports.

Phase 2 – Cell quality control and filtering: The additional steps used to isolate, capture and sequence single cells bring additional biases that need to be identified and controlled. These include how many transcripts were sequenced and aligned, and how many genes were counted and quantified above a threshold. Additional steps include examining cell expression levels of rRNA as a measure of depletion success, and mitochondrial gene expression as an indicator of cell stress (Zhao et al., 2002). Additionally, genes directly involved with the cell cycle are measured and cell states predicted, providing further biological insights and filtering criteria (Scialdone et al., 2015). Depending on the sequencing protocol and technology, other factors such as the proportion of reads mapping to spiked in controls and detecting when more than one cell was captured within a droplet (doublets), should be taken into consideration (Brennecke et al., 2013; McGinnis et al., 2019). Filtering based on these parameters can be either set with predetermined thresholds, or can use statistical approaches that are calculated

based on the protocol and the underlying cell expression profiles (Lun et al., 2016b; McCarthy et al., 2017).

Phase 3 – Library normalisation, removing confounding effects and imputation: Normalising and removing confounding effects from a scRNA-seq derived count matrix helps to account for any technological and biological variability that has occurred, thereby allowing a more accurate comparison of the expression levels across all of the cells (Stegle et al., 2015). Technical variability can be introduced from steps such as isolation, amplification and different batches of sequencing; whereas biological variation includes differences such as cell size, cell cycle, and the amount of RNA per cell (Stegle et al., 2015; Vallejos et al., 2016).

The normalisation approach to remove differences between library sizes will depend on the experimental setup if UMIs, or spike-ins were or weren't used (Ding et al., 2017). Consideration should also be given to the statistical method depending on the amount of variability in the dataset. Normalisation methods can be used on either bulk or single cell datasets, which include transcripts per kilobase million (TPM) and TMM (Hwang et al., 2018). Single cell specific methods allow for the increased variability that appears between cells, such as quantile regression (Bacher et al., 2017) and deconvoluting size factors from clusters of (Lun et al., 2016a) cells.

Additional software is required to remove technical confounding effects resulting from different reagents, isolation methods and the different batches of sequenced cells. Using the underlying metadata of the experiment (batch etc), software can identify and remove sources of variation from the expression data that is not related to the biological signal of interest. Different approaches include using spike-ins, regression tactics, housekeeping genes and control samples/cells; from software such as RUVSeq (Risso et al., 2014) and Combat (Johnson et al., 2007b). The best way to determine which normalisation method to use, or if confounding effects need to be removed, is through the use of visualisation tools such as RLE

plots and or PCA plots. RLE plots are useful to compare the effectiveness of different library normalisations, and PCA plots can help highlight when additional confounding effects are having an impact on expected clustering outcomes.

Due to the small amounts of RNA per cell and often low efficiency in capture rates, many genes that are potentially being expressed are not captured. This is called the dropout effect, with recent imputation software aiming to provide *in silico* corrections (Eraslan et al., 2019; Li and Li, 2018). Since we do not know what are technical artefacts and when transcripts are truly absent, imputation is a difficult challenge and current methods will likely introduce false-positives in downstream analyses (Andrews and Hemberg, 2018a).

Phase 4 – Dimensionality reduction, clustering and feature selection: Large datasets accompany scRNA-seq experiments, and it is often beneficial to apply a form of dimensionality reduction to significantly reduce noise, making the data easier to visualise in a two- or three-dimensional space. Methods include PCA and t-distributed stochastic neighbour embedding (tSNE) plots, and depending on the dataset, often a log-transformation will be applied to further increase resolution and biological relevance (Luecken and Theis, 2019). Resulting scatterplots will display each cell and will usually be coloured based on experimental data, such as time point or cell type. Clustering cells based on the similarity of their expression profiles can be performed by supervised or unsupervised approaches. Supervised methods require experimental input and will try to cluster based on these parameters, such as CellAssign (Zhang et al., 2019) and Garnett (Pliner et al., 2019). Unsupervised clustering solely relies on expression profiles, which can be particularly challenging when working with scRNA-seq datasets that contain high levels of technical and biological noise (Kiselev et al., 2019). Unsupervised methods include hierarchical clustering, k-means clustering and graph-based clustering (Hwang et al., 2018).

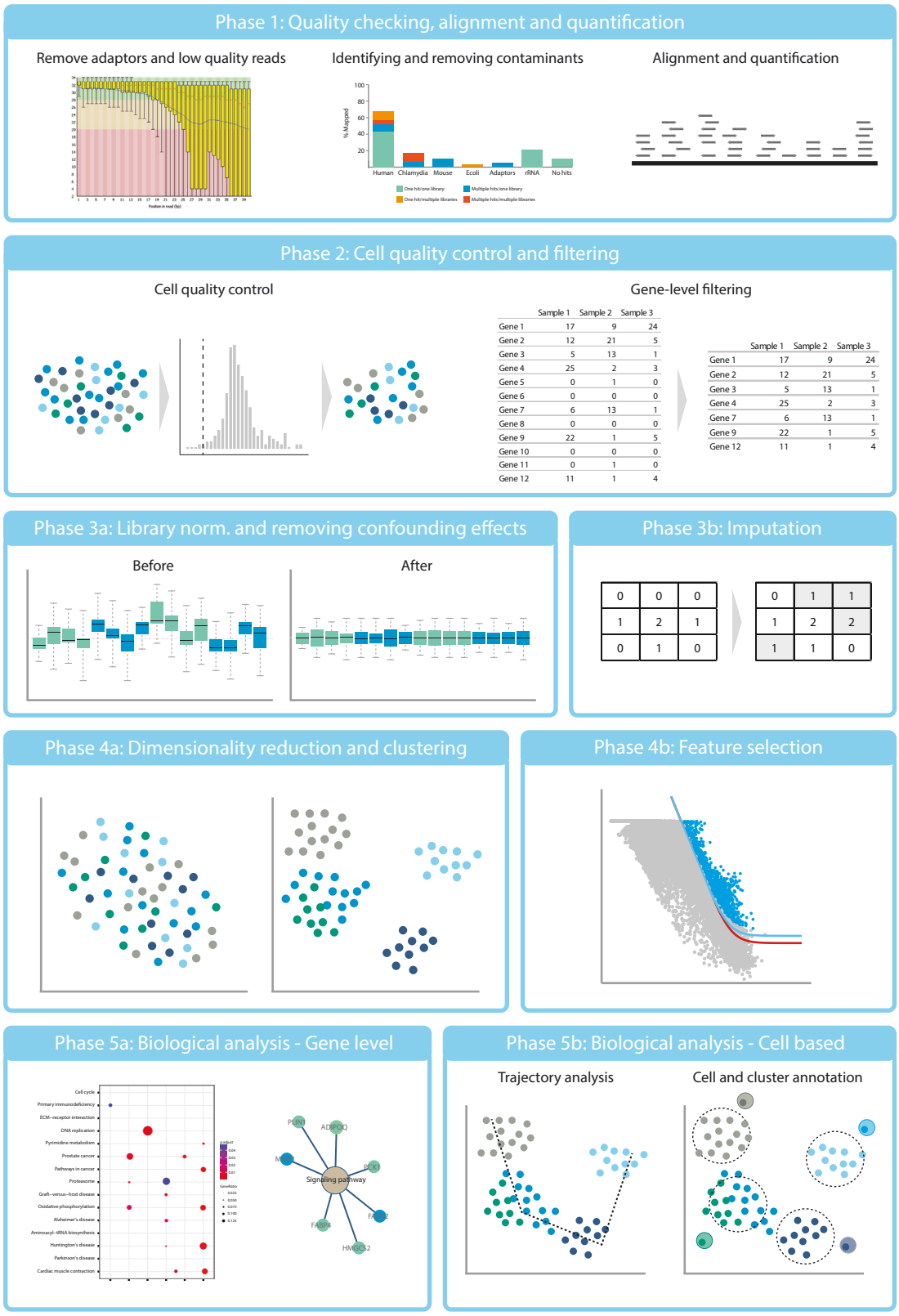
To understand which genes are contributing to the separation of the clusters, or between different biological conditions of interest, feature selection can be an important step. It can also be used to remove genes that contribute to technical noise from downstream analysis (Luecken and Theis, 2019). Different approaches exist that detect for highly variable genes (HVG) or for marker genes between conditions or clusters. Underlying methods can compare mean expression differences such as Seurat (Satija et al., 2015), or are based on statistical models such as a M3Drop, which uses a michaelis-menten curve (Andrews and Hemberg, 2018b).

Phase 5 – Biological analysis: Biological interpretation can be separated into either gene level analysis or cell-based analysis. Gene level analysis includes differential expression (DE), regulatory networks and gene-set or pathway analysis. All of these methods can be applied to bulk or scRNA-seq experiments and are outlined in the transcriptomics section (*Phase 5 – Differential expression and biological analysis*). However, many existing bulk DE software performs poorly when a large number of zero counts are present (from events such as dropouts) (Luecken and Theis, 2019). This has led to scRNA-seq specific DE software such as SCDE (Kharchenko et al., 2014) and MAST (Finak et al., 2015), accounting for differences in the underlying distributions and generally lower counts. However, improvements can still be made as many of the tools generally don't allow complex experimental designs, they instead only allow simple comparisons such as the differences between clusters or experimental conditions.

Cell-based analyses are specific to scRNA-seq experiments, with > 200 tools currently available (Zappia et al., 2018). A major advantage is the ability to identify cell types and sub-populations from a heterogeneous population of cells (Kolodziejczyk et al., 2015). Coupled with reference databases such as the Human Cell Atlas (Regev et al., 2017) or the Allen Brain

Atlas (Sunkin et al., 2013), clusters of cells and individual cells can be compared and classified, potentially identifying rare and disease-related cell subsets.

Trajectory analysis (pseudotemporal ordering and inference) can help to understand the various lifecycle and differentiation profiles of each cell (Bacher and Kendziorski, 2016). For example, at the unique snapshot in time when each cell was lysed and sequenced; each cell may have been undergoing transformation into a different cell type, or undergoing cellular division, or even apoptosis (Leng et al., 2015; Stegle et al., 2015). Depending on the huge variety of biological pathways a cell can take, will ultimately alter its transcriptional profile (Linnarsson, 2015). Trajectory analysis groups cells that have similar expression profiles to generate small clusters (similar to sub-population software). Once a number of clusters have been identified, a pathway will be created linking up each of the clusters. This creates a two-dimensional pathway of the varying cellular states in a mock time setting, or ‘pseudotime’ (Trapnell et al., 2014).



	Sample 1	Sample 2	Sample 3
Gene 1	17	9	24
Gene 2	12	21	5
Gene 3	5	13	1
Gene 4	25	2	3
Gene 5	0	1	0
Gene 6	0	0	0
Gene 7	6	13	1
Gene 8	0	0	0
Gene 9	22	1	5
Gene 10	0	0	0
Gene 11	0	1	0
Gene 12	11	1	4

	Sample 1	Sample 2	Sample 3
Gene 1	17	9	24
Gene 2	12	21	5
Gene 3	5	13	1
Gene 4	25	2	3
Gene 7	6	13	1
Gene 9	22	1	5
Gene 12	11	1	4

0	0	0
1	2	1
0	1	0

0	1	1
1	2	2
1	1	0

Figure 1.6: Bioinformatic analysis of single cell RNA-seq data

Phase 1: Quality control helps to remove sequencing adaptors and low-quality reads, in addition to identifying potential contaminant reads. Processed reads can be aligned to either a reference genome or transcriptome. If a genome is used, then the aligner may need to be capable of determining alternatively spliced transcripts. The quantification of aligned reads helps to remove ambiguous reads that map to multiple locations or overlapping features. **Phase 2)** A cell quality control step will use thresholds from metrics such as read depth and mapping rates to filter out low quality cells. Within the remaining cells, genes will also be filtered based on their expression levels. **Phase 3a)** Normalising for differences in library size between cells can be performed with some bulk RNA-seq approaches or single cell specific tools. Relative log expression (RLE) plots are useful way to visualise comparisons between methods. Confounding effects are introduced mostly from technical factors, such as different reagents and sequencing machinery. Different tools exist and are usually performed after library normalisation. **Phase 3b)** Due to the low amount of DNA/RNA per cell, single cell sequencing techniques generally cannot capture all of the associated material. As a result, some genes that are expressed are not captured, resulting in a dropout effect. Imputation software aims to correct for these errors through statistical methods. **Phase 4a)** Dimensionality reduction reduces the complexity of the combined expression matrix of all cells, providing two or three dimensions at a time to help visualise cellular clustering. **Phase 4b)** Feature selection helps improve the signal to noise ratio which can enhance downstream analyses such as clustering and pseudotime analysis. **Phase 5)** Biological analysis can occur at the gene level or cell-based level, identifying statistically relevant genes and pathways, in addition to predicting cell trajectories and annotating clusters.

This technique is useful in identifying genes that regulate cellular states (Linnarsson, 2015), and genes that are involved in switch like mechanisms controlling differentiation (Tang et al., 2010). Additionally, trajectory analysis has been used to highlight key stages within infection-associated settings (Kunz et al., 2018). Different trajectory analysis software packages are available, such as Monocle (Trapnell et al., 2014) and Slingshot (Street et al., 2018).

Compositional analysis is a further cell-based approach where clusters can be analysed in terms of their compositional structure and the proportions of cells within. For example, in response to *Salmonella* infection, an increase in the proportion of enterocytes was identified in the mouse intestinal epithelium (Haber et al., 2017).

1.3.4. Epigenetics

1.3.4.1. Overview

Epigenetic regulation was defined to explain the occurrence of phenotype changes without any genotype changes. The name epigenetics has been attributed to Conrad Waddington (1905–1975) who theorised that developmental processes occur as a series of 'decisions' that could be represented as 'valleys' and 'forks' within a landscape (**Figure 1.7**) (Waddington, 1956). Through a small manipulation of this landscape, one channel would be more favoured than another, showing how the same genome (represented by the starting sphere) can lead to different phenotypes. Today, epigenetic mechanisms are associated with DNA and RNA and attributed to wide range of developmental processes, immunological disorders and diseases (Kiefer, 2007; Portela and Esteller, 2010). Epigenetic mechanisms have been identified not just in mammalian cells, but are found in many organisms including plants, bacteria and viruses (Pikaard and Mittelsten Scheid, 2014; Willbanks et al., 2016).

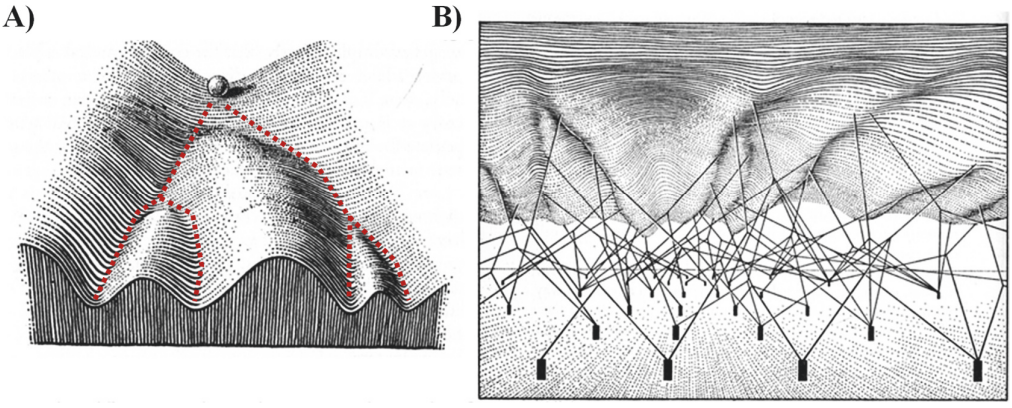


Figure 1.7: Waddington's developmental landscape and epigenetic interactions

A) The path of the sphere corresponds to the developmental trajectory based on the landscape and underlying influences. **B)** The proposed complex system of interactions that define the epigenetic landscape. The black rectangles represent genes, the lines represent

the chemical signals from each gene. The landscape is controlled by these genetic interactions and chemical signals, influencing the path the sphere takes. Adapted from (Waddington, 1957).

Although the term *epi* refers to features associated above genomic elements, structural adaptations associated with chromatin structure also fall under the same epigenetic term. Regulation mainly includes DNA modifications which include methylation, post-translational histone modifications, chromatin remodelling and non-coding RNAs (Zhang and Cao, 2019). Sequencing protocols have been developed to capture each of these specific modifications (**Figure 1.8**).

1.3.4.2. Regulatory mechanisms and associated sequencing protocols

Methylation

DNA methylation is a process where methyl (CH_3) groups have been added to cytosine bases at specific regions throughout a genome, altering gene expression. Hypermethylated bases can block transcription and silence gene expression, whereas the removal of the methyl group (hypomethylation) can activate gene expression. DNA methylation arrays can identify the locations of these methyl groups, however, the underlying microarrays are limited to the number of probes the array can target (Wilhelm-Benartzi et al., 2013). For whole genome methylation studies, bisulphite sequencing is more commonly used, due to its high throughput nature and ability to capture methylation patterns genome-wide. DNA is treated with sodium bisulphite, resulting in unmethylated cytosines converting to uracil, and thus identifiable (Li and Tollefsbol, 2011). CpG sites (cytosine and guanine appearing consecutively on the same strand) are of importance when examining methylation patterns, as the cytosine belonging to the CG-pair is frequently methylated. When CpG sites occur frequently within a genomic

region, they are called CpG islands, and are commonly used as probes in array-based studies (Bibikova et al., 2011).

Chromatin accessibility

Eukaryotic DNA is highly condensed within the nucleus of each cell. This is achieved by tightly wrapping DNA around histone proteins, which are referred to as nucleosomes. This occurs across the whole genome, with the resulting condensed fibres called chromatin (**Figure 1.8**). The regulation of genes and cellular processes occur when chromatin fibres become less condensed (open chromatin), providing accessibility to transcription factors and regulatory binding machinery (Klemm et al., 2019).

The process of condensing and recondensing is a consistently occurring dynamic process. A simple example might be a small region of DNA which encodes a gene that is required to be transcribed. A more complex example might be a large gene that undergoes alternative splicing events which results in only the first part of the gene being transcribed. Furthermore, the genes promoter may also be influenced by an enhancer/silencer region at a different location within the genome. These are only a few examples of how an incredibly diverse number of regulatory mechanisms are directly associated with the accessibility of chromatin, and how examining regions of open and closed chromatin can assist in uncovering these events.

Examining chromatin accessibility to identify regions of open and closed chromatin can be achieved from protocols that include Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq), using a Micrococcal Nuclease (MNase-seq), and formaldehyde-assisted isolation of regulatory elements (FAIRE-seq). FAIRE-seq involves crosslinking DNA with formaldehyde, where samples are then lysed and sonicated. After phenol/chloroform extraction, the aqueous layer is purified and sequenced, identifying

regions of open chromatin (Simon et al., 2012). FAIRE-seq is a slightly older method that contains a higher background signal and shows a higher affinity for enhancer regions compared to more recent methods such as ATAC-seq (Tsompana and Buck, 2014). MNase-seq utilises a micrococcal nuclease digestion, followed by sequencing. Nucleosomes protect associated DNA from digestion by the nuclease, whereby unveiling areas of the genome occupied by nucleosomes and regulatory factors. The resulting fragments reveal nucleosome location information, and provide an indirect method to probe chromatin accessibility (Cui and Zhao, 2012). Although an accurate method, a large number of cells are required and preparation steps need to be carried out precisely to allow for accurate and reproducible results (Tsompana and Buck, 2014). ATAC-seq uses Tn5 transposases which preferentially attach to areas of open chromatin. The transposases simultaneously fragment and add adaptors at these regions, which are ultimately sequenced and reveal nucleosome-free (open chromatin) regions genome-wide (Buenrostro et al., 2013). Advantages of ATAC-seq include a simple protocol, high sensitivity and high resolution (Klemm et al., 2019), making it the current preferred method to examine chromatin accessibility.

Histone modifications

Histone proteins (as discussed above) facilitate in condensing chromatin fibres by acting as spools which DNA can wrap around. The histone core is comprised of an octamer of grouped proteins, with only two (H3 and H4) containing long tails that protrude from the nucleosome (Mariño-Ramírez et al., 2005). Due to the major role histones play with condensing DNA, modifications directly affect gene expression. Although modifications are associated with some core proteins, they are predominantly located within the protruding tails and include methylation, acetylation, phosphorylation and ubiquitination (Mersfelder and Parthun, 2006).

Chromatin immunoprecipitation followed by sequencing (ChIP–seq) is a protocol that can map DNA binding proteins and histone modifications genome-wide. The method uses

antibodies to immunoprecipitate proteins crosslinked to chromatin (non-histone ChIP), or to modified nucleosomes (histone ChIP) (Johnson et al., 2007a). Antibodies specific to the protein or histone modification are used to enrich the crosslinked regions, identifying different regulatory factors such as TF binding sites and histone modifications. Currently, ChIP–seq is the preferred method to examine histone modifications and associated DNA-binding proteins due to the flexibility of the protocol, the sensitivity and specificity and high quality of data (Park, 2009).

Chromosome conformation

Many regulatory elements such as enhancers and silencers are often placed distally to the genomic loci they can control, with some even located on different chromosomes (Ong and Corces, 2011). Regulation can occur when open chromatin from both regions comes into close proximity through mechanisms including looping, tracking and linking (Khan and Zhang, 2015). Therefore, there is an inherent advantage of combining chromatin accessibility dynamics with additional sequencing approaches that can examine the 3D structure of genomes such as Hi-C (Belton et al., 2012) and Chromatin Interaction Analysis by Paired-End Tag sequencing (ChIA-PET) (Li et al., 2014). Results from both analyses can identify genome-wide events such as enhancer-promoter interactions to be identified more accurately and precisely.

Single-cell approaches

Recently, epigenetic-based single cell approaches have been developed, such as scChIP-seq (Rotem et al., 2015), scATAC-seq (Buenrostro et al., 2015) and scBS-seq (Smallwood et al., 2014). These methods and others allow epigenetic mechanisms to be identified in single cells instead of from populations of cells where the regulatory effects are often averaged. As described in the scRNA-seq section, single cell approaches can identify subsets of cells with

differing expression profiles that may provide novel biological insights that would typically be obscured by bulk sequencing methods. This is still a developing field, and as a result, refinements are still needed to increase the resolution and capture rates, as well as lower costs to increase sequencing depths (Rahmani et al., 2019; Schwartzman and Tanay, 2015; Shema et al., 2019). Additionally, more computational tools need to be developed or existing tools modified specifically for single cell methods; again, similar to what has been seen in scRNA-seq, where underlying distributions and statistical assumptions are potentially not valid when examining populations of single cells.

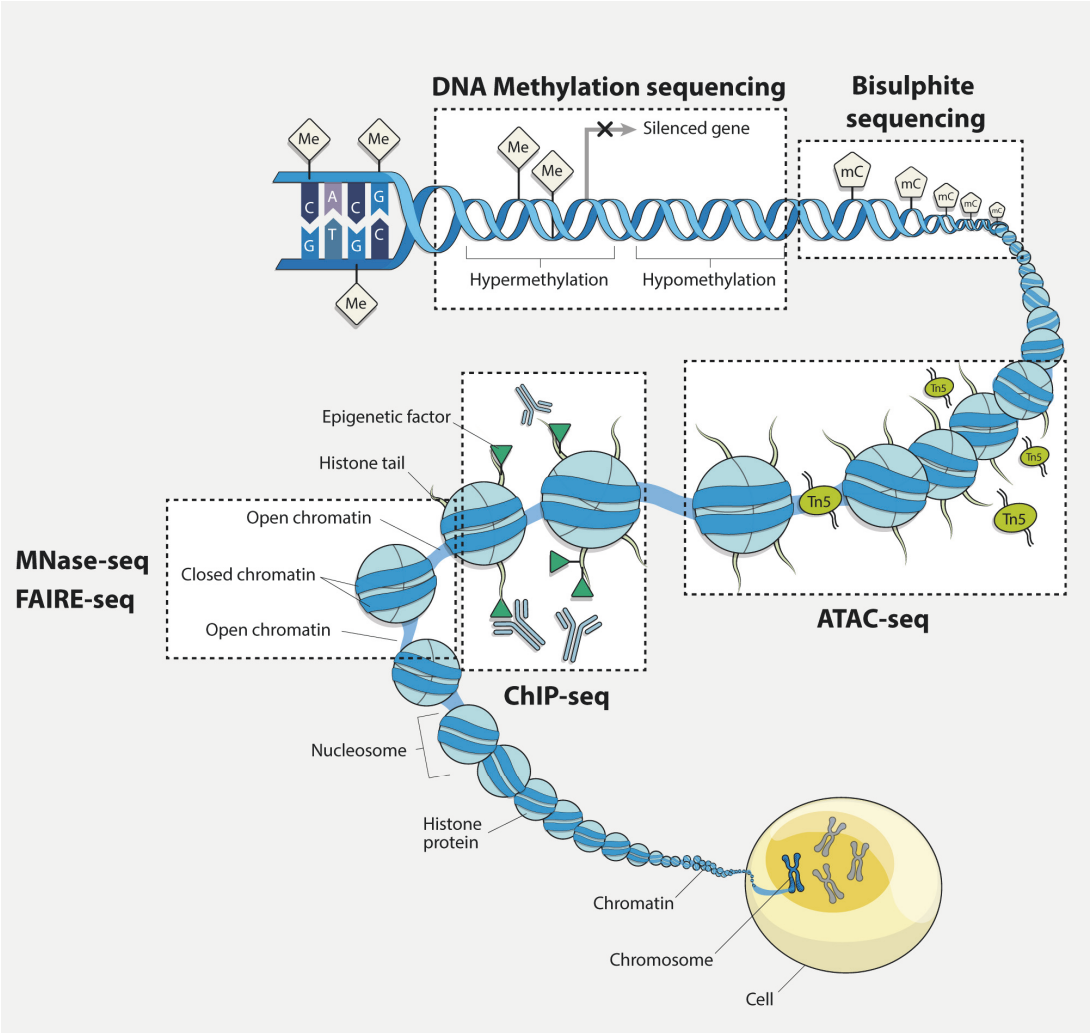


Figure 1.8: Ways in which epigenetic changes can alter gene expression

Epigenetic factors can modify DNA in various ways that alter gene expression. Common technologies to examine these modifications include methylation arrays, which identify methylated cytosine bases. Hypermethylation involves the addition of a methyl group which can silence gene expression, while hypomethylation can activate gene expression. Bisulfite sequencing involves treating DNA with sodium bisulphite, causing unmethylated cytosines to convert to uracil. Sequencing can then reveal where the methylated cytosine bases are throughout a genome. Assay for transposase-accessible chromatin using sequencing (ATAC-seq) uses a Tn5 transposase to attach to areas of open chromatin. These enzymes insert sequencing adaptors in open areas, which are cut out, purified, amplified and sequenced, revealing regions of open chromatin. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) uses antibodies to immunoprecipitate proteins crosslinked to chromatin. Sequencing can identify where different regulatory factors bind throughout the genome. Formaldehyde-assisted isolation of regulatory elements (FAIRE-seq) involves crosslinking DNA with formaldehyde, where samples are then lysed and sonicated. After phenol/chloroform extraction, the aqueous layer is purified and sequenced, identifying regions of open chromatin. Micrococcal nuclease sequencing (MNase-seq) sequencing utilises a micrococcal nuclease digestion, followed by sequencing. This distinguishes nucleosome positioning, allowing regions of closed chromatin to be identified along with locations of regulatory DNA-binding proteins. Modified from (Nature-Research, 2019).

1.3.4.3. Bacterial epigenetics

Bacterial gene expression was initially thought to be a simple process compared to eukaryotic systems, where mechanisms such as alternative splicing and an array of non-coding genes help regulate transcription. More recently however, this simplistic view has begun to shift as we now understand that bacterial systems have RNA-based biological mechanisms and functions that rival eukaryotes. These include metabolite-sensing riboswitches, RNA thermometers, and a wide range of small non-coding RNA (sRNA) (Hor et al., 2018). Furthermore, one-third of *Escherichia coli* operons contain internal promoters and terminators, generating multiple transcription units (TUs) with differing expression patterns (Conway et al., 2014). Unfortunately, the limiting factor to further explore these mechanisms is a lack of usable, reproducible and genome-wide methods.

Bacteria also have the ability to alter epigenetic marks and machinery for their own benefit within hosts. Their range of activity is surprisingly diverse, where bacterial effector proteins from species including *Anaplasma phagocytophilum*, *Shigella flexneri* and *Mycobacterium tuberculosis*, will hijack cellular signalling pathways, alter histone modifications to silence host defence genes, and control chromatin complexes (Bierne et al., 2012; Zhang and Cao, 2019). As discussed in previous sections, the intracellular developmental cycle of *Chlamydia* has delayed the functional characterisation of effector proteins, thus hindering possible discoveries of chlamydial specific epigenetic alterations. Due to this restriction, limited epigenetic-based studies have been published. An initial study identified a novel methyltransferase (NUE) from *C. trachomatis* acting as a Type III secretion system (T3SS) effector with methyltransferase activity. NUE enters the host nucleus and methylates eukaryotic histones H2B, H3 and H4 *in vitro* (Pennini et al., 2010), indicating that *Chlamydia* may have the capacity to directly modulate host cell epigenetic regulation. However, NUE targets and any subsequently affected pathways remain uncharacterised.

More recently, genome-wide methylation studies have emerged, examining differentially methylated genomic regions from within the host (Kessler et al., 2019; Rajić et al., 2017; Xiong et al., 2019). An advantage of these studies was they didn't need to examine chlamydial effector proteins directly, but instead look at potential infection-induced changes. Results identified numerous differentially methylated regions (DMR) within fallopian tube organoids from *in vitro C. trachomatis* infections, and from lung tissues of *C. pneumoniae*-infected patients, suggesting the likelihood of chlamydial-derived epigenetic modulators which are thus potential candidates for drug targets. However, as highlighted by Kessler et al., 2019, more complex infection models or controls are needed to separate chlamydial-induced effects from infection-induced effects. While a lot more work is still required to decipher what epigenetic-based mechanisms *Chlamydia* affect, it is exciting to see new sequencing approaches being applied and new discoveries made.

Although different epigenetic-based protocols and pipelines exist, the following bioinformatic analysis is focused on chromatin accessibility as used in Chapter 3, and is applicable to protocols including MNase-seq, ATAC-seq and FAIRE-seq.

1.3.4.4. Bioinformatic analysis

Phase 1 & 2 – QC and Alignment: Reads are initially quality checked and prepared (see *Phase 1* from the genome sequencing section) followed by alignment to an annotated genome. In theory, any aligner will work as the fragmented areas of chromatin are not subjected to alternative splicing. Due to their speed, ease of use and reproducibility, Bowtie2 (Langmead and Salzberg, 2012) and BWA (Li and Durbin, 2009) are the most commonly used aligners. After alignment, further QC from software such as deepTools (Ramírez et al., 2014) can help evaluate the quality of each sample. Metrics include the number of aligned reads, overall alignment rates, and the similarity of replicates within the same conditions. Duplicate reads

that may have arisen during library construction and reads mapping to blacklist regions (problematic regions with high signals) should be removed as recommended from the encyclopaedia of DNA elements (ENCODE) guidelines (Landt et al., 2012).

Phase 3 – Peak calling: Determining how peaks are represented is an important step as most downstream analysis is reliant on their correct identification and determination. Considerations include the sequencing approach and any associated biases, such as FAIRE-seq which generates broader peaks compared to ATAC-seq, which generates more defined and generally narrower peaks over regulatory regions (Meyer and Liu, 2014). If single end sequencing was performed, quality checked reads can be directly used to identify peaks, while paired end reads need to be combined first. This involves merging the 5' and 3' tags that represent the same peak, into a single peak (**Figure 1.9**). Different merging or centering approaches exist depending on the software, such as MACS2 which extends the reads in the 3' direction to the fragment length obtained from an underlying statistical model (Zhang et al., 2008). A further software-specific parameter that influences peak calling is defining the difference between what constitutes a background signal and a significantly enriched region (Meyer and Liu, 2014). Although default thresholds can be used, the alignment metrics and depth of sequencing should direct if this needs to be changed. Many peak calling software tools exist (> 30), with different implementations and algorithms to identify significant peaks (Thomas et al., 2016). These include spatial clustering approaches such as SLICER (Zang et al., 2009) and F-Seq (Boyle et al., 2008), while MACS2 (Zhang et al., 2008) and CSAW use a sliding window approach throughout a genome to identify peaks (Lun and Smyth, 2016).

Phase 4 – Normalisation and filtering: After peak calling, normalisation is needed to account for differences in sequencing depth across samples which may affect peak widths and heights. The first step is to count the reads under each peak by defining peak boundaries. After repeating this for all samples, a count matrix similar to RNA-seq exists with samples as

columns and genomic regions (peaks) as rows. To help identity outlier samples and removing low abundant regions, peaks from all samples can be combined to create a consensus peak set (Stark and Brown, 2019). A threshold can be set to remove peaks if they do not appear in a particular amount of samples. Also, the fraction of reads in peaks (FRiP) can be calculated from the consensus peak set, highlighting samples with low enrichment and thus possible outliers (Landt et al., 2012). With the filtered count matrix, normalisation approaches are again similar to RNA-seq with options including TMM and RPKM (Meyer and Liu, 2014; Stark and Brown, 2019). More specific options include using peak widths, heights, and/or the location of the highest point of the peak, which is required for some experimental outcomes (Ross-Innes et al., 2012; Wang et al., 2018).

Phase 5 – Differential comparisons: Further advantages of using a count matrix is the ability to use RNA-seq based tools to identify differential chromatin accessibility between samples. Packages such as Diffbind combine commonly used DE software (edgeR and DESeq2), allowing simple comparisons, or more complex designs requiring features such as blocking factors (Stark and Brown, 2019). Differential results generate p-values and fold-changes, making it possible to infer open and closed states of chromatin from any protocol. However, this can potentially introduce some bias from what the protocol captures to the distribution of fold-changes. For example, FAIRE-seq and ATAC-seq were designed to identify open chromatin, while MNase-seq identifies closed chromatin (Tsompana and Buck, 2014).

Phase 6 – Peak annotation: The annotation of significant peaks requires their overlap with known biological features such as promoters, introns, intergenic regions, TSSs and TTSs. Different annotating options exists that can give conflicting results, especially with large peaks that overlap more than one feature. For example, underlying software will allow you to select which part of the peak to use (left, right, or middle) that is associated to or overlaps the closest feature (Zhu et al., 2010). This is a known challenge, especially when the peak covers

multiple features, or the significant region contains more than one identifiable peak (Salmon-Divon et al., 2010). Often highly significant peaks are viewed in a genome browser and adjusted accordingly. However, common analyses identify hundreds or thousands of peaks, and viewing each annotation isn't realistic. Therefore, in many outcomes a small percent of annotated peaks are likely false-positives. Annotating software includes ChIPpeakAnno (Zhu et al., 2010), Homer (Heinz et al., 2010) and GREAT (McLean et al., 2010).

In humans, < 2% of the genome is encoded with protein coding genes (Lander et al., 2001), often resulting in many peaks being assigned to unannotated intergenic regions. Initially, these intergenic regions were labelled 'junk DNA' (Wong et al., 2000), but have continuously been associated with a range of non-coding regulatory features such as enhancers, silencers and insulators (Kolovos et al., 2012; Thurman et al., 2012). Features are often highly associated with specific cell lines and tissues, and therefore cannot be included in a general analysis (Gao et al., 2016). To overcome this, specific databases, methods and software exist containing experimentally validated and *in silico* predicted features that can be compared against. Databases include EnhancerAtlas (Gao et al., 2016) and VISTA (Visel et al., 2006), while software and methods from (Huang et al., 2019) and (Doni Jayavelu et al., 2018) can identify silencers and enhancers, predominantly within intergenic regions.

Phase 7 – Motif identification and annotation: The underlying regions from significant peaks can also be inspected for putative transcription factor (TF) binding sites. Regions can be scanned and either compared against online databases to identify 'known' TF motifs, or novel binding sites can be discovered from *de novo* approaches. Software includes MEME (Bailey et al., 2009), Homer (Heinz et al., 2010) and CompleteMotif (Kuttippurathu et al., 2011). A limitation to *de novo* discovery is during the annotation step, where predicted TFs are assigned based on the closest match to internal lists. To help confirm these annotations (which are often biased or restricted based on the underlying lists), motifs of interest should

be compared to find overlapping matches from multiple online databases such as JASPAR (Mathelier et al., 2014), TRANSFAC (Matys et al., 2006) and UniPROBE (Newburger and Bulyk, 2009).

Phase 8 – Biological analysis: Once peaks have been annotated to their nearest genomic feature and putative TFs identified, their biological interpretation can be further examined using pathway and gene set enrichment analyses as outlined in the transcriptomics section (*Phase 5 – Differential expression and biological analysis*).

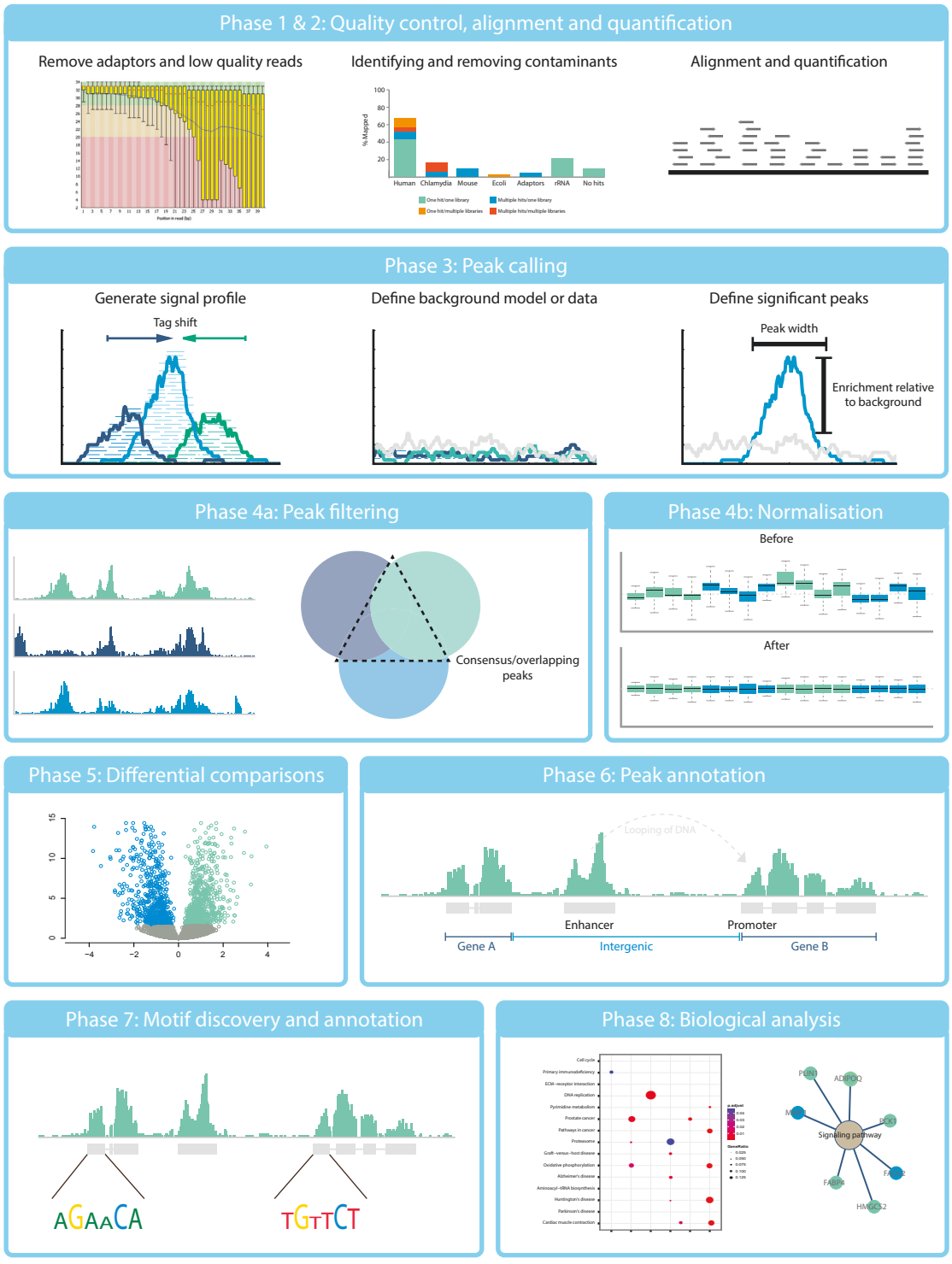


Figure 1.9: Bioinformatic analysis of chromatin accessibility data

Phase 1: Quality control steps help to remove sequencing adaptors and low-quality reads, in addition to identifying potential contaminant reads. **Phase 2)** Resulting reads are aligned

to a reference genome to identify regulatory regions of chromatin. **Phase 3)** Peak calling is performed in three steps. First, 5' and 3' tags are shifted or merged together to identify specific regions. The second step involves defining what constitutes the background or non-significant data. The final step defines peaks widths and fold-changes relative to a background signal. **Phase 4)** Filtering peaks that either only appear in a limited number of samples or are comprised of minimal reads are removed. This helps when normalising for differences in library size. Normalisation methods are based on RNA-seq and can include peak widths and their heights. Relative log expression (RLE) plots are useful to visualise which method is best suited to a particular dataset. **Phase 5)** Software allowing the comparison of experimental conditions are based on RNA-seq methods but can identify regions of open and closed chromatin relative to each condition. **Phase 6)** Significant peaks or peaks of interest can be annotated based on their proximity to genomic features such as enhancers, promoters, introns and exons. **Phase 7)** Motifs can be searched for within selected peaks, identifying binding sites for putative transcription factors. **Phase 8)** Genes from associated peaks and/or transcription factors can be examined for their biological significance through pathway analysis, gene set enrichment and gene-to-gene interactions.

1.4. Thesis summary

1.4.1. Overview

Chlamydia trachomatis is an obligate intracellular bacterial pathogen with multiple disease outcomes in humans. Due to the experimental barriers created by its intracellular developmental cycle, genomic manipulation until recently has been minimal and remains restrictive, limiting our knowledge of infection-based biological processes. This thesis details bioinformatic analyses of multiple novel next generation sequencing datasets, examining the host epithelial cell response to *C. trachomatis* using *in vitro* models of infection over different time courses. All Results chapters were designed to be exploratory analyses, providing new methods to explore disease mechanisms and ultimately their applicability for further use in more advanced settings, such as *in vivo* models and *ex vivo* clinical samples.

1.4.2. Chapter summary and knowledge gaps

Chapter 3 – Examining chromatin accessibility dynamics

Through the use of Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq), chromatin accessibility of open and closed regions of DNA was identified from host cells. Two conditions of infected and mock-infected cells were examined to distinguish infection-relevant responses. Four key time points throughout the development cycle were examined (1, 12, 24 and 48 hours), with three biological replicates for each condition. This study is among a small number of chromatin accessibility studies examining bacterial infections and the first examining host responses to chlamydial infection. Analyses identified both conserved and distinct temporal changes genome-wide. Differentially accessible chromatin regions were linked to genomic features and genes associated with metabolism, apoptosis, intracellular

signalling, cell-cell adhesion, re-direction of host cell nutrients and immune responses. Transcription factors within the same regions across the developmental cycle identified different Krüppel-like-factors (KLFs) which are ubiquitously expressed in reproductive tissues and associated with a variety of uterine pathologies; identifying a novel association with chlamydial infection.

Chapter 4 – Exploring cellular heterogeneity from individually infected host cells

The transcriptional responses from individually infected host cells were captured using single cell RNA-seq (scRNA-seq). This pilot dataset provided a more detailed resolution than traditional bulk RNA-seq approaches that average expression over many cells. 264 cells were captured and equally divided into infected and mock-infected cells from three early infection associated time points (3, 6 and 12 hours). This study is among a small number of scRNA-seq experiments examining bacterial infections and the first examining host responses to chlamydial infection. Analyses highlighted infection-specific host cell biology, including two distinct clusters separating 3 hour cells from 6 and 12 hours, confirming that host-cell responses to infection can be distinguished by time. Pseudotime analysis identified a possible infection-specific cellular trajectory for *Chlamydia*-infected cells, and differential expression identified temporally expressed genes involved with cell cycle regulation, innate immune responses, cytoskeletal components, lipid biosynthesis and cellular stress. Overall, this pilot dataset and analyses highlighted the complex nature of infections, providing considerations for future single-cell-based experiments.

Chapter 5 – Examining different MOIs and depletion methods

This chapter captured host and chlamydial transcripts from two time points (1 and 24 hours) using RNA-seq. Within each time point, three MOIs (0.1, 1 and 10) were generated. Within each condition, six biological replicates were used, with three subjected to rRNA depletion and three subjected to rRNA depletion plus the depletion of polyadenylated transcripts. This

study is amongst a growing number of host-pathogen RNA-seq analyses and the second examining chlamydial infection. It also belongs to limited group of research examining MOIs and depletion methods, particularly in chlamydial research. Analyses identified that combing depletion methods increases the capture rate of chlamydial transcripts, but impacts host-cell expression. When the MOI is increased to 10, sequence capture rates significantly increase at both times, and are more beneficial for capturing chlamydial transcripts. Comparative analysis between MOIs show increased expression of inflammatory and immune-based genes, while chlamydial expression is more dynamic relative to the developmental stage. Overall, this work will help influence future NGS-based experimental designs by increasing capture rates and highlighting the impact different MOIs have, whereby achieving more specific infection-related biological outcomes.

1.5. References

Abdelrahman, Y., Ouellette, S.P., Belland, R.J., and Cox, J.V. (2016). Polarized Cell Division of *Chlamydia trachomatis*. PLoS pathogens *12*, e1005822.

AbdelRahman, Y.M., and Belland, R.J. (2005). The chlamydial developmental cycle. FEMS Microbiology Reviews *29*, 949-959.

Adiconis, X., Haber, A.L., Simmons, S.K., Levy Moonshine, A., Ji, Z., Busby, M.A., Shi, X., Jacques, J., Lancaster, M.A., Pan, J.Q.*, et al.* (2018). Comprehensive comparative analysis of 5′-end RNA-sequencing methods. Nat Methods *15*, 505-511.

Al-Rifai, K.M.J. (1988). Trachoma through history. International Ophthalmology *12*, 9-14.

Alhakami, H., Mirebrahim, H., and Lonardi, S. (2017). A comparative evaluation of genome assembly reconciliation tools. Genome Biol *18*, 93-93.

Alvesalo, J., Greco, D., Leinonen, M., Raitila, T., Vuorela, P., and Auvinen, P. (2008). Microarray Analysis of a *Chlamydia pneumoniae*-Infected Human Epithelial Cell Line by Use of Gene Ontology Hierarchy. The Journal of infectious diseases *197*, 156-162.

Amor, B. (1983). Chlamydia and Reiter's syndrome. British journal of rheumatology *22*, 156-160.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166-169.

Andersson, P., Harris, S.R., Smith, H.M.B.S., Hadfield, J., O'Neill, C., Cutcliffe, L.T., Douglas, F.P., Asche, L.V., Mathews, J.D., Hutton, S.I.*, et al.* (2016). *Chlamydia trachomatis* from Australian Aboriginal people with trachoma are polyphyletic composed of multiple distinctive lineages. Nature Communications *7*, 10688.

Andrews, S. (2010). FastQC. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Andrews, T.S., and Hemberg, M. (2018a). False signals induced by single-cell imputation. F1000Res *7*, 1740.

Andrews, T.S., and Hemberg, M. (2018b). M3Drop: dropout-based feature selection for scRNASeq. Bioinformatics *35*, 2865-2867.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M.*, et al.* (2004). UniProt: the Universal Protein knowledgebase. Nucleic acids research *32*, D115-119.

Armour, C.D., Castle, J.C., Chen, R., Babak, T., Loerch, P., Jackson, S., Shah, J.K., Dey, J., Rohl, C.A., Johnson, J.M.*, et al.* (2009). Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. Nat Methods *6*, 647-649.

Avital, G., Avraham, R., Fan, A., Hashimshony, T., Hung, D.T., and Yanai, I. (2017). scDual-Seq: mapping the gene regulatory program of *Salmonella* infection by host and pathogen single-cell RNA-sequencing. Genome Biol *18*, 200.

Avraham, R., Haseley, N., Brown, D., Penaranda, C., Jijon, H.B., Trombetta, J.J., Satija, R., Shalek, A.K., Xavier, R.J., Regev, A.*, et al.* (2015). Pathogen Cell-to-Cell Variability Drives Heterogeneity in Host Immune Responses. Cell *162*, 1309-1321.

Bacher, R., Chu, L.-F., Leng, N., Gasch, A.P., Thomson, J.A., Stewart, R.M., Newton, M., and Kendziorski, C. (2017). SCnorm: robust normalization of single-cell RNA-seq data. Nat Methods *14*, 584.

Bacher, R., and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biol *17*, 63-63.

Bachmann, N.L., Fraser, T.A., Bertelli, C., Jelocnik, M., Gillett, A., Funnell, O., Flanagan, C., Myers, G.S., Timms, P., and Polkinghorne, A. (2014). Comparative genomics of koala, cattle and sheep strains of Chlamydia pecorum. BMC Genomics *15*, 667.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic acids research *37*, W202-W208.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D.*, et al.* (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol *19*, 455-477.

Bastidas, R.J., Elwell, C.A., Engel, J.N., and Valdivia, R.H. (2013). Chlamydial intracellular survival strategies. Cold Spring Harb Perspect Med *3*, a010256-a010256.

Bavoil, P.M. (2014). What's in a word: the use, misuse, and abuse of the word "persistence" in *Chlamydia* biology. Frontiers in cellular and infection microbiology *4*, 27.

Bebear, C., and de Barbeyrac, B. (2009). Genital *Chlamydia trachomatis* infections. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases *15*, 4-10.

Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. Methods (San Diego, Calif) *58*, 268-276.

Besemer, J., and Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic acids research *33*, W451-W454.

Besser, J., Carleton, H.A., Gerner-Smidt, P., Lindsey, R.L., and Trees, E. (2018). Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases *24*, 335-341.

Betts-Hampikian, H.J., and Fields, K.A. (2010). The Chlamydial Type III Secretion Mechanism: Revealing Cracks in a Tough Nut. Frontiers in microbiology *1*, 114-114.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L.*, et al.* (2011). High density DNA methylation array with single CpG site resolution. Genomics *98*, 288-295.

Bierne, H., Hamon, M., and Cossart, P. (2012). Epigenetics and bacterial infections. Cold Spring Harb Perspect Med *2*, a010272.

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

Bommana, S., and Polkinghorne, A. (2019). Mini Review: Antimicrobial Control of Chlamydial Infections in Animals: Current Practices and Issues. Frontiers in Microbiology *10*.

Bossel Ben-Moshe, N., Hen-Avivi, S., Levitin, N., Yehezkel, D., Oosting, M., Joosten, L.A.B., Netea, M.G., and Avraham, R. (2019). Predicting bacterial infection outcomes using single cell RNA-sequencing analysis of human immune cells. Nature Communications *10*, 3266.

Boyle, A.P., Guinney, J., Crawford, G.E., and Furey, T.S. (2008). F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics *24*, 2537-2538.

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

Brennecke, P., Anders, S., Kim, J.K., Kolodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C.*, et al.* (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods *10*, 1093-1095.

Brinkworth, A.J., Wildung, M.R., and Carabeo, R.A. (2018). Genomewide Transcriptional Responses of Iron-Starved *Chlamydia trachomatis* Reveal Prioritization of Metabolic Precursor Synthesis over Protein Translation. mSystems *3*, e00184-00117.

Brunelle, B.W., Nicholson, T.L., and Stephens, R.S. (2004). Microarray-based genomic surveying of gene polymorphisms in *Chlamydia trachomatis*. Genome Biol *5*, R42-R42.

Brunelle, B.W., and Sensabaugh, G.F. (2006). The ompA gene in *Chlamydia trachomatis* differs in phylogeny and rate of evolution from other regions of the genome. Infection and immunity *74*, 578-585.

Brunham, R.C., and Rey-Ladino, J. (2005). Immunology of Chlamydia infection: implications for a *Chlamydia trachomatis* vaccine. Nature reviews Immunology *5*, 149-161.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods *10*, 1213-1218.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486.

Bumgarner, R. (2013). Overview of DNA microarrays: types, applications, and their future. Curr Protoc Mol Biol *Chapter 22*, Unit-22.21.

Burton, M.J. (2009). Prevention, treatment and rehabilitation. Community Eye Health *22*, 33-35.

Burton, M.J., and Mabey, D.C.W. (2009). The Global Burden of Trachoma: A Review. PLOS Neglected Tropical Diseases *3*, e460.

Byrne, G.I. (2010). *Chlamydia trachomatis* Strains and Virulence: Rethinking Links to Infection Prevalence and Disease Severity. The Journal of infectious diseases *201*, S126-S133.

Caldwell, H.D., Wood, H., Crane, D., Bailey, R., Jones, R.B., Mabey, D., Maclean, I., Mohammed, Z., Peeling, R., Roshick, C.*, et al.* (2003). Polymorphisms in *Chlamydia trachomatis* tryptophan synthase genes differentiate between genital and ocular isolates. The Journal of clinical investigation *111*, 1757-1769.

Campbell, M.S., Holt, C., Moore, B., and Yandell, M. (2014). Genome Annotation and Curation Using MAKER and MAKER-P. Current protocols in bioinformatics *48*, 4.11.11-14.11.39.

Carlson, J.H., Hughes, S., Hogan, D., Cieplak, G., Sturdevant, D.E., McClarty, G., Caldwell, H.D., and Belland, R.J. (2004). Polymorphisms in the *Chlamydia trachomatis* cytotoxin locus associated with ocular and genital isolates. Infection and immunity *72*, 7063-7072.

Carlson, J.H., Whitmire, W.M., Crane, D.D., Wicke, L., Virtaneva, K., Sturdevant, D.E., Kupko, J.J., 3rd, Porcella, S.F., Martinez-Orengo, N., Heinzen, R.A.*, et al.* (2008). The *Chlamydia trachomatis* plasmid is a transcriptional regulator of chromosomal genes and a virulence factor. Infection and immunity *76*, 2273-2283.

CDC (2014). Recommendations for the laboratory-based detection of *Chlamydia trachomatis* and Neisseria gonorrhoeae--2014. MMWR Recomm Rep *63*, 1-19.

CDC (2015a). 2015 STD Treatment Guidelines: *Chlamydia* Infections.

CDC (2015b). 2015 STD Treatment Guidelines: Lymphogranuloma Venereum (LGV).

CDC (2017). Sexually Transmitted Disease Surveillance 2017.

CDC (2019a). *Chlamydia* - CDC Fact Sheet (Detailed).

CDC (2019b). Hygiene-related Diseases: Trachoma.

Chernesky, M.A. (2005). The laboratory diagnosis of *Chlamydia trachomatis* infections. Can J Infect Dis Med Microbiol *16*, 39-44.

Cocchiaro, J.L., and Valdivia, R.H. (2009). New insights into *Chlamydia* intracellular survival mechanisms. Cell Microbiol *11*, 1571-1578.

Conway, T., Creecy, J.P., Maddox, S.M., Grissom, J.E., Conkle, T.L., Shadid, T.M., Teramoto, J., San Miguel, P., Shimada, T., Ishihama, A.*, et al.* (2014). Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. mBio *5*, e01442-01414.

Cui, K., and Zhao, K. (2012). Genome-wide approaches to determining nucleosome occupancy in metazoans using MNase-Seq. Methods Mol Biol *833*, 413-419.

Darville, T. (2005). *Chlamydia trachomatis* infections in neonates and young children. Seminars in pediatric infectious diseases *16*, 235-244.

Darville, T., and Hiltke, T.J. (2010). Pathogenesis of Genital Tract Disease Due to *Chlamydia trachomatis*. The Journal of infectious diseases *201*, S114-S125.

Davies, B., Turner, K.M.E., Frolund, M., Ward, H., May, M.T., Rasmussen, S., Benfield, T., and Westh, H. (2016). Risk of reproductive complications following *chlamydia* testing: a population-based retrospective cohort study in Denmark. The Lancet Infectious diseases *16*, 1057-1064.

Dawson, C.R., and Schachter, J. (1978). Sexually Transmitted Chlamydial Eye Infections Are Not Trachoma. JAMA *239*, 1790-1791.

De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. Bioinformatics *34*, 2666-2669.

de la Maza, L.M., Zhong, G., and Brunham, R.C. (2017). Update on *Chlamydia trachomatis* Vaccinology. Clinical and Vaccine Immunology *24*, e00543-00516.

De Puysseleyr, L., De Puysseleyr, K., Braeckman, L., Morre, S.A., Cox, E., and Vanrompay, D. (2017). Assessment of *Chlamydia suis* Infection in Pig Farmers. Transboundary and emerging diseases *64*, 826-833.

Depledge, D.P., Srinivas, K.P., Sadaoka, T., Bready, D., Mori, Y., Placantonakis, D.G., Mohr, I., and Wilson, A.C. (2019). Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. Nature Communications *10*, 754.

Derré, I. (2015). *Chlamydiae* interaction with the endoplasmic reticulum: contact, function and consequences. Cell Microbiol *17*, 959-966.

Ding, B., Zheng, L., and Wang, W. (2017). Assessment of Single Cell RNA-Seq Normalization Methods. G3 (Bethesda) *7*, 2039-2045.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2012). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., Amselem, J., Bouri, L., Bocs, S., Klopp, C.*, et al.* (2018). Ten steps to get started in Genome Assembly and Annotation. F1000Res *7*, ELIXIR-148.

Doni Jayavelu, N., Jajodia, A., Mishra, A., and Hawkins, R.D. (2018). An atlas of silencer elements for the human and mouse genomes. bioRxiv, 252304.

Elwell, C., Mirrashidi, K., and Engel, J. (2016). *Chlamydia* cell biology and pathogenesis. Nat Rev Microbiol *14*, 385-400.

Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. Nature protocols *2*, 953-971.

Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., and Theis, F.J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nature Communications *10*, 390.

Everett, K.D., Bush, R.M., and Andersen, A.A. (1999). Emended description of the order *Chlamydiales*, proposal of *Parachlamydiaceae* fam. nov. and *Simkaniaceae* fam. nov., each

containing one monotypic genus, revised taxonomy of the family *Chlamydiaceae*, including a new genus and five new species, and standards for the identification of organisms. International journal of systematic bacteriology *49 Pt 2*, 415-440.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047-3048.

Feibel, R.M. (2011). Fred Loe, MD, and the History of TrachomaFred Loe, MD, and the History of Trachoma. JAMA Ophthalmology *129*, 503-508.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M.*, et al.* (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol *16*, 278-278.

Gao, T., He, B., Liu, S., Zhu, H., Tan, K., and Qian, J. (2016). EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. Bioinformatics *32*, 3543-3551.

García-Campos, M.A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2015). Pathway Analysis: State of the Art. Frontiers in Physiology *6*.

Gaston, J.S.H. (2000). Immunological basis of *chlamydia* induced reactive arthritis. Sexually transmitted infections *76*, 156.

Gentleman R, C.V., Huber W, Hahne F (2011). genefilter: Methods for Filtering Genes from Microarray Experiments. R package version 1.34.0.

Grieshaber, S., Grieshaber, N., Yang, H., Baxter, B., Hackstadt, T., and Omsland, A. (2018). Impact of Active Metabolism on *Chlamydia trachomatis* Elementary Body Transcript Profile and Infectivity. Journal of Bacteriology *200*, e00065-00018.

Grieshaber, S.S., Grieshaber, N.A., and Hackstadt, T. (2003). *Chlamydia trachomatis* uses host cell dynein to traffic to the microtubule-organizing center in a p50 dynamitin-independent process. Journal of cell science *116*, 3793-3802.

Griffiths, E., and Gupta, R.S. (2002). Protein signatures distinctive of chlamydial species: horizontal transfers of cell wall biosynthesis genes glmU from archaea to *chlamydiae* and

murA between *chlamydiae* and *Streptomyces*. Microbiology (Reading, England) *148*, 2541-2549.

Griffiths, E., Ventresca, M.S., and Gupta, R.S. (2006). BLAST screening of chlamydial genomes to identify signature proteins that are unique for the *Chlamydiales*, *Chlamydiaceae*, *Chlamydophila* and *Chlamydia* groups of species. BMC Genomics *7*, 14-14.

Gruber, A.J., Schmidt, R., Gruber, A.R., Martin, G., Ghosh, S., Belmadani, M., Keller, W., and Zavolan, M. (2016). A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. Genome research *26*, 1145-1159.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072-1075.

Haber, A.L., Biton, M., Rogel, N., Herbst, R.H., Shekhar, K., Smillie, C., Burgin, G., Delorey, T.M., Howitt, M.R., Katz, Y.*, et al.* (2017). A single-cell survey of the small intestinal epithelium. Nature *551*, 333-339.

Hackstadt, T., Rockey, D.D., Heinzen, R.A., and Scidmore, M.A. (1996). *Chlamydia trachomatis* interrupts an exocytic pathway to acquire endogenously synthesized sphingomyelin in transit from the Golgi apparatus to the plasma membrane. EMBO J *15*, 964-977.

Hadfield, J., Harris, S.R., Seth-Smith, H.M.B., Parmar, S., Andersson, P., Giffard, P.M., Schachter, J., Moncada, J., Ellison, L., Vaulet, M.L.G.*, et al.* (2017). Comprehensive global genome dynamics of *Chlamydia trachomatis* show ancient diversification followed by contemporary mixing and recent lineage expansion. Genome research *27*, 1220-1229.

Hafner, L.M., Wilson, D.P., and Timms, P. (2014). Development status and future prospects for a vaccine against *Chlamydia trachomatis* infection. Vaccine *32*, 1563-1571.

Haggerty, C.L., Gottlieb, S.L., Taylor, B.D., Low, N., Xu, F., and Ness, R.B. (2010). Risk of sequelae after *Chlamydia trachomatis* genital infection in women. The Journal of infectious diseases *201 Suppl 2*, S134-155.

Halberstaedter, L. (1907). Ueber Zelleinschlusse parasitarer Natur beim Trachom. Arb K Gesundh Amt *26*, 44-47.

Harris, J.K., Kelley, S.T., Spiegelman, G.B., and Pace, N.R. (2003). The genetic core of the universal ancestor. Genome research *13*, 407-412.

Harris, S.R., Clarke, I.N., Seth-Smith, H.M.B., Solomon, A.W., Cutcliffe, L.T., Marsh, P., Skilton, R.J., Holland, M.J., Mabey, D., Peeling, R.W*., et al.* (2012). Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. In Nat Genet, pp. 413-419, S411.

Haugland, S., Thune, T., Fosse, B., Wentzel-Larsen, T., Hjelmevoll, S.O., and Myrmel, H. (2010). Comparing urine samples and cervical swabs for Chlamydia testing in a female population by means of Strand Displacement Assay (SDA). BMC women's health *10*, 9.

Hebenstreit, D. (2012). Methods, Challenges and Potentials of Single Cell RNA-seq. Biology *1*, 658-667.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell *38*, 576-589.

Hobolt-Pedersen, A.-S., Christiansen, G., Timmerman, E., Gevaert, K., and Birkelund, S. (2009). Identification of *Chlamydia trachomatis* CT621, a protein delivered through the type III secretion system to the host cell cytoplasm and nucleus. FEMS immunology and medical microbiology *57*, 46-58.

Hooppaw, A.J., and Fisher, D.J. (2016). A Coming of Age Story: *Chlamydia* in the Post-Genetic Era. Infection and immunity *84*, 612.

Hor, J., Gorski, S.A., and Vogel, J. (2018). Bacterial RNA Biology on a Genome Scale. Molecular cell *70*, 785-799.

Horn, M., Collingro, A., Schmitz-Esser, S., Beier, C.L., Purkhold, U., Fartmann, B., Brandt, P., Nyakatura, G.J., Droege, M., Frishman, D*., et al.* (2004). Illuminating the evolutionary history of *chlamydiae*. Science (New York, NY) *304*, 728-730.

Huang, D., Petrykowska, H.M., Miller, B.F., Elnitski, L., and Ovcharenko, I. (2019). Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. Genome research *29*, 657-667.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols *4*, 44-57.

Humphrys, M.S., Creasy, T., Sun, Y., Shetty, A.C., Chibucos, M.C., Drabek, E.F., Fraser, C.M., Farooq, U., Sengamalay, N., Ott, S.*, et al.* (2013). Simultaneous Transcriptional Profiling of Bacteria and Their Host Cells. PloS one *8*, e80597.

Hwang, B., Lee, J.H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Experimental & Molecular Medicine *50*, 96.

Hybiske, K., and Stephens, R.S. (2007). Mechanisms of host cell exit by the intracellular bacterium *Chlamydia*. Proceedings of the National Academy of Sciences of the United States of America *104*, 11430-11435.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2013). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods *11*, 163.

Janda, J.M., and Abbott, S.L. (2007). 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. Journal of clinical microbiology *45*, 2761-2764.

Johnson, C.M., and Fisher, D.J. (2013). Site-specific, insertional inactivation of incA in *Chlamydia trachomatis* using a group II intron. PloS one *8*, e83989.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007a). Genome-wide mapping of in vivo protein-DNA interactions. Science (New York, NY) *316*, 1497-1502.

Johnson, W.E., Li, C., and Rabinovic, A. (2007b). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics (Oxford, England) *8*, 118-127.

Ju, X., Li, D., and Liu, S. (2019). Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. Nature Microbiology.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research *28*, 27-30.

Kang, Y., McMillan, I., Norris, M.H., and Hoang, T.T. (2015). Single prokaryotic cell isolation and total transcript amplification protocol for transcriptomic analysis. Nature protocols *10*, 974-984.

Kang, Y., Norris, M.H., Zarzycki-Siek, J., Nierman, W.C., Donachie, S.P., and Hoang, T.T. (2011). Transcript amplification from single bacterium for transcriptome analysis. Genome research *21*, 925-935.

Kapoor, S. (2008). Re-emergence of lymphogranuloma venereum. Journal of the European Academy of Dermatology and Venereology : JEADV *22*, 409-416.

Kari, L., Goheen, M.M., Randall, L.B., Taylor, L.D., Carlson, J.H., Whitmire, W.M., Virok, D., Rajaram, K., Endresz, V., McClarty, G.*, et al.* (2011). Generation of targeted *Chlamydia trachomatis* null mutants. Proceedings of the National Academy of Sciences of the United States of America *108*, 7189-7193.

Kawasaki, E.S. (2004). Microarrays and the gene expression profile of a single cell. Annals of the New York Academy of Sciences *1020*, 92-100.

Kazar, J., Gillmore, J.D., and Gordon, F.B. (1971). Effect of Interferon and Interferon Inducers on Infections with a Nonviral Intracellular Microorganism, *Chlamydia trachomatis*. Infection and immunity *3*, 825-832.

Kessler, M., Hoffmann, K., Fritsche, K., Brinkmann, V., Mollenkopf, H.-J., Thieck, O., Teixeira da Costa, A.R., Braicu, E.I., Sehouli, J., Mangler, M.*, et al.* (2019). Chronic *Chlamydia* infection in human organoids increases stemness and promotes age-dependent CpG methylation. Nature Communications *10*, 1194.

Khan, A., and Zhang, X. (2015). dbSUPER: a database of super-enhancers in mouse and human genome. Nucleic acids research *44*, D164-D171.

Khan, A.R., Pervez, M.T., Babar, M.E., Naveed, N., and Shoaib, M. (2018). A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective. Evol Bioinform Online *14*, 1176934318758650-1176934318758650.

Kharchenko, P.V., Silberstein, L., and Scadden, D.T. (2014). Bayesian approach to single-cell differential expression analysis. Nat Methods *11*, 740.

Kieckens, E., Van den Broeck, L., Van Gils, M., Morre, S., and Vanrompay, D. (2018). Co-Occurrence of *Chlamydia suis* DNA and *Chlamydia suis*-Specific Antibodies in the Human Eye. Vector borne and zoonotic diseases (Larchmont, NY).

Kiefer, J.C. (2007). Epigenetics in development. Developmental dynamics : an official publication of the American Association of Anatomists *236*, 1144-1156.

Kirby-Institute (2018). HIV, viral hepatitis and sexually transmissible infections in Australia: annual surveillance report 2018 (Sydney: Kirby Institute, UNSW Sydney).

Kiselev, V.Y., Andrews, T.S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. Nature Reviews Genetics *20*, 273-282.

Kleba, B., and Stephens, R.S. (2008). Chlamydial effector proteins localized to the host cell cytoplasmic compartment. Infection and immunity *76*, 4842-4850.

Klemm, S.L., Shipony, Z., and Greenleaf, W.J. (2019). Chromatin accessibility and the regulatory epigenome. Nature Reviews Genetics *20*, 207-220.

Kokes, M., Dunn, J.D., Granek, J.A., Nguyen, B.D., Barker, J.R., Valdivia, R.H., and Bastidas, R.J. (2015). Integrating chemical mutagenesis and whole-genome sequencing as a platform for forward and reverse genetic analysis of *Chlamydia*. Cell host & microbe *17*, 716-725.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015). The technology and biology of single-cell RNA sequencing. Molecular cell *58*, 610-620.

Kolovos, P., Knoch, T.A., Grosveld, F.G., Cook, P.R., and Papantonis, A. (2012). Enhancers and silencers: an integrated and simple model for their function. Epigenetics & chromatin *5*, 1-1.

Kunz, D.J., Gomes, T., and James, K.R. (2018). Immune Cell Dynamics Unfolded by Single-Cell Technologies. Front Immunol *9*, 1435-1435.

Kuttippurathu, L., Hsing, M., Liu, Y., Schmidt, B., Maskell, D.L., Lee, K., He, A., Pu, W.T., and Kong, S.W. (2011). CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments. Bioinformatics *27*, 715-717.

Łabaj, P.P., Leparc, G.G., Linggi, B.E., Markillie, L.M., Wiley, H.S., and Kreil, D.P. (2011). Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. Bioinformatics *27*, i383-i391.

LaBrie, S.D., Dimond, Z.E., Harrison, K.S., Baid, S., Wickstrum, J., Suchland, R.J., and Hefty, P.S. (2019). Transposon Mutagenesis in *Chlamydia trachomatis* Identifies CT339 as a ComEC Homolog Important for DNA Uptake and Lateral Gene Transfer. MBio *10*.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P.*, et al.* (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome research *22*, 1813-1831.

Lange, F.D., Jones, K., Ritte, R., Brown, H.E., and Taylor, H.R. (2017). The impact of health promotion on trachoma knowledge, attitudes and practice (KAP) of staff in three work settings in remote Indigenous communities in the Northern Territory. PLoS neglected tropical diseases *11*, e0005503-e0005503.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods *9*, 357-359.

Larner, A.J. (2004). Ophthalmological observations made during the mid-19th-century European encounter with Africa. Archives of ophthalmology (Chicago, Ill : 1960) *122*, 267-272.

Leng, N., Chu, L.-F., Barry, C., Li, Y., Choi, J., Li, X., Jiang, P., Stewart, R.M., Thomson, J.A., and Kendziorski, C. (2015). Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. Nat Methods *12*, 947-950.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, G., Cai, L., Chang, H., Hong, P., Zhou, Q., Kulakova, E.V., Kolchanov, N.A., and Ruan, Y. (2014). Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. BMC Genomics *15*, S11.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094-3100.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, W.V., and Li, J.J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun *9*, 997.

Li, Y., and Tollefsbol, T.O. (2011). DNA methylation detection: bisulfite genomic sequencing analysis. Methods Mol Biol *791*, 11-21.

Liang, P., Rosas-Lemus, M., Patel, D., Fang, X., Tuz, K., and Juarez, O. (2018). Dynamic energy dependency of *Chlamydia trachomatis* on host cell metabolism during intracellular growth: Role of sodium-based energetics in chlamydial ATP generation. The Journal of biological chemistry *293*, 510-522.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923-930.

Lindner, R., and Friedel, C.C. (2012). A Comprehensive Evaluation of Alignment Algorithms in the Context of RNA-Seq. PloS one *7*, e52403.

Linnarsson, S. (2015). Sequencing Single Cells Reveals Sequential Stem Cell States. Cell stem cell *17*, 251-252.

Lischer, H.E.L., and Shimizu, K.K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. BMC Bioinformatics *18*, 474.

Liu, B., Liu, Y., Zang, T., and Wang, Y. (2019). deSALT: fast and accurate long transcriptomic read alignment with de Bruijn graph-based index. bioRxiv, 612176.

Liu, H., Bebu, I., and Li, X. (2010). Microarray probes and probe sets. Front Biosci (Elite Ed) *2*, 325-338.

Loman, N.J., and Quinlan, A.R. (2014). Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics *30*, 3399-3401.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550-550.

Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular Systems Biology *15*, e8746.

Lun, A., Bach, K., and Marioni, J.C. (2016a). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol *17*, 75.

Lun, A., McCarthy, D., and Marioni, J. (2016b). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor [version 2; peer review: 3 approved, 2 approved with reservations]. F1000Res *5*.

Lun, A.T., and Smyth, G.K. (2016). csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. Nucleic acids research *44*, e45.

Lun, A.T.L., Calero-Nieto, F.J., Haim-Vilmovsky, L., Gottgens, B., and Marioni, J.C. (2017). Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. Genome research *27*, 1795-1806.

Mabey, D., and Peeling, R.W. (2002). Lymphogranuloma venereum. Sexually transmitted infections *78*, 90-92.

Macaulay, I.C., and Voet, T. (2014). Single Cell Genomics: Advances and Future Perspectives. PLOS Genetics *10*, e1004126.

Marić, J., Sović, I., Križanović, K., Nagarajan, N., and Šikić, M. (2019). Graphmap2 - splice-aware RNA-seq mapper for long reads. bioRxiv, 720458.

Mariño-Ramírez, L., Kann, M.G., Shoemaker, B.A., and Landsman, D. (2005). Histone structure and nucleosome stability. Expert Rev Proteomics *2*, 719-729.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011 *17*, 3.

Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C.Y., Chou, A., Ienasescu, H.*, et al.* (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic acids research *42*, D142-147.

Matsumoto, A. (2019). Chapter 2: Structural Characteristics of Chlamydial Bodies. in Microbiology Of Chlamydia (CRC Press).

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K.*, et al.* (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic acids research *34*, D108-110.

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

McGinnis, C.S., Murrow, L.M., and Gartner, Z.J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Systems *8*, 329-337.e324.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology *28*, 495-501.

Meijer, A., Kwakkel, G.J., de Vries, A., Schouls, L.M., and Ossewaarde, J.M. (1997). Species identification of *Chlamydia* isolates by analyzing restriction fragment length polymorphism of the 16S-23S rRNA spacer region. Journal of clinical microbiology *35*, 1179-1183.

Menon, S., Timms, P., Allan, J.A., Alexander, K., Rombauts, L., Horner, P., Keltz, M., Hocking, J., and Huston, W.M. (2015). Human and Pathogen Factors Associated with <span class="named-content genus-species" id="named-content-

1">*Chlamydia trachomatis*-Related Infertility in Women. Clinical microbiology reviews *28*, 969.

Mersfelder, E.L., and Parthun, M.R. (2006). The tale beyond the tail: histone core domain modifications and the regulation of chromatin structure. Nucleic acids research *34*, 2653-2662.

Meyer, C.A., and Liu, X.S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nature Reviews Genetics *15*, 709.

Meyer, T. (2016). Diagnostic Procedures to Detect *Chlamydia trachomatis* Infections. Microorganisms *4*, 25.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P.D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic acids research *45*, D183-D189.

Miyairi, I., Ramsey, K.H., and Patton, D.L. (2010). Duration of Untreated Chlamydial Genital Infection and Factors Associated with Clearance: Review of Animal Studies. The Journal of infectious diseases *201*, S96-S103.

Mockler, T.C., Chan, S., Sundaresan, A., Chen, H., Jacobsen, S.E., and Ecker, J.R. (2005). Applications of DNA tiling arrays for whole-genome analysis. Genomics *85*, 1-15.

Mohammadpour, M., Abrishami, M., Masoumi, A., and Hashemi, H. (2016). Trachoma: Past, present and future. J Curr Ophthalmol *28*, 165-169.

Molloy, S. (2014). Chlamydia and the cytoskeleton. Nature Reviews Microbiology *12*, 792.

Moulder, J.W. (1966). The relation of the psittacosis group (*Chlamydiae*) to bacteria and viruses. Annual review of microbiology *20*, 107-130.

Mpiga, P., and Ravaoarinoro, M. (2006). *Chlamydia trachomatis* persistence: an update. Microbiological research *161*, 9-19.

Mueller, K.E., Wolf, K., and Fields, K.A. (2016). Gene Deletion by Fluorescence-Reported Allelic Exchange Mutagenesis in <span class="named-content genus-

species" id="named-content-1">*Chlamydia trachomatis*</span&gt. mBio *7*, e01817-01815.

Muramatsu, M.K., Brothwell, J.A., Stein, B.D., Putman, T.E., Rockey, D.D., and Nelson, D.E. (2016). Beyond Tryptophan Synthase: Identification of Genes That Contribute to *Chlamydia trachomatis* Survival during Gamma Interferon-Induced Persistence and Reactivation. Infection and immunity *84*, 2791-2801.

Nans, A., Ford, C., and Hayward, R.D. (2015). Host-pathogen reorganisation during host cell entry by *Chlamydia trachomatis*. Microbes Infect *17*, 727-731.

Nature-Research (2019). Hitting the mark in cancer epigenetics (https://www.nature.com/magazine-assets/d42473-019-00204-6/d42473-019-00204-6.pdf: Nature-Research).

Newburger, D.E., and Bulyk, M.L. (2009). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic acids research *37*, D77-82.

Newman, L., Rowley, J., Vander Hoorn, S., Wijesooriya, N.S., Unemo, M., Low, N., Stevens, G., Gottlieb, S., Kiarie, J., and Temmerman, M. (2015). Global Estimates of the Prevalence and Incidence of Four Curable Sexually Transmitted Infections in 2012 Based on Systematic Review and Global Reporting. PloS one *10*, e0143304.

Nguyen, B.D., and Valdivia, R.H. (2012). Virulence determinants in the obligate intracellular pathogen *Chlamydia trachomatis* revealed by forward genetic approaches. Proceedings of the National Academy of Sciences of the United States of America *109*, 1263-1268.

Nunes, A., and Gomes, J.P. (2014). Evolution, phylogeny, and molecular epidemiology of *Chlamydia*. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases *23*, 49-64.

Oakeshott, P., Kerry, S., Aghaizu, A., Atherton, H., Hay, S., Taylor-Robinson, D., Simms, I., and Hay, P. (2010). Randomised controlled trial of screening for *Chlamydia trachomatis* to prevent pelvic inflammatory disease: the POPI (prevention of pelvic infection) trial. BMJ *340*, c1642.

Omsland, A., Sixt, B.S., Horn, M., and Hackstadt, T. (2014). Chlamydial metabolism revisited: interspecies metabolic variability and developmental stage-specific physiologic activities. FEMS Microbiology Reviews *38*, 779-801.

Ong, C.T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nature reviews Genetics *12*, 283-293.

Ostaszewska, I., Zdrodowska-Stefanow, B., Darewicz, B., Darewicz, J., Badyda, J., Pucilo, K., Bulhak, V., and Szczurzewski, M. (2000). Role of *Chlamydia trachomatis* in epididymitis. Part II: Clinical diagnosis. Medical science monitor : international medical journal of experimental and clinical research *6*, 1119-1121.

Ouellette, S.P., Hatch, T.P., AbdelRahman, Y.M., Rose, L.A., Belland, R.J., and Byrne, G.I. (2006). Global transcriptional upregulation in the absence of increased translation in *Chlamydia* during IFNgamma-mediated host cell tryptophan starvation. Molecular microbiology *62*, 1387-1401.

Panzetta, M.E., Valdivia, R.H., and Saka, H.A. (2018). *Chlamydia* Persistence: A Survival Strategy to Evade Antimicrobial Effects in-vitro and in-vivo. Frontiers in Microbiology *9*.

Park, J.S., Helble, J.D., Lazarus, J.E., Yang, G., Blondel, C.J., Doench, J.G., Starnbach, M.N., and Waldor, M.K. (2019). A FACS-Based Genome-wide CRISPR Screen Reveals a Requirement for COPI in *Chlamydia trachomatis* Invasion. iScience *11*, 71-84.

Park, P.J. (2009). ChIP–seq: advantages and challenges of a maturing technology. Nature Reviews Genetics *10*, 669-680.

Pennini, M.E., Perrinet, S., Dautry-Varsat, A., and Subtil, A. (2010). Histone Methylation by NUE, a Novel Nuclear Effector of the Intracellular Pathogen *Chlamydia trachomatis*. PLoS pathogens *6*, e1000995.

Phillips, S., Quigley, B.L., and Timms, P. (2019). Seventy Years of *Chlamydia* Vaccine Research – Limitations of the Past and Directions for the Future. Frontiers in Microbiology *10*.

Pightling, A.W., Petronella, N., and Pagotto, F. (2014). Choice of Reference Sequence and Assembler for Alignment of Listeria monocytogenes Short-Read Sequence Data Greatly Influences Rates of Error in SNP Analyses. PloS one *9*, e104579.

Pikaard, C.S., and Mittelsten Scheid, O. (2014). Epigenetic regulation in plants. Cold Spring Harb Perspect Biol *6*, a019315-a019315.

Pliner, H.A., Shendure, J., and Trapnell, C. (2019). Supervised classification enables rapid annotation of cell atlases. Nat Methods *16*, 983-986.

Poran, A., Nötzel, C., Aly, O., Mencia-Trinchant, N., Harris, C.T., Guzman, M.L., Hassane, D.C., Elemento, O., and Kafsack, B.F.C. (2017). Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites. Nature *551*, 95.

Portela, A., and Esteller, M. (2010). Epigenetic modifications and human disease. Nature biotechnology *28*, 1057-1068.

Poston, T.B., Gottlieb, S.L., and Darville, T. (2017). Status of vaccine research and development of vaccines for *Chlamydia trachomatis* infection. Vaccine.

Prost, A., and Négrel, A.D. (1989). Water, trachoma and conjunctivitis. Bull World Health Organ *67*, 9-18.

Putman, T., Hybiske, K., Jow, D., Afrasiabi, C., Lelong, S., Cano, M.A., Wu, C., and Su, A.I. (2019). ChlamBase: a curated model organism database for the Chlamydia research community. Database (Oxford) *2019*, baz041.

Quainoo, S., Coolen, J.P.M., van Hijum, S.A.F.T., Huynen, M.A., Melchers, W.J.G., van Schaik, W., and Wertheim, H.F.L. (2017). Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. Clinical Microbiology Reviews *30*, 1015.

Rahman, M.U., Hudson, A.P., and Schumacher, H.R., Jr. (1992). *Chlamydia* and Reiter's syndrome (reactive arthritis). Rheumatic diseases clinics of North America *18*, 67-79.

Rahmani, E., Schweiger, R., Rhead, B., Criswell, L.A., Barcellos, L.F., Eskin, E., Rosset, S., Sankararaman, S., and Halperin, E. (2019). Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. Nature Communications *10*, 3417.

Rajić, J., Inic-Kanada, A., Stein, E., Dinić, S., Schuerer, N., Uskoković, A., Ghasemian, E., Mihailović, M., Vidaković, M., Grdović, N.*, et al.* (2017). *Chlamydia trachomatis* Infection Is Associated with E-Cadherin Promoter Methylation, Downregulation of E-Cadherin

Expression, and Increased Expression of Fibronectin and α-SMA-Implications for Epithelial-Mesenchymal Transition. Frontiers in cellular and infection microbiology *7*, 253-253.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic acids research *42*, W187-W191.

Rang, F.J., Kloosterman, W.P., and de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol *19*, 90.

Read, T.D., Brunham, R.C., Shen, C., Gill, S.R., Heidelberg, J.F., White, O., Hickey, E.K., Peterson, J., Utterback, T., Berry, K.*, et al.* (2000). Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic acids research *28*, 1397-1406.

Redgrove, K.A., and McLaughlin, E.A. (2014). The Role of the Immune Response in *Chlamydia trachomatis* Infection of the Male Genital Tract: A Double-Edged Sword. Front Immunol *5*, 534-534.

Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M.*, et al.* (2017). The Human Cell Atlas. Elife *6*.

Reid, A.J., Talman, A.M., Bennett, H.M., Gomes, A.R., Sanders, M.J., Illingworth, C.J.R., Billker, O., Berriman, M., and Lawniczak, M.K.N. (2018). Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. eLife *7*, e33105.

Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nature biotechnology *32*, 896-902.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol *11*, R25.

Rohde, G., Straube, E., Essig, A., Reinhold, P., and Sachse, K. (2010). Chlamydial zoonoses. Dtsch Arztebl Int *107*, 174-180.

Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R., Dunning, M.J., Brown, G.D., Gojis, O., Ellis, I.O., Green, A.R.*, et al.* (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature *481*, 389-393.

Rotem, A., Ram, O., Shoresh, N., Sperling, R.A., Goren, A., Weitz, D.A., and Bernstein, B.E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nature biotechnology *33*, 1165-1172.

Sachse, K., Bavoil, P.M., Kaltenboeck, B., Stephens, R.S., Kuo, C.C., Rossello-Mora, R., and Horn, M. (2015). Emendation of the family *Chlamydiaceae*: proposal of a single genus, *Chlamydia*, to include all currently recognized species. Systematic and applied microbiology *38*, 99-103.

Saka, H.A., Thompson, J.W., Chen, Y.S., Kumar, Y., Dubois, L.G., Moseley, M.A., and Valdivia, R.H. (2011). Quantitative proteomics reveals metabolic and pathogenic properties of *Chlamydia trachomatis* developmental forms. Molecular microbiology *82*, 1185-1203.

Saliba, A.-E., Li, L., Westermann, A.J., Appenzeller, S., Stapels, D.A.C., Schulte, L.N., Helaine, S., and Vogel, J. (2016). Single-cell RNA-seq ties macrophage polarization to growth rate of intracellular *Salmonella*. Nature Microbiology *2*, 16206.

Salmela, L., and Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. Bioinformatics *30*, 3506-3514.

Salmon-Divon, M., Dvinge, H., Tammoja, K., and Bertone, P. (2010). PeakAnalyzer: Genome-wide annotation of chromatin binding and modification loci. BMC Bioinformatics *11*, 415.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nature biotechnology *33*, 495-502.

Schachter, J. (1977). The Expanding Clinical Spectrum of Infections with *Chlamydia trachomatis*. Sexually Transmitted Diseases *4*, 116-118.

Schiex, T., Moisan, A., and Rouzé, P. (2000). EuGene: an eukaryotic gene finder that combines several sources of evidence. Paper presented at: International Conference on Biology, Informatics, and Mathematics (Springer).

Schwartzman, O., and Tanay, A. (2015). Single-cell epigenomics: techniques and emerging applications. Nature Reviews Genetics *16*, 716.

Scialdone, A., Natarajan, K.N., Saraiva, L.R., Proserpio, V., Teichmann, S.A., Stegle, O., Marioni, J.C., and Buettner, F. (2015). Computational assignment of cell-cycle stage from single-cell transcriptome data. Methods (San Diego, Calif) *85*, 54-61.

Scidmore, M.A., Fischer, E.R., and Hackstadt, T. (1996). Sphingolipids and glycoproteins are differentially trafficked to the *Chlamydia trachomatis* inclusion. The Journal of cell biology *134*, 363-374.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068-2069.

Seth-Smith, H.M.B., Harris, S.R., Persson, K., Marsh, P., Barron, A., Bignell, A., Bjartling, C., Clark, L., Cutcliffe, L.T., Lambden, P.R.*, et al.* (2009). Co-evolution of genomes and plasmids within *Chlamydia trachomatis* and the emergence in Sweden of a new variant strain. BMC Genomics *10*, 239.

Shang, J., Zhu, F., Vongsangnak, W., Tang, Y., Zhang, W., and Shen, B. (2014). Evaluation and Comparison of Multiple Aligners for Next-Generation Sequencing Data Analysis. BioMed Research International *2014*, 16.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature reviews Genetics *14*, 618-630.

Shema, E., Bernstein, B.E., and Buenrostro, J.D. (2019). Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. Nat Genet *51*, 19-25.

Shuman, H.A., and Silhavy, T.J. (2003). The art and design of genetic screens: *Escherichia coli*. Nature reviews Genetics *4*, 419-431.

Siezen, R.J., Wilson, G., and Todt, T. (2010). Prokaryotic whole-transcriptome analysis: deep sequencing and tiling arrays. Microb Biotechnol *3*, 125-130.

Sigalova, O., Chaplin, A.V., Bochkareva, O.O., Shelyakin, P.V., Filaretov, V.A., Akkuratov, E.E., Burskaya, V., and Gelfand, M.S. (2019). *Chlamydia* pan-genomic analysis reveals balance between host adaptation and selective pressure to genome reduction. bioRxiv, 506121.

Sigar, I.M., Schripsema, J.H., Wang, Y., Clarke, I.N., Cutcliffe, L.T., Seth-Smith, H.M.B., Thomson, N.R., Bjartling, C., Unemo, M., Persson, K.*, et al.* (2014). Plasmid deficiency in urogenital isolates of *Chlamydia trachomatis* reduces infectivity and virulence in a mouse model. Pathogens and disease *70*, 61-69.

Simon, J.M., Giresi, P.G., Davis, I.J., and Lieb, J.D. (2012). Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. Nature protocols *7*, 256-267.

Smallwood, S.A., Lee, H.J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S.R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods *11*, 817-820.

Song, L., Carlson, J.H., Whitmire, W.M., Kari, L., Virtaneva, K., Sturdevant, D.E., Watkins, H., Zhou, B., Sturdevant, G.L., Porcella, S.F.*, et al.* (2013). *Chlamydia trachomatis* plasmid-encoded Pgp4 is a transcriptional regulator of virulence-associated genes. Infection and immunity *81*, 636-644.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics *24*, 637-644.

Stark, D., van Hal, S., Hillman, R., Harkness, J., and Marriott, D. (2007). Lymphogranuloma venereum in Australia: anorectal *Chlamydia trachomatis* serovar L2b in men who have sex with men. Journal of clinical microbiology *45*, 1029-1031.

Stark, R., and Brown, G. (2019). DiffBind Vignette: Differential binding analysis of ChIP-Seq peak data.

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. Nature reviews Genetics.

Stegle, O., Teichmann, S.A., and Marioni, J.C. (2015). Computational and analytical challenges in single-cell transcriptomics. Nature reviews Genetics *16*, 133-145.

Stephens, R.S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R.L., Zhao, Q.*, et al.* (1998). Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science (New York, NY) *282*, 754-759.

Stephens, R.S., Myers, G., Eppinger, M., and Bavoil, P.M. (2009). Divergence without difference: phylogenetics and taxonomy of *Chlamydia* resolved. FEMS immunology and medical microbiology *55*, 115-119.

Stephens, R.S., Tam, M.R., Kuo, C.C., and Nowinski, R.C. (1982). Monoclonal antibodies to *Chlamydia trachomatis*: antibody specificities and antigen characterization. The Journal of Immunology *128*, 1083.

Stoner, B.P., and Cohen, S.E. (2015). Lymphogranuloma Venereum 2015: Clinical Presentation, Diagnosis, and Treatment. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America *61 Suppl 8*, S865-873.

Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics *19*, 477.

Sunkin, S.M., Ng, L., Lau, C., Dolbeare, T., Gilbert, T.L., Thompson, C.L., Hawrylycz, M., and Dang, C. (2013). Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic acids research *41*, D996-D1008.

Svensson, V., Natarajan, K.N., Ly, L.-H., Miragaia, R.J., Labalette, C., Macaulay, I.C., Cvejic, A., and Teichmann, S.A. (2017). Power analysis of single-cell RNA-sequencing experiments. Nat Methods *14*, 381.

Svensson, V., Vento-Tormo, R., and Teichmann, S.A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. Nature protocols *13*, 599-604.

Szreter, S. (2019). The Hidden Affliction: Sexually Transmitted Infections and Infertility in History (University of Rochester Press).

Taborisky, J. (1952). Historic and ethnologic factors in the distribution of trachoma. American journal of ophthalmology *35*, 1305-1311.

Tan, G., Opitz, L., Schlapbach, R., and Rehrauer, H. (2019). Long fragments achieve lower base quality in Illumina paired-end sequencing. Sci Rep *9*, 2856-2856.

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K., and Surani, M.A. (2010). Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell stem cell *6*, 468-478.

Tanizawa, Y., Fujisawa, T., and Nakamura, Y. (2017). DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. Bioinformatics *34*, 1037-1039.

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Current protocols in bioinformatics *Chapter 4*, Unit 4.10.

Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F.J., Del Risco, H., Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K.*, et al.* (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome research.

Taylor-Brown, A., Vaughan, L., Greub, G., Timms, P., and Polkinghorne, A. (2015). Twenty years of research into Chlamydia-like organisms: a revolution in our understanding of the biology and pathogenicity of members of the phylum *Chlamydiae*. Pathogens and disease *73*, 1-15.

The-Gene-Ontology-Consortium (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic acids research *47*, D330-d338.

The-Papyrus-Ebers (1932). The Papyrus Ebers. Translated from the German Version. JAMA *98*, 842-842.

Thomas, R., Thomas, S., Holloway, A.K., and Pollard, K.S. (2016). Features that define the best ChIP-seq peak calling algorithms. Briefings in Bioinformatics *18*, 441-450.

Thomson, N.R., Yeats, C., Bell, K., Holden, M.T., Bentley, S.D., Livingstone, M., Cerdeno-Tarraga, A.M., Harris, B., Doggett, J., Ormond, D.*, et al.* (2005). The *Chlamydophila abortus* genome sequence reveals an array of variable proteins that contribute to interspecies variation. Genome research *15*, 629-640.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B.*, et al.* (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75-82.

Thygeson, P. (1939). The Treatment of Trachoma with Sulfanilamide: A Report of 28 Cases. Transactions of the American Ophthalmological Society *37*, 395-403.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature biotechnology *32*, 381-386.

Tsompana, M., and Buck, M.J. (2014). Chromatin accessibility: a window into the genome. Epigenetics & Chromatin *7*, 33.

Valdivia, R.H. (2008). *Chlamydia* effector proteins and new insights into chlamydial cellular microbiology. Current opinion in microbiology *11*, 53-59.

Vallejos, C.A., Richardson, S., and Marioni, J.C. (2016). Beyond comparisons of means: understanding changes in gene expression at the single-cell level. Genome Biol *17*, 70.

Vasileva, H., Butcher, R., Pickering, H., Sokana, O., Jack, K., Solomon, A.W., Holland, M.J., and Roberts, C.h. (2018). Conjunctival transcriptome profiling of Solomon Islanders with active trachoma in the absence of *Chlamydia trachomatis* infection. Parasites & Vectors *11*, 104.

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L.A. (2006). VISTA Enhancer Browser—a database of tissue-specific human enhancers. Nucleic acids research *35*, D88-D92.

Vivancos, A.P., Güell, M., Dohm, J.C., Serrano, L., and Himmelbauer, H. (2010). Strand-specific deep sequencing of the transcriptome. Genome research *20*, 989-999.

Waddington, C.H. (1956). Genetic Assimilation of the Bithorax Phenotype. Evolution *10*, 1-13.

Waddington, C.H. (1957). The strategy of the genes: a discussion of some aspects of theoretical biology (Allen & Unwin).

Wang, B., Zhang, L., Zhang, T., Wang, H., Zhang, J., Wei, J., Shen, B., Liu, X., Xu, Z., and Zhang, L. (2013). *Chlamydia pneumoniae* infection promotes vascular smooth muscle cell migration through a Toll-like receptor 2-related signaling pathway. Infection and immunity *81*, 4583-4591.

Wang, J., Chen, L., Chen, Z., and Zhang, W. (2015). RNA-seq based transcriptomic analysis of single bacterial cells. Integrative biology : quantitative biosciences from nano to macro *7*, 1466-1476.

Wang, J., Zibetti, C., Shang, P., Sripathi, S.R., Zhang, P., Cano, M., Hoang, T., Xia, S., Ji, H., Merbs, S.L.*, et al.* (2018). ATAC-Seq analysis reveals a widespread decrease of chromatin accessibility in age-related macular degeneration. Nature Communications *9*, 1364.

Wang, S.P., Kuo, C.C., Barnes, R.C., Stephens, R.S., and Grayston, J.T. (1985). Immunotyping of *Chlamydia trachomatis* with monoclonal antibodies. The Journal of infectious diseases *152*, 791-800.

Wang, Y., Kahane, S., Cutcliffe, L.T., Skilton, R.J., Lambden, P.R., and Clarke, I.N. (2011). Development of a Transformation System for *Chlamydia trachomatis*: Restoration of Glycogen Biosynthesis by Acquisition of a Plasmid Shuttle Vector. PLoS pathogens *7*, e1002258.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics *10*, 57-63.

White, J.A. (2009). Manifestations and management of lymphogranuloma venereum. Current opinion in infectious diseases *22*, 57-66.

WHO (2016a). WHO | Trachoma control: a guide for programme managers. WHO.

WHO (2016b). WHO | WHO guidelines for the treatment of *Chlamydia trachomatis* (World Health Organization).

WHO (2019). Trachoma.

Wick, R.R., Judd, L.M., Gorrie, C.L., and Holt, K.E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLOS Computational Biology *13*, e1005595.

Wilhelm-Benartzi, C.S., Koestler, D.C., Karagas, M.R., Flanagan, J.M., Christensen, B.C., Kelsey, K.T., Marsit, C.J., Houseman, E.A., and Brown, R. (2013). Review of processing and analysis methods for DNA methylation array data. British Journal of Cancer *109*, 1394-1402.

Willbanks, A., Leary, M., Greenshields, M., Tyminski, C., Heerboth, S., Lapinska, K., Haskins, K., and Sarkar, S. (2016). The Evolution of Epigenetics: From Prokaryotes to Humans and Its Biological Consequences. Genet Epigenet *8*, 25-36.

Wingett, S.W., and Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. F1000Res *7*, 1338-1338.

Wong, G.K., Passey, D.A., Huang, Y., Yang, Z., and Yu, J. (2000). Is "junk" DNA mostly intron DNA? Genome research *10*, 1672-1678.

Wu, D.C., Yao, J., Ho, K.S., Lambowitz, A.M., and Wilke, C.O. (2018). Limitations of alignment-free tools in total RNA-seq quantification. BMC Genomics *19*, 510.

Xiong, W.M., Xu, Q.P., Xiao, R.D., Hu, Z.J., Cai, L., and He, F. (2019). Genome-wide DNA methylation and RNA expression profiles identified RIPK3 as a differentially methylated gene in *Chlamydia pneumoniae* infection lung carcinoma patients in China. Cancer management and research *11*, 5785-5797.

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics *13*, 329-342.

Young, M.D., Wakefield, M.J., Smyth, G.K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol *11*, R14.

Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics *25*, 1952-1958.

Zappia, L., Phipson, B., and Oshlack, A. (2018). Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLOS Computational Biology *14*, e1006245.

Zhang, A.W., O'Flanagan, C., Chavez, E.A., Lim, J.L.P., Ceglia, N., McPherson, A., Wiens, M., Walters, P., Chan, T., Hewitson, B.*, et al.* (2019). Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. Nat Methods *16*, 1007-1015.

Zhang, Q., and Cao, X. (2019). Epigenetic regulation of the innate immune response to infection. Nature Reviews Immunology *19*, 417-432.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.

Zhao, Q., Wang, J., Levichkin, I.V., Stasinopoulos, S., Ryan, M.T., and Hoogenraad, N.J. (2002). A mitochondrial specific stress response in mammalian cells. EMBO J *21*, 4411-4419.

Zhong, G. (2017). Chlamydial Plasmid-Dependent Pathogenicity. Trends Microbiol *25*, 141-152.

Zhu, L.J., Gazin, C., Lawson, N.D., Pages, H., Lin, S.M., Lapointe, D.S., and Green, M.R. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC bioinformatics *11*, 237.

Chapter 2

Research methodology

2.1. Laboratory-based methods and materials

2.1.1. Chlamydial tissue culture and infections

Human epithelial type 2 (HEp-2) cells (American Type Culture Collection, ATCC No. CCL-23) were grown as monolayers in 6 x 100 mm tissue culture dishes until cells were 90% confluent. To harvest EBs for the subsequent infections, additional monolayers were grown and infected with *C. trachomatis* serovar E in sucrose phosphate glutamate (SPG) as previously outlined (Tan et al., 2009). The resulting EBs and cell lysates were then harvested and used to infect new HEp2 monolayers.

Infections for each dataset used the previously prepared HEp2 monolayers, infecting with *C. trachomatis* serovar E in 3.5 mL SPG buffer as previously described (Tan et al., 2009); using centrifugation to synchronise infections. The ratio of EBs was 1:1 (MOI 1) for the FAIRE-seq and scRNA-seq samples, whereas bulk-RNA-seq samples contained three different MOIs (0.1, 1 and 10). Mock-infected samples were generated using the same HEp-2 monolayers without the presence of any EBs. To remove non-viable or dead EBs, each sample was incubated at 25ºC for 2 hours, and washed twice in SPG. Cell monolayers were incubated at 37ºC with 5% CO2, including the addition of 10 ml of fresh medium (DMEM+10% FBS, 25 µg/ml gentamycin, 1.25 µg/ml Fungizone). After each infection time point, the infected and mock-infected dishes were harvested by scraping and resuspending in 150 µl sterile PBS. Resuspended samples were stored at −80°C.

2.1.2. Library preparation and sequencing

The detailed methods for each sequencing approach applied in this thesis (FAIRE-seq, scRNA-seq and bulk RNA-seq) are available from each chapter. All sequencing and library preparations were performed at the Genome Resource Center, Institute for Genome Sciences, University of Maryland School of Medicine.

2.1.2.1. FAIRE-seq

Infected and mock-infected samples were prepared by crosslinking with formaldehyde, followed by cell lysis and sonication to generate an average DNA fragment size of approximately 300-400 bp. Phenol-chloroform extraction separated cross-linked DNA (representing fragments of open chromatin) to the aqueous layer, which were then extracted and purified; all steps were performed as previously described (Simon et al., 2012). Libraries were prepared in triplicate from infected and mock-infected samples at 1, 12, 24 and 48 hours, using the Illumina TruSeq Sample Prep kit. The subsequent libraries were sequenced on an Illumina HiSeq2000 using the 100 bp paired-end protocol.

2.1.2.2. scRNA-seq

Single cells were obtained from monolayers grown to 3, 6 and 12 hours, from both infected and mock-infected conditions. Single cells were isolated and collected using the C1 Single-Cell Auto Prep IFC microfluidic chip (Fluidigm). A balanced design was applied across three 96-well plates, with each plate comprising all three time points and both conditions. Each plate contained the same design, but in different configurations, to minimise any plate or experiment-related confounding effects. Libraries were constructed using Illumina's Nextera XT library prep kit per the recommended Fluidigm protocol and were sequenced on an Illumina HiSeq 4000, using the 150 bp paired-end protocol.

2.1.2.3. Dual RNA-seq

Two kits were used simultaneously (Human/Mouse/Rat and Gram-negative) to deplete samples of both human and gram-negative bacterial rRNA. Each sample was equally separated, with one half further subjected to poly-A depletion. This removed host-derived poly-A transcripts, thus enriching samples for bacterial transcripts. Magnetic beads were used to bind to poly-A mRNAs and were extracted from the solution with a magnet. The mRNA libraries were prepared from depleted samples at 1 and 24 hours post infection, using the TruSeq RNA Sample Prep kit. The resulting libraries were sequenced on an Illumina HiSeq2000 using the 100 bp paired-end protocol.

2.2. Bioinformatic methods

Each of the three experimental chapters contain detailed bioinformatic analysis specific to their sequencing approach and can be referred to separately for the full analyses. The following subsections contain bioinformatic analysis overlapping all three sequencing approaches, highlighting different software and the underlying methods across common steps.

With the ever-increasing amount of tools available for each task, software packages were chosen based on the following criteria: 1) good documentation with examples; 2) reproducibility; and 3) is consistently being updated.

2.2.1. Quality control

FastQC (Andrews, 2010) was used to check the quality of raw sequencing reads, in addition to confirmation of the removal of adaptors, low quality reads and other sequencing-based metrics. FastQ Screen (Wingett and Andrews, 2018) was used to align the trimmed reads against a pre-selection of indexed genomes, looking for sources of contamination. This highlighted the 3 hour mock-infected cells from scRNA-seq, which aligned to *Escherichia coli* and *Mus musculus* genomes.

Throughout each of the quality control steps and subsequent analysis outlined below, MultiQC (Ewels et al., 2016) was used to collate output from software into single reports. This was especially useful for scRNA-seq with over 300 cells, bulk RNA-seq (32 replicates) and FAIRE-seq (24 replicates).

2.2.2. Mapping and feature counting

Both the scRNA-seq and bulk RNA-seq datasets were generated from sequenced transcripts, allowing the same underlying mapping and feature counting software to be used. Mapping the transcripts to the human genome was completed with the splice-aware aligner STAR (Dobin et al., 2013), allowing alternatively spliced transcripts to be aligned correctly. In addition, the bulk RNA-seq dataset required alignment of transcripts to the chlamydial genome. As bacteria including *Chlamydia* do not typically undergo alternative splicing-based mechanisms to their operonic structures, the non-splice-aware aligner Bowtie2 (Langmead and Salzberg, 2012) was used. FeatureCounts (Liao et al., 2014) was used to count the aligned transcripts within genomic features (genes), discarding transcripts that aligned to more than one gene, or aligned within an area where two or more genes overlapped.

Sequenced open chromatin fragments from FAIRE-seq were aligned to the human genome using Bowtie2 (Langmead and Salzberg, 2012), as alternative splicing events are not able to be determined. Annotation was performed with Homer (Heinz et al., 2010), providing a range of genomic features that included promoters, introns, exons, and intragenic regions. To identify enhancers, intragenic regions were uploaded to two online databases, dbSuper (Khan and Zhang, 2015) and Enhancer atlas (Gao et al., 2016); both containing experimentally validated enhancers from HeLa cells (S3 and S4).

2.2.3. Filtering, normalisation and confounding effects

The majority of NGS experiments capture DNA and RNA fragments from an entire genome. The challenge is identifying key biological signals from background signals, such as low expressed genes that can disrupt the underlying statistical models (Chen et al., 2016). To reduce the background signal, genomic regions/genes or are discarded if the number of aligned reads are less than a pre-determined threshold. To further increase statistical power, confounding effects that include variability in the library sizes, are identified and significantly reduced or removed.

Counted features from the scRNA-seq dataset were stored in a count matrix containing rows of genes and columns representing each cell. Filtering lowly expressed genes (counts needed to be greater than zero in four or more cells) reduced the number of genes from 58k to 23k. To normalise the variation in library sizes, the single cell specific deconvolution method from Scran (Lun et al., 2016) was used. Next, a custom filter was created to remove low quality cells, consisting of four steps: 1) total mapped reads should be greater than 1,000,000; 2) total features greater than 5,000; 3) expression from mitochondrial genes less than 20% (identifying stressed and damaged cells (Zhao et al., 2002)); and 4) expression from rRNA

genes less than 10% (confirming rRNA removal was successful). RUVSeq (Risso et al., 2014) was then used to remove confounding effects resulting from sequencing different batches of cells and other biological factors, such as the cell cycle.

In the bulk RNA-seq dataset, human and chlamydial reads were separated by time point due to vast differences in library sizes between MOIs 0.1 and 10. The human count matrix was filtered using genefilter (Gentleman et al., 2018) ensuring that expression was greater than 50 in at least 3 replicates (reduced to 15k genes at 1 hour and 10k genes at 24 hours). The chlamydial count matrix containing only ~1,000 genes had a more relaxed filtering condition with expression above 10 in at least 3 replicates (reduced to 457 genes at 1 hour and 986 genes at 24 hours). To normalise for differences in library size, the trimmed mean of M-values (TMM) method (Robinson and Oshlack, 2010) was used for both species.

In the FAIRE-seq dataset, aligned reads corresponded to peaks representing areas of open chromatin. MACS2 (Zhang et al., 2008) was used to highlight statistically significant enriched genomic regions (peaks) genome-wide. Peaks from each time point were grouped, creating four consensus peak sets, each with six samples. Low abundant peaks were removed if they did not appear in at least two replicates. Further 'low coverage' peaks were removed if they contained < 3 mapped reads after normalising for library size.

2.2.4. Differential comparisons

All three experimental designs were created to identify infection-specific mechanisms by comparing an infected condition against an equivalent mock-infected condition.

The scRNA-seq comparisons examined differences between infected and mock-infected cells at 6 and 12 hours, but not 3 hours due to the loss of the mock-infected cells. In addition, comparisons between the two main clusters of cells (as identified through unsupervised

clustering) were also examined. Due to the limitation that differential expression software for single-cell data only allow for basic comparisons (i.e. cluster A vs cluster B), edgeR (Robinson et al., 2010) (predominantly used for bulk RNA-seq) was used, allowing for more complex experimental designs to be incorporated. With guidance from one of the developers (Professor Gordon Smyth, Walter and Eliza Hall Institute of Medical Research), a model was created comparing the two main clusters that took into consideration infected and mock-infected cells, and the absence of the 3 hour mock-infected cells.

To compare differences across MOIs from the bulk RNA-seq dataset, edgeR (Robinson et al., 2010) was also used. To increase the statistical significance by using all six replicates at each MOI, the depletion methods were added as a blocking factor to each comparative model. Using the consensus peak sets from the FAIRE-seq dataset, differential comparisons between infected and mock-infected peaks were identified using the inbuilt DESeq2 method from within Diffbind (Stark and Brown, 2011). Although DESeq2 was developed for RNA-seq experiments, the underlying count matrix was designed in a similar format allowing comparisons to be made; with columns as individual replicates, rows as genomic features, and populated with the number of aligned fragments. Positive fold changes therefore correspond to an increase in open chromatin, and negative fold changes to a decrease in open chromatin.

2.2.5. Figure creation and manipulation

Base R and GGplot2 (Wickham, 2009) were used to create the majority of graphs and related graphical output. Other numerical-based figures were created using R-based packages, for example PCA and volcano plots from PCAtools (Blighe and Lewis, 2018), and heatmaps from Pheatmap (Kolde, 2011).

Adobe Photoshop and Illustrator CC were used to create and manipulate individual figure components, including lifecycle diagrams and genomic annotation. Illustrator was used as the underlying tool to organise and collate sub-elements for each main figure.

2.2.6. Computational resources

All raw files, quality control steps, mapping, and other analysis that required extensive computational resources, were run on the ARCLab high performance computing cluster operated by the University of Technology Sydney. The underlying nodes were built on various distributions of RedHat Linux, with the majority of bioinformatic-based software installed using Bioconda (Grüning et al., 2018) where available. The Intersect (https://intersect.org.au) cloud-based data repository 'SpaceShuttle' was used to store all raw and processed files, using Aspera (https://asperasoft.com) for file transfers.

2.3. References

Andrews, S. (2010). FastQC A Quality Control tool for High Throughput Sequence Data.

Blighe, K., and Lewis, M. (2018). PCAtools: everything Principal Components Analysis. https://githubcom/kevinblighe/PCAtools.

Chen, Y., Lun, A.T.L., and Smyth, G.K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; peer review: 5 approved]. F1000Res *5*.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047-3048.

Gao, T., He, B., Liu, S., Zhu, H., Tan, K., and Qian, J. (2016). EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. Bioinformatics *32*, 3543-3551.

Gentleman, R., Carey, V., Huber, W., and Hahne, F. (2018). genefilter: methods for filtering genes from microarray experiments - R package version 1.64.0.

Grüning, B., Dale, R., Sjödin, A., Chapman, B.A., Rowe, J., Tomkins-Tinch, C.H., Valieris, R., Köster, J., and Team, T.B. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods *15*, 475-476.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell *38*, 576-589.

Khan, A., and Zhang, X. (2015). dbSUPER: a database of super-enhancers in mouse and human genome. Nucleic acids research *44*, D164-D171.

Kolde, R. (2011). Package 'pheatmap'. http://cranr-projectorg/web /packages/pheatmap/pheatmappdf.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods *9*, 357-359.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923-930.

Lun, A.T.L., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res *5*, 2122-2122.

Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nature biotechnology *32*, 896-902.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol *11*, R25-R25.

Simon, J.M., Giresi, P.G., Davis, I.J., and Lieb, J.D. (2012). Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. Nature protocols *7*, 256-267.

Stark, R., and Brown, G. (2011). DiffBind differential binding analysis of ChIP-Seq peak data, Vol 100.

Tan, C., Hsia, R.-c., Shou, H., Haggerty, C.L., Ness, R.B., Gaydos, C.A., Dean, D., Scurlock, A.M., Wilson, D.P., and Bavoil, P.M. (2009). *Chlamydia trachomatis*-Infected Patients Display Variable Antibody Profiles against the Nine-Member Polymorphic Membrane Protein Family. Infection and immunity *77*, 3218 LP-3226.

Wickham, H. (2009). Ggplot2: Elegant Graphics for Data Analysis, 2nd edn (Springer Publishing Company, Incorporated).

Wingett, S.W., and Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. F1000Res *7*, 1338-1338.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol *9*, R137-R137.

Zhao, Q., Wang, J., Levichkin, I.V., Stasinopoulos, S., Ryan, M.T., and Hoogenraad, N.J. (2002). A mitochondrial specific stress response in mammalian cells. EMBO J *21*, 4411-4419.

Chapter 3

Chromatin accessibility dynamics of *Chlamydia*-infected epithelial cells

3.1. Abstract

Chlamydia are Gram-negative, obligate intracellular bacterial pathogens responsible for a broad spectrum of human and animal diseases. In humans, *Chlamydia trachomatis* is the most prevalent bacterial sexually transmitted infection worldwide and is the causative agent of trachoma (infectious blindness) in disadvantaged populations. Over the course of its developmental cycle, *Chlamydia* extensively remodels its intracellular niche and parasitises the host cell for nutrients, with substantial resulting changes to the host cell transcriptome and proteome. However, little information is available on the impact of chlamydial infection on the host cell epigenome and global gene regulation. Regions of open eukaryotic chromatin correspond to nucleosome-depleted regions, which in turn are associated with regulatory functions and transcription factor binding.

We applied Formaldehyde-Assisted Isolation of Regulatory Elements enrichment followed by sequencing (FAIRE-seq) to generate temporal chromatin maps of *C. trachomatis*-infected human epithelial cells *in vitro* over the chlamydial developmental cycle. We detected both conserved and distinct temporal changes to genome-wide chromatin accessibility associated with *C. trachomatis* infection. The observed differentially accessible chromatin regions, may help shape the host cell response to infection. These regions and motifs were linked to genomic features and genes associated with immune responses, re-direction of host cell nutrients, intracellular signalling, cell-cell adhesion, extracellular matrix, metabolism and apoptosis.

This work provides another perspective to the complex response to chlamydial infection, and will inform further studies of transcriptional regulation and the epigenome in *Chlamydia*-infected human cells and tissues.

3.2. Introduction

Members of the genus *Chlamydia* are Gram-negative, obligate intracellular bacterial pathogens responsible for a broad spectrum of human and animal diseases (Schachter et al., 1973). In humans, *Chlamydia trachomatis* is the most prevalent bacterial sexually transmitted infection (STI) (Reyburn, 2016), causing substantial reproductive tract disease globally (Menon et al., 2015), and is the causative agent of trachoma (infectious blindness) in disadvantaged populations (Burton, 2007). All members of the genus exhibit a unique biphasic developmental cycle where the non-replicating infectious elementary bodies (EBs) invade host cells and differentiate into replicating reticulate bodies (RBs) within a membrane-bound vacuole, escaping phagolysomal fusion (Fields and Hackstadt, 2002). *Chlamydia* actively modulates host cell processes to establish this intracellular niche, using secreted effectors and other proteins to facilitate invasion, internalisation and replication, while countering host defence strategies (Betts-Hampikian and Fields, 2010; Dautry-Varsat et al., 2004). At the end of the developmental cycle, RBs condense into EBs, which are released from the host cell by lysis or extrusion to initiate new infections (Hybiske and Stephens, 2007).

Bacterial interactions with mammalian cells can induce dynamic transcriptional responses from the cell, either through bacterial modulation of host cell processes or from innate immune signalling cascades and other cellular responses (Alonso and Garcia-del Portillo, 2004; Brunham and Rey-Ladino, 2005; Ribet and Cossart, 2015). In addition, effector proteins specifically targeting the nucleus (nucleomodulins) can influence cell physiology and directly interfere with transcriptional machinery including chromatin remodelling, DNA replication and repair (Bierne and Cossart, 2012). Host cell epigenetic-mediated transcriptional regulatory changes, including histone modifications, DNA methylation, chromatin accessibility, RNA splicing, and non-coding RNA expression (Bierne et al., 2012;

Grabiec and Potempa, 2018; Hamon and Cossart, 2008) may also be arbitrated by bacterial proteins and effectors. Consistent with host cell interactions with other bacterial pathogens, *C. trachomatis* infection alters host cell transcription over the course of its developmental cycle (Humphrys et al., 2013) and may also modulate the host cell epigenome. For example, NUE (NUclear Effector), a *C. trachomatis* type III secreted effector with methyltransferase activity, enters the host nucleus and methylates eukaryotic histones H2B, H3 and H4 *in vitro* (Pennini et al., 2010). However, the ultimate gene targets of NUE activity or the affected host transcriptional networks are uncharacterised, as is the influence of chlamydial infection on the host cell epigenome in general.

Genetic information in eukaryotes is compactly organised within the nucleus of each cell in highly ordered structures composed of DNA and proteins, designated chromatin. Cellular processes occur when chromatin fibres become less condensed, providing areas of open chromatin which allow transcription to proceed. Areas of open chromatin are associated with active DNA regulatory elements, including promoters, enhancers, silencers, and insulators. Chromatin accessibility is also relevant to alternative splicing, alternative promoter usage and alternative polyadenylation, where different forms of RNA are generated from the same gene (Reyes and Huber, 2018). Thus, the underlying structures (introns, exons, TSS and TTS) can be differentially used and thus differentially accessed. To examine the impact of chlamydial infection on host cell chromatin dynamics, we applied FAIRE-seq (Formaldehyde-Assisted Isolation of Regulatory Elements sequencing) (Simon et al., 2012) to *C. trachomatis*-infected HEp-2 epithelial cells and time-matched mock-infected cells, spanning the chlamydial developmental cycle (1, 12, 24 and 48 hours post infection). FAIRE protocols rely on the variable crosslinking efficiency of DNA to nucleosomes by formaldehyde, where nucleosome-bound DNA is more efficiently crosslinked. DNA fragments that are not crosslinked are subsequently enriched in the aqueous phase during phenol-chloroform

extraction. These fragments represent regions of open chromatin, which in turn can be associated with regulatory factor binding sites. In FAIRE-seq, libraries are generated from these enriched fragments, followed by sequencing and read mapping to a reference genome (Simon et al., 2012), allowing patterns of chromatin accessibility to be identified (Giresi et al., 2007). We identify infection-responsive changes in chromatin accessibility over the chlamydial developmental cycle, and identify several candidate host transcription factors that may be relevant to the cellular response to chlamydial infection.

3.3. Methods

3.3.1. Cell culture, infection and experimental design

HEp-2 cells (American Type Culture Collection, ATCC No. CCL-23) were grown as monolayers in 6 x 100mm TC dishes until 90% confluent. Monolayers were infected with *C. trachomatis* serovar E in sucrose-phosphate-glutamate (SPG) as previously described (Tan et al., 2009). Additional monolayers were mock-infected with SPG only. The infection was allowed to proceed 48 hours prior to EB harvest, as previously described (Tan et al., 2009). *C. trachomatis* EBs and mock-infected cell lysates were subsequently used to infect fresh HEp-2 monolayers. Fresh monolayers were infected with *C. trachomatis* serovar E in 3.5 mL SPG buffer for an MOI ~ 1 as previously described (Tan et al., 2009), using centrifugation to synchronize infections. Infections and subsequent culture were performed in the absence of cycloheximide or DEAE dextran. A matching number of HEp-2 monolayers were also mock-infected using uninfected cell lysates. Each treatment was incubated at 25°C for 2h and subsequently washed twice with SPG to remove dead or non-viable EBs. 10 mL fresh medium (DMEM + 10% FBS, 25μg/ml gentamycin, 1.25μg/ml Fungizone) was added and cell monolayers incubated at 37°C with 5% CO_2. Three biological replicates of infected and mock-

infected dishes per time were harvested post-infection by scraping and resuspending cells in 150µL sterile PBS. Resuspended cells were stored at -80°C.

We note that the experimental design used here cannot distinguish *Chlamydia*-mediated effects from infection-specific or non-specific host cell responses. Further experiments with inactivated *Chlamydia* or selected gene knock-outs or knock-downs will help to elucidate the extent of specific *Chlamydia*-mediated interference with the host cell epigenome. We also note that the use of *in vitro* immortalized HEp-2 epithelial cells means that, despite their utility and widespread use in chlamydial research, the full diversity of host cell responses that are likely to be found within *in vivo* infections will not be captured.

3.3.2. FAIRE enrichment and sequencing

Formaldehyde-crosslinking of cells, sonication, DNA extraction of FAIRE-enriched fractions and Illumina library preparation was performed as previously described (Simon et al., 2012). Libraries were prepared in triplicate from infected and mock-infected samples at 1, 12, 24 and 48 hours (24 samples), using the Illumina TruSeq Sample Prep kit, and were sequenced on the Illumina 2500 platform (101 bp paired-end read protocol) at the Genome Resource Centre, Institute for Genome Sciences, University of Maryland School of Medicine. Sequence data is available from the NCBI GEO archive GSE132448.

3.3.3. Bioinformatic analyses

Raw sequencing reads were trimmed and quality checked using Trimmomatic (0.36) (Bolger et al., 2014) and FastQC (0.11.5) (Andrews, 2010). Trimmed reads were aligned to the human genome (GRCh 38.87) using Bowtie2 (2.3.2) (Langmead and Salzberg, 2012) with additional

parameters of 'no mismatches' and '–very-sensitive-local'. Duplicate reads were removed using Picard tools (2.10.4) (Wysoker et al., 2017). Additional replicate quality control was performed using deepTools (2.5.3) (Ramírez et al., 2014) and in-house scripts.

Peak calling of open chromatin regions was performed using MACS2 (2.1.1) (Zhang et al., 2008) in paired-end mode, with additional parameters of '–no-model –broad –q 0.05' and MACS2 predicted extension sizes. Care was taken to ensure parameters were best suited for FAIRE-seq data, particularly as peaks are generally broader than other methods as well as exhibiting a slightly higher background signal (Tsompana and Buck, 2014). All replicates were called separately, with significant peaks determined against the software-predicted background signal. Any peaks that fell within ENCODE blacklisted regions (regions exhibiting ultra-high signal artefacts) (Kundaje, 2016), or were located on non-standard chromosomes such as (ChrMT and ChrUn) were removed.

Consensus peak sets were created by combining significant peaks from the infected and mock-infected replicates for each time using Diffbind (Ross-Innes et al., 2012). Peaks were removed if they appeared in less than two replicates. Reads were counted under each peak within each consensus peak set; the resulting read depths were normalised to their relative library sizes. The resulting count matrices from each consensus peak set were used to look at the differences in chromatin accessibility between infected and mock-infected replicates at each time using the built in DESeq2 method of Diffbind (FDR < 0.05). This created a list of differential chromatin accessible regions with fold-changes relative to the mock-infected conditions. Although the FAIRE protocol only detects open chromatin regions, corresponding patterns of closed chromatin to be identified from regions with negative fold-changes.

Annotation of the set of differential chromatin accessible regions was performed with Homer (v4.9) (Heinz et al., 2010) and separated into three main categories: Intragenic, Promoter and Intergenic. Intergenic: located >1kbp upstream of the transcriptional start site (TSS), or

downstream from the transcription termination site (TTS); Promoter: located within 1kb upstream or 100bp downstream of the TSS (all promoter regions taken from RefSeq); and, Intragenic: annotated to a 3'UTR, 5'UTR, intron, exon, TTS, miRNA, ncRNA or a pseudogene. When regions overlapped multiple features, the resulting annotation was ordered by promoter, intragenic feature, then intergenic regions. To identify enhancers, all intergenic regions were compared against enhancer regions from HeLa cells using Hacer (Wang et al., 2018), Enhancer-atlas (Gao et al., 2016) and dbSuper (Khan and Zhang, 2016). The use of Hacer allowed enhancers from ENCODE and FANTOM5 to also be used. All enhancer regions were converted from hg19 to hg38 using the UCSC LiftOver tool (Hinrichs et al., 2006).

Motif analysis was performed with Homer (Heinz et al., 2010). Target sequences were regions with significant differential chromatin accessibility as identified by DESeq2, while the number of background sequences were software-determined randomly selected regions throughout the human genome (excluding target regions and normalised for GC content). Additional parameters included using a hypergeometric distribution, searching for motifs between 8-16 bp long and allowing for four mismatches as recommended by the software. To confirm motif significance within the 120 conserved regions, the number of background sequences was varied, and only motifs that appeared across a consensus of values were used. Motif enrichment was also performed with Homer (Heinz et al., 2010), followed by filtering and assessment of human tissue specificity of the enriched transcription factors (TF). Time-specific TFs filtering was controlled by p-value < 0.01 and >5% of target sequences, due to the large number of results. From the 120 conserved regions, TFs were filtered based on a p-value < 0.05, due to the lower number of input regions and resulting hits. For significant *de novo* TFs, motif matrices were compared against the Jaspar (Khan et al., 2018) and TomTom

(Gupta et al., 2007) databases, where enriched TFs were discarded unless the Homer annotation matched top hits in either database, and were also human-tissue specific.

All identified TFs from the conserved and time-specific regions were examined against relevant gene expression data to ensure that each TF is expressed in HEp2 cells (unpublished data). To confirm their relevance during infection, TF expression was also examined across a range of different times (0.5, 1.5, 3, 6, 12, 24, 30 and 48 hours), cell lines (HEp2, HeLa and endocervical) and across different *C. trachomatis*-based infection models (Ohmer et al., 2019; Xiang et al., 2019; Zadora et al., 2019) and (unpublished data).

To identify lncRNAs (long non-coding RNA) affecting cell growth and proliferation in HeLa cells, supplementary data was obtained from (Liu et al., 2018) . To confirm suitable candidates, genes with a screen score > 2 were kept, as outlined in their methods.

The annotation of small groups of genes or singular genes was performed manually through numerous online databases including NCBI (Ncbi Resource Coordinators, 2016), Uniprot (The UniProt Consortium, 2016), WikiGenes (Hoffmann, 2008) and GeneCards (Stelzer et al., 2016). Genes were annotated this way as pathway analyses are typically designed to work with large gene lists and when run with low numbers of genes, pathway-associated p-values were generally not significant or the results were biased on subsets of specific genes. Gene Ontology analysis was performed on the 48-hour time-specific differential chromatin regions due to a higher number of input genes. The underlying bioinformatic code used to analyse the data throughout this manuscript can be viewed here: https://github.com/reganhayward/Manuscripts-code.

3.4. Results and Discussion

3.4.1. Chromatin accessibility landscapes of *Chlamydia*-infected and mock-infected cells

We applied FAIRE-seq to *C. trachomatis* serovar E-infected and mock-infected human HEp-2 epithelial cells in triplicate at 1, 12, 24, and 48 hours post-infection (hpi). Following initial quality control measures, a single *C. trachomatis*-infected replicate was identified as an outlier and was removed from further analysis. The remaining replicates were mapped to the human genome (GRCh38), resulting in 52,584,839 mapped reads for mock-infected replicates and 98,802,927 mapped reads for *Chlamydia*-infected replicates (151,387,766 in total) (**Table 3.1**).

Table 3.1: Summary of mapped reads, separated by time and condition

	Mock-infected		Infected	
Time	**Mean**	**S.D**	**Mean**	**S.D**
1	2,603,472	± 417,306	2,686,613	± 554,905
12	6,328,838	± 2,952,657	6,437,002	± 2,511,144
24	3,841,611	± 3,818,015	9,903,858	± 2,394,999
48	6,034,896	± 1,553,435	14,802,374	± 8,475,785
Mapped reads per condition	**52,584,839**		**98,802,927**	
Total mapped reads			**151,387,766**	

Significant peaks, representing regions of open chromatin, were subsequently identified from these mapped reads. Each peak file was examined in IGV to ensure peaks were dispersed genome-wide without discernible chromosomal biases (**Supplementary File 3.1**). The total

number of significant peaks from each replicate varied across the examined times and conditions, ranging between 1,759 and 17,450 peaks (**Figure 3.1A**).

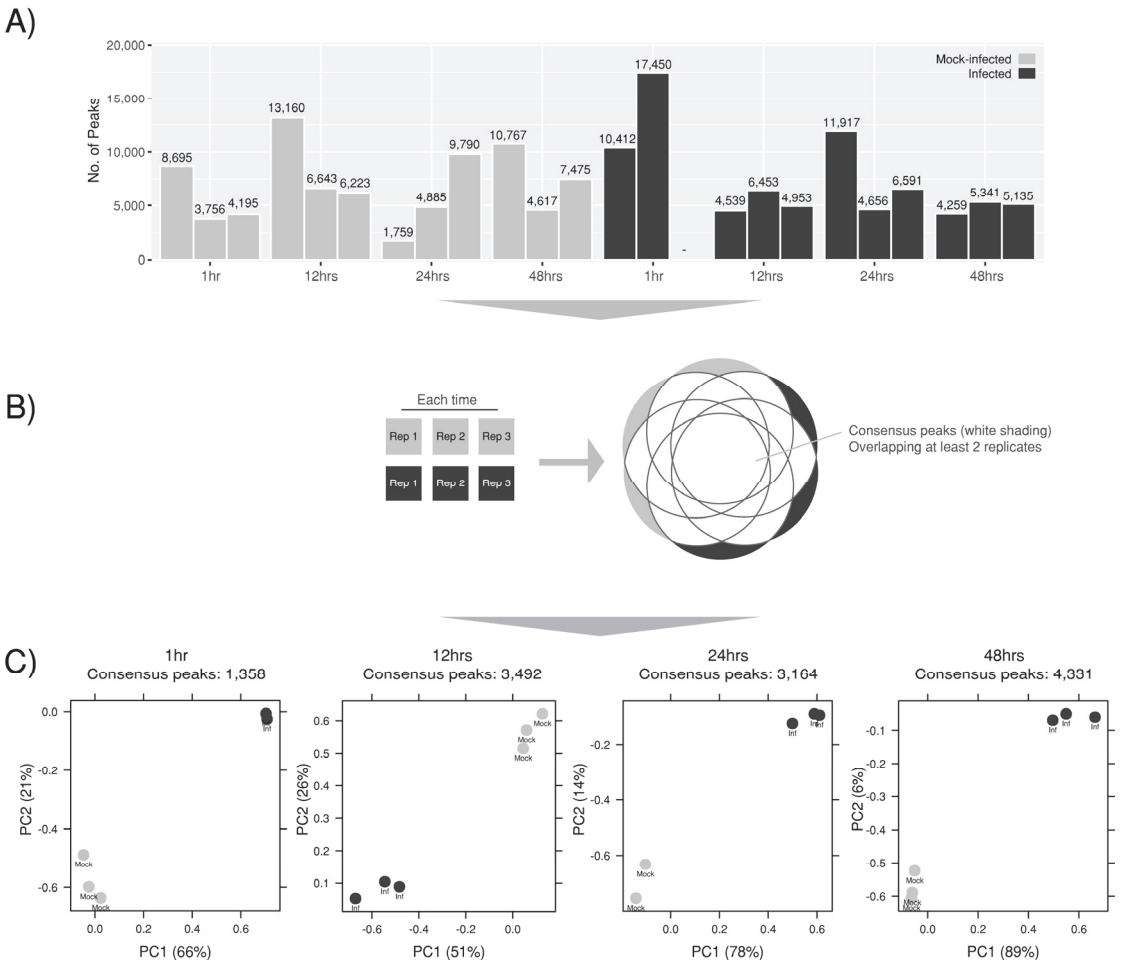


Figure 3.1: Identifying significant peaks and creating consensus peaksets

A) Significant peaks per replicate (p-value < 0.05). **B)** Consensus peaks were created for each time by combining significant peaks from *Chlamydia*-infected and mock-infected conditions, retaining peaks which appeared in > 2 replicates. **C)** PCA plots demonstrating tight clustering within each consensus peak set grouping infected and mock-infected replicates.

Diffbind (Ross-Innes et al., 2012) was used to group and filter peaks at each time post infection by removing regions with low coverage or any regions that were not represented across a consensus of replicates (**Figure 3.1B**). After normalisation for library size, principal component analysis (PCA) of the consensus peak sets led to the removal of one further outlier at 24 hours (mock-infected). The remaining peak sets exhibit tight clustering between mock-infected and infected conditions respectively at each time (**Figure 3.1C**). Total consensus peak numbers increased across the chlamydial developmental cycle, independent of the total mapped reads over time.

3.4.2. *C. trachomatis* infection is associated with temporal changes to chromatin accessibility in host cells

We identified genomic regions with significant differences in chromatin accessibility between infected and mock-infected conditions throughout the development cycle (FDR<0.05). The resulting set of differential chromatin accessible regions identifies both open and closed chromatin (relative to mock-infection). The total number of significant differentially accessible regions rose over the development cycle, with the number of regions increasing (3.6x) from 1 hpi (864) to 48 hpi (3,128) (**Figure 3.2A**). Open chromatin regions predominate over closed chromatin regions at each time (86-99%), suggesting that host cell transcription and regulatory activity increases in response to infection. We also find that closed chromatin regions increase over time, but at a much lower frequency. This may be related to the underlying FAIRE protocol that enriches open chromatin.

At 12 hours, the number of significant differentially accessible regions was lower (8%), compared to the other times (64% at 1 hpi, 43% at 24 hpi and 72% at 48 hpi). The number of mapped reads was similar for all 12 hour replicates across conditions, and similar to other times, suggesting minimal bias from the variability of the underlying mapped reads (**Table**

3.1) and significant peaks (**Figure 3.1A**). In addition, each replicate had consistent peak coverage across the human genome (**Supplementary Figure 3.1**). Furthermore, 12 hour peak annotation is similar to other times (**Figure 3.3B**), and the distribution of peaks around the TSS are within promoter regions, as also seen at 48 hours (**Figure 3.3D**). Thus, in the absence of any discernible bias, the lower number of significant differentially accessible regions at 12 hours may reflect a lower efficiency of formaldehyde crosslinking, or that this time in the course of chlamydial infection is relatively quiescent.

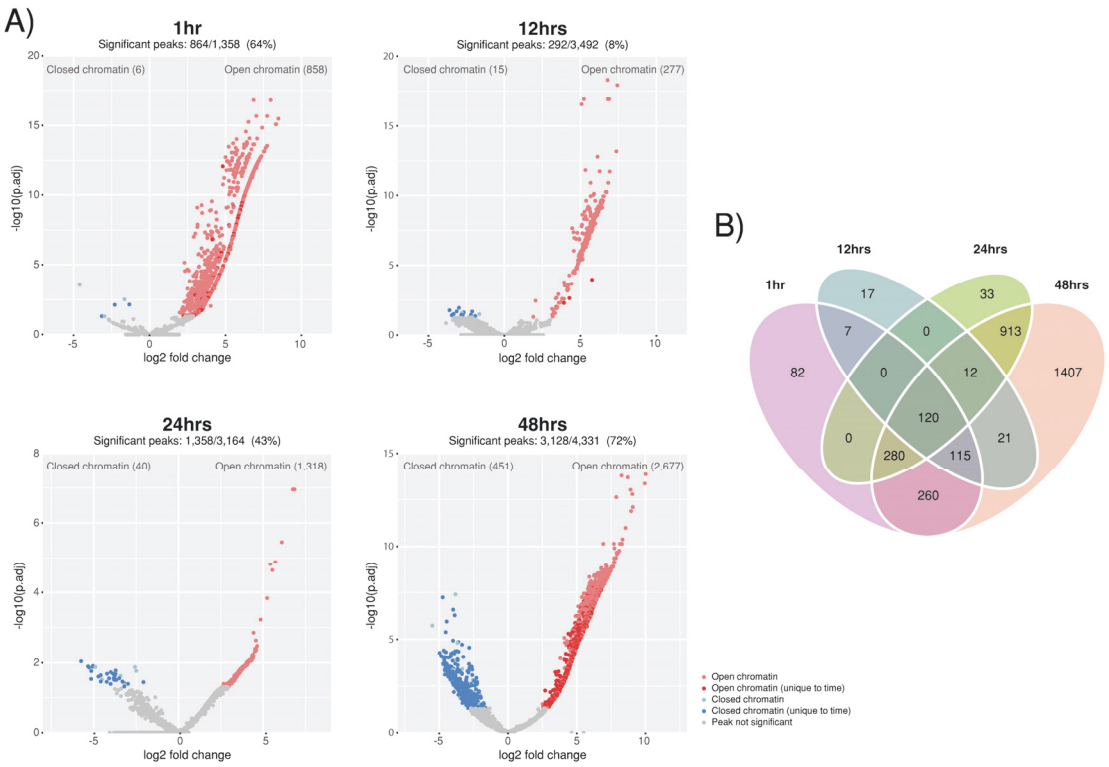


Figure 3.2: Changes in chromatin accessibility throughout the chlamydial developmental cycle

A) Volcano plots highlighting changes in chromatin accessibility between infected and mock-infected conditions. Regions of closed chromatin are represented as blue dots, while open

chromatin regions are red dots. Peaks unique to a specific time have darker shading. Percentages above the plots show the proportion of consensus peaks with significant changes of chromatin accessibility between conditions (FDR < 0.05). **B)** Unique and conserved regions of differential chromatin accessibility across the developmental cycle.

120 differentially accessible chromatin regions are common at all examined times (**Figure 3.2B**), indicating a conserved response to chlamydial infection-associated events or general disruption of cellular homeostasis, irrespective of infection progression. Conversely, unique sets of differentially accessible regions are found at each time post-infection, highlighting the dynamism of the cellular response to infection over time, particularly at 48 hpi (**Figure 3.2B**). Differential chromatin accessible regions were annotated based on four categories as described in the **Methods** and portrayed in (**Figure 3.3A**). It should be noted that although the enhancer region displayed in the figure is upstream of an associated promoter, they can appear anywhere throughout the genome. Enhancers often interact with the promoter region of genes through looping of DNA (**Figure 3.3A**), but can also interact through tracking, linking and relocation mechanisms (Khan and Zhang, 2016). Most infection-associated differential chromatin accessible regions were annotated to either intergenic or intragenic regions (**Figure 3.3B**). Intergenic regions spanned considerable distances upstream and downstream from the closest gene (**Figure 3.3C**), while enhancers that were identified from within these regions appear much closer to the TSS. Intragenic regions were predominantly (>90%) annotated to intronic regions (**Supplementary File 3.1**), consistent with other chromatin accessibility studies (Gaulton et al., 2010; He et al., 2014), and the overall distribution of protein-coding genes within the human genome (Gregory, 2005). The distribution of differential chromatin-accessible regions around TSSs (+/- 5kb) at 12 and 48 hpi show that many regions are in close proximity to TSSs, with many regions directly

upstream. Due to our strict classification of overlapping RefSeq–based promoters (-1,000 to 100 bp from TSS), we are confident these directly represent infection-specific promoters. At 24 hpi we also see a large number of regions directly upstream of the TSS, but also an increase of regions and variability further up and downstream. At 1 hpi the regions exhibit a slight bi-modal distribution (Bimodality coefficient 0.67), with fewer regions directly surrounding the TSS (**Figure 3.3D**). The increased number of regions not immediately surrounding TSSs at 1 and 24 hpi is suggestive of additional regulatory mechanisms such as different transcription initiation sites or that differential intron/exon usage may be contributing to or influencing the regulatory response to infection-associated events.

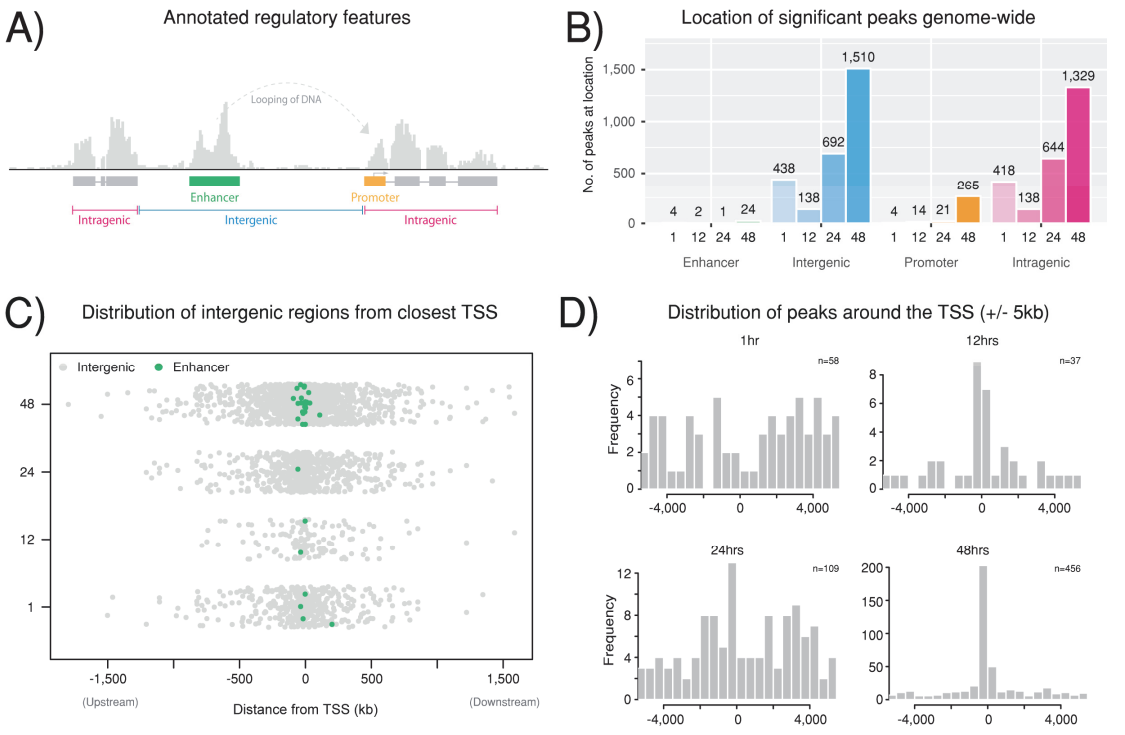


Figure 3.3: Annotation of significant peaks

A) Example illustration of annotating significant differential peaks to enhancer, promoter, intragenic or intergenic regions. **B)** Number of peaks per annotated category, separated by

time. **C)** All intergenic peaks plotted based on their proximity to the TSS of the closest gene. All enhancers were identified from within these regions and are coloured green. **D)** Frequency distribution of significant peaks and their proximity to the TSS of their associated genes (+/- 5KB).

3.4.3. Differential chromatin accessibility at promoter regions

The proportion of all differentially accessible regions mapping to promoter regions is 0.5% (4) at 1 hpi, 4.8% (14) at 12 hpi, (1.5% (21) at 24 hpi and (8.5% (265) at 48 hpi (**Figure 3.4A**). Notably, 48 hpi exhibits a >10-fold increase in the number of significant regions compared to 24 hpi, with the majority of regions showing a reduction in chromatin accessibility, likely representing down-regulation of promoter-associated genes (**Figure 3.4A**). The large number of differentially accessible chromatin regions within promoters at 48 hours is a likely reflection of the diversity of events occurring at this late stage of the developmental cycle, including apoptosis, necrosis, lysis and cellular stress. Associated 48 hpi genes are linked with heat-shock stress (DNAJB1, DNAJB5, DNAJC21 and HSPA1B), cell defence (ILF2, MAP2K3 and STAT2), and cell stress/apoptosis (ATF3, PPM1B, GAS5, BAG1 and TMBIM6). ATP7A, which has a promoter exhibiting an increase in chromatin accessibility, is a key regulator of copper transport into phagosomes as part of a host cell response to intracellular infection (Hodgkinson and Petris, 2012; Ladomersky et al., 2017).

Fifteen promoter-specific differentially accessible regions are found at two or more times. Two promoter regions are associated with genes encoding sorting nexin 16 (SNX16) and oligosaccharyltransferase complex subunit (OSTC) respectively (**Figure 3.4B**). The promoter region of OSTC exhibits increased chromatin accessibility at 24 and 48 hours; OSTC is linked to cellular stress responses (Parnas et al., 2015). Conversely, SNX16 shows a reduction in

chromatin accessibility at both 1 and 48 hpi. Sorting nexins are a family of phosphatidylinositol binding proteins sharing a common PX domain that are involved in intracellular trafficking. Sorting nexins are a key component of retromer, a highly conserved protein complex that recycles host protein cargo from endosomes to plasma membranes or the Golgi (Seaman, 2012).

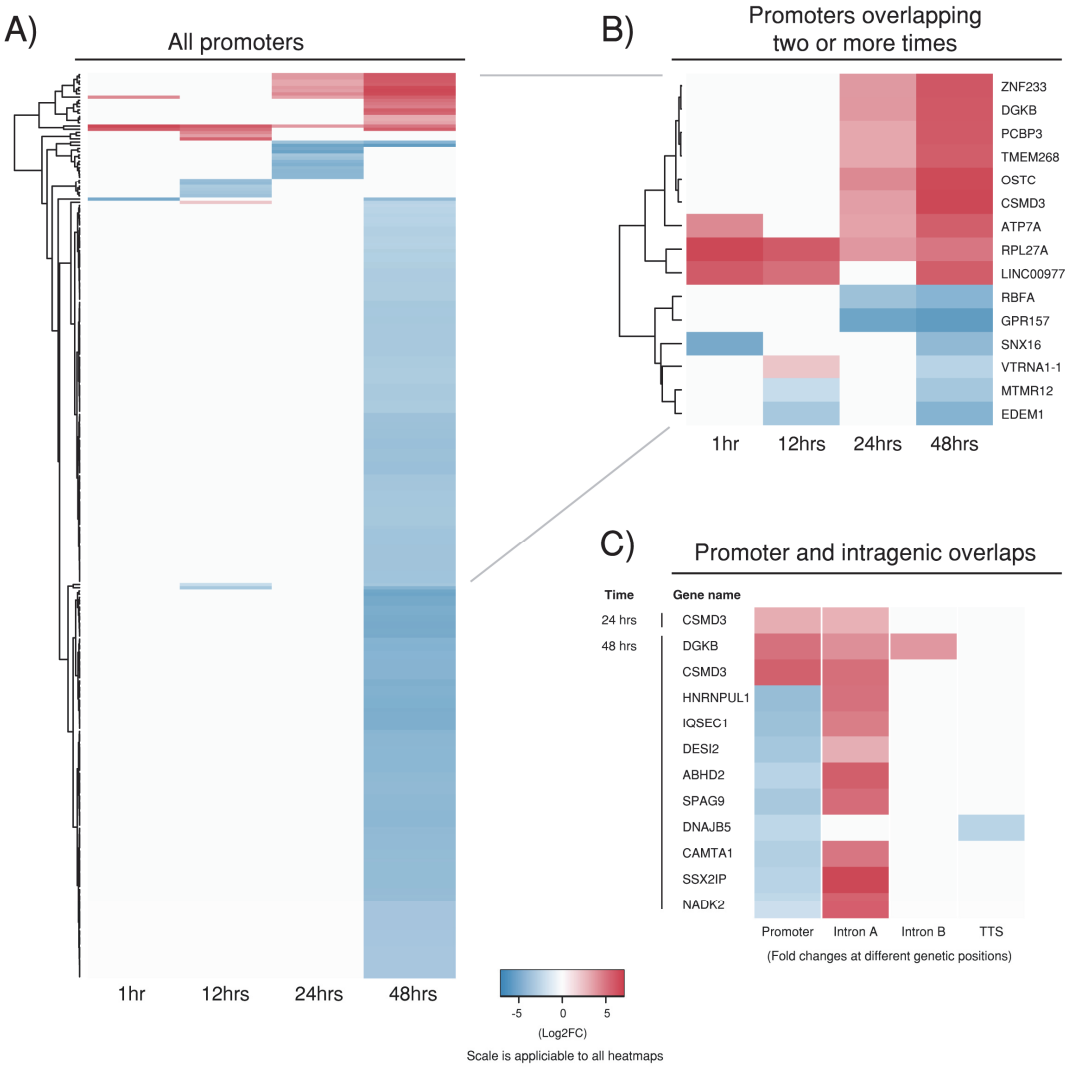


Figure 3.4: Differential chromatin accessibility within promoter regions

Heatmaps of significant differential peaks that were annotated to a promoter region. **A)** All promoter regions from each time post-infection. **B)** Promoters overlapping two or more times post-infection. Red and blue shading indicates fold-changes, while grey indicates no significant peaks. **C)** Genes which contained significant differential peaks within a promoter region and also within one or more intragenic regions.

Retromer is targeted by several intracellular pathogens, including *Chlamydia*, as a key strategy for intracellular survival (Elwell and Engel, 2018). The *C. trachomatis* effector protein, IncE, binds to sorting nexins 5 and 6, disrupting retromer-mediated host trafficking pathways (Elwell and Engel, 2018) and potentially perturbing the endolysomal-mediated bacterial destruction capacity of the host cell (Paul et al., 2017). However, SNX16 is a unique member of this family, containing a coiled-coil domain in addition to a PX domain, and is not associated with retromer (Xu et al., 2017). SNX16 is instead associated with the recycling and trafficking of E-cadherin (Xu et al., 2017), which mediates cell-cell adhesion in epithelial cells, and is associated with a diversity of tissue specific processes, including fibrosis and epithelial-mesenchymal transition (EMT) (Schneider and Kolligs, 2015). Separately, *C. trachomatis* infection has been shown to downregulate E-cadherin expression via increased promotor methylation, potentially contributing to EMT-like changes (Rajic et al., 2017). Thus, downregulation of SNX16, as inferred by the observed reduction in promotor-associated chromatin accessibility may contribute to chlamydial fibrotic scarring outcomes. In other bacterial pathogens, modulation of E-cadherin is a known virulence mechanism where it is degraded by proteases, such as HtrA, disrupting tight and adherens junctions to facilitate invasion through the epithelial barrier (Backert et al., 2017; Boehm et al., 2018). Although chlamydial HtrA has been detected outside the inclusion and in exported blebs (Wu et al., 2011), E-cadherin has not yet been identified as a chlamydial HtrA target. Nevertheless,

HtrA has been shown to be critical for *in vivo* chlamydial infections, indicating that this functionality may be revealed in the future (Gloeckl et al., 2013).

All promoter-regulated genes were overlapped with genes containing differentially accessible intragenic peaks (**Figure 3.4C**). Of these 12 genes, only one gene (CSMD3) appeared at more than one time point. All genes exhibited regulation at intronic regions apart from DNAJB5 (involved in heat-stress as indicated above), which exhibited an increase in chromatin accessibility at its promoter and TTS. DGKB was the only gene to exhibit regulation at its promoter and more than one intronic region, each with decreased chromatin accessibility. DGKB is a diacylglycerol kinase that metabolises 1,2,diacylglycerol (DAG) to produce phosphatidic acid (PA), a key precursor in the biosynthesis of triacylglycerols and phospholipids, and a major signalling molecule (Topham and Prescott, 1999). *Chlamydia* obtains and redirects host-derived lipids through multiple pathways (Yao et al., 2015a), and as further identified in the enriched time-specific GO section below. Regulation occurring in patterns like these is generally associated with alternative-splicing mechanisms. Unfortunately, there is limited annotation of the alternatively-spliced transcripts from these genes, particularly in an infection-based setting. Their identification could provide suitable targets for follow-up studies, or overlapped with RNA-seq or expression-based studies that allow the capture alternative-splicing events and their resulting transcripts/proteins, allowing these events to be examined in more detail.

3.4.4. Differential chromatin accessibility from enhancer-regulated genes

Changes in chromatin accessibility of regions overlapping tissue-specific enhancers from a range of online databases were examined, identifying 316 enhancer and 13 “super-enhancer” regulated genes (**Figure 3.5A-B**). The super-enhancers used are defined as clusters of

transcriptional enhancers that drive cell-type-specific gene expression, are crucial to cell identity, and can contain disease-associated sequence variations (Hnisz et al., 2013). Each enhancer can regulate more than one gene, explaining the substantial increase in the number of enhancer-associated genes reported here (**Figure 3.5A**), compared to just the enhancer regions that were reported earlier (**Figure 3.3B**). The majority of super-enhancers exhibited a decrease in chromatin accessibility, and were associated with genes mediating energy production (SDHB and CDHC), cell protection (IER3) and the stress-regulated polyubiquitin gene UBC (ubiquitin C) (Bianchi et al., 2018), which is discussed in further detail below.

Only one super-enhancer regulated gene appeared across three times (SGK1 at 1, 12 and 48 hours) and exhibited an increase in chromatin accessibility. This serum/glucocorticoid regulated kinase has been associated with coordinating a range of different cellular processes that are crucial to reproductive activities, with deregulation resulting in reproductive disorders such as pregnancy loss, infertility and endometriosis (Lou et al., 2017).

The majority of enhancer regions were identified at 48 hours (78%) and predominantly exhibited decreased chromatin accessibility (**Figure 3.5A**). Enrichment of closed chromatin regions identified the biological process *"long-chain fatty acid biosynthetic process (GO:0042759)"* with the greatest significance. Decreased regulation of long-chain fatty acids (lauric acid and capric acid) has been shown to be an effective way to inactivate *C. trachomatis* (Bergsson et al., 1998). Reduced expression of other long-chain fatty acids such as oleic acid has negative impacts regarding the inclusion membrane, as it cannot be synthesised and is directly required from the host (Yao et al., 2015b).

We also identify the molecular function *"type I transforming growth factor beta receptor binding (GO:0034713)"*. Transforming growth factor beta (TGF-β) is a multifunctional cytokine involved with roles in both host defence and immunopathogenesis (Williams et al., 1996). As a result, TGF-β is regarded as an important signalling marker during and after

infection (Sharkey et al., 2012; Ziklo et al., 2019). A reduction in chromatin accessibility at enhancer regions regulating many of the underlying genes is surprising, considering the universal role that TGF-β plays during infection as previously described.

Four enhancer-regulated genes were identified (CROCCP2, LA16c-321D4.2, LINC00514 and RP5-1173A5.1) by overlapping lncRNAs affecting cell growth and proliferation in HeLa cells (Liu et al., 2018). All enhancer regions exhibited decreased chromatin accessibility (log_2FC ranging between -3.3 to -4.6). Due to the lack of annotation of many lncRNAs, understanding their specific functions still remains challenging. However, as identified in the associated CRISPR-based study, these lncRNAs directly affect the survival of HeLa cells, further highlighting the complex nature involved with chlamydial infections, but also identifying a novel direction to explore the role of non-coding RNAs as seen in other pathogenic bacteria (Duval et al., 2017; Ortega et al., 2014).

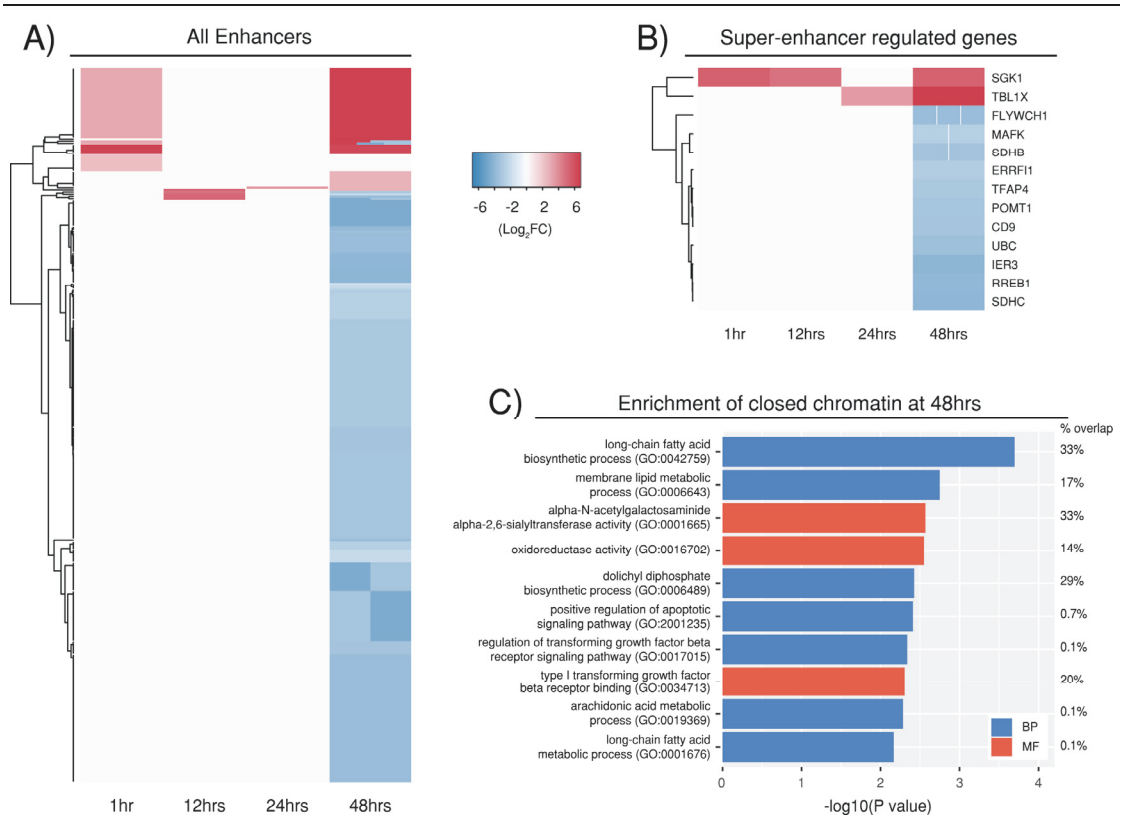


Figure 3.5: Differential chromatin accessibility within enhancer regions

Significant differential peaks annotated as intergenic were compared against tissue-specific enhancers. **A)** All enhancer regions across each of the four times. **B)** Super-enhancer regulated genes. Red and blue shading indicate fold-changes, while grey indicates that no significant peaks were associated with that enhancer at that time. Some enhancer regions contain more than one peak, explaining why there are multiple fold-changes at some times. **C)** Gene Ontology enrichment from the large number of enhancer-associated genes at 48 hours.

3.4.5. Conserved host responses to infection over the chlamydial developmental cycle

Differential chromatin accessible regions that are present at all four times during infection demonstrate a conserved host cell response to chlamydial infection (**Figure 3.2B**). Time-specific differential chromatin accessibility is also evident over the chlamydial developmental cycle (**Figure 3.2B**). To investigate the conserved host cell response, we focused upon 58 of the 120 differential chromatin accessible regions (intragenic, promoter or enhancer regions) identified above, excluding the likely ambiguous intergenic regions (**Figure 3.6A**). 56 were within intronic regions, one within a 3'UTR (FECH) and one within a promoter region (RPL27A). Only 54 of these 58 significant differentially accessible regions show a decrease in overall chromatin accessibility. However, these same regions also exhibit increased chromatin accessibility at different intragenic locations at 48 hpi, further highlighting the potential for infection-related alternative splicing mechanisms (**Figure 3.6A**). The remaining conserved differentially accessible regions were associated with genes involved in infection-relevant cellular processes, including C8A as part of the complement cascade, and lipase activity from LIPI that is essential for chlamydial replication (Cocchiaro et al., 2008); while multiple genes (HDAC2, HNRNPUL1, NCOA7 and YAP1) are known transcriptional regulators. We also examined any differential chromatin accessible regions that appeared across three times. This identified further effects of infection on the complement cascade. Key components of the membrane attack complex (MAC) and complement activation pathways exhibit increased differential chromatin accessibility (C8B at 1, 12 and 24 hours and CFHR5 at 24 and 48 hours). Conversely, C6 exhibits decreased chromatin accessibility at 48 hours.

All conserved regions were also examined for commonly occurring motifs and their associated transcription factors (TFs) to identify potential master-regulators of infection (**Figure 3.6B**). Four TFs were identified (ETS1, POU3F2, TFAP4 and PKNOX1), containing

binding sites within 49 different intergenic and intragenic regions. An increase in chromatin accessibility (positive fold-changes) was seen at all binding sites and across all time points. Functional annotation of the TFs identified that TFAP4 (Transcription Factor AP-4) functions as an activator of gene-expression of both cellular and viral genes, that can also form functional dimers to control transcriptional networks during cellular differentiation (Hu et al., 1990). ETS1 (ETS Proto-Oncogene 1, Transcription Factor) also functions as an activator and is able to directly control expression levels of cytokine and chemokine genes within a wide array of cellular dynamics and contents (Wasylyk et al., 2002; Yordy et al., 2004). Interestingly, each of these TFs have highlighted that a large proportion of the binding sites fall within un-annotated intergenic regions. Their direct association with infection-based mechanisms across all four time points highlights their relevance, and could be interesting targets for future studies.

For each of the four TFs identified above, we isolated significant regions containing the associated motif. Regions that overlapped intergenic features were then compared against gene expression data to examine expression changes during different infection settings **(Supplementary Figure 3.2)**. POU3F2 was unable to be compared as both regions overlapped intergenic regions that could not be compared. The comparison of fold changes from the remaining three TFs shows a mean decrease in regulation across the 9 different datasets, with the greatest changes occurring at 3 hours post infection. Many of the genes do exhibit an increase in regulation as described above, but due to the complexities from sampling at different time points and from different tissues, the regulation does vary considerably.

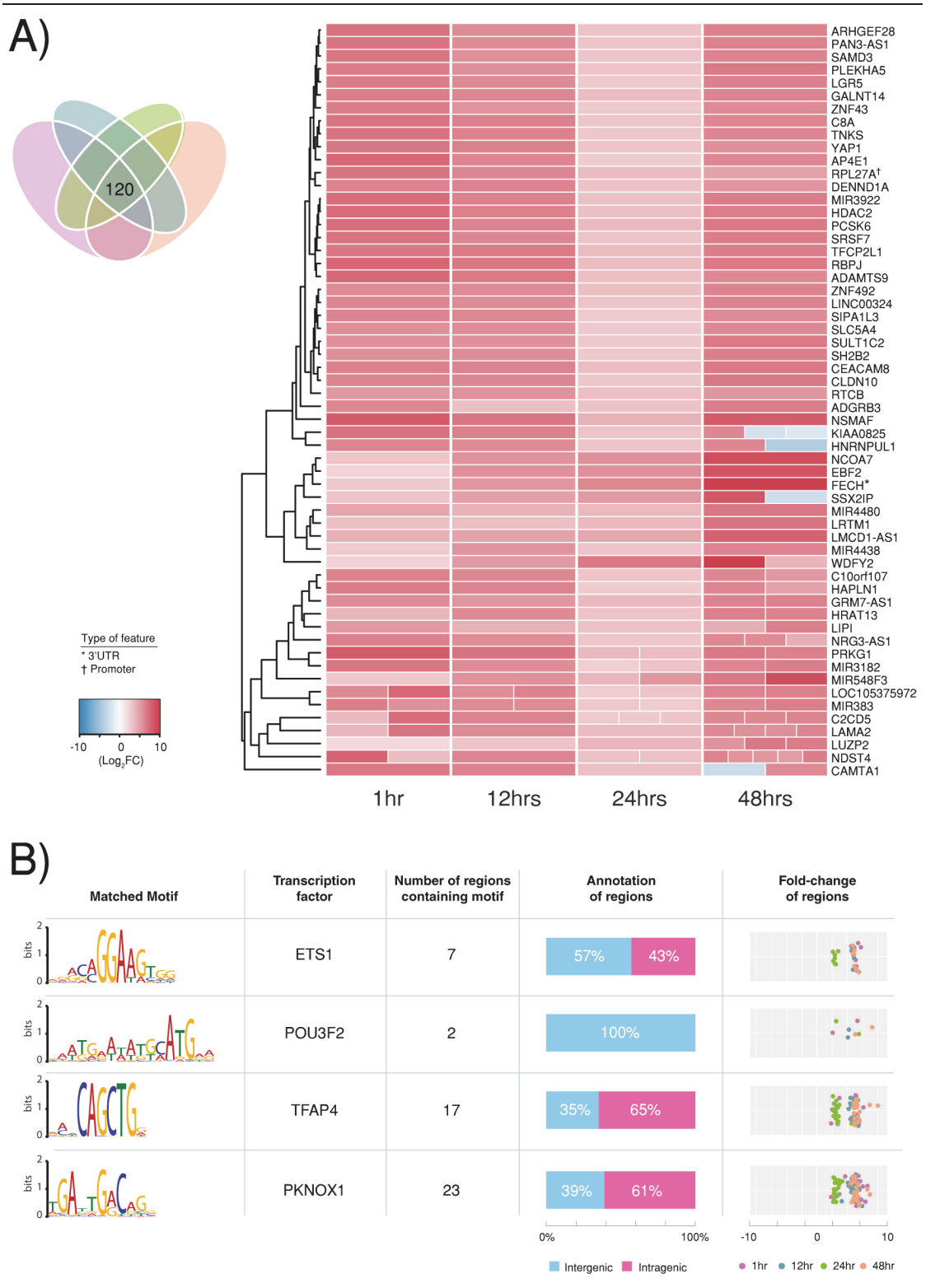


Figure 3.6: Conserved host cell response to infection

A) 120 differentially accessible regions found in all four times were extracted, representing a conserved host cell response to infection. Intergenic regions were removed due to the ambiguity of annotating to the closest feature. If a gene contained more than one peak within a specific time, the different fold changes are split out evenly within the column at that time. **B)** Significant motifs, enriched transcription factors (TFs) and associated information based on the associated chromatin accessibility within these conserved regions.

3.4.6. Time-specific host responses to infection over the chlamydial developmental cycle

We identified unique differentially accessible regions across the chlamydial developmental cycle (**Figure 3.7A**). At 1, 12 and 24 hpi, there are a relatively small number of significant differential chromatin accessible regions. In contrast, 48 hpi exhibits over 1,400 regions, further reflecting the diverse processes associated with the end of the *in vitro* developmental cycle as indicated previously. As mentioned in the conserved section above, we focused on differential chromatin accessibility within promoters, enhancers and intragenic regions (50 at 1 hpi, 17 at 12 hpi, 27 at 24 hpi and 866 at 48 hpi) (**Figure 3.7B, Supplementary File 3.2)**.

We illustrate the *in vitro C. trachomatis* developmental cycle into three stages (early, mid and late) outlining key biological events, and giving a visual representation of the expected biology from the times that were extracted (**Figure 3.7C**). Due to the limited number of differential regions at the first three times, each of the genes were individually annotated through multiple online sources and grouped into biological sub-categories.

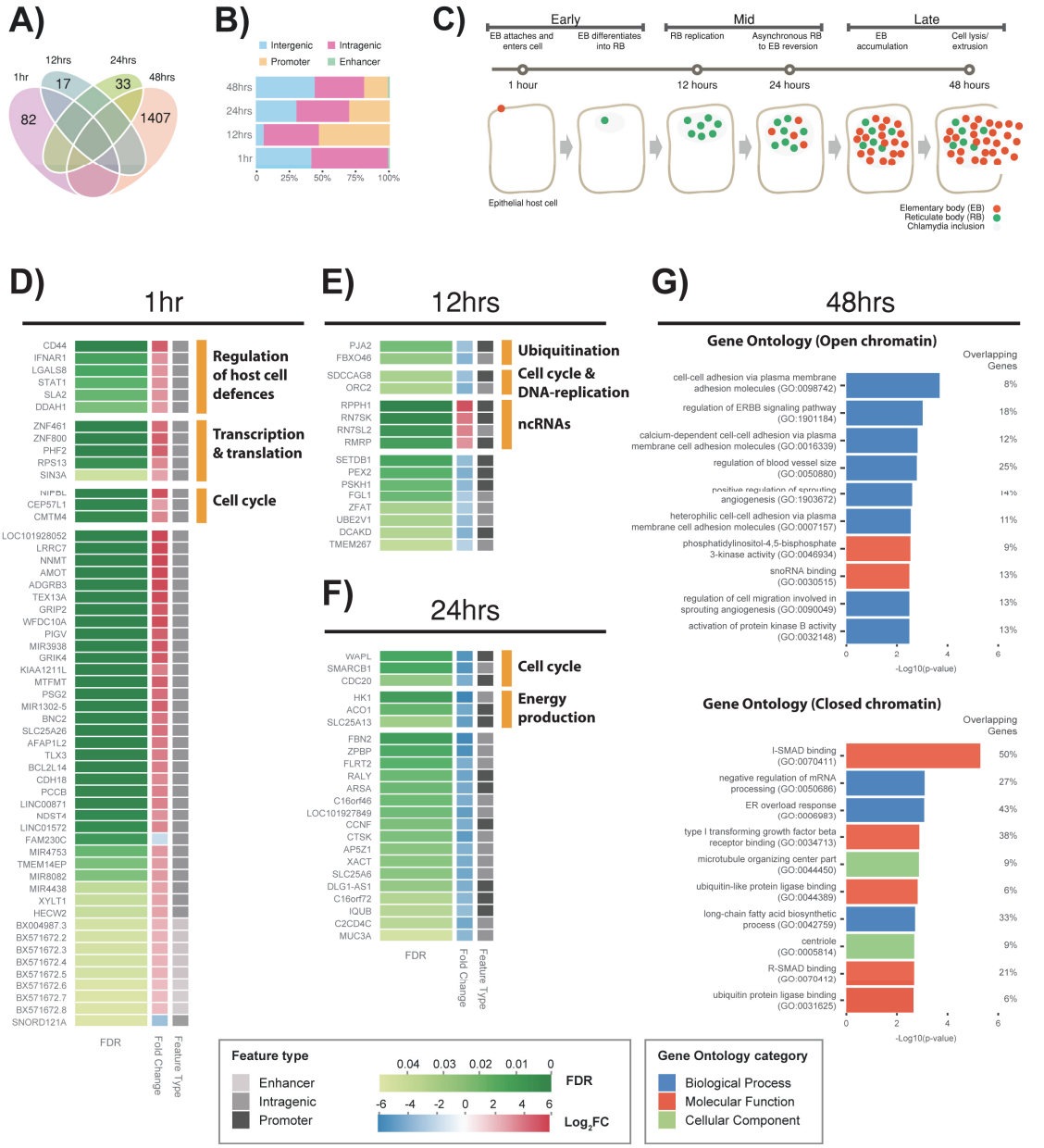


Figure 3.7: Enrichment of time-specific differential chromatin regions

A) The numbers of significant differential chromatin accessible regions at each time. **B)** The annotation of each of these regions. **C)** The *in vitro C. trachomatis* developmental cycle separated in to three stages, representing known biological events from the times that were examined. **D)** Annotated time-specific differential chromatin regions associated with 1 hour, 12 hours **E)**, and 24 hours **F)**. Where genes have been grouped into annotated categories,

multiple underlying sources were used for verification. **G)** At 48 hours, a substancial increase in genes allowed Gene Ontology (GO) enrichment. All three GO categories were enriched, with the top ten p-values across the categories displayed.

At 1 hpi, increased chromatin accessibility was associated with a variety of genes involved in the regulation of host cell defences (CD44, IFNAR1, LGALS8, STAT1, SLA2 and DDAH1), transcription and translation (ZNF461, ZNF800, PHF2, RPS13 and SIN3A), the cell cycle (NIPBL, CEP57L1 and CMTM4) and BCL2L14 (Apoptosis facilitator Bcl-2-like protein 14) a member of the Bcl-2 Family of proteins that are linked to apoptosis (Guo et al., 2001) (**Figure 3.7D**). At 12 hours, four ncRNAs were identified (RPPH1, RN7SK, RN7SL2 and RMRP) that are involved in RNA processing, signalling and transcriptional regulation (Baer et al., 1990; Egloff et al., 2018; Hermanns et al., 2005; Ullu and Weiner, 1984). The remaining genes at 12 hours exhibited decreased chromatin accessibility, encompassing the cell cycle and DNA replication (SDCCAG8 and ORC2), and ubiquitination (PJA2 and FBXO46) (**Figure 3.7E**). At 24 hours, all genes were associated with decreased chromatin accessibility and were grouped into two sub-categories: cell cycle (WAPL, SMARCB1 and CDC20) and energy production (HK1, ACO1 and SLC25A13) (**Figure 3.7F**).

3.4.7. Increased changes to differential chromatin accessibility at the end of the developmental cycle

With the increased number of differential chromatin regions at 48 hours, GO enrichment was performed and separated by regions exhibiting an increase in open chromatin (positive fold-changes) and regions exhibiting an increase in closed chromatin (negative fold-changes) (**Figure 3.7G**). Significantly enriched ontologies associated with regions of increased chromatin accessibility include the ErbB signalling pathway *(GO:1901184),* which is linked

to a wide range of cellular functions including growth, proliferation and apoptosis. ErbB transmembrane receptors are also often exploited by bacterial pathogens for host cell invasion (Ho et al., 2017). Notably, epidermal growth factor receptor (EGFR), a member of the ErbB family, is the target receptor for *C. pneumoniae* Pmp21 as an EGFR-dependent mechanism of host cell entry (Mölleken et al., 2013). The *C. trachomatis* Pmp21 ortholog, PmpD, also has adhesin-like functions (Paes et al., 2018), however the host ligands are unknown. Nevertheless, EGFR inhibition results in small, immature *C. trachomatis* inclusions, with calcium mobilisation and F-actin assembly disrupted (Patel et al., 2014), indicating the functional importance of EGFR and the ErbB signaling pathway for *C. trachomatis* attachment and development.

Three enriched biological processes share the term '*cell-cell adhesion via plasma membrane adhesion molecules*' (*GO:0098742, GO:0016339 and GO:0007157*). Several genes common to these categories with infection-responsive differential chromatin accessibility are associated with cadherins (CDH4, CDH12, CDH17, CDH20, FAT4 and PTPRD), which are calcium-dependent transmembrane glycoproteins associated with the actin cytoskeleton and an essential structural component to maintain cells bind together (Wallis et al., 1996). Disruption of cadherin function has been described in *C. trachomatis* infection, and is linked to the alteration of adherens junctions and the induction of EMT events that may underlie chlamydial fibrotic outcomes (Igietseme et al., 2018; Rajic et al., 2017). Altered chromatin accessibility for a further cadherin-relevant locus was apparent in the promoter region of SNX16 (see above), suggesting that alteration or disruption of cadherin regulation is a key feature of chlamydial infection. The molecular function of snoRNA binding was a surprise addition as there is limited information of small nuclear RNA regulation from bacterial infections. The lipid-based ontology '*Membrane lipid biosynthetic process* (*GO:0030148*) was also associated with regions of open chromatin. *Chlamydia* scavenges a range of host-

cell-derived metabolites for intracellular growth and survival, particularly lipids (Elwell and Engel, 2012; van Ooij et al., 2000).

Significantly enriched ontologies associated with regions of decreased chromatin accessibility include the '*I-Smad (inhibition of Smad) binding*, *(GO:0070411)'*. I-Smads (Inhibitory-Smads) are one of three sub-types of Smads that inhibit intracellular signalling of TGF-β by various mechanisms including receptor-mediated inhibition (Miyazawa and Miyazono, 2017). This coincides with the appearance of *'Type 1 transforming growth factor beta receptor binding (GO:0034713)'*. In addition, four genes (SMAD2, DDX5, SMURF1 and SMAD6) are associated with closed chromatin and *'R-Smad binding (GO:0005814)'*, which are part of the R-Smad sub-family that regulates TGF-β signalling directly (Attisano and Tuen Lee-Hoeflich, 2001; Takimoto et al., 2010). TGF-β induces I-Smad expression, and has been hypothesised to be a central component of dysregulated fibrotic processes in *Chlamydia*-infected cells, provoking runaway positive feedback loops that generate excessive ECM deposition and proteolysis, potentially leading to inflammation and scarring (Humphrys et al., 2013).

We also identify over ten genes localised within the cellular component *'Microtubule organising centre (GO:0044450)'*. Dynein-based motor proteins have been shown to move the chlamydial inclusion via the internal microtubule network to the MTOC (Microtubule-Organizing Centre); the close proximity to the MTOC is thought to facilitate the transfer of host vesicular cargo to the chlamydial inclusion (Grieshaber et al., 2003).

Two similar ontologies *'Ubiquitin-like protein ligase binding (GO:0044389)'* and *'Ubiquitin protein ligase binding (GO:0031625)'* are involved in ubiquitination and protein quality control. The eukaryotic ubiquitination modification marks proteins for degradation and regulates cell signalling of a variety of cellular processes, including innate immunity and vesicle trafficking (Zhou and Zhu, 2015). The deposition of ubiquitin onto intracellular

pathogens is a conserved mechanism found in a diverse range of hosts (Manzanillo et al., 2013). In *Chlamydia*, host cell ubiquitin systems can mark chlamydial inclusions for subsequent destruction (Haldar et al., 2016) and there is emerging evidence that various *Chlamydia* species, using secreted effectors and other proteins, are able to subvert or avoid these host ubiquitination marks for intracellular survival (Haldar et al., 2016; Misaghi et al., 2006). Our observation of decreased chromatin accessibility of numerous ubiquitination genes, further highlighting the complex role of ubiquitination in chlamydial infection.

3.4.8. Identification of transcription factor binding motifs

Transcription factor (TFs) binding sites were identified from motifs within the significant differential chromatin accessible regions at each time post-infection (**Supplementary File 3.3**). Eleven of the most significant TF motifs are shown in (**Table 3.2**), spanning across the development cycle.

Table 3.2: Motifs and enriched transcription factors

Target sequences are significant differential peaks and background sequences are randomly selected throughout the genome to determine significance. A star (*) denotes a de-novo motif where various sources were used to annotate the corresponding transcription factor.

Time	Motif	P.value	Target sequences with Motif (%)	Background sequences with Motif (%)	Transcription factor
1		1e-13	10.53	3.84	**IRF3***
24		1e-12	17.45	9.78	**Homeobox***

48		1e-28	7.67	1.82	**Sp1(Zf)**
		1e-22	6.30	1.58	**KLF9(Zf)**
		1e-21	7.58	2.40	**KLF3(Zf)**
		1e-15	32.58	23.46	**MEF2C***
		1e-13	9.81	4.90	**KLF6(Zf)**
		1e-10	6.30	2.87	**KLF10(Zf)**
		1e-7	11.06	7.18	**KLF5(Zf)**
		1e-7	10.45	6.71	**NFYB**
		1e-6	5.37	2.87	**E2F3**

IRF3 (Interferon Regulatory Factor) motifs are enriched at 1 hpi; IRF3 is a key transcriptional regulator of type I interferon (IFN)-dependent innate immune responses and is induced by chlamydial infection. The type I IFN response to chlamydial infection can induce cell death or enhance the susceptibility of cells to pro-death stimuli (Di Paolo et al., 2013), but may also be actively dampened by *Chlamydia* (Gyorke and Nagarajan, 2018; Sixt et al., 2017). Specificity Protein 1 (Sp1) is a zinc-finger TF that binds to a wide range of promoters with GC-rich motifs. Sp1 may activate or repress transcription in a variety of cellular processes that include responses to physiological and pathological stimuli, cell differentiation, growth, apoptosis, immune responses, response to DNA damage and chromatin remodelling (Deniaud et al., 2009; Tan and Khachigian, 2009).

XCPE1 (X Core Promoter Element 1) is an activator-dependent core promoter that drives RNA polymerase II transcription. It's found in approximately 1% of core promoters in human genes, particularly in TATA-less promoters (Tokusumi et al., 2007). The heterotrimeric TF NFY(CAAT), or "CAAT-binding factor" binds specifically to the human H ferritin promoter on the B site. Transcription of the gene is controlled by two promoter elements A and B

located upstream of the TSS (Faniello et al., 1999). Element A binds Sp1 and typically controls about 50% of transcription. Element B is recognised by the CAAT sequence on the non-coding strand, increases transcription in differentiating cells and is also the binding site of cAMP signalling (Bevilacqua et al., 1995; Bevilacqua et al., 1997; Bevilacqua et al., 1992).

The majority of TF motifs enriched at 48 hours correspond to Krüppel-like-factors (KLFs). KLFs are zinc-finger TFs in the same family as Sp1, which is also enriched at 48 hours. The members of this large family orchestrate a range of paracrine and autocrine regulatory circuits and are ubiquitously expressed in reproductive tissues (Simmen et al., 2015). Dysregulation of KLFs and their dynamic transcriptional networks is associated with a variety of uterine pathologies (Simmen et al., 2015). We find motif enrichment for five distinct KLFs (KLF3, KLF5, KLF6, KLF9 and KLF10) at 48 hours, in addition to further KLFs at 12 (KLF3, KLF4, KLF6, KLF9), 24 hours (KLF 10) and 48 hours (KLF 4) when relaxing the initial filtering steps (**Supplementary File 3.3**). KLF5 is a transcriptional activator found in various epithelial tissues and is linked to regulation of inflammatory signalling, cell proliferation, survival and differentiation (Dong and Chen, 2009). KLF6 is also a transcriptional activator ubiquitously expressed across a range of tissues and plays a crucial role in regulating genes involved with tissue development, differentiation, cell cycle control, and proliferation (Bieker, 2001). Target genes include collagen α1, keratin 4, TGFβ type I and II receptors, and others (Chiambaretta et al., 2006). KLF3 is primarily associated as a strong transcriptional repressor associated with adipogenesis and lipid metabolism (Pearson et al., 2011), with expression rates varying across different tissues and cell types (Bieker, 2001). KLF9 and 10 also act as transcriptional repressors but are ubiquitously expressed across a wider range of tissues (Swamynathan, 2010). KLF9 is a tumour suppressor (Sun et al., 2014) and regulates inflammation, while KLF10 has a major role in TGF-β-linked inhibition of cell proliferation, inflammation and initiating apoptosis (Subramaniam et al., 2010).

We identify that the majority of TFs at 48 hours are KLFs exhibiting a range of transcriptional regulation. However, as outlined above, most KLFs have been examined in cancer or other disease models with limited insight into pathogen-mediated infections. Of particular interest is KLF10, which acts as a repressor to the multifunctional cytokine TGF-β; which plays an important role in host resistance and cell immunity by acting with various interferons and interleukins (Papadakis et al., 2015; Sarmento et al., 2015). By increasing the expression of this TFs, it could provide an additional avenue for *Chlamydia* to regulate the host immune system.

Histone deacetylases (HDACs) modify the core histones of the nucleosome, providing an important function in transcriptional regulation (de Ruijter et al., 2003), and many bacterial pathogens subvert HDACs to suppress host defences (Grabiec and Potempa, 2018). KLF9, and 10 share the co-factor Sin3A (SIN3 Transcription Regulator Family Member A) (Swamynathan, 2010), which is also a core component of the chromatin-modifying complex mediating transcriptional repression (Cowley et al., 2005). The Sin3a/HDAC complex is made up of two histone deacetylases HDAC1 and HDAC2. HDAC2 has increased chromatin accessibility at all four time points, and HDAC9 has increased chromatin accessibility at 1, 24 and 48 hours, further supporting the potential for histone modifications to be a component of the host cell response to chlamydial infection, or to be targets of chlamydial effectors (Pennini et al., 2010).

Due to strict filtering and thresholds to determine the final set of TF motifs, we acknowledge that we have likely missed many weak binding sites that may also be relevant to infection. However, as this analysis is amongst a limited number of studies examining the impacts of chromatin accessibility in response to infection, we have only shown the most significant results, and further analyses can be run using the large amount of supplementary data available.

All enriched TFs discussed above (**Table 3.2**), were compared against gene expression studies to ensure that each TF is expressed in HEp2 and HeLa cells from similar times (data not shown). To compare each TFs regulatory control on target genes, we examined the genes underlying significant regions associated with each motif, comparing fold-change differences across different infection-based environments (**Supplementary Figure 3.3**).

Surprisingly, we do not see a consensus of affected gene expression from any TF. For example, previous studies have indicated that KLF3 is a strong transcriptional repressor, but here, we see a range of fold-changes in the corresponding genes, with two of the three datasets showing only a slight decrease in mean expression (**Supplementary Figure 3.3E**). We attribute these differences to the variability as previously discussed in overlaying different infection-based datasets. However, this does highlight possible further roles for these TFs in an infection-based setting, as many of the studies outlining each TF activating or repressing roles were taken from more common disease models such as cancer. Therefore, these results further highlight the diverse and complex mechanisms associated with chlamydial infection.

3.4.9. Challenges and questions associated with analysing chromatin accessibility

As identified earlier, the number of protein coding genes in the human genome is < 5%, with the majority of regions encoding regulatory features and mechanisms (Shabalina and Spiridonov, 2004). We see this displayed in this data, particularly in the large number of intergenic regions identified (49%) that are likely regulatory in nature. Additional prediction-based software analyses would likely reduce the number of intergenic regions by predicting additional features such as silencers, insulators and possibly more enhancers. Although extremely useful in the right setting, we chose not to run prediction-based tools as we wanted the results to be less speculative and only highlight significant infection-relevant events.

Throughout this manuscript, we have made the assumption that open and closed chromatin are directly associated with an increase or decrease in gene expression respectively. Due to the snapshot-based capture of current sequencing-based approaches, we identify that this may not always be valid. For example, some regions of open chromatin may be in the process of being closed, and therefore would not exhibit an increase in expression. Additionally, open chromatin regions can facilitate the binding of a transcriptional repressor, again resulting in decreased expression. Due to the challenges in overlapping gene expression studies from the same cell line, time points, chlamydial species and experimental conditions; overlapping all identified genes identified here with their matching expression patterns was not practical. Therefore, we recommend future chromatin accessibility-based studies take this into consideration and complement future studies with matching gene expression data.

3.5. Conclusions

We describe comprehensive changes to chromatin accessibility upon chlamydial infection in epithelial cells *in vitro* using FAIRE-seq. We identify both conserved and time-specific infection-responsive changes to a variety of features and regulatory elements over the course of the chlamydial developmental cycle that may shape the host cell response to infection, including promotors, enhancers, and transcription factor motifs. Some of these changes are associated with genomic features and genes known to be relevant to chlamydial infection, including innate immunity and complement, acquisition of host cell lipids and nutrients, intracellular signalling, cell-cell adhesion, metabolism and apoptosis.

Host cell chromatin accessibility changes are evident over the entire chlamydial developmental cycle, with a large proportion of all chromatin accessibility changes at 48 hours post infection. This likely reflects the confluence of late stages of developmental cycle

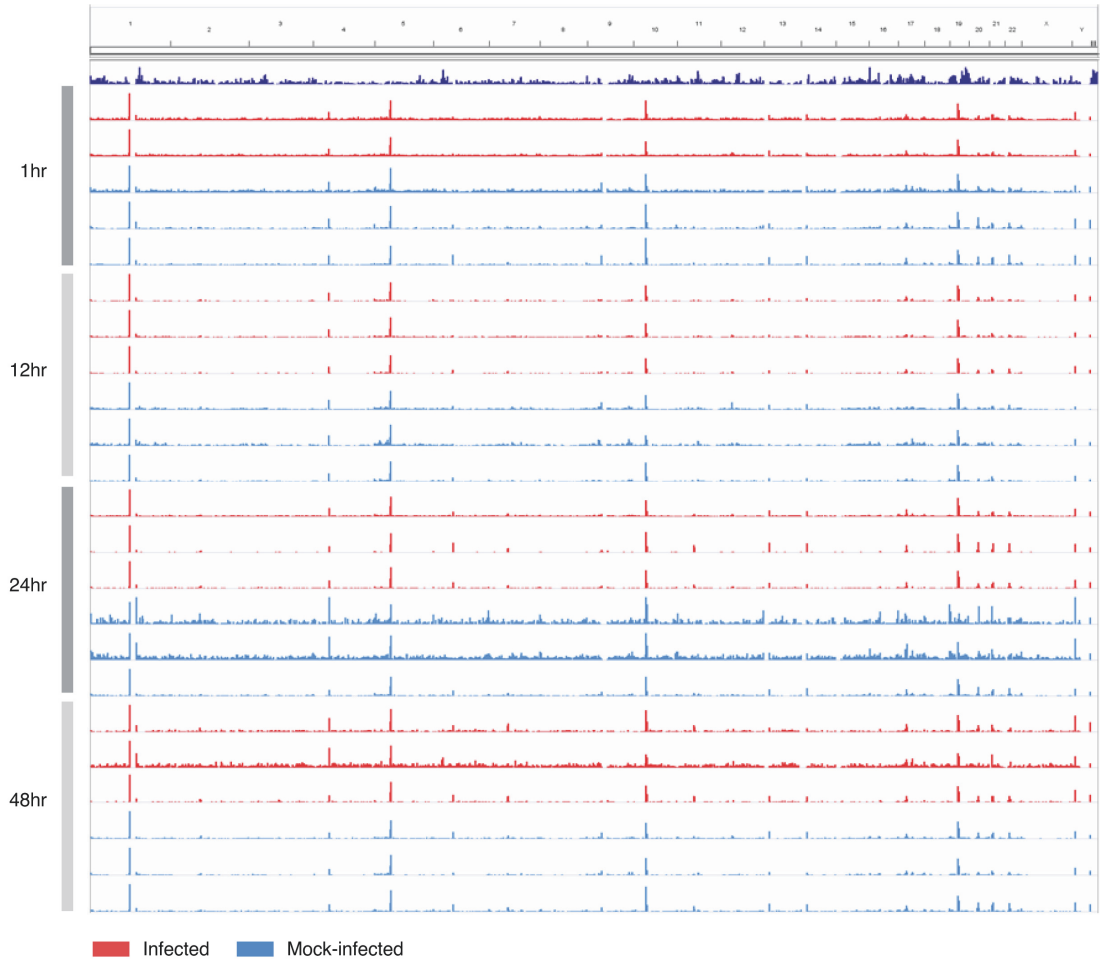
events, however significant changes to chromatin accessibility are readily apparent as early as 1-hour post infection. We find altered chromatin accessibility in several gene regions, ontologies and TF motifs associated with ECM moieties, particularly cadherins and their interconnected regulatory pathways, and Smad signalling. Disruption of the ECM is thought to be a central component of dysregulated fibrotic processes that may underpin the inflammatory scarring outcomes of chlamydial infection (Humphrys et al., 2013), and our data further highlights a central role of the ECM in epithelial cell responses to infection. We also identify factors that have not been previously described in the context of chlamydial infection, notably the enrichment of the KLF family of transcription factor motifs within differential chromatin accessible regions in the latter stages of infection. Dysregulation of the biologically complex KLFs and their transcriptional networks is linked to several reproductive tract pathologies in both men and women (Simmen et al., 2015), thus our discovery of enriched KLF binding motifs in response to chlamydial infection is compelling, given the scale and burden of chlamydial reproductive tract disease globally (Menon et al., 2015).

We also identify limitations and considerations for future studies. Specifically, including gene expression data that overlaps similar times (and generated from the same cell cultures), will help to characterise which chromatin accessibility events are directly related to infection-specific changes in gene expression.

In summary, this is the first genome-scale analysis of the impact of chlamydial infection on the human epithelial cell epigenome, encompassing the chlamydial developmental cycle at early, mid and late times. This has yielded a novel perspective of the complex host epithelial cell response to infection, and will inform further studies of transcriptional regulation and epigenomic regulatory elements in *Chlamydia*-infected human cells and tissues. Examination of the multifaceted human epigenome, and its potential subversion by *Chlamydia*, using *in*

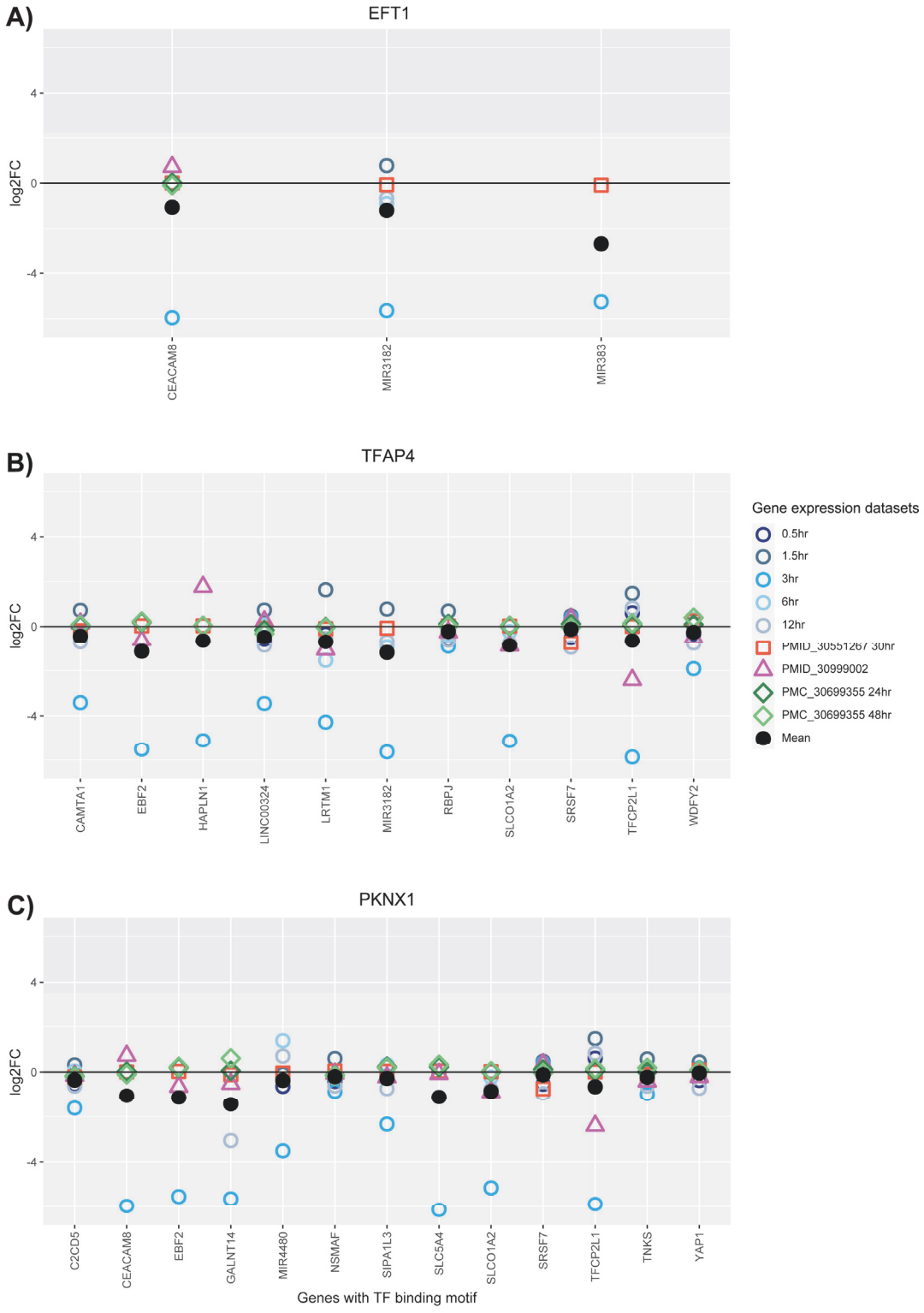
vivo mouse models of infection and *ex vivo* human reproductive tract tissues, will continue to shed light on how the host cell response contributes to infection outcomes.

3.6. Supplementary figures



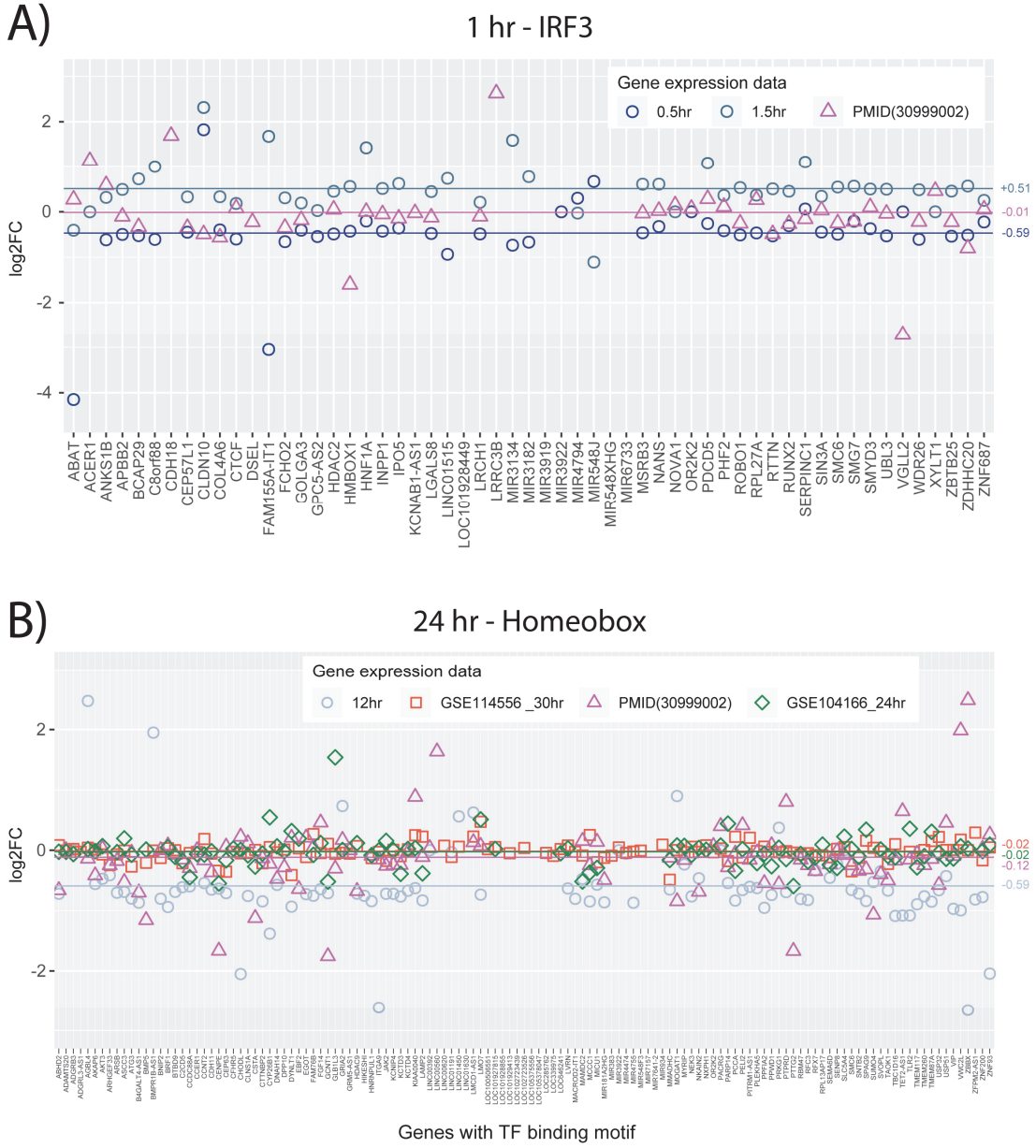
Supplementary Figure 3.1. Genome coverage plots

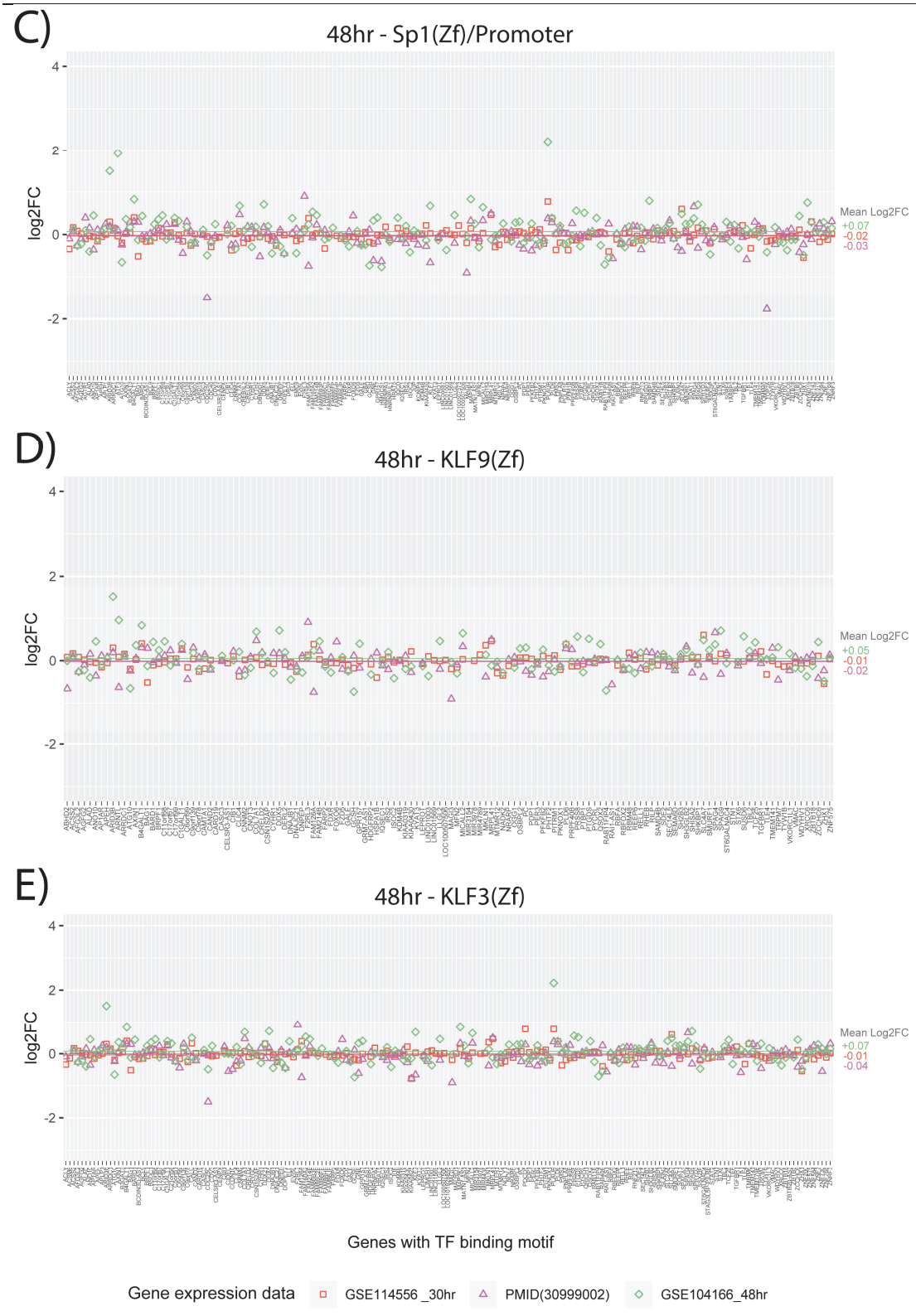
Significant peaks from each replicate as determined by MACS2. Screenshots are from IGV (Integrative Genomics Viewer) showing that all replicates contain significant peaks genome-wide (human genome) without any visual chromosomal bias.

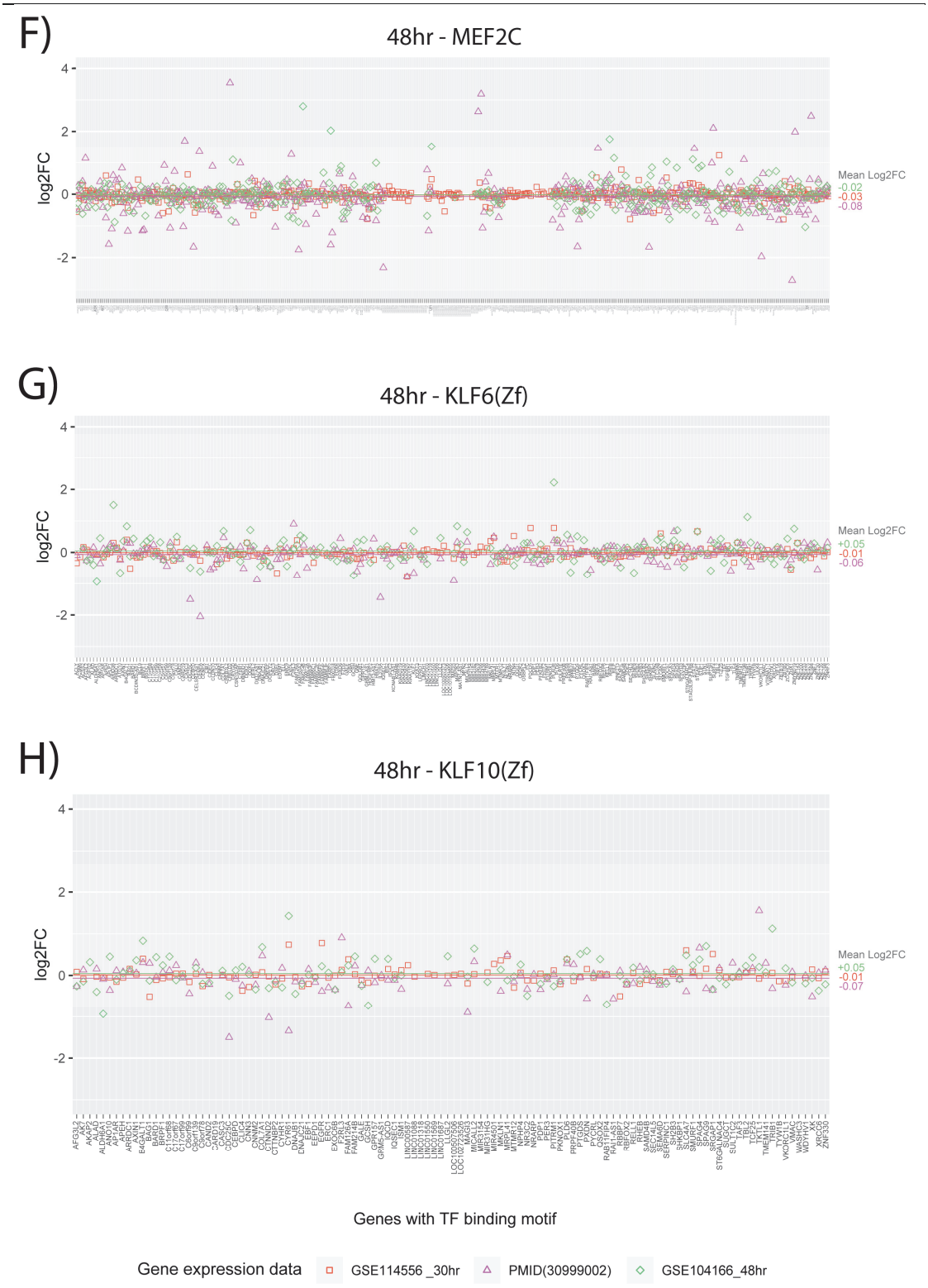


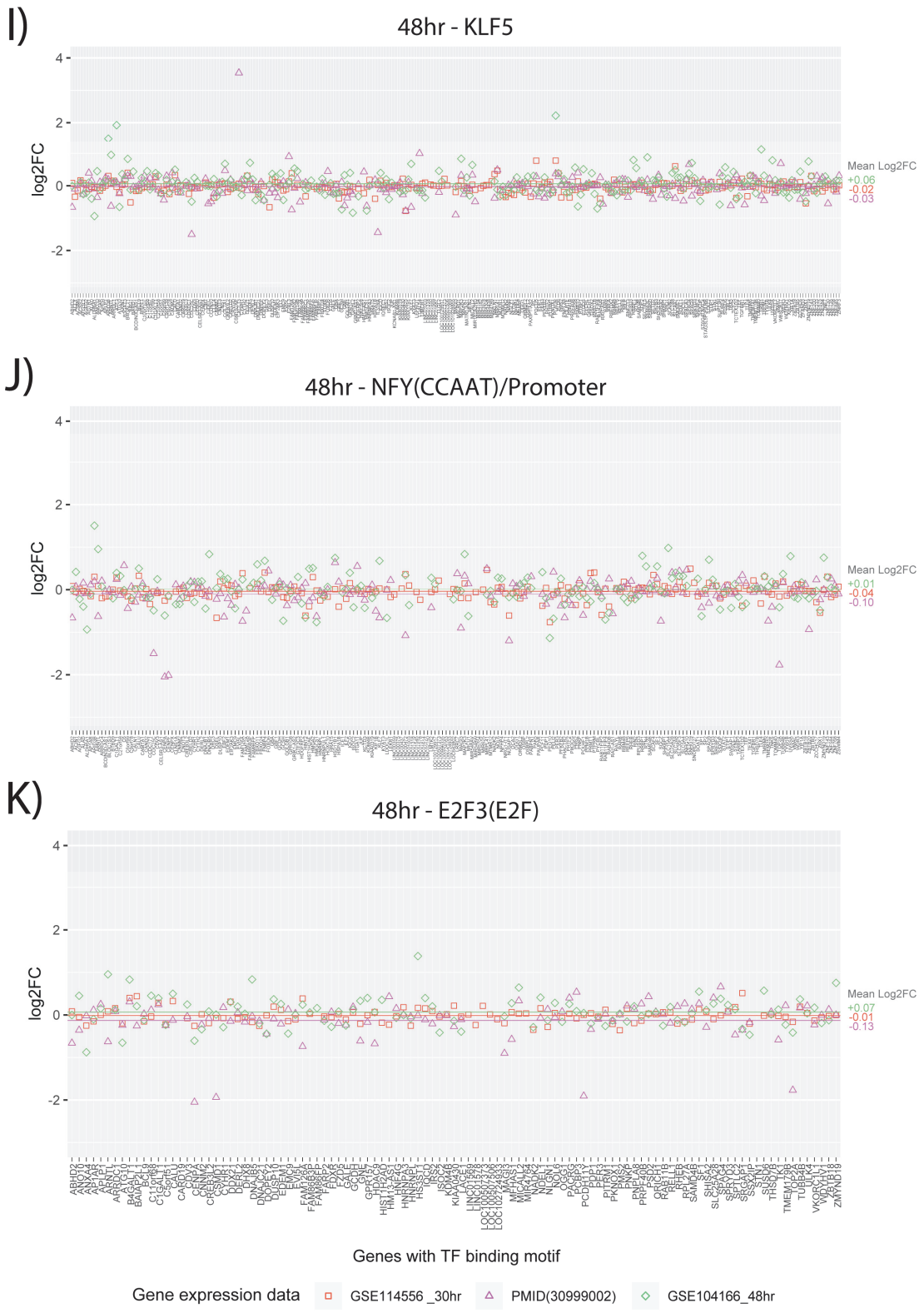
Supplementary Figure 3.2. Conserved transcription factor expression

Motifs associated with each transcription factor (TF) as identified within the conserved regions. Genes associated with these regions were compared against relevant gene expression data to identify their level of regulation during infection. The TF POU3F2 was not able to be compared as the motif was only identified within intergenic regions that could not be overlapped. **A)** ETS1 TF. **B)** TFAP4 TF. **C)** PKNOX1 TF.


A)
1 hr - IRF3
Gene expression data
0.5hr
1.5hr
PMID(30999002)
log2FC
+0.51
-0.01
-0.59
B)
24 hr - Homeobox
Gene expression data
12hr
GSE114556 _30hr
PMID(30999002)
GSE104166_24hr
log2FC
-0.02
-0.02
-0.12
-0.59
Genes with TF binding motif


C)
48hr - Sp1(Zf)/Promoter
log2FC
Mean Log2FC
+0.07
-0.02
-0.03
D)
48hr - KLF9(Zf)
log2FC
Mean Log2FC
+0.05
-0.01
-0.02
E)
48hr - KLF3(Zf)
log2FC
Mean Log2FC
+0.07
-0.01
-0.04
Genes with TF binding motif
Gene expression data
GSE114556 _30hr
PMID(30999002)
GSE104166_48hr


F)
48hr - MEF2C
log2FC
Mean Log2FC
-0.02
-0.03
-0.08
G)
48hr - KLF6(Zf)
log2FC
Mean Log2FC
+0.05
-0.01
-0.06
H)
48hr - KLF10(Zf)
log2FC
Mean Log2FC
+0.05
-0.01
-0.07
Genes with TF binding motif
Gene expression data
GSE114556 _30hr
PMID(30999002)
GSE104166_48hr


I)
48hr - KLF5
log2FC
Mean Log2FC
+0.06
-0.02
-0.03
J)
48hr - NFY(CCAAT)/Promoter
log2FC
Mean Log2FC
+0.01
-0.04
-0.10
K)
48hr - E2F3(E2F)
log2FC
Mean Log2FC
+0.07
-0.01
-0.13
Genes with TF binding motif
Gene expression data
GSE114556 _30hr
PMID(30999002)
GSE104166_48hr

Supplementary Figure 3.3. Time specific transcription factor expression

Motifs associated with each transcription factor (TF) (**Table 3.2**) were identified within significant differentially accessible regions. Genes associated with these regions were compared against relevant gene expression data to identify their level of regulation during infection. **A)** IRF3 TF from 1 hour. **B)** Homeobox TF from 24 hours. **C-K**) Nine TFs identified at 48 hours.

3.7. Supplementary files

Supplementary File 3.1 Annotation of all significant peaks

Annotation of all the significant peaks, with tabs separating genomic features and fold-change regulation.

Supplementary File 3.1.xlsx

Supplementary File 3.2 Time specific regions

The list of time-specific differential chromatin accessible regions. It should be noted that some genes in these lists are repeated at each time due to multiple peaks occurring at an annotated interval, that enhancers can affect more than one gene, and single genes can be affected by more than one enhancer.

Supplementary File 3.2.xlsx

Supplementary File 3.3 Complete list of motifs and transcription factors

The complete list of significant motifs and enriched transcription factors.

Supplementary File 3.3.xlsx

3.8. References

Alonso, A., and Garcia-del Portillo, F. (2004). Hijacking of eukaryotic functions by intracellular bacterial pathogens. International microbiology : the official journal of the Spanish Society for Microbiology *7*, 181-191.

Andrews, S. (2010). FastQC: A Quality Control tool for High Throughput Sequence Data.

Attisano, L., and Tuen Lee-Hoeflich, S. (2001). The Smads. Genome Biol *2*, reviews3010.3011.

Backert, S., Schmidt, T.P., Harrer, A., and Wessler, S. (2017). Exploiting the Gastric Epithelial Barrier: Helicobacter pylori's Attack on Tight and Adherens Junctions. Current topics in microbiology and immunology *400*, 195-226.

Baer, M., Nilsen, T.W., Costigan, C., and Altman, S. (1990). Structure and transcription of a human gene for H1 RNA, the RNA component of human RNase P. Nucleic acids research *18*, 97-103.

Bergsson, G., Arnfinnsson, J., Karlsson, S.M., Steingrímsson, Ó., and Thormar, H. (1998). In Vitro Inactivation of *Chlamydia trachomatis* by Fatty Acids and Monoglycerides. Antimicrobial Agents and Chemotherapy *42*, 2290-2294.

Betts-Hampikian, H.J., and Fields, K.A. (2010). The Chlamydial Type III Secretion Mechanism: Revealing Cracks in a Tough Nut. Frontiers in microbiology *1*, 114.

Bevilacqua, M.A., Faniello, M.C., D'Agostino, P., Quaresima, B., Tiano, M.T., Pignata, S., Russo, T., Cimino, F., and Costanzo, F. (1995). Transcriptional activation of the H-ferritin gene in differentiated Caco-2 cells parallels a change in the activity of the nuclear factor Bbf. The Biochemical journal *311 (Pt 3)*, 769-773.

Bevilacqua, M.A., Faniello, M.C., Quaresima, B., Tiano, M.T., Giuliano, P., Feliciello, A., Avvedimento, V.E., Cimino, F., and Costanzo, F. (1997). A common mechanism underlying the E1A repression and the cAMP stimulation of the H ferritin transcription. The Journal of biological chemistry *272*, 20736-20741.

Bevilacqua, M.A., Giordano, M., D'Agostino, P., Santoro, C., Cimino, F., and Costanzo, F. (1992). Promoter for the human ferritin heavy chain-encoding gene (FERH): structural and functional characterization. Gene *111*, 255-260.

Bianchi, M., Crinelli, R., Arbore, V., and Magnani, M. (2018). Induction of ubiquitin C (UBC) gene transcription is mediated by HSF1: role of proteotoxic and oxidative stress. FEBS Open Bio *8*, 1471-1485.

Bieker, J.J. (2001). Kruppel-like factors: three fingers in many pies. The Journal of biological chemistry *276*, 34355-34358.

Bierne, H., and Cossart, P. (2012). When bacteria target the nucleus: the emerging family of nucleomodulins. Cell Microbiol *14*, 622-633.

Bierne, H., Hamon, M., and Cossart, P. (2012). Epigenetics and bacterial infections. Cold Spring Harb Perspect Med *2*, a010272.

Boehm, M., Simson, D., Escher, U., Schmidt, A.M., Bereswill, S., Tegtmeyer, N., Backert, S., and Heimesaat, M.M. (2018). Function of Serine Protease HtrA in the Lifecycle of the Foodborne Pathogen *Campylobacter jejuni*. European journal of microbiology & immunology *8*, 70-77.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England) *30*, 2114-2120.

Brunham, R.C., and Rey-Ladino, J. (2005). Immunology of *Chlamydia* infection: implications for a *Chlamydia trachomatis* vaccine. Nat Rev Immunol *5*, 149-161.

Burton, M.J. (2007). Trachoma: an overview. British Medical Bulletin *84*, 99-116.

Chiambaretta, F., Nakamura, H., De Graeve, F., Sakai, H., Marceau, G., Maruyama, Y., Rigal, D., Dastugue, B., Sugar, J., Yue, B.Y.*, et al.* (2006). Kruppel-like factor 6 (KLF6) affects the promoter activity of the alpha1-proteinase inhibitor gene. Investigative ophthalmology & visual science *47*, 582-590.

Cocchiaro, J.L., Kumar, Y., Fischer, E.R., Hackstadt, T., and Valdivia, R.H. (2008). Cytoplasmic lipid droplets are translocated into the lumen of the *Chlamydia trachomatis*

parasitophorous vacuole. Proceedings of the National Academy of Sciences *105*, 9379-9384.

Cowley, S.M., Iritani, B.M., Mendrysa, S.M., Xu, T., Cheng, P.F., Yada, J., Liggitt, H.D., and Eisenman, R.N. (2005). The mSin3A Chromatin-Modifying Complex Is Essential for Embryogenesis and T-Cell Development. Molecular and Cellular Biology *25*, 6990-7004.

Dautry-Varsat, A., Balana, M.E., and Wyplosz, B. (2004). Chlamydia--host cell interactions: recent advances on bacterial entry and intracellular development. Traffic (Copenhagen, Denmark) *5*, 561-570.

de Ruijter, A.J., van Gennip, A.H., Caron, H.N., Kemp, S., and van Kuilenburg, A.B. (2003). Histone deacetylases (HDACs): characterization of the classical HDAC family. The Biochemical journal *370*, 737-749.

Deniaud, E., Baguet, J., Chalard, R., Blanquier, B., Brinza, L., Meunier, J., Michallet, M.-C., Laugraud, A., Ah-Soon, C., Wierinckx, A.*, et al.* (2009). Overexpression of Transcription Factor Sp1 Leads to Gene Expression Perturbations and Cell Cycle Inhibition. PloS one *4*, e7035.

Di Paolo, Nelson C., Doronin, K., Baldwin, Lisa K., Papayannopoulou, T., and Shayakhmetov, Dmitry M. (2013). The Transcription Factor IRF3 Triggers "Defensive Suicide" Necrosis in Response to Viral and Bacterial Pathogens. Cell Reports *3*, 1840-1846.

Dong, J.T., and Chen, C. (2009). Essential role of KLF5 transcription factor in cell proliferation and differentiation and its implications for human diseases. Cellular and molecular life sciences : CMLS *66*, 2691-2706.

Duval, M., Cossart, P., and Lebreton, A. (2017). Mammalian microRNAs and long noncoding RNAs in the host-bacterial pathogen crosstalk. Seminars in Cell & Developmental Biology *65*, 11-19.

Egloff, S., Studniarek, C., and Kiss, T. (2018). 7SK small nuclear RNA, a multifunctional transcriptional regulatory RNA with gene-specific features. Transcription *9*, 95-101.

Elwell, C., and Engel, J. (2018). Emerging Role of Retromer in Modulating Pathogen Growth. Trends in microbiology *26*, 769-780.

Elwell, C.A., and Engel, J.N. (2012). Lipid acquisition by intracellular *Chlamydiae*. Cell Microbiol *14*, 1010-1018.

Faniello, M.C., Bevilacqua, M.A., Condorelli, G., de Crombrugghe, B., Maity, S.N., Avvedimento, V.E., Cimino, F., and Costanzo, F. (1999). The B subunit of the CAAT-binding factor NFY binds the central segment of the Co-activator p300. The Journal of biological chemistry *274*, 7623-7626.

Fields, K.A., and Hackstadt, T. (2002). The Chlamydial Inclusion: Escape from the Endocytic Pathway. Annual Review of Cell and Developmental Biology *18*, 221-245.

Gao, T., He, B., Liu, S., Zhu, H., Tan, K., and Qian, J. (2016). EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. Bioinformatics *32*, 3543-3551.

Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., Bosco, D.*, et al.* (2010). A map of open chromatin in human pancreatic islets. Nature genetics *42*, 255-259.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., and Lieb, J.D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome research *17*, 877-885.

Gloeckl, S., Ong, V.A., Patel, P., Tyndall, J.D.A., Timms, P., Beagley, K.W., Allan, J.A., Armitage, C.W., Turnbull, L., Whitchurch, C.B.*, et al.* (2013). Identification of a serine protease inhibitor which causes inclusion vacuole reduction and is lethal to *Chlamydia trachomatis*. Molecular microbiology *89*, 676-689.

Grabiec, A.M., and Potempa, J. (2018). Epigenetic regulation in bacterial infections: targeting histone deacetylases. Critical Reviews in Microbiology *44*, 336-350.

Gregory, T.R. (2005). Synergy between sequence and size in large-scale genomics. Nature reviews Genetics *6*, 699-708.

Grieshaber, S.S., Grieshaber, N.A., and Hackstadt, T. (2003). *Chlamydia trachomatis* uses host cell dynein to traffic to the microtubule-organizing center in a p50 dynamitin-independent process. Journal of cell science *116*, 3793-3802.

Guo, B., Godzik, A., and Reed, J.C. (2001). Bcl-G, a novel pro-apoptotic member of the Bcl-2 family. The Journal of biological chemistry *276*, 2780-2785.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol *8*, R24-R24.

Gyorke, C.E., and Nagarajan, U. (2018). Interferon-Independent Protection by Interferon Regulatory Factor 3. The Journal of Immunology *200*, 114.125-114.125.

Haldar, A.K., Piro, A.S., Finethy, R., Espenschied, S.T., Brown, H.E., Giebel, A.M., Frickel, E.-M., Nelson, D.E., and Coers, J. (2016). *Chlamydia trachomatis* Is Resistant to Inclusion Ubiquitination and Associated Host Defense in Gamma Interferon-Primed Human Epithelial Cells. mBio *7*, e01417-01416.

Hamon, M.A., and Cossart, P. (2008). Histone modifications and chromatin remodeling during bacterial infections. Cell host & microbe *4*, 100-109.

He, Y., Carrillo, J.A., Luo, J., Ding, Y., Tian, F., Davidson, I., and Song, J. (2014). Genome-wide mapping of DNase I hypersensitive sites and association analysis with gene expression in MSB1 cells. Frontiers in genetics *5*, 308-308.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell *38*, 576-589.

Hermanns, P., Bertuch, A.A., Bertin, T.K., Dawson, B., Schmitt, M.E., Shaw, C., Zabel, B., and Lee, B. (2005). Consequences of mutations in the non-coding RMRP RNA in cartilage-hair hypoplasia. Human molecular genetics *14*, 3723-3740.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F.*, et al.* (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res *34*, D590-598.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934-947.

Ho, J., Moyes, D.L., Tavassoli, M., and Naglik, J.R. (2017). The Role of ErbB Receptors in Infection. Trends Microbiol *25*, 942-952.

Hodgkinson, V., and Petris, M.J. (2012). Copper homeostasis at the host-pathogen interface. The Journal of biological chemistry *287*, 13549-13555.

Hoffmann, R. (2008). A wiki for the life sciences where authorship matters. Nat Genet *40*, 1047-1051.

Hu, Y.F., Luscher, B., Admon, A., Mermod, N., and Tjian, R. (1990). Transcription factor AP-4 contains multiple dimerization domains that regulate dimer specificity. Genes & development *4*, 1741-1752.

Humphrys, M.S., Creasy, T., Sun, Y., Shetty, A.C., Chibucos, M.C., Drabek, E.F., Fraser, C.M., Farooq, U., Sengamalay, N., Ott, S.*, et al.* (2013). Simultaneous Transcriptional Profiling of Bacteria and Their Host Cells. PloS one *8*, e80597.

Hybiske, K., and Stephens, R.S. (2007). Mechanisms of host cell exit by the intracellular bacterium Chlamydia. Proceedings of the National Academy of Sciences of the United States of America *104*, 11430-11435.

Igietseme, J.U., Omosun, Y., Nagy, T., Stuchlik, O., Reed, M.S., He, Q., Partin, J., Joseph, K., Ellerson, D., George, Z.*, et al.* (2018). Molecular Pathogenesis of Chlamydia Disease Complications: Epithelial-Mesenchymal Transition and Fibrosis. Infection and immunity *86*.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Cheneby, J., Kulkarni, S.R., Tan, G.*, et al.* (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic acids research *46*, D260-d266.

Khan, A., and Zhang, X. (2016). dbSUPER: a database of super-enhancers in mouse and human genome. Nucleic acids research *44*, D164-171.

Kundaje, A. (2016). A comprehensive collection of signal artifact blacklist regions in the human genome. ENCODE. [hg19-blacklist-README.pdf].

Ladomersky, E., Khan, A., Shanbhag, V., Cavet, J.S., Chan, J., Weisman, G.A., and Petris, M.J. (2017). Host and Pathogen Copper-Transporting P-Type ATPases Function Antagonistically during *Salmonella* Infection. Infection and immunity *85*, e00351-00317.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*.

Liu, Y., Cao, Z., Wang, Y., Guo, Y., Xu, P., Yuan, P., Liu, Z., He, Y., and Wei, W. (2018). Genome-wide screening for functional long noncoding RNAs in human cells by Cas9 targeting of splice sites. Nature biotechnology *36*, 1203-1210.

Lou, Y., Hu, M., Mao, L., Zheng, Y., and Jin, F. (2017). Involvement of serum glucocorticoid-regulated kinase 1 in reproductive success. FASEB journal : official publication of the Federation of American Societies for Experimental Biology *31*, 447-456.

Manzanillo, P.S., Ayres, J.S., Watson, R.O., Collins, A.C., Souza, G., Rae, C.S., Schneider, D.S., Nakamura, K., Shiloh, M.U., and Cox, J.S. (2013). The ubiquitin ligase parkin mediates resistance to intracellular pathogens. Nature *501*, 512.

Menon, S., Timms, P., Allan, J.A., Alexander, K., Rombauts, L., Horner, P., Keltz, M., Hocking, J., and Huston, W.M. (2015). Human and Pathogen Factors Associated with *Chlamydia trachomatis*-Related Infertility in Women. Clinical microbiology reviews *28*, 969-985.

Misaghi, S., Balsara, Z.R., Catic, A., Spooner, E., Ploegh, H.L., and Starnbach, M.N. (2006). *Chlamydia trachomatis*-derived deubiquitinating enzymes in mammalian cells during infection. Molecular microbiology *61*, 142-150.

Miyazawa, K., and Miyazono, K. (2017). Regulation of TGF-beta Family Signaling by Inhibitory Smads. Cold Spring Harb Perspect Biol *9*.

Mölleken, K., Becker, E., and Hegemann, J.H. (2013). The *Chlamydia pneumoniae* Invasin Protein Pmp21 Recruits the EGF Receptor for Host Cell Entry. PLoS pathogens *9*, e1003325.

Ncbi Resource Coordinators (2016). Database resources of the National Center for Biotechnology Information. Nucleic acids research *44*, D7-19.

Ohmer, M., Tzivelekidis, T., Niedenfuhr, N., Volceanov-Hahn, L., Barth, S., Vier, J., Borries, M., Busch, H., Kook, L., Biniossek, M.L.*, et al.* (2019). Infection of HeLa cells with *Chlamydia trachomatis* inhibits protein synthesis and causes multiple changes to host cell pathways. Cell Microbiol *21*, e12993.

Ortega, Á.D., Quereda, J.J., Pucciarelli, M.G., and García-del Portillo, F. (2014). Non-coding RNA regulation in pathogenic bacteria located inside eukaryotic cells. Frontiers in cellular and infection microbiology *4*.

Paes, W., Dowle, A., Coldwell, J., Leech, A., Ganderton, T., and Brzozowski, A. (2018). The *Chlamydia trachomatis* PmpD adhesin forms higher order structures through disulphide-mediated covalent interactions. PloS one *13*, e0198662.

Papadakis, K.A., Krempski, J., Reiter, J., Svingen, P., Xiong, Y., Sarmento, O.F., Huseby, A., Johnson, A.J., Lomberk, G.A., Urrutia, R.A.*, et al.* (2015). Krüppel-like factor KLF10 regulates transforming growth factor receptor II expression and TGF-β signaling in CD8+ T lymphocytes. Am J Physiol Cell Physiol *308*, C362-C371.

Parnas, O., Jovanovic, M., Eisenhaure, Thomas M., Herbst, Rebecca H., Dixit, A., Ye, Chun J., Przybylski, D., Platt, Randall J., Tirosh, I., Sanjana, Neville E.*, et al.* (2015). A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. Cell *162*, 675-686.

Patel, A.L., Chen, X., Wood, S.T., Stuart, E.S., Arcaro, K.F., Molina, D.P., Petrovic, S., Furdui, C.M., and Tsang, A.W. (2014). Activation of epidermal growth factor receptor is required for *Chlamydia trachomatis* development. BMC Microbiol *14*, 277.

Paul, B., Kim, H.S., Kerr, M.C., Huston, W.M., Teasdale, R.D., and Collins, B.M. (2017). Structural basis for the hijacking of endosomal sorting nexin proteins by *Chlamydia trachomatis*. Elife *6*.

Pearson, R.C., Funnell, A.P., and Crossley, M. (2011). The mammalian zinc finger transcription factor Kruppel-like factor 3 (KLF3/BKLF). IUBMB life *63*, 86-93.

Pennini, M.E., Perrinet, S., Dautry-Varsat, A., and Subtil, A. (2010). Histone Methylation by NUE, a Novel Nuclear Effector of the Intracellular Pathogen *Chlamydia trachomatis*. PLOS Pathogens *6*, e1000995.

Rajic, J., Inic-Kanada, A., Stein, E., Dinic, S., Schuerer, N., Uskokovic, A., Ghasemian, E., Mihailovic, M., Vidakovic, M., Grdovic, N.*, et al.* (2017). *Chlamydia trachomatis* Infection Is Associated with E-Cadherin Promoter Methylation, Downregulation of E-Cadherin Expression, and Increased Expression of Fibronectin and alpha-SMA-Implications for Epithelial-Mesenchymal Transition. Frontiers in cellular and infection microbiology *7*, 253.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic acids research *42*, W187-W191.

Reyburn, H. (2016). WHO Guidelines for the Treatment of *Chlamydia trachomatis*. WHO *340*, c2637-c2637.

Reyes, A., and Huber, W. (2018). Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. Nucleic acids research *46*, 582-592.

Ribet, D., and Cossart, P. (2015). How bacterial pathogens colonize their hosts and invade deeper tissues. Microbes Infect *17*, 173-183.

Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R., Dunning, M.J., Brown, G.D., Gojis, O., Ellis, I.O., Green, A.R.*, et al.* (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature *481*, 389-393.

Sarmento, O.F., Svingen, P.A., Xiong, Y., Xavier, R.J., McGovern, D., Smyrk, T.C., Papadakis, K.A., Urrutia, R.A., and Faubion, W.A. (2015). A novel role for KLF14 in T regulatory cell differentiation. Cellular and molecular gastroenterology and hepatology *1*, 188-202.e184.

Schachter, J., Storz, J., Tarizzo, M.L., and Bögel, K. (1973). *Chlamydiae* as agents of human and animal diseases. Bull World Health Organ *49*, 443-449.

Schneider, M.R., and Kolligs, F.T. (2015). E-cadherin's role in development, tissue homeostasis and disease: Insights from mouse models: Tissue-specific inactivation of the adhesion protein E-cadherin in mice reveals its functions in health and disease. BioEssays : news and reviews in molecular, cellular and developmental biology *37*, 294-304.

Seaman, M.N.J. (2012). The retromer complex – endosomal protein recycling and beyond. Journal of cell science *125*, 4693-4702.

Shabalina, S.A., and Spiridonov, N.A. (2004). The mammalian transcriptome and the function of non-coding DNA sequences. Genome Biol *5*, 105.

Sharkey, D.J., Macpherson, A.M., Tremellen, K.P., Mottershead, D.G., Gilchrist, R.B., and Robertson, S.A. (2012). TGF-β Mediates Proinflammatory Seminal Fluid Signaling in Human Cervical Epithelial Cells. The Journal of Immunology *189*, 1024-1035.

Simmen, R.C., Heard, M.E., Simmen, A.M., Montales, M.T., Marji, M., Scanlon, S., and Pabona, J.M. (2015). The Kruppel-like factors in female reproductive system pathologies. J Mol Endocrinol *54*, R89-r101.

Simon, J.M., Giresi, P.G., Davis, I.J., and Lieb, J.D. (2012). Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. Nat Protoc *7*, 256-267.

Sixt, B.S., Bastidas, R.J., Finethy, R., Baxter, R.M., Carpenter, V.K., Kroemer, G., Coers, J., and Valdivia, R.H. (2017). The *Chlamydia trachomatis* Inclusion Membrane Protein CpoS Counteracts STING-Mediated Cellular Surveillance and Suicide Programs. Cell host & microbe *21*, 113-121.

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T.I., Nudel, R., Lieder, I., Mazor, Y.*, et al.* (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Current protocols in bioinformatics *54*, 1.30.31-31.30.33.

Subramaniam, M., Hawse, J.R., Rajamannan, N.M., Ingle, J.N., and Spelsberg, T.C. (2010). Functional role of KLF10 in multiple disease processes. BioFactors (Oxford, England) *36*, 8-18.

Sun, J., Wang, B., Liu, Y., Zhang, L., Ma, A., Yang, Z., Ji, Y., and Liu, Y. (2014). Transcription factor KLF9 suppresses the growth of hepatocellular carcinoma cells in vivo and positively regulates p53 expression. Cancer letters *355*, 25-33.

Swamynathan, S.K. (2010). Krüppel-like factors: three fingers in control. Human genomics *4*, 263-270.

Takimoto, T., Wakabayashi, Y., Sekiya, T., Inoue, N., Morita, R., Ichiyama, K., Takahashi, R., Asakawa, M., Muto, G., Mori, T.*, et al.* (2010). Smad2 and Smad3 are redundantly essential for the TGF-beta-mediated regulation of regulatory T plasticity and Th1 development. Journal of immunology (Baltimore, Md : 1950) *185*, 842-855.

Tan, C., Hsia, R.-c., Shou, H., Haggerty, C.L., Ness, R.B., Gaydos, C.A., Dean, D., Scurlock, A.M., Wilson, D.P., and Bavoil, P.M. (2009). *Chlamydia trachomatis*-infected patients display variable antibody profiles against the nine-member polymorphic membrane protein family. Infection and immunity *77*, 3218-3226.

Tan, N.Y., and Khachigian, L.M. (2009). Sp1 Phosphorylation and Its Regulation of Gene Transcription. Molecular and Cellular Biology *29*, 2483-2488.

The UniProt Consortium (2016). UniProt: the universal protein knowledgebase. Nucleic Acids Research *45*, D158-D169.

Tokusumi, Y., Ma, Y., Song, X., Jacobson, R.H., and Takada, S. (2007). The New Core Promoter Element XCPE1 (X Core Promoter Element 1) Directs Activator-, Mediator-, and TATA-Binding Protein-Dependent but TFIID-Independent RNA Polymerase II Transcription from TATA-Less Promoters. Molecular and Cellular Biology *27*, 1844-1858.

Topham, M.K., and Prescott, S.M. (1999). Mammalian diacylglycerol kinases, a family of lipid kinases with signaling functions. J Biol Chem *274*, 11447-11450.

Tsompana, M., and Buck, M.J. (2014). Chromatin accessibility: a window into the genome. Epigenetics & Chromatin *7*, 33.

Ullu, E., and Weiner, A.M. (1984). Human genes and pseudogenes for the 7SL RNA component of signal recognition particle. EMBO J *3*, 3303-3310.

van Ooij, C., Kalman, L., van, I., Nishijima, M., Hanada, K., Mostov, K., and Engel, J.N. (2000). Host cell-derived sphingolipids are required for the intracellular growth of *Chlamydia trachomatis*. Cell Microbiol *2*, 627-637.

Wallis, J., Moore, R., Smith, P., and Walsh, F.S. (1996). Cadherins: A review of structure and function. In Biomembranes: A Multi-Volume Treatise, A.G. Lee, ed. (JAI), pp. 127-157.

Wang, J., Dai, X., Berry, L.D., Cogan, J.D., Liu, Q., and Shyr, Y. (2018). HACER: an atlas of human active enhancers to interpret regulatory variants. Nucleic Acids Research *47*, D106-D112.

Wasylyk, C., Schlumberger, S.E., Criqui-Filipe, P., and Wasylyk, B. (2002). Sp100 interacts with ETS-1 and stimulates its transcriptional activity. Mol Cell Biol *22*, 2687-2702.

Williams, D.M., Grubbs, B.G., Park-Snyder, S., Rank, R.G., and Bonewald, L.F. (1996). Activation of latent transforming growth factor beta during *Chlamydia trachomatis*-induced murine pneumonia. Research in microbiology *147*, 251-262.

Wu, X., Lei, L., Gong, S., Chen, D., Flores, R., and Zhong, G. (2011). The chlamydial periplasmic stress response serine protease cHtrA is secreted into host cell cytosol. BMC microbiology *11*, 87-87.

Wysoker, A., Tibbetts, K., and Fennell, T. (2017). Picard tools. http://picardsourceforgenet.

Xiang, M., Zhang, W., Wen, H., Mo, L., Zhao, Y., and Zhan, Y. (2019). Comparative transcriptome analysis of human conjunctiva between normal and conjunctivochalasis persons by RNA sequencing. Experimental eye research *184*, 38-47.

Xu, J., Zhang, L., Ye, Y., Shan, Y., Wan, C., Wang, J., Pei, D., Shu, X., and Liu, J. (2017). SNX16 Regulates the Recycling of E-Cadherin through a Unique Mechanism of Coordinated Membrane and Cargo Binding. Structure (London, England : 1993) *25*, 1251-1263.e1255.

Yao, J., Cherian, P.T., Frank, M.W., and Rock, C.O. (2015a). *Chlamydia trachomatis* Relies on Autonomous Phospholipid Synthesis for Membrane Biogenesis. The Journal of biological chemistry *290*, 18874-18888.

Yao, J., Dodson, V.J., Frank, M.W., and Rock, C.O. (2015b). *Chlamydia trachomatis* Scavenges Host Fatty Acids for Phospholipid Synthesis via an Acyl-Acyl Carrier Protein Synthetase. The Journal of biological chemistry *290*, 22163-22173.

Yordy, J.S., Li, R., Sementchenko, V.I., Pei, H., Muise-Helmericks, R.C., and Watson, D.K. (2004). SP100 expression modulates ETS1 transcriptional activity and inhibits cell invasion. Oncogene *23*, 6654-6665.

Zadora, P.K., Chumduri, C., Imami, K., Berger, H., Mi, Y., Selbach, M., Meyer, T.F., and Gurumurthy, R.K. (2019). Integrated Phosphoproteome and Transcriptome Analysis Reveals *Chlamydia*-Induced Epithelial-to-Mesenchymal Transition in Host Cells. Cell Rep *26*, 1286-1302.e1288.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biology *9*, R137.

Zhou, Y., and Zhu, Y. (2015). Diversity of bacterial manipulation of the host ubiquitin pathways. Cellular microbiology *17*, 26-34.

Ziklo, N., Huston, W.M., Taing, K., and Timms, P. (2019). High expression of IDO1 and TGF-β1 during recurrence and post infection clearance with *Chlamydia trachomatis*, are independent of host IFN-γ response. BMC Infectious Diseases *19*, 218.

Chapter 4

Early transcriptional landscapes of *Chlamydia trachomatis*-infected epithelial cells at single-cell resolution

4.1. Abstract

Chlamydia are Gram-negative obligate intracellular bacterial pathogens responsible for a variety of disease in humans and animals worldwide. *C. trachomatis* causes trachoma in disadvantaged populations, and is the most common bacterial sexually transmitted infection in humans, causing reproductive tract disease. Antibiotic therapy successfully treats diagnosed chlamydial infections; however asymptomatic infections are common. High-throughput transcriptomic approaches have explored chlamydial gene expression and infected host cell gene expression. However, these were performed on large cell populations, averaging gene expression profiles across all cells sampled and potentially obscuring biologically relevant subsets of cells. We generated a pilot dataset, applying single cell RNA-seq (scRNA-seq) to *C. trachomatis* infected and mock-infected epithelial cells to assess the utility, pitfalls and challenges of single cell approaches applied to chlamydial biology, and to potentially identify early host cell biomarkers of chlamydial infection. 264 time-matched *C. trachomatis*-infected and mock-infected HEp-2 cells were collected and subjected to scRNA-seq. After quality control, 200 cells were retained for analysis. Two distinct clusters distinguished 3-hour cells from 6- and 12-hours. Pseudotime analysis identified a possible infection-specific cellular trajectory for *Chlamydia*-infected cells, while differential expression analyses found temporal expression of metallothioneins and genes involved with cell cycle regulation, innate immune responses, cytoskeletal components, lipid biosynthesis and cellular stress. We find that changes to the host cell transcriptome at early times of *C. trachomatis* infection are readily discernible by scRNA-seq, supporting the utility of single cell approaches to identify host cell biomarkers of chlamydial infection, and to further deconvolute the complex host response to infection.

4.2. Introduction

Chlamydia are Gram-negative obligate intracellular bacterial pathogens that cause disease in humans and a wide variety of animals. In humans, *Chlamydia trachomatis* typically infects cells within the ocular and genital mucosa, causing the most prevalent bacterial sexually transmitted infections (STI) (Reyburn, 2016), inducing acute and chronic reproductive tract diseases that impact all socioeconomic groups, and trachoma in disadvantaged populations (Burton and Mabey, 2009). Disease outcomes arise from complex inflammatory cascades and immune-mediated host processes that can lead to tissue damage and fibrotic scarring in the upper genital tract or the conjunctiva (Menon et al., 2015; Taylor et al., 2014). Reproductive tract disease outcomes include pelvic inflammatory disease (PID), preterm delivery, ectopic pregnancy, hydrosalpinx, tubal factor infertility (TFI) and chronic pelvic pain in women, as well as epididymitis, testicular pain and infertility in men. Antibiotic therapy with azithromycin or doxycycline successfully treats diagnosed infections, however asymptomatic infections are common (Ali et al., 2015; Hafner et al., 2014). Without overt symptoms that lead individuals to seek primary health care, antibiotic interventions are not able to be employed. Asymptomatic infection rates are estimated to exceed diagnosed infection rates by at least 4.3-fold (Ali et al., 2015).

Chlamydia have a unique biphasic developmental cycle with distinct morphological forms. The cycle begins with attachment and entry of the infectious elementary bodies (EBs) into host cells, typically mucosal epithelial cells. After entry, EBs reside within membrane-bound vacuoles that escape phagolysomal fusion (Scidmore et al., 1996). Differentiation into the replicating reticulate bodies (RBs) occurs within the first 2-3 hours, followed by continued growth of the inclusion accommodating the increased number of RBs. Over the course of infection, *Chlamydia* parasitises and modifies the host cell by deploying type III effectors and other secreted proteins (Valdivia, 2008), which also facilitate invasion, internalisation, and

replication, while countering host cell defences (Bastidas et al., 2013; Saka et al., 2011). At the end of the developmental cycle, RBs asynchronously transition back into EBs (~20-44 hours) and, through either extrusion or host cell lysis (~48-70 hours), are released to repeat the cycle (Elwell et al., 2016).

Chlamydial transcriptomes have been examined over the developmental cycle, in EBs and RBs, in different chlamydial species (Abdelrahman et al., 2011; Albrecht et al., 2011; Albrecht et al., 2010; Belland et al., 2003). Epithelial cell transcriptomes responding to plasmid-bearing/plasmid-less *C. trachomatis* has been characterized by microarray (Porcella et al., 2015). Dual RNA-seq (Humphrys et al., 2013; Marsh et al., 2018) has allowed the transcriptomes of both *C. trachomatis* and infected epithelial cells to be profiled simultaneously, identifying previously unrecognised early chlamydial gene expression and complex host cell responses (Humphrys et al., 2013). However in these studies to date, the derived transcriptional profiles represent averaged gene expression over the population of cells sampled (Hebenstreit, 2012). Subsets of cells with dominant gene expression profiles can skew the analysis (Łabaj et al., 2011), possibly obscuring other potentially important cell subsets and their transcriptional profiles (Liu and Trapnell, 2016; Saliba et al., 2014). By examining the expression profiles of individual cells, single cell RNA sequencing (scRNA-seq) can minimise these biases, enabling a deeper understanding of population heterogeneity, cell states and interactions, and gene regulation (Kolodziejczyk et al., 2015; Regev et al., 2017). scRNA-seq and other single cell methods have been instrumental in discovering new cell types (Regev et al., 2017) and advancing the understanding of many disease states (Sandberg, 2013), particularly tumour heterogeneity (Patel et al., 2014; Tirosh et al., 2016), hematopoiesis (Kowalczyk et al., 2015) and embryonic development (Yan et al., 2013). Applications of scRNA-seq to pathogen-infected cells are more limited so far, but are exemplified by studies that show the heterogeneity of macrophage responses to *Salmonella*

enterica serovar Typhimurium infection (Saliba et al., 2016), the high degree of cell-cell transcriptional variation induced by influenza virus infection (Russell et al., 2018), and the characterization of lymph node-derived innate responses to bacterial, helminth and fungal pathogens (Blecher-Gonen et al., 2019).

Here we explore the application of single cell analysis methodologies to *Chlamydia*-infected cells, with the goals of identifying host cell developmental-stage biomarkers, and to assess the utility of these methodologies for deciphering chlamydial biology in cells and tissues. We generated a pilot scRNA-seq dataset of time-matched infected and mock-infected HEp-2 epithelial cells *in vitro* encompassing the early chlamydial developmental cycle (3, 6 and 12 hours). We show that infection responsive changes to the early host cell transcriptome are readily discernible by scRNA-seq, supporting the potential for host derived infection biomarkers.

4.3. Results

4.3.1. Single cell capture, library construction, quality assessment and filtering

Chlamydia-infected (*C. trachomatis* serovar E*,* MOI~1) and time-matched mock-infected cells spanning three times post-infection were captured using the Fluidigm C1 microfluidic instrument and workflows (**Figure 4.1A**). We obtained 80 single cells at 3 hours (48 *Chlamydia*-infected, 32 mock-infected), 96 cells at 6 hours (48 *Chlamydia*-infected, 48 mock-infected) and 88 cells at 12 hours (40 *Chlamydia*-infected, 48 mock-infected) for an initial total of 264 cells (**Figure 4.1B**). Following Illumina library construction and sequencing, the raw sequencing reads from these 264 cells were demultiplexed using DeML (Renaud et al., 2015), yielding 1.03 billion sequence reads (**Supplementary Figure 1**). Single cell datasets

were removed from subsequent analyses if they contained less than 1 million reads after trimming and alignment, and less than 5,000 counted features (genes). Further quality assessment measures ensured that sequence reads mapped across all chromosomes and that the majority of reads mapped to protein-coding genes (**Supplementary Figure 2**).

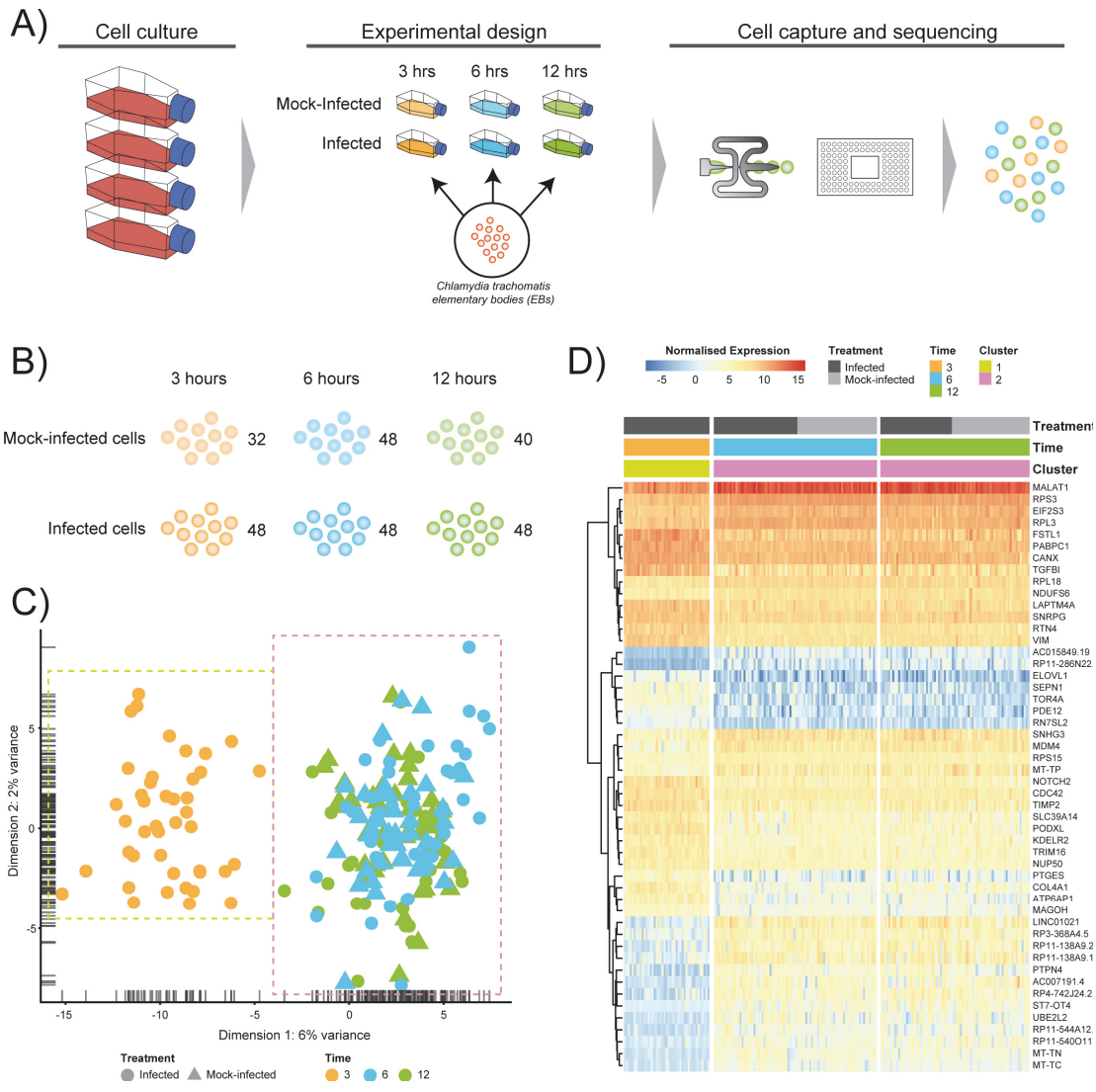


Figure 4.1: Experimental design and analysis

A) Cell culture using HEp2 epithelial cell monolayers used to grow and harvest *Chlamydia trachomatis* E elementary bodies (EBs). Fresh monolayers were infected with EBs, (MOI ~1) using centrifugation to synchronize infections. Experimental design time-matched *Chlamydia*-infected and mock-infected cell monolayers at 3, 6 and 12 hours, prior to capture and scRNA-seq library preparation on the Fluidigm C1 instrument. **B)** Numbers of captured and sequenced single cells by experimental condition and time. **C)** After quality control steps, unsupervised clustering identifies two primary clusters. Cluster 1 contains all 3 hour cells, while cluster 2 contains all 6 and 12 hour cells. **D)** Putative marker genes grouped by hierarchical clustering.

Single cell datasets were pooled and subjected to additional quality assessment steps, including examining rRNA as a measure of depletion success and mitochondrial gene expression as an indicator of cell stress (Zhao et al., 2002), as both are potential sources of bias (**Supplementary Figure 3**). During quality control, the mock-infected cells at 3 hours failed to pass cut-offs, and were excluded from further downstream analysis (**Supplementary Figure 3**). After all quality measures, datasets from 200 high quality single cells remained across the three times: 43 *Chlamydia*-infected cells at 3 hours; 82 6 hour cells (42 *Chlamydia*-infected, 40 mock-infected) and 75 12 hour cells (36 *Chlamydia*-infected, 39 mock-infected).

4.3.2. Removal of confounding effects

To normalise by library size, Scran's single-cell specific method was used to deconvolute library size factors from cell clusters (Lun et al., 2016). We applied RUVSeq (Risso et al., 2014) to identify and remove further confounding effects, including differences between batches of sequenced cells. Reduction of variation was confirmed in relative log expression

(RLE) plots (**Supplementary Figure 4A**). Density curve plots further show the effect of removing variability from the raw counts, after library size normalisation, and after removing further confounding effects (**Supplementary Figure 4B**). The PCA bi-plot (**Supplementary Figure 4C**) shows the structure of the data and grouping of the cells based on their transcriptional profiles following these steps. By examining the underlying variables driving PC1 variation, we found that total read counts and time post infection account for 99% of the total variation (**Supplementary Figure 4C**), confirming that most variation is not from experimental factors. In addition, doublets (where at least two cells are captured into the same well) can skew the resulting expression profiles, adding a further confounding factor. Although the C1 platform uses integrated fluidic circuits (IFCs) to isolate single cells, it has been associated with a doublet rate as high as 25% (Wang et al., 2019). Due to this high reported rate, we ran different tools to identify doublets, confirming that our data had minimal detected doublets (**Supplementary Figure 5**).

4.3.3. Cell cycle classification

Due to the constraints imposed by chlamydial infection within *in vitro* tissue culture and, given the potential for cell-cell variability despite infection synchronization, we expected to observe a range of cell cycle stages in our data (**Figure 4.2**). Two of the three stages (G1 and G2/M) show more than double the number of cells from 3 to 6 hours, while DNA synthesis (S) is the only cell cycle stage with a decrease in the number of cells from 3 to 12 hours. However, despite these trends, no distinct cell cycle clusters are apparent (**Figure 4.2B)**. In addition, there is no clustering between cell cycle state and time post-infection, or infection condition (infected *vs* mock-infected). Although we identify cell cycle stage as a likely confounding effect (Barron and Li, 2016) that was removed from our subsequent analyses, it may be relevant to the infection and growth strategies of *Chlamydia*. For example, while

infected cells can still grow and divide, the burden of infection causes these cells to proliferate more slowly than uninfected cells, resulting in dividing cells which may be more or less susceptible to infection (Balsara et al., 2006). Additionally, chlamydial infectivity has been related to distinct cell cycle phases, where infection can modulate cell cycle parameters (Johnson et al., 2009).

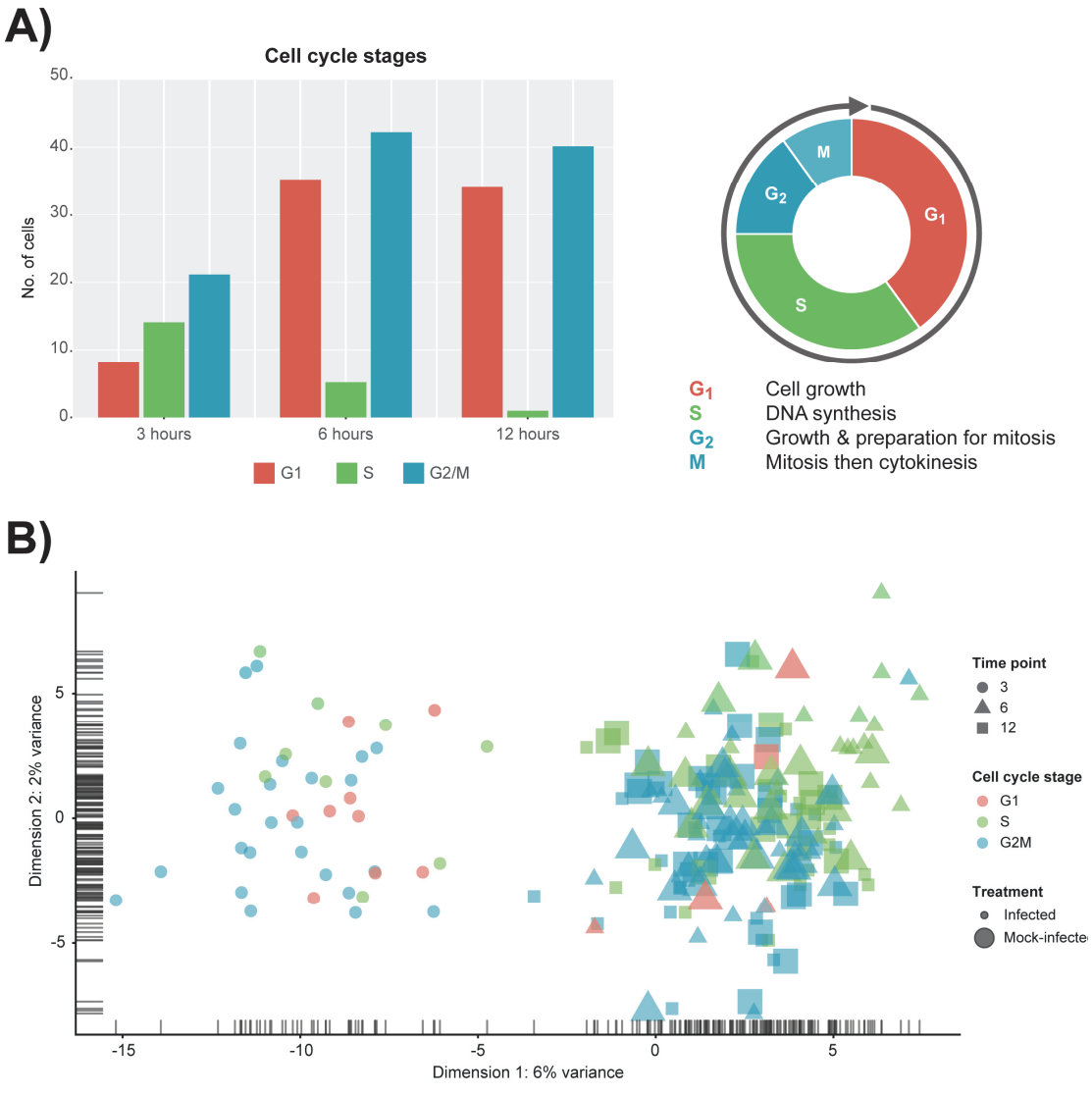


Figure 4.2: Cell cycle classification

A) Cell cycle classification of single cells after removing outliers. **B)** PCA plot examining cell-cycle related trends by time-point and infection status.

4.3.4. Clustering demonstrates transcriptional heterogeneity of infected epithelial cells over the early chlamydial developmental cycle

Unsupervised clustering identified two distinct clusters across the three time points (**Figure 4.1C** and **Supplementary Figure 6**). Cluster 1 contains only 3-hour infected cells, while cluster 2 contains a mixture of cells from 6 and 12 hours, with no clear separation between infected and uninfected conditions. We used k-nearest neighbour smoothing (kNN-smoothing) to further reduce scRNA-seq-specific noise within the expression matrix (Wagner et al., 2018), which is a common occurrence from effects such as dropouts (Gong et al., 2018). The resulting PCA plot recapitulated the clusters identified above, indicating that the previous clustering result was not influenced by noise-related factors. Additional clustering analyses were performed to identify any sub-populations within each cluster on the basis of experimental factors such as time or infection status (**Supplementary Figure 7**). *Chlamydia*-infected cells again clustered into two main groups, closely matching the overall clustering that separated the 3 hours cells from 6 and 12 hours, with no further sub-clustering evident.

4.3.5. Pseudotime analysis over the early chlamydial developmental cycle

Unsupervised clustering demonstrates that both infected and uninfected cells have minimal cluster separation at 6 and 12 hours **(Figure 4.1C)**. We applied pseudotime analysis to further deconvolute cellular trajectories that may follow a time course or biological mechanisms such

as differentiation or infection (Ji and Ji, 2016; Lönnberg et al., 2017). Pseudotime analysis of *Chlamydia*-infected cells alone predicted 3 distinct cell states (**Figure 4.3A**). Cell state 1 contained 3 hour cells, state 2 contained a mixture of 3, 6 and 12 hour cells, and state 3 contains a mixture of 6 and 12 hour cells (**Figure 4.3B**). The line connecting the 3 cell states (minimum spanning tree) does not provide a realistic linear trajectory of the infection course from 3-12 hours.

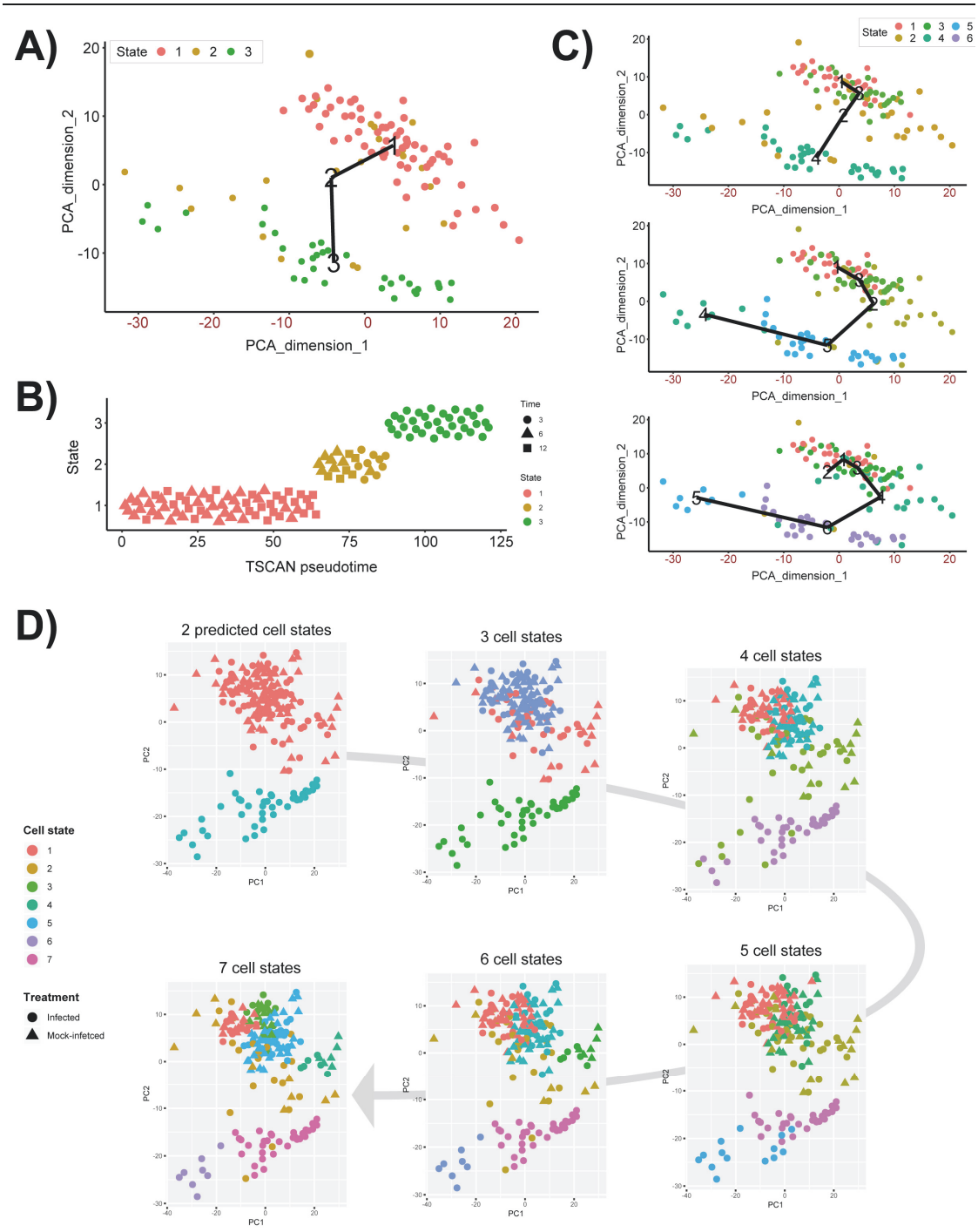


Figure 4.3: Pseudotime analysis

A) Pseudotime analysis of infected cells predicts three cell states. The minimum spanning tree (black line) is uninformative and not a true indication of an expected infection trajectory encompassing all cells from 3 to 12 hours. **B)** Each cell ordered throughout the

predicted pseudotime and separated by cell state. **C)** Manually increasing the number of cell states to six appears to show a more realistic infection trajectory with a wider number of cells, in addition to showing start and end points correlating to 3 and 12 hour cells. **D)** When all cells are used, two cell states are predicted that support the initial clustering outcomes. When the number of cell states is manually increased, smaller subsets appear, providing a finer resolution.

Manually increasing the number of states does provide a more realistic trajectory (**Figure 4.3C**); however, the similarity of 6 and 12 hour cells (both mock-infected and infected states) will require more cells to accurately capture any putative sub-stages of infection.

We overlaid the predicted cell cycle states from pseudotime analysis for each cell to identify any shared characteristics of infected cells with mock-infected cells, which could classify cells that were either not infected or had unproductive infections. This analysis identified only two cell states (**Figure 4.3A-C**), recapitulating the initial clustering results. By manually increasing the number of cell states to 7, smaller sub-clusters within the two main clusters became evident (**Figure 4.3D**). However, we still observe a mixture of infected and mock-infected cells within each sub-cluster (albeit with small numbers of cells), further highlighting the transcriptional similarity between infected cells at 6 and 12 hours in this dataset.

4.3.6. Differentially expressed genes in *Chlamydia*-infected and mock-infected cells highlight infection mechanisms

Subsets of genes with significant expression differences between the two primary clusters were examined in order to identify any putative host cell marker genes that distinguish different times post-infection (**Figure 4.1D**). At 6 and 12 hours, genes associated with processes governing RNA and protein metabolism (*RPS3*, *RPL3*, *RPS15*, *RPL18*, *PABPC1*, *MAGOH* and *EIF2S3*) predominate, with most showing increased expression. Increased expression of vimentin (*VIM*), a type III intermediate filament (IF) present in the cytoskeleton and involved in maintaining cell shape and integrity (Mak and Brüggemann, 2016), distinguishes the 3 hour cluster.

We further examined differentially expressed (DE) genes firstly by comparing infected and mock-infected cells at 6 and 12 hours respectively (cluster 2), and secondly by comparing the 3 hour infected cells (cluster 1) against cluster 2, as the 3 hour mock-infected cells were removed after initial quality control steps. At 6 hours, 44 DE genes were identified (13 up-regulated and 21 down-regulated) (**Figure 4.4A**), including three up-regulated metallothionein (MT) genes (*MT1E*, *MT2A* and *MT1X*). MT up-regulation occurs in response to intracellular zinc concentration increases, reactive oxygen species (ROS) and proinflammatory cytokines (Rice et al., 2016). Intracellular zinc concentrations are an integral component of immunity and inflammation, and zinc deficiency results in an increased susceptibility to infection (Subramanian Vignesh and Deepe, 2017). MTs may also have a role in protecting against DNA damage and in apoptosis, as well as regulating gene expression during the cell cycle (Cherian and Apostolova, 2000), which are likely to be relevant at 6 hours post infection. Down-regulated pathways at 6 hours were dominated by three genes *HSP90AA1* (Heat Shock Protein 90 Alpha Family Class A Member 1), *TUBB* (Tubulin Beta

Class I), and *TUBA4A* (Tubulin Alpha 4a), which are linked to the cell cycle, specifically centrosome maturation and microtubule assembly mediating mitosis (**Figure 4.4B**).

At 12 hours, there is an increase in DE genes (245) with 98 up-regulated and 147 down-regulated (**Figure 4.4A**). We continue to see up-regulated genes that are likely part of a continued immune response to infection, including two MTs (*MT1M* and *MT1E*), *TRIM25* (Tripartite Motif Containing 25), *ISG15* (*ISG15* Ubiquitin Like Modifier), *HLA-A* (Major Histocompatibility Complex, Class II, DR Beta 1), *IFIT3* (Interferon Induced Protein With Tetratricopeptide Repeats 3), *OASL* (2'-5'-Oligoadenylate Synthetase Like), *IL6* (Interleukin 6), and genes associated with cholesterol and fatty acid synthesis (**Figure 4.4B**).

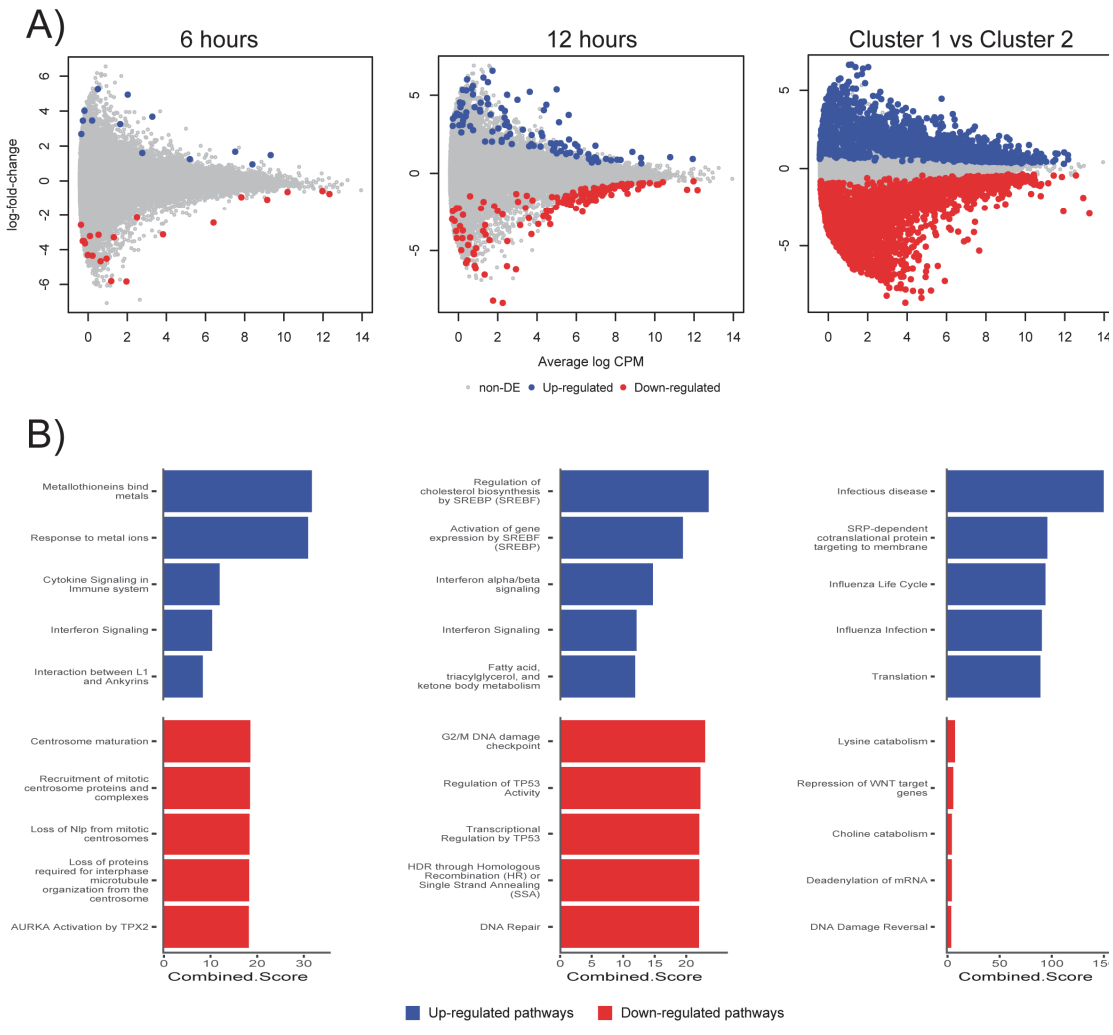


Figure 4.4: Differentially expressed genes and enriched pathways

A) Differentially expressed genes from infected and mock-infected cells at 6 hours and 12 hours. When comparing cells from cluster 1 against cluster 2, a more complex experimental design was needed that took into consideration the variety of underlying cells. **B)** Enriched pathways from Reactome using differentially expressed genes from **A**.

The exploitation of a variety of host lipids by *Chlamydia* to subvert intracellular signalling, survival and growth is well established (Cocchiaro et al., 2008; Elwell and Engel, 2012; Kumar and Valdivia, 2008). All down-regulated pathways at 12 hours indicate that *Chlamydia*-infected cells are exhibiting stress responses. DNA damage as part of the cell cycle, and repair pathways are enriched, possibly representing a continuation of infection stresses at 6 hours, and likely indicative of further *Chlamydia*-induced interruption of the cell cycle. Notably, two p53 associated pathways were enriched from associated genes. p53 expression tightly controls the cell cycle and is modulated in response to activities including cell stress, DNA damage, as well as bacterial infection (Zaika et al., 2015). *Chlamydia*-induced down-regulation of p53 may help to protect infected cells against death-inducing host responses, thus allowing chlamydial survival (Siegl et al., 2014). Only four DE genes from 6 and 12 hours overlap. The two up-regulated genes were *DUSP5* (Dual Specificity Phosphatase 5) and *MT1E* (metallothionein 1E), while the two down-regulated genes were *TUBA4A* and *HSP90AA1*.

Comparing cluster 1 (3-hour infected cells) against cluster 2 (DE genes from 6 and 12 hours) demonstrates a substantial number of DE genes (2,291 up and 3,487 down-regulated) (**Figure 4.4A**). Although the model attempted to account for the loss of the 3 hour mock-infected cells, we note a proportion of these DE genes may not be related to a productive chlamydial infection as a result. The down-regulated pathways have low combined scores (a combination

of p-values and z-scores) compared to the up-regulated pathways, which may be explained by the large number of down-regulated non-coding RNAs (ncRNAs), which are typically not incorporated into the underlying enrichment analyses. Three of the up-regulated pathways are associated with infection (*Infectious disease*, *Influenza life cycle* and *Influenza infection*) (**Figure 4.4B**), demonstrating that general infection mechanisms are the key differences between these temporally defined clusters.

4.4. Methods

4.4.1. Cell culture and infection

HEp-2 cells (American Type Culture Collection, ATCC No. CCL-23) were grown as monolayers until 90% confluent. Monolayers were infected with *C. trachomatis* serovar E in SPG as previously described (Tan et al., 2009). Additional monolayers were mock-infected with SPG only. The infection was allowed to proceed 48 hours prior to EB harvest, as previously described (Tan et al., 2009). *C. trachomatis* EBs and mock-infected cell lysates were subsequently used to infect fresh HEp-2 monolayers. Fresh monolayers were infected with *C. trachomatis* serovar E in 3.5 mL SPG buffer for an MOI ~ 1 as previously described (Tan et al., 2009), using centrifugation to synchronize infections. Infections and subsequent culture were performed in the absence of cycloheximide or DEAE dextran. A matching number of HEp-2 monolayers were also mock-infected and synchronised as above using uninfected cell lysates. Each treatment was incubated at 25°C for 2h and subsequently washed twice with SPG to remove dead or non-viable EBs. 10 mL fresh medium (DMEM + 10% FBS, 25 μg/ml gentamycin, 1.25 μg/ml Fungizone) was added and cell monolayers incubated at 37°C with 5% CO_2. Three biological replicates of infected and mock-infected dishes per

time were harvested post-infection into single cell populations by trypsin in sterile PBS prior to immediate single cell capture and library preparation.

4.4.2. Library preparation and sequencing

A Fluidigm C1 instrument was used for cell capture. This instrument uses microfluidics on IFCs to capture single cells, lyse and prepare cDNA, using 96 well plates as input. Only polyadenylated fragments are captured from each cell, typically restricting analysis to eukaryotic mRNA. Cell lysis, reverse transcription, and cDNA amplification were performed using the C1 Single-Cell Auto Prep IFC. The SMARTer Ultra Low RNA Kit (Clontech) was used for cDNA synthesis and Illumina NGS libraries were constructed using the Nextera XT DNA Sample Prep kit. The resulting 264 single cell libraries were sequenced using the Illumina HiSeq 4000 platform (150bp paired-end reads) across three batches. Each plate was designed with a balanced distribution of time points and conditions.

4.4.3. Pre-processing and quality control

Raw sequencing reads were demultiplexed using DeML (Renaud et al., 2015) with default settings. Trim Galore (v.0.4.3) (Krueger, 2012) was used to trim adaptors and low quality reads. Confirmation of the removal of adaptors, low quality reads and other quality control measurements was performed with FastQC (v.0.11.5) (Andrews, 2010). Reads were subsequently aligned to the human genome version (GRCh 38.87) with STAR (v.2.5.1a) (Dobin et al., 2013) retaining paired and unpaired mapped reads that were merged in to a single BAM file.

FastQ-Screen (v.0.11.1) (Wingett and Andrews, 2018) was used to screen for sources of contamination across all cells. This output and low mapping rates confirmed the removal of all 3-hour uninfected cells, due to extremely low mapping rates to the Human genome and high mapping rates to other organisms. Features of the remaining cells were counted with FeatureCounts (v.1.5.0-p3) (Liao et al., 2014). MultiQC (v.1.0) (Ewels et al., 2016) was used throughout the previous steps, combining output from each piece of software to easily make comparisons between batches and across time points.

4.4.4. Identifying outlier cells based on filtering

Counted features were imported into Scater (v.1.5.11) (McCarthy et al., 2017), where subsequent quality control reduced the total number of cells from 264 to 200. The filter settings were comprised of four steps: 1) total mapped reads should be greater than 1,000,000; 2) total features greater than 5,000; 3) expression from mitochondrial genes less than 20% of total expression; and 4) expression from rRNA genes comprise less than 10%.

4.4.5. Removing confounding effects

Cell cycle classification was performed using Cyclone (Lun et al., 2016) prior to filtering out low abundance genes, as recommended. To account for the differences in library sizes between cells, the deconvolution method from Scran (v.1.6.0) (Lun et al., 2016) was used. Further confounding effects such as cell cycle and sequencing batch effects were removed using the RUVs method of RUVSeq (v.1.12.0) (Risso et al., 2014) using k=4. Doublet detection was performed using Scrublet (v.0.1) (Wolock et al., 2019) and DoubletDetection (v.2.4) (Gayoso and Shor, 2019).

4.4.6. Clustering

Unsupervised clustering was performed using the Single-Cell Consensus Clustering (SC3) package (v.1.10.1) (Kiselev et al., 2017). Two clusters (k) were chosen based on automatic prediction by SC3 after iterating through a range of k (2:10). Higher values of K were examined; however, two clusters remained the best fit to this data as assessed by various internal plots including consensus matrices, silhouette plots and cluster stability plots. To further confirm the two clusters, the KNN-Smoothing (Wagner et al., 2018) (K-nearest neighbour smoothing) function (v.1) was also applied to different transformations of the library normalised data.

4.4.7. Pseudotime analysis

TSCAN (v.1.16.0) (Ji and Ji, 2016) was used to perform pseudotime analysis. When all cells were analysed, the "pseudocount" and "minexpr_value" flags were set to 0.5 in pre-processing to allow more features to be selected, resulting in an increase of cells with assigned cell states, especially when the number of states was manually increased. The default pre-processing settings were used to examine infected cells alone.

4.4.8. Differential expression

Most scRNA-seq differential expression software only allows for direct comparisons, such as cluster comparisons. As our experimental design examined both infected and mock-infected cells and, due to the loss of the 3 hour mock cells following QC measures, we used edgeR (v.3.24.3) (Robinson et al., 2010), as it provides better functionality for more complex comparisons than most single-cell specific tools. Initial comparisons were between infected

and mock infected cells at 6 and 12 hours. The use of edgeR allowed the comparison between 3 hour infected cells (cluster 1) and the remaining cells (cluster 2), taking into consideration the differences between infected and mock infected cells from 6 and 12 hours. In addition to including the RUVSeq factors of unwanted variation to the model matrix, the dispersion trend was estimated using "locfit" and "mixed.df" flags set to true. Resulting p-values were adjusted using a false discovery rate (FDR) < 0.05 and separated based on their respective fold-changes. Significant genes were examined using enrichR (Chen et al., 2013), with enriched pathways from Reactome sorted by their combined scores (a calculation of p-values and z-scores) (Croft et al., 2011).

4.4.9. Availability of supporting data and materials

The data set supporting the results of this article is available in the GEO repository, GSE132525.

4.5. Discussion

To better understand bacterial pathogenesis and resulting disease outcomes, it is critical to understand functional changes to specific cell populations of infected and neighbouring cells, and recruited immune cells in the infected tissue context. This is especially relevant for *Chlamydia* which, due to its obligate intracellular niche and distinct morphologies, has long been refractory to research. As a result, many infection and disease processes at the cellular and tissue level remain largely unknown or poorly characterised *in vivo*. Gene expression profiling of *Chlamydia*-infected cells by microarray (Porcella et al., 2015), dual RNA-seq (Humphrys et al., 2013; Marsh et al., 2018) or other genome-scale analyses (Chapters 3 and 5) are powerful techniques to help deconvolute these interactions and processes. However, these and similar genome-scale analyses of infected cells have typically been performed on bulk cell populations, i.e.; infected cell monolayers *in vitro*, or selectively sorted/purified subsets of cell populations. Such bulk cell approaches can potentially miss cell-cell variability, or cells that contribute to overlapping phenotypic characteristics, potentially masking critical biological heterogeneity as irrelevant signals from non-participating cells that can skew the average. This may influence the understanding of multifactorial and dynamic processes, such as inflammation and fibrosis during ascending chlamydial infection. Single cell approaches can potentially alleviate some of these concerns, but also provide new challenges.

We describe the first application of scRNA-seq to *Chlamydia*-infected cells. This pilot dataset, comprising 264 single infected and mock-infected cells encompassing three early times of the *in vitro* chlamydial developmental cycle, was designed to examine the feasibility and pitfalls of single cell approaches to investigate chlamydial biology and to ultimately identify host-derived transcriptional biomarkers of chlamydial infection. After quality

assessment and filtering measures, we retained 200 high quality, *C. trachomatis*-infected and mock-infected cells.

We note that the experimental design used here will not distinguish *Chlamydia*-mediated effects from infection-specific or non-specific epithelial cell responses. In addition, the *in vitro* infections used as the source of the single cells are centrifugation-synchronized in order to minimise the degree of heterogeneity at each infection time to enable more accurate examination of temporal effects. Despite this, lag time of differentiating EBs to RBs between distinct cells may still influence host responses mediated by temporally expressed/secreted chlamydial factors. Given that the minimal chlamydial generation time during exponential growth has been estimated as 2.6 to 4.6 hours (Lambden et al., 2006; Wilson et al., 2004), it is plausible that cells at each time may cluster with an earlier or later time.

Clustering, pseudotime and cell state prediction analyses demonstrated that *Chlamydia*-treated cells at 3 hours are readily distinguishable from *Chlamydia*-treated and mock-infected cells at 6 and 12 hours. Curiously, cells at 6 and 12 hours clustered together and could not be further deconvoluted from each other, possibly showing that host cell transcription at these times is broadly similar. A recent FAIRE-seq analysis of *Chlamydia*-infected epithelial cells, examining patterns of host cell chromatin accessibility over the developmental cycle (Chapter 3), found that 12 hours post infection was relatively quiescent in terms of host cell transcriptional activity. This finding is reflected by our scRNA-seq analyses here and may be extended to 6 hours post-infection. In addition, both *Chlamydia*-infected and mock-infected cells at 6 and 12 hours clustered together. One interpretation of this phenomenon is that these early infection times represent a period where the ongoing establishment of the inclusion and chlamydial division after initial entry and infection events is largely cryptic to the host cell as manifested by transcriptional processes.

However, limitations inherent to the experimental design and the technology used here may also influence our results, and should inform future single cell experiments. We used an MOI~1, based upon our previous work with bulk dual RNA-seq (Humphrys et al., 2013), which typically results in highly infected HEp-2 monolayers (95%+) when using *C. trachomatis* serovar E. When combined with the closed nature of the integrated fluidic circuits used to capture individual cells, the early infection times, and the destructive nature of single cell RNA-seq, using a lower MOI may have led to populations of both infected and uninfected cells to be sampled. In bulk transcriptomic experiments, any distinct signal from a small number of uninfected cells will be largely overwhelmed. In contrast, uninfected cells in single cell experiments may have an outsized effect, particularly if the total number of cells sampled is insufficient. In addition, the population of infecting EBs may include differentially viable EBs, leading to a divergence of transcriptional profiles between cells with productive infections, compared to cells that are initially infected with non-viable EBs that will not proceed to a productive infection. Similarly, any putative "neighbour" effect of uninfected cells next to infected cells may lead to distinct transcriptional profiles. Given the relatively low number of single cells sampled and the early infection times examined, these factors are a potential source of bias in these pilot experiments. With these limitations in mind, it may be more accurate to describe the *Chlamydia*-infected cells in these experiments as "*Chlamydia*-exposed".

Design of *in vitro* single cell experiments in the *Chlamydia*-infected cell context will benefit firstly from a higher MOI to ensure maximal productive infection. Secondly, collection of much higher numbers of single cells for scRNA-seq and other single cell genome-scale measurements are now possible, minimising the potential for introduced sampling biases that arise from low infected cell numbers. Thirdly, the inclusion of additional controls, such as UV-inactivated EBs or opsonized latex beads, will allow host cell transcriptional responses

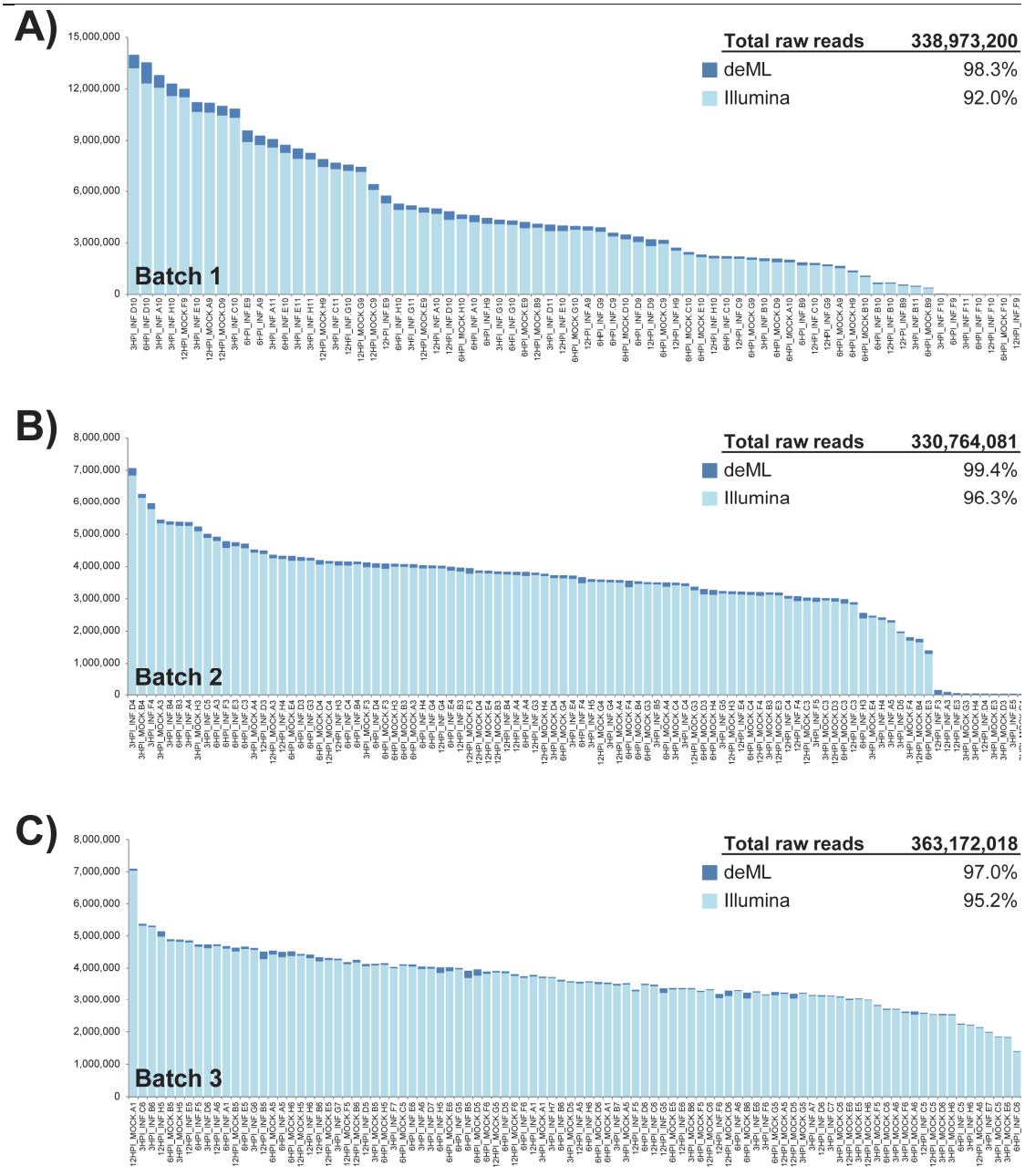
to productive versus non-productive infection events to be examined, as well as separating host cell infection-specific processes from non-specific phagocytic responses. Finally, moving away from poly(A) capture library construction to random hexamers instead will allow pathogen transcripts to simultaneously be interrogated in the single cell mode, as recently applied to *Salmonella typhimurium*-infected cells (Avital et al., 2017).

A range of cell cycle states were observed in our data. We attempted to remove these effects as potential confounders through bioinformatic means. *Chlamydia*-infected cells are still able to undergo mitosis, however mitosis-related defects do occur during chlamydial infection. These include an increase in supernumerary centrosomes, abnormal spindle poles, and chromosomal segregation defects, and result in a heavily burdened cell that proliferates more slowly (Grieshaber et al., 2006; Knowlton et al., 2011). Cells that recover from infection are still likely to contain chromosome instabilities, which can then be passed down to uninfected daughter cells (Grieshaber et al., 2006). This may be manifested in our data as we see a number of pathways related to the cell cycle that are down-regulated. While this could be an off-target effect of infection that does not benefit *Chlamydia*, interference with the cell cycle may constitute an infection strategy, as *in vivo* cells will be at different cell cycle stages and thus some may be more or less susceptible to infection. Nevertheless, future *in vitro* investigations of chlamydial infection should attempt to explore and/or mitigate these effects through cell cycle arrest strategies (Johnson et al., 2009) prior to infection and/or single cell separation.

Differential expression comparing cells between clusters identified both conserved and temporally specific gene expression over the times examined. Comparison of these differentially expressed genes with other published *Chlamydia*-infected cell transcriptomic datasets (Humphrys et al., 2013; Porcella et al., 2015) showed little overlap (data not shown), most likely as a consequence of an accumulation of technical differences that make direct

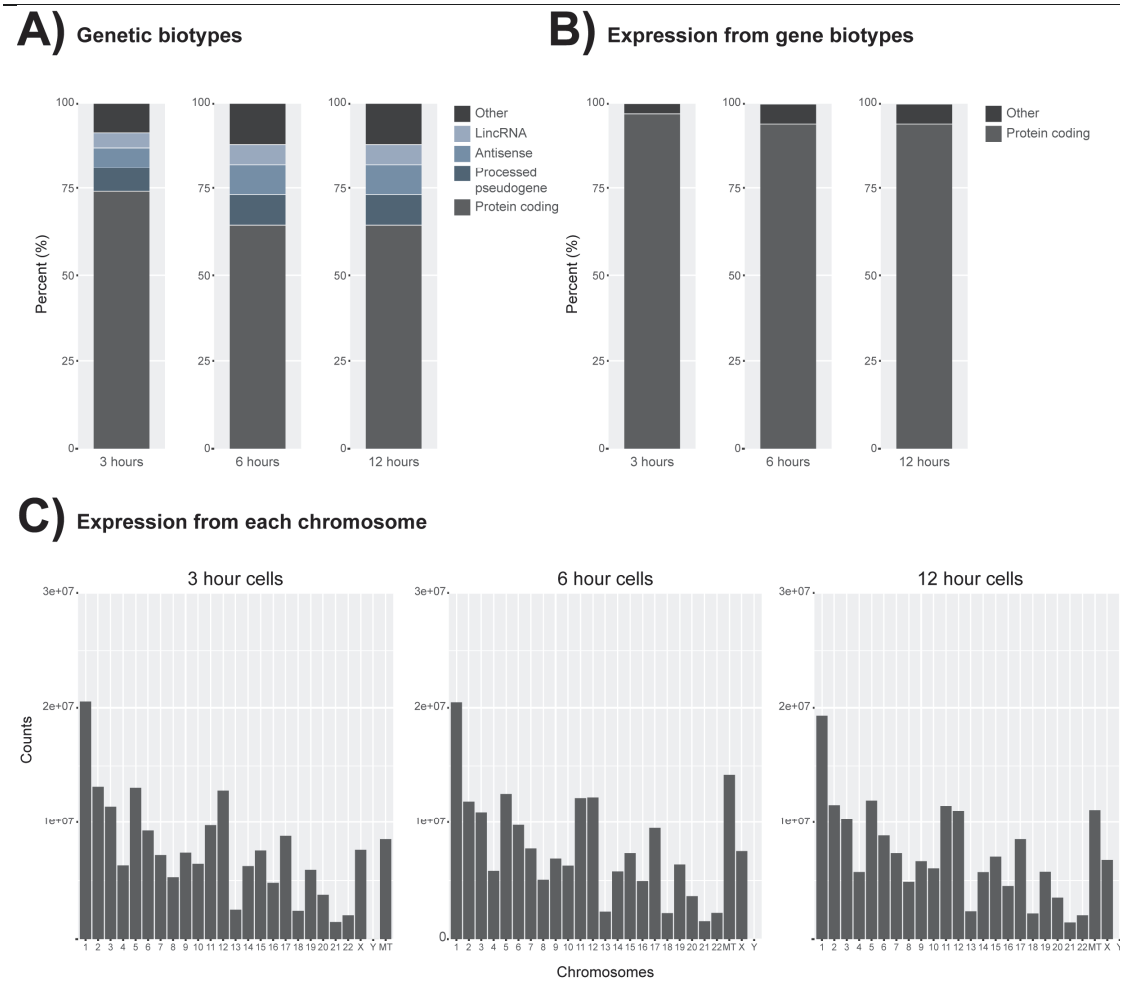
comparisons difficult, including the relatively small numbers of single cells sampled here, different times post-infection, different MOIs, and different chlamydial serovars and *in vitro* cell lines. Single cell RNA-seq approaches may also benefit from parallel bulk RNA-seq approaches from the same input material to cross-check, compare and validate. Nevertheless, many of the identified pathways and genes are directly relevant to known chlamydial infection processes, including metallothioneins, innate immune processes, cytoskeletal components, lipid biosynthesis and cellular stress. These analyses demonstrate that, despite the limitations of this pilot dataset, distinct host cell transcriptional responses to infection are readily discernible by single cell approaches, even at the early stages of the chlamydial developmental cycle, yielding robust data and confirming that host cell-derived transcriptional biomarkers of chlamydial infection are identifiable. Thus, single cell genome-scale approaches applied to *Chlamydia*-infected and neighbouring cells, recruited immune cells from inflammatory processes, and structural cells obtained from clinical swabs or *ex vivo* tissues, are likely to lend significant insight to the complex processes that underpin chlamydial infection and the associated inflammatory disease outcomes.

4.6. Supplementary figures



Supplementary Figure 1. Demuxing results comparing Illumina and DeML

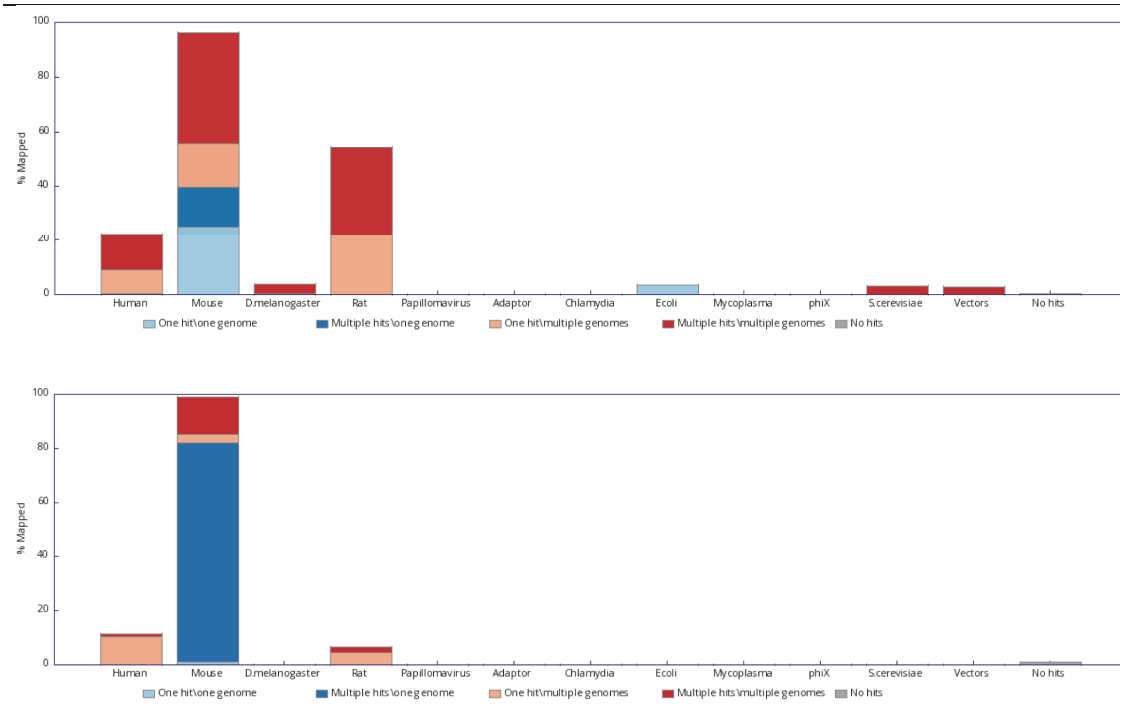
Demuxing comparison between the standard Illumina software and DeML. 1.8-5.5% additional reads were recovered over each cell batch **A-C**.



Supplementary Figure 2. Biotype distribution

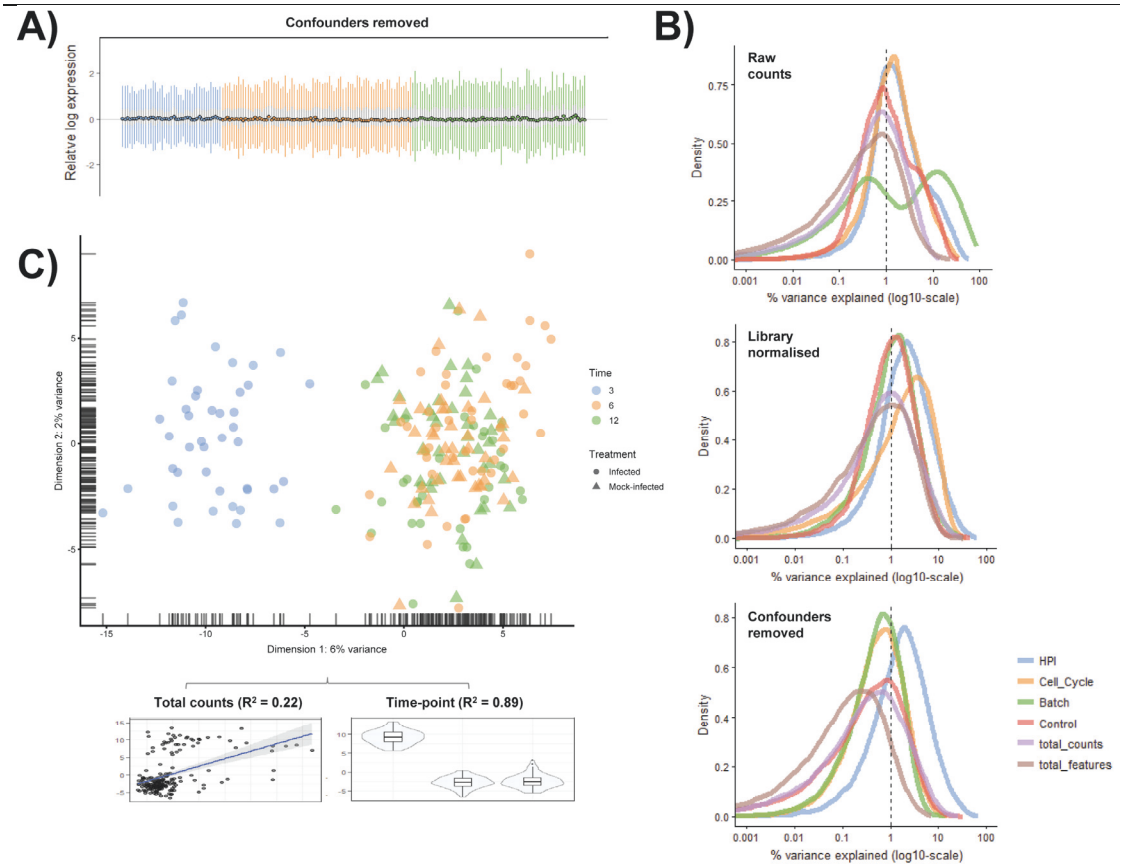
A) Distribution of different gene biotypes by time. **B)** Total gene expression by biotype.

C) Total gene expression by chromosome location.



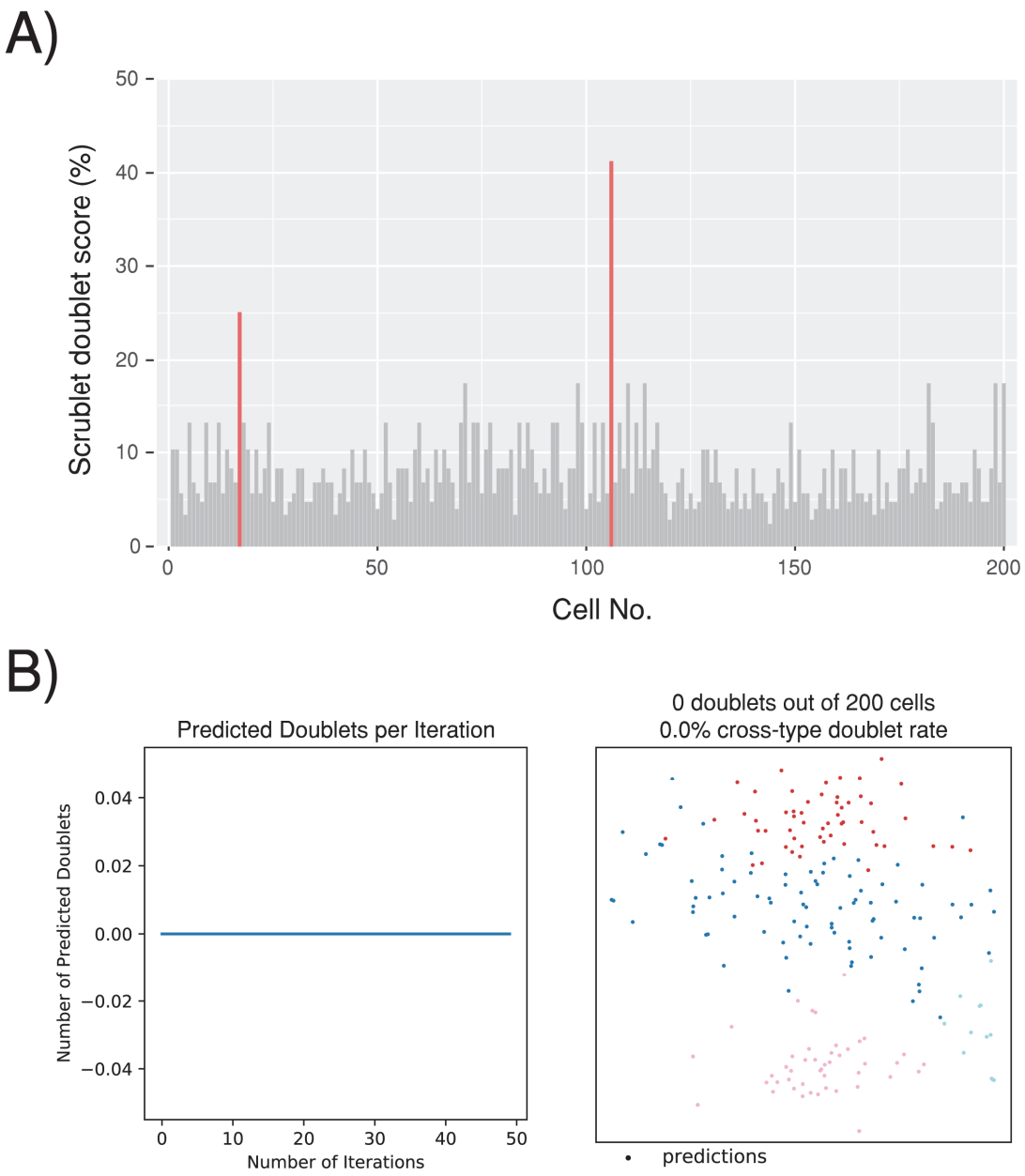
Supplementary Figure 3. Contamination of 3-hour mock-infected cells

Examples of two 3-hour mock-infected cells with unusual mapping to 12 different genomes, indicating cross-contamination. All 3 hour mock-infected cells showed similar profiles and were removed from further downstream analysis.



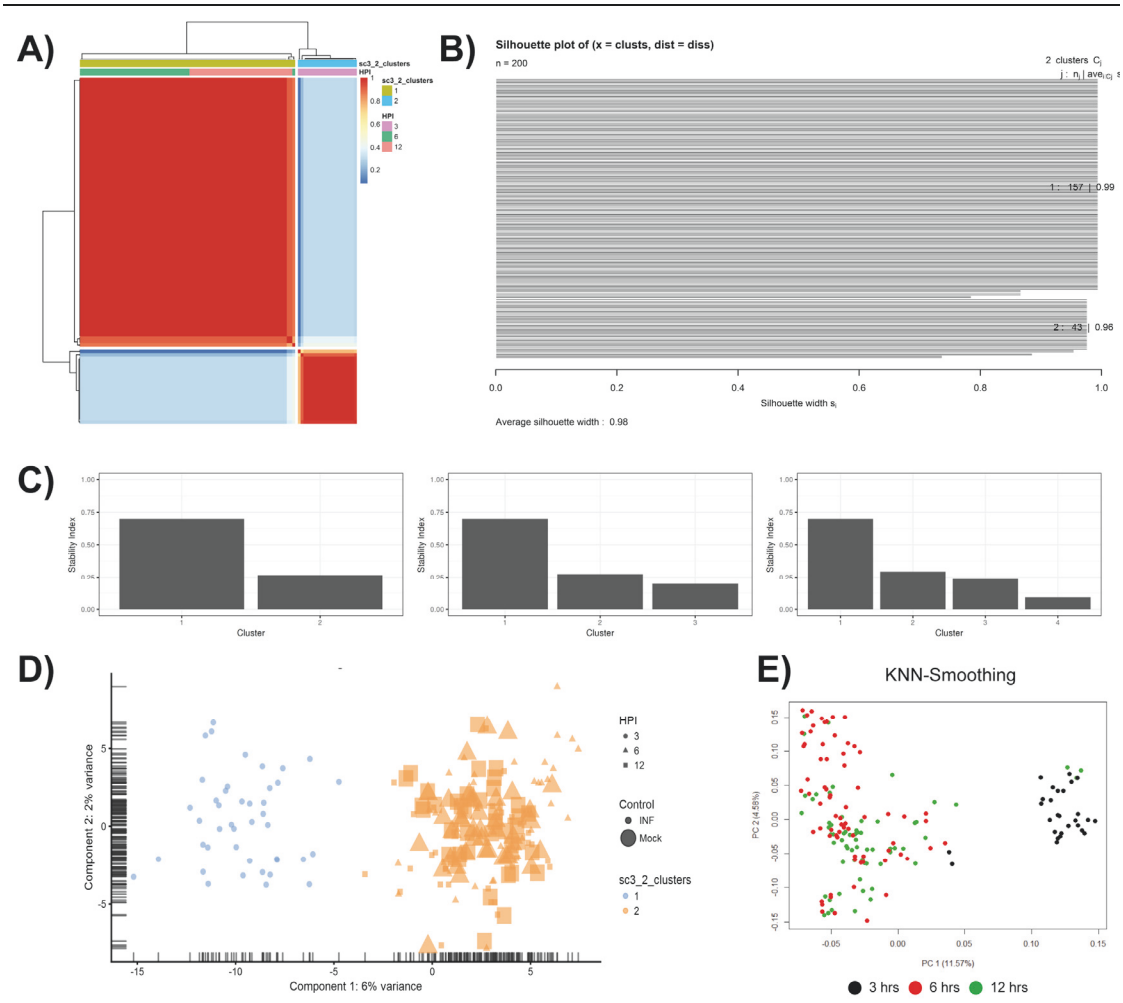
Supplementary Figure 4. Identifying and removing confounding effects

A) Relative log expression (RLE) plot of gene expression from all cells after removal of confounding effects using RUVSeq. **B)** Density curve distribution showing variability associated to key variables from raw counts, after library normalisation, and after using RUVSeq respectively. **C)** PCA plot demonstrating two-dimensional structure of cell expression profiles after removing confounding effects. Two variables (Total features and Time-point) account for 99% of the variability at component 1 (PC1).



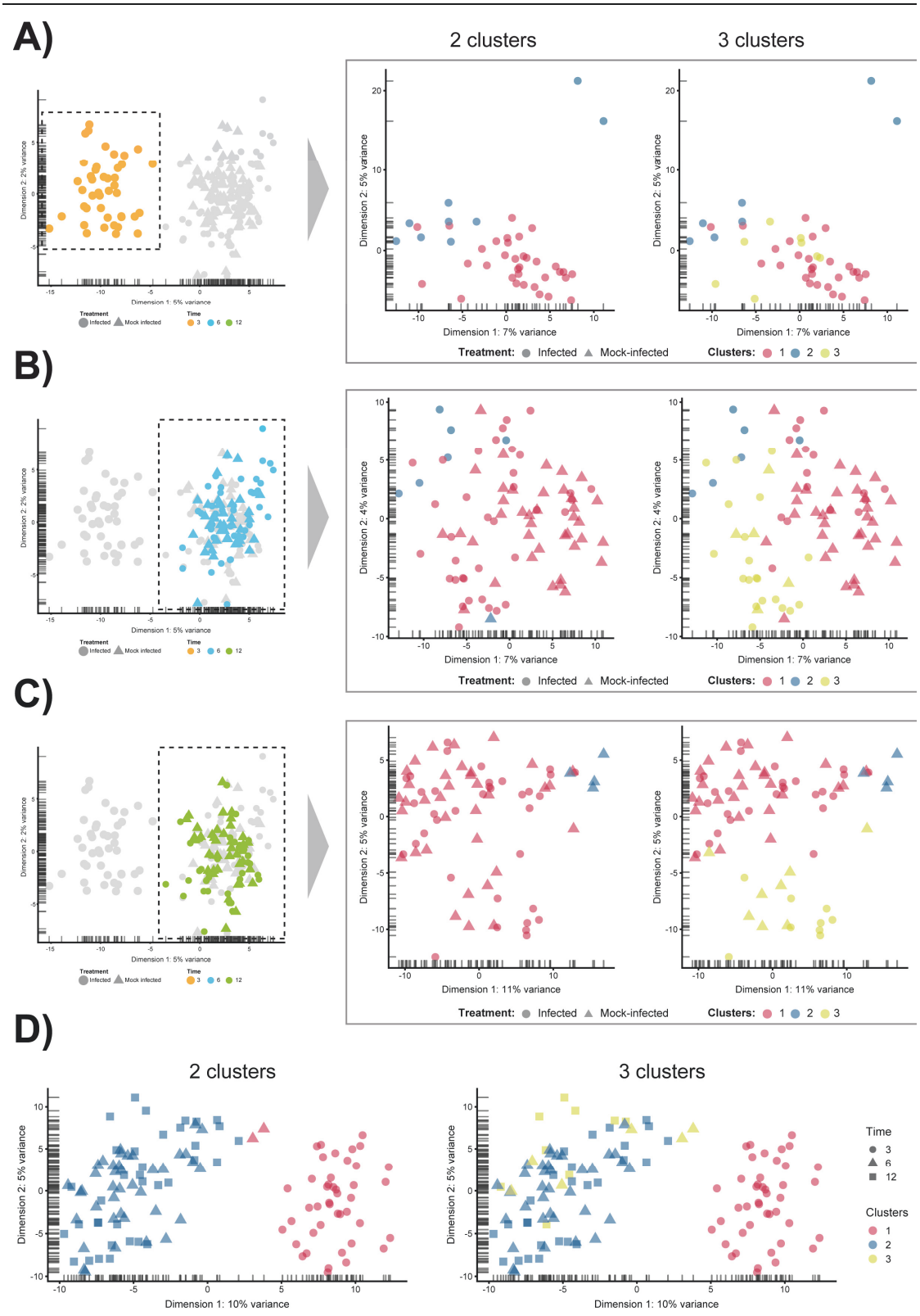
Supplementary Figure 5. Doublet detection

A) Scrublet identifies two cells as possible doublets, red bars. **B)** DoubletDetector found no cells exhibited doublet characteristics.



Supplementary Figure 6. Clustering

A) SC3 consensus matrix predicted 2 clusters, dark red colouring. **B)** Silhouette plot of the consensus matrix (100% indicates perfect clustering). **C)** Cluster stability plots showing that as the number of clusters increases past two, cluster stability decreases. **D)** PCA plot of the two predicted clusters, coloured by time-point, sized by infection status and shaped by cluster. **E)** PCA plot following kNN-smoothing on the expression matrix.



Supplementary Figure 7. Sub-clustering

The four comparisons shown here were created by manually selecting two and three clusters to examine any sub-clustering events not automatically detected. **A)** 3 hour cells - no sub-clustering evident. **B)** 6 hour cells - no apparent sub-clustering with two clusters; three clusters do display more of a separation (between blue and green), while infected and mock-infected cells cannot be distinguished. **C)** 12 hour cells - some separation evident with 3 clusters, but infection state is not distinguishable. **D)** Extracting only infected cells show a clear separation of 3 hour cells, but not 6 and 12 hours cells

4.7. References

Abdelrahman, Y.M., Rose, L.A., and Belland, R.J. (2011). Developmental expression of non-coding RNAs in *Chlamydia trachomatis* during normal and persistent growth. Nucleic acids research *39*, 1843-1854.

Albrecht, M., Sharma, C.M., Dittrich, M.T., Muller, T., Reinhardt, R., Vogel, J., and Rudel, T. (2011). The transcriptional landscape of *Chlamydia pneumoniae*. Genome Biol *12*, R98-R98.

Albrecht, M., Sharma, C.M., Reinhardt, R., Vogel, J., and Rudel, T. (2010). Deep sequencing-based discovery of the *Chlamydia trachomatis* transcriptome. Nucleic acids research *38*, 868-877.

Ali, H., Cameron, E., Drovandi, C.C., McCaw, J.M., Guy, R.J., Middleton, M., El-Hayek, C., Hocking, J.S., Kaldor, J.M., Donovan, B.*, et al.* (2015). A new approach to estimating trends in *chlamydia* incidence. Sexually transmitted infections *91*, 513-519.

Andrews, S. (2010). FastQC A Quality Control tool for High Throughput Sequence Data.

Avital, G., Avraham, R., Fan, A., Hashimshony, T., Hung, D.T., and Yanai, I. (2017). scDual-Seq: mapping the gene regulatory program of *Salmonella* infection by host and pathogen single-cell RNA-sequencing. Genome Biol *18*, 200-200.

Balsara, Z.R., Misaghi, S., Lafave, J.N., and Starnbach, M.N. (2006). *Chlamydia trachomatis* infection induces cleavage of the mitotic cyclin B1. Infection and immunity *74*, 5602-5608.

Barron, M., and Li, J. (2016). Identifying and removing the cell-cycle effect from single-cell RNA-Sequencing data. Sci Rep *6*, 33892-33892.

Bastidas, R.J., Elwell, C.A., Engel, J.N., and Valdivia, R.H. (2013). Chlamydial intracellular survival strategies. Cold Spring Harb Perspect Med *3*, a010256-a010256.

Belland, R.J., Zhong, G., Crane, D.D., Hogan, D., Sturdevant, D., Sharma, J., Beatty, W.L., and Caldwell, H.D. (2003). Genomic transcriptional profiling of the developmental cycle of

Chlamydia trachomatis. Proceedings of the National Academy of Sciences of the United States of America *100*, 8478-8483.

Blecher-Gonen, R., Bost, P., Hilligan, K.L., David, E., Salame, T.M., Roussel, E., Connor, L.M., Mayer, J.U., Bahar Halpern, K., Tóth, B.*, et al.* (2019). Single-Cell Analysis of Diverse Pathogen Responses Defines a Molecular Roadmap for Generating Antigen-Specific Immunity. Cell Systems *8*, 109-121.e106.

Burton, M.J., and Mabey, D.C.W. (2009). The Global Burden of Trachoma: A Review. PLoS Neglected Tropical Diseases *3*, e460-e460.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics *14*, 128-128.

Cherian, M.G., and Apostolova, M.D. (2000). Nuclear localization of metallothionein during cell proliferation and differentiation. Cellular and molecular biology (Noisy-le-Grand, France) *46*, 347-356.

Cocchiaro, J.L., Kumar, Y., Fischer, E.R., Hackstadt, T., and Valdivia, R.H. (2008). Cytoplasmic lipid droplets are translocated into the lumen of the *Chlamydia trachomatis* parasitophorous vacuole. Proceedings of the National Academy of Sciences *105*, 9379 LP-9384.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B.*, et al.* (2011). Reactome: a database of reactions, pathways and biological processes. Nucleic acids research *39*, D691-D697.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England) *29*, 15-21.

Elwell, C., Mirrashidi, K., and Engel, J. (2016). *Chlamydia* cell biology and pathogenesis. Nat Rev Microbiol *14*, 385-400.

Elwell, C.A., and Engel, J.N. (2012). Lipid acquisition by intracellular *Chlamydiae*. Cell Microbiol *14*, 1010-1018.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047-3048.

Gayoso, A., and Shor, J. (2019). GitHub: DoubletDetection.

Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N., and Garry, D.J. (2018). DrImpute: imputing dropout events in single cell RNA sequencing data. BMC bioinformatics *19*, 220-220.

Grieshaber, S.S., Grieshaber, N.A., Miller, N., and Hackstadt, T. (2006). *Chlamydia trachomatis* causes centrosomal defects resulting in chromosomal segregation abnormalities. Traffic (Copenhagen, Denmark) *7*, 940-949.

Hafner, L.M., Wilson, D.P., and Timms, P. (2014). Development status and future prospects for a vaccine against *Chlamydia trachomatis* infection. Vaccine *32*, 1563-1571.

Hebenstreit, D. (2012). Methods, Challenges and Potentials of Single Cell RNA-seq. Biology *1*, 658-667.

Humphrys, M.S., Creasy, T., Sun, Y., Shetty, A.C., Chibucos, M.C., Drabek, E.F., Fraser, C.M., Farooq, U., Sengamalay, N., Ott, S.*, et al.* (2013). Simultaneous Transcriptional Profiling of Bacteria and Their Host Cells. PloS one *8*, e80597-e80597.

Ji, Z., and Ji, H. (2016). TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic acids research *44*, e117-e117.

Johnson, K.A., Tan, M., and Sütterlin, C. (2009). Centrosome abnormalities during a *Chlamydia trachomatis* infection are caused by dysregulation of the normal duplication pathway. Cell Microbiol *11*, 1064-1073.

Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N., Reik, W., Barahona, M., Green, A.R.*, et al.* (2017). SC3: consensus clustering of single-cell RNA-seq data. Nat Methods *14*, 483-483.

Knowlton, A.E., Brown, H.M., Richards, T.S., Andreolas, L.A., Patel, R.K., and Grieshaber, S.S. (2011). *Chlamydia trachomatis* infection causes mitotic spindle pole defects independently from its effects on centrosome amplification. Traffic (Copenhagen, Denmark) *12*, 854-866.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015). The Technology and Biology of Single-Cell RNA Sequencing. Molecular cell *58*, 610-620.

Kowalczyk, M.S., Tirosh, I., Heckl, D., Rao, T.N., Dixit, A., Haas, B.J., Schneider, R.K., Wagers, A.J., Ebert, B.L., and Regev, A. (2015). Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome research *25*, 1860-1872.

Krueger, F. (2012). Trim Galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/).

Kumar, Y., and Valdivia, R.H. (2008). Actin and intermediate filaments stabilize the *Chlamydia trachomatis* vacuole by forming dynamic structural scaffolds. Cell host & microbe *4*, 159-169.

Łabaj, P.P., Leparc, G.G., Linggi, B.E., Markillie, L.M., Wiley, H.S., and Kreil, D.P. (2011). Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. Bioinformatics *27*, i383-i391.

Lambden, P.R., Pickett, M.A., and Clarke, I.N. (2006). The effect of penicillin on *Chlamydia trachomatis* DNA replication. Microbiology (Reading, England) *152*, 2573-2578.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923-930.

Liu, S., and Trapnell, C. (2016). Single-cell transcriptome sequencing: recent advances and remaining challenges. F1000Res *5*.

Lönnberg, T., Svensson, V., James, K.R., Fernandez-Ruiz, D., Sebina, I., Montandon, R., Soon, M.S.F., Fogg, L.G., Nair, A.S., Liligeto, U.N.*, et al.* (2017). Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves T_H1/T_{FH} fate bifurcation in malaria. Science Immunology *2*, eaal2192-eaal2192.

Lun, A.T.L., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor [version 2; referees: 3 approved, 2 approved with reservations]. F1000Res *5*.

Mak, T.N., and Brüggemann, H. (2016). Vimentin in Bacterial Infections. Cells *5*, 18-18.

Marsh, J.W., Hayward, R.J., Shetty, A.C., Mahurkar, A., Humphrys, M.S., and Myers, G.S.A. (2018). Bioinformatic analysis of bacteria and host cell dual RNA-sequencing experiments. Briefings in bioinformatics *19*, 1115-1129.

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

Menon, S., Timms, P., Allan, J.A., Alexander, K., Rombauts, L., Horner, P., Keltz, M., Hocking, J., and Huston, W.M. (2015). Human and Pathogen Factors Associated with *Chlamydia trachomatis*-Related Infertility in Women. Clinical microbiology reviews *28*, 969-985.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L.*, et al.* (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science (New York, NY) *344*, 1396 LP-1401.

Porcella, S.F., Carlson, J.H., Sturdevant, D.E., Sturdevant, G.L., Kanakabandi, K., Virtaneva, K., Wilder, H., Whitmire, W.M., Song, L., and Caldwell, H.D. (2015). Transcriptional profiling of human epithelial cells infected with plasmid-bearing and plasmid-deficient *Chlamydia trachomatis*. Infection and immunity *83*, 534-543.

Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M.*, et al.* (2017). The Human Cell Atlas. eLife *6*.

Renaud, G., Stenzel, U., Maricic, T., Wiebe, V., and Kelso, J. (2015). deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. Bioinformatics *31*, 770-772.

Reyburn, H. (2016). WHO Guidelines for the Treatment of *Chlamydia trachomatis*. WHO *340*, c2637-c2637.

Rice, J.M., Zweifach, A., and Lynes, M.A. (2016). Metallothionein regulates intracellular zinc signaling during CD4(+) T cell activation. BMC immunology *17*, 13-13.

Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nature biotechnology *32*, 896-902.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Russell, A.B., Trapnell, C., and Bloom, J.D. (2018). Extreme heterogeneity of influenza virus infection in single cells. eLife *7*, e32303-e32303.

Saka, H.A., Thompson, J.W., Chen, Y.-S., Kumar, Y., Dubois, L.G., Moseley, M.A., and Valdivia, R.H. (2011). Quantitative proteomics reveals metabolic and pathogenic properties of *Chlamydia trachomatis* developmental forms. Molecular microbiology *82*, 1185-1203.

Saliba, A.-E., Westermann, A.J., Gorski, S.A., and Vogel, J. (2014). Single-cell RNA-seq: advances and future challenges. Nucleic acids research *42*, 8845-8860.

Saliba, A.E., Li, L., Westermann, A.J., Appenzeller, S., Stapels, D.A.C., Schulte, L.N., Helaine, S., and Vogel, J. (2016). Single-cell RNA-seq ties macrophage polarization to growth rate of intracellular *Salmonella*. Nature Microbiology *2*, 1-8.

Sandberg, R. (2013). Entering the era of single-cell transcriptomics in biology and medicine. Nat Methods *11*, 22-22.

Scidmore, M.A., Rockey, D.D., Fischer, E.R., Heinzen, R.A., and Hackstadt, T. (1996). Vesicular interactions of the *Chlamydia trachomatis* inclusion are determined by chlamydial early protein synthesis rather than route of entry. Infection and immunity *64*, 5366 LP-5372.

Siegl, C., Prusty, Bhupesh K., Karunakaran, K., Wischhusen, J., and Rudel, T. (2014). Tumor Suppressor p53 Alters Host Cell Metabolism to Limit *Chlamydia trachomatis* Infection. Cell Reports *9*, 918-929.

Subramanian Vignesh, K., and Deepe, G.S., Jr. (2017). Metallothioneins: Emerging Modulators in Immunity and Infection. International journal of molecular sciences *18*.

Tan, C., Hsia, R.-c., Shou, H., Haggerty, C.L., Ness, R.B., Gaydos, C.A., Dean, D., Scurlock, A.M., Wilson, D.P., and Bavoil, P.M. (2009). *Chlamydia trachomatis*-Infected

Patients Display Variable Antibody Profiles against the Nine-Member Polymorphic Membrane Protein Family. Infection and Immunity *77*, 3218 LP-3226.

Taylor, H.R., Burton, M.J., Haddad, D., West, S., and Wright, H. (2014). Trachoma. Lancet (London, England) *384*, 2142-2152.

Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth 2nd, M.H., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G.*, et al.* (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science (New York, NY) *352*, 189-196.

Valdivia, R.H. (2008). *Chlamydia* effector proteins and new insights into chlamydial cellular microbiology. Current opinion in microbiology *11*, 53-59.

Wagner, F., Yan, Y., and Yanai, I. (2018). K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. bioRxiv, 217737-217737.

Wang, Y.J., Schug, J., Lin, J., Wang, Z., Kossenkov, A., and Kaestner, K.H. (2019). Comparative analysis of commercially available single-cell RNA sequencing platforms for their performance in complex human tissues. bioRxiv, 541433-541433.

Wilson, D.P., Mathews, S., Wan, C., Pettitt, A.N., and McElwain, D.L.S. (2004). Use of a quantitative gene expression assay based on micro-array techniques and a mathematical model for the investigation of chlamydial generation time. Bulletin of mathematical biology *66*, 523-537.

Wingett, S.W., and Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. F1000Res *7*, 1338-1338.

Wolock, S.L., Lopez, R., and Klein, A.M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Systems *8*, 281-291.e289.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J.*, et al.* (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nature structural & molecular biology *20*, 1131-1139.

Zaika, A.I., Wei, J., Noto, J.M., and Peek, R.M. (2015). Microbial Regulation of p53 Tumor Suppressor. PLoS pathogens *11*, e1005099-e1005099.

Zhao, Q., Wang, J., Levichkin, I.V., Stasinopoulos, S., Ryan, M.T., and Hoogenraad, N.J. (2002). A mitochondrial specific stress response in mammalian cells. EMBO J *21*, 4411-4419.

Chapter 5

Comparative analysis using different MOIs from *Chlamydia*-infected epithelial cells

5.1. Introduction

Chlamydia trachomatis is a Gram-negative, obligate intracellular bacterial pathogen that causes disease within humans (Schachter and Caldwell, 1980). Infection typically occurs within ocular and urogenital mucosal epithelial cells. Ocular infections cause trachoma (infectious blindness), typically in disadvantaged communities, and is the leading cause of preventable blindness worldwide (Burton and Mabey, 2009; Reyburn, 2016); while genital infections are the most prevalent sexually transmitted infection (STI) worldwide (Reyburn, 2016), and if untreated can lead to complex disease outcomes including ectopic pregnancy and infertility (Brunham et al., 1986; Menon et al., 2015).

Chlamydial species are distinguishable from other bacterial pathogens by their unique biphasic developmental cycle that alternates between the infectious non-replicating elementary body (EB) and the replicative reticulate body (RB) (AbdelRahman and Belland, 2005). Infection begins with EBs attaching to the host cell and entering through endocytosis, forming separate membrane bound inclusions for each successfully internalised EB (Elwell et al., 2016). The formed inclusions are able to escape phagolysosomal fusion (Scidmore et al., 2003), thereby providing a niche for survival within the host cell. Within the first 2-3 hours the EB differentiates into an RB, which continues to multiply; during this time we also see continued growth of the inclusion to accommodate the increase of RBs (AbdelRahman and Belland, 2005). Proteins are released from the inclusion using a Type III secretion system (T3SS) that facilitates in countering host defences and retrieving host-based nutrients for growth, replication and survival (Betts-Hampikian and Fields, 2010; Elwell et al., 2016; Saka and Valdivia, 2010). At around 20-44 hours, RBs asynchronously transition into EBs, then through either host cell lysis or extrusion (~48-70 hours), are released and able to infect new cells (Hybiske and Stephens, 2007).

Throughout the course of an infection, a host cell will try to counter and eliminate a bacterial threat through innate and adaptive immune responses, in addition to a variety of mechanisms corresponding to the stage of infection. For example, at early time points, defence mechanisms include repelling EB attachment and entry (Vats et al., 2007); while at later time points eliminating effector proteins is necessary to disrupt bacterial growth and replication (Bastidas et al., 2013). Snapshots of gene expression from both the host and pathogen when examined, can help to uncover and identify these underlying processes.

RNA-sequencing (RNA-seq) is an established method to capture transcripts within a population of cells from any organism (Kukurba and Montgomery, 2015). However, the resulting composition of transcripts is abundant in ribosomal (r)RNA (~ 90%), which in the majority of cases does not provide any meaningful biological context (O'Neil et al., 2013). To overcome this, rRNA depletion is carried out via specialised kits before sequencing libraries are prepared, enriching samples for more biologically relevant transcripts. Other depletion/selection methods exist that can modify sequencing libraries for a variety of biological-based purposes (Heyer et al., 2019; Teder et al., 2018). For example, polyA depletion can remove the majority of host-based transcripts, enriching the library for bacterial reads (Kumar et al., 2016).

Dual species transcriptomic experiments (dual RNA-seq) allow multiple organisms to be analysed from within the same sample, such as host and bacterial transcripts during an infection (Westermann et al., 2012). To date, there have been numerous RNA-seq experiments separately examining chlamydial biology and infection from either a host-centric or chlamydial-specific point of view. Host cells have been predominantly from human and mouse tissues or from *in vitro* models of infection, while chlamydial species include *C. trachomatis* (Albrecht et al., 2010; Belland et al., 2003a; Belland et al., 2003b; Grieshaber et al., 2018), *C. pneumoniae* (Beaulieu et al., 2015; Wang et al., 2013), *C. muridarum* (Johnson

et al., 2018; O'Connell et al., 2011), *C. caviae* (Wali et al., 2014), *C. psittaci* (Paul et al., 2018) and *C. pecorum* (Phillips et al., 2019). Only one dual-RNA-seq experiment has analysed both the host and chlamydial transcripts simultaneously (Humphrys et al., 2013). Their depletion technique removed rRNAs in all samples, followed by subjecting half of these libraries to polyA depletion to further enrich chlamydial transcripts. Although two depletion methods were used, it is uncertain if this did increase the abundance of chlamydial transcripts.

Host-based RNA-seq experiments in an infection setting will typically try and achieve a ratio of 1 infectious bacterial entity per host cell. This ratio is referred to as the multiplicity of infection (MOI), with an MOI of 1 indicating a 1:1 ratio, and is frequently used to assess baseline changes in both organisms without any directional bias. RNA-seq and microarray experiments that have focused on chlamydial infection have utilised a range of MOIs ranging from 1 (Yeung et al., 2017) to 100 (Belland et al., 2003b; Wang et al., 2013); with higher ratios helping to exaggerate and highlight the chlamydial impact. However, too high an MOI and the whole monolayer of cells dies before the infection can proceed. In addition, a higher MOI has been shown to reduce the length of the development cycle due to the underlying stress this places on host cells, but also has the ability to rapidly produce large numbers of infectious progeny (Lyons et al., 2005; Miyairi et al., 2006).

In this experiment, both host and chlamydial gene expression were examined applying dual-RNA-seq to *in vitro C. trachomatis*-infected HEp-2 epithelial cells. The first aim was to understand the influence different MOIs have on sequence capture rates, but also the transcriptional variation from *Chlamydia* and the host-cell. The second aim attempted to improve the enrichment of chlamydial reads by comparing different RNA depletion methods. To address these questions, two time points were chosen covering the chlamydial developmental cycle (1 and 24 hours), with each time point split into three MOIs (0.1, 1 and 10), each in triplicate. Each of these biological replicates (16 samples) were split in half,

where one library was prepared solely with rRNA depletion, while the second was prepared with rRNA depletion followed by polyA depletion (**Figure 5.1**).

5.2. Methods

5.2.1. Cell culture and infection

Human epithelial type 2 (HEp-2) cells (American Type Culture Collection, ATCC No. CCL-23) were grown as monolayers in 6x 100 mm tissue culture dishes until cells were 90% confluent. To harvest EBs for the subsequent infections, additional monolayers were grown and infected with *C. trachomatis* serovar E in sucrose phosphate glutamate (SPG) as previously outlined (Tan et al., 2009). The resulting EBs and cell lysates were then harvested and used to infect new HEp2 monolayers (**Figure 5.1A**).

Infections for each dataset used the previously prepared HEp2 monolayers, infecting with *C. trachomatis* serovar E in 3.5 mL SPG buffer as previously outlined (Tan et al., 2009); infections were synchronised using centrifugation. EBs were introduced into monolayers from three MOIs (0.1, 1 and 10) using 1:10 dilutions beginning from an MOI of 10. EBs were quantified as previously described (Humphrys et al., 2013). To remove non-viable or dead EBs, each sample was incubated at 25ºC for 2 hours, and washed twice in SPG. Cell monolayers were incubated at 37ºC with 5% CO2, including the addition of 10 mL fresh medium (DMEM+10% FBS, 25 μg/ml gentamycin, 1.25 μg/ml Fungizone). After each infection time point, the infected and uninfected dishes were harvested by scraping and resuspending in 150 μL sterile PBS. Any resuspended samples were stored at −80°C.

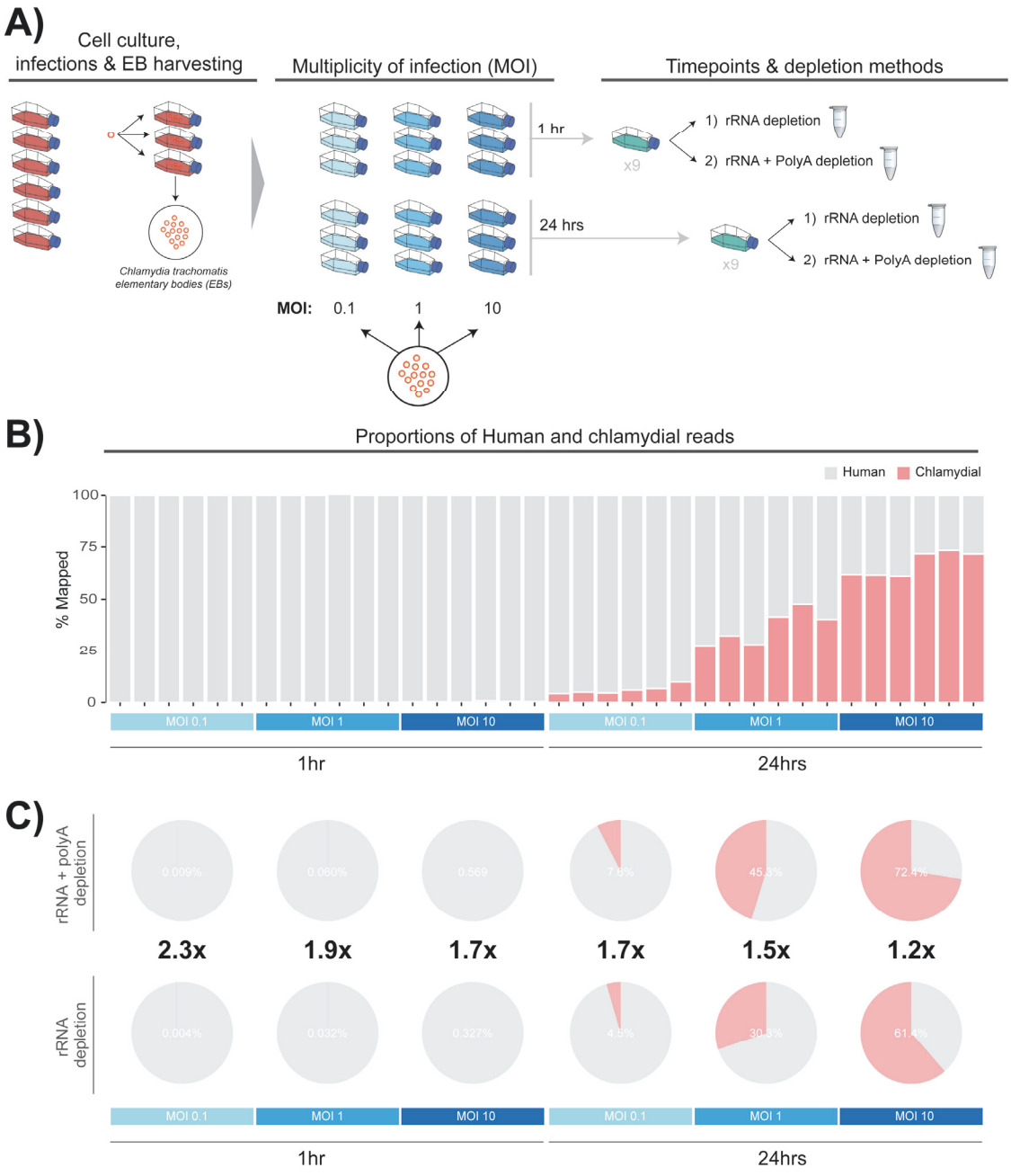


Figure 5.1: Experimental process and design

A) The process of growing cell cultures and harvesting (elementary bodies) EBs to use for downstream experiments is a time-consuming process spanning multiple days. The resulting EBs were used for three different infection ratios (multiplicity of infection) of 0.1, 1 and 10. After infections, samples were left for 1 hour and 24 hours. Each replicate was then prepared

with rRNA or rRNA plus PolyA depletion, generating 32 samples in total. **B)** Showing the percent of Human and chlamydial reads across the experimental design. **C)** By combining rRNA depletion and polyA depletion, we were able to increase the capture efficiency of chlamydial transcripts at both time points and across MOIs.

5.2.2. Library preparation and sequencing

Ribo-Zero rRNA Removal kits (Human/Mouse/Rat and Gram-negative) were used to deplete samples of both human and gram-negative bacterial rRNA. Equivalent volumes from each kit were combined, thereby allowing the removal of bacterial and human rRNA simultaneously within each sample. Each sample was equally separated, with one half subjected to polyA depletion by the Poly(A) Purist Mag purification kit (Ambion), whereby removing host-based polyA transcripts to allow the enrichment of bacterial transcripts. Magnetic beads were used to bind to polyA mRNAs and were extracted from the solution with a magnet. Samples with combined depletion methods were further purified using Zymo-Spin IC columns (Zymo Research) before being re-combined for library construction.

The mRNA libraries were prepared from depleted samples as previously stated at 1 and 24 hours post infection, using the TruSeq RNA Sample Prep kit (Illumina, San Diego, CA) per the manufacturer's protocol with IGS-specific optimisations. Adapters and indexes (6 bp) were ligated to the double-stranded cDNA, which was subsequently purified with AMPure XT beads (Beckman Coulter Genomics, Danvers, MA) between enzymatic reactions and size selection steps (~250 to 300 bp). The resulting libraries were sequenced on an Illumina HiSeq2000 using the 100 bp paired-end protocol at the Genome Resource Centre, Institute for Genome Sciences, University of Maryland School of Medicine.

5.2.3. Bioinformatic analysis

Sequencing reads were trimmed and quality checked using Trim Galore (0.45) (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and FastQC (0.11.5) (Andrews, 2010). Host reads were aligned to the human genome (GRCh 38.87) using STAR (2.5.2b) (Dobin et al., 2013), while chlamydial reads were aligned to the *Chlamydia trachomatis* (serovar E, Charm001) genome using Bowtie2 (2.3.2) (Langmead and Salzberg, 2012) with additional parameters of '1 mismatch' and '–very-sensitive-local'. Samtools (1.6) (Li et al., 2009) was used to remove duplicate reads in addition to only keeping mapped reads in both the host and chlamydial BAM files. To remove reads that mapped to both genomes, we first extracted the mapped reads back into paired-end fastq files using bedtools (2.26.0) (Quinlan and Hall, 2010). Reads were then aligned using the initial mapping software to the reciprocal genomes. Any reads that mapped to both genomes were removed from the originating BAM files using the "FilterSamReads" command from Picard tools (2.10.4) (Wysoker et al., 2013). Additional quality control metrics were examined using Bamtools (2.5.1) (Barnett et al., 2011), MultiQC (1.2) (Ewels et al., 2016) and various in-house scripts.

Features (genes) were counted using featureCounts (1.5.0-p1) (Liao et al., 2014) with additional parameters of "-Q 10 -p -C". Genefilter (1.64.0) ((Gentleman et al., 2018) was used to filter out genes with low counts, where host genes were retained if expression > 50 in at least three samples. To accommodate the vast differences in expression between host and chlamydial reads, a separate filter was used retaining chlamydial genes with expression > 10 in at least three samples. Chlamydial and host reads were further separated by time point due to the large amount of variability in expression between an MOI of 0.1 at 1 hour to an MOI of 10 at 24 hours. Once separated, library normalisation was performed using the trimmed mean of M-values (TMM) method (Robinson and Oshlack, 2010).

To identify outliers, four PCA bi-plots were generated from library normalised counts using PCATools (0.99.13) (Blighe and Lewis, 2018), where eigenvalues from PC1 and PC2 for each replicate were calculated and used to highlight outlier samples if an eigenvalue was > |3| standard deviations from the mean within that group. If an outlier was removed, eigenvalues were recalculated and the process repeated until no further outliers were detected. To determine the underlying variation at each principal component, the "plotloadings" function within PCATools (Blighe and Lewis, 2018) was used.

Differential expression was performed with edgeR (3.24.3) (Robinson et al., 2010), adding the difference between the depletion methods as a blocking factor, whereby allowing MOI and time point comparisons to utilise all six replicates to increase significance. Host DE genes were uploaded and enriched for KEGG pathways using the Enrichr database (Kuleshov et al., 2016). Relevant host pathways were determined using the combined score with a cutoff of > 50. Combined scores were calculated by adding together the combined scores comparing MOIs 0.1 vs 1, and 1 vs 10. Enrichment of chlamydial DE genes was performed using the 'UniProt Keywords' feature of STRING (11.0) (Szklarczyk et al., 2018).

5.3. Results

5.3.1. Quantifying expression differences between host and chlamydial reads

Dual RNA-seq was applied to *C. trachomatis* serovar E-infected human HEp-2 epithelial cells in triplicate at 1 and 24 hours post-infection (hpi). Within each time point, three MOIs were used (0.1, 1 and 10), in addition to two depletion methods 1) rRNA depletion, 2) rRNA depletion and polyA depletion; totalling 36 samples across the experimental design.

Capture rates of chlamydial reads at early time points is challenging due to limited biological activity, where the majority of transcripts in a sample (>99%) will be associated with the host (Humphrys et al., 2013). We increased the sequencing depth at 1 hour (> 6 fold) to try and capture more chlamydial reads; generating 391,847,337 mapped reads at 1 hour compared to 63,710,236 mapped reads at 24 hours. Even with this greater depth of sequencing at 1 hour, the number of chlamydial reads was still quite low; especially at an MOI of 0.1 with an average of 1,407 reads across the six replicates. However, as the MOI increases, we do see an increase in chlamydial transcripts, with average mapped reads of 10,392 (MOI 1), and 55,426 (MOI 10) (**Figure 5.2A**). At 24 hours we see an expected increase in the number of chlamydial reads, following a similar trend with 1 hour, where the number of mapped reads increases as the MOI increases (**Figure 5.2B**). The number of mapped host reads tends to vary more than the chlamydial reads, particularly between depletion methods of the same MOI (**Figure 5.2C-D**). This is likely due to the increased variety of host transcripts resulting from post-transcriptional modifications, such as polyadenylation, which does not occur in bacterial systems.

When examining the proportions of host and chlamydial reads together across the experimental design, we see that 1 hour is dominated by host reads, while at 24 hours we see a gradual increase of chlamydial reads as the MOI increases. Surprisingly, at 24 hours with an MOI of 10, the proportion of chlamydial reads across all replicates is over 60% (**Figure 5.1B**).

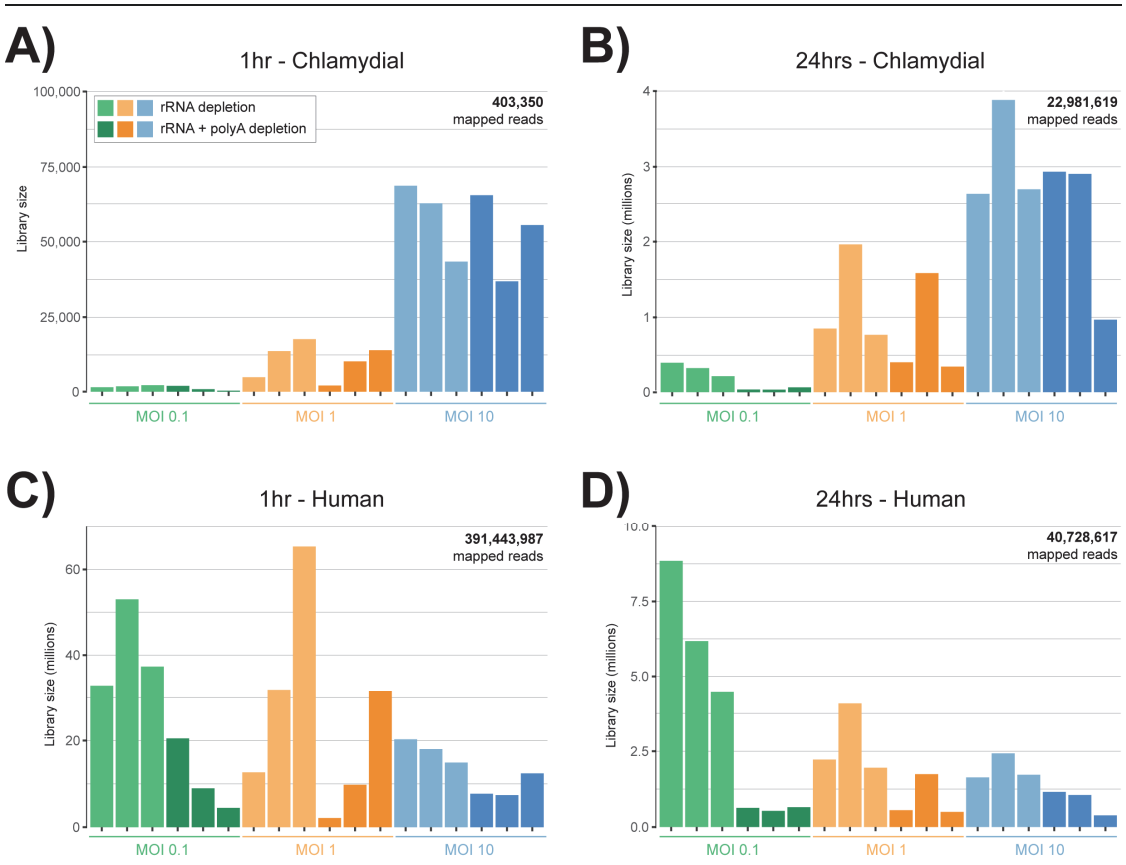


Figure 5.2: Human and chlamydial mapped reads

The number of mapped sequence reads to both human and chlamydial genomes. Green bars represent an MOI of 0.1, orange 1, and blue 10. Light shaded colours represent the rRNA depletion method, while darker shades represent rRNA and polyA depletion methods combined. **A)** Low numbers of chlamydial mapped reads at lower MOIs, but a substantial increase at an MOI of 10. **B)** At 24 hours the number of captured transcripts dramatically increases from 1 hour, but follows a similar distribution with a spike of reads at an MOI of 10. **C)** At early time points, chlamydial genes are expressed at low quantities and are often challenging to capture. The substantial increase in mapped host reads at 1 hour was a result of increasing sequencing depth with an aim of capturing more chlamydial-based transcripts. **D)** As the MOI increases at 24 hours, the number of mapped host reads declines.

5.3.2. Combining depletion methods increases yield of bacterial transcripts

We attempted to capture chlamydial reads by combining two depletion methods (rRNA depletion and polyA depletion). The addition of polyA depletion should theoretically remove any polyadenylated host transcripts, thereby increasing the relative amount of chlamydial transcripts available to be captured and sequenced.

Overall, we see an increase in chlamydial reads when combining depletion methods. Even at 1 hour, when there are limited transcripts circulating within the cell, we still see an average increase of 2.0x. At 24 hours, when more chlamydial transcripts are being expressed, we see an average increase of 1.5x more reads. Interestingly, at 24 hours as the MOI increases, the capture efficiency begins to decline slightly from 1.7x to 1.2x (**Figure 5.1C**).

5.3.3. Differences in chlamydial expression between depletion methods

PCA bi-plots were created to compare the expression profiles across replicates from both depletion methods. At 1 hour, we see minimal separation at an MOI of 1 and 10 compared to 0.1 where replicates appear separated and not grouped by depletion method as expected (**Figure 5.3A**). However, none of the replicates were considered outliers using a robust statistical approach (see **Methods 5.2.3**). We therefore attribute this variability to the low number of chlamydial reads present at an MOI of 0.1 as identified earlier. At 24 hours a distinct separation between depletion methods within each MOI can be easily visualised (**Figure 5.3B**).

To understand if the variability between depletion methods is driven by a small subset of highly expressed genes, or an assortment of genes, we extracted the top 5% of genes driving the underlying variation at PC1 and PC2 for each MOI (**Figure 5.3C-D**). At both time points,

we see subsets of genes specific to each MOI, indicating that each MOI may be inducing a different chlamydial response. In addition, overlapping genes highlight that the variation between depletion methods was also captured and overlaps considerably. Therefore, the inclusion of polyA depletion increases bacterial reads and does not seem to be driven by small subsets of highly expressed transcripts, but allows for a wide array of transcripts to be captured.

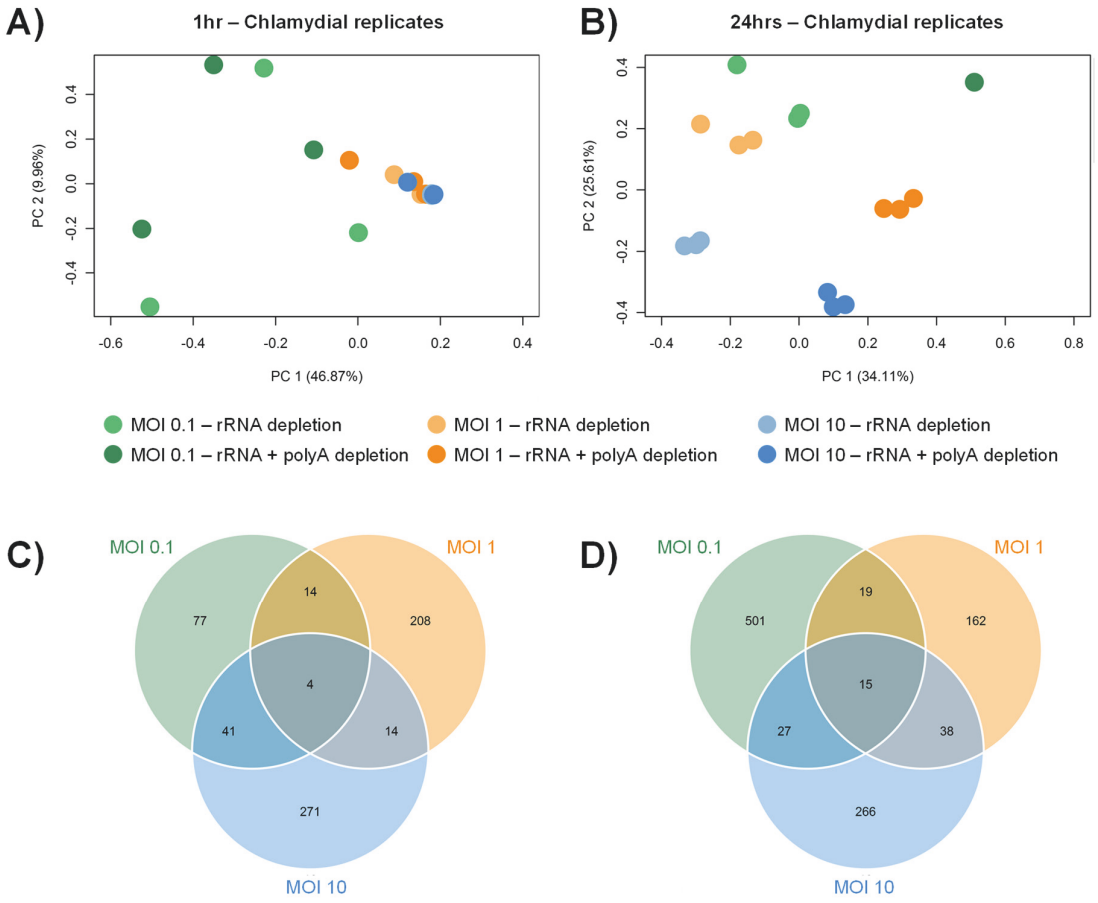


Figure 5.3: Chlamydial-based expression differences between depletion methods

A) At 1 hour, minimal separation is seen apart from at an MOI of 0.1 where replicates appear to not group or cluster together. **B)** At 24 hours, replicates group together within the same depletion method, while separation is seen between depletion methods at each MOI.

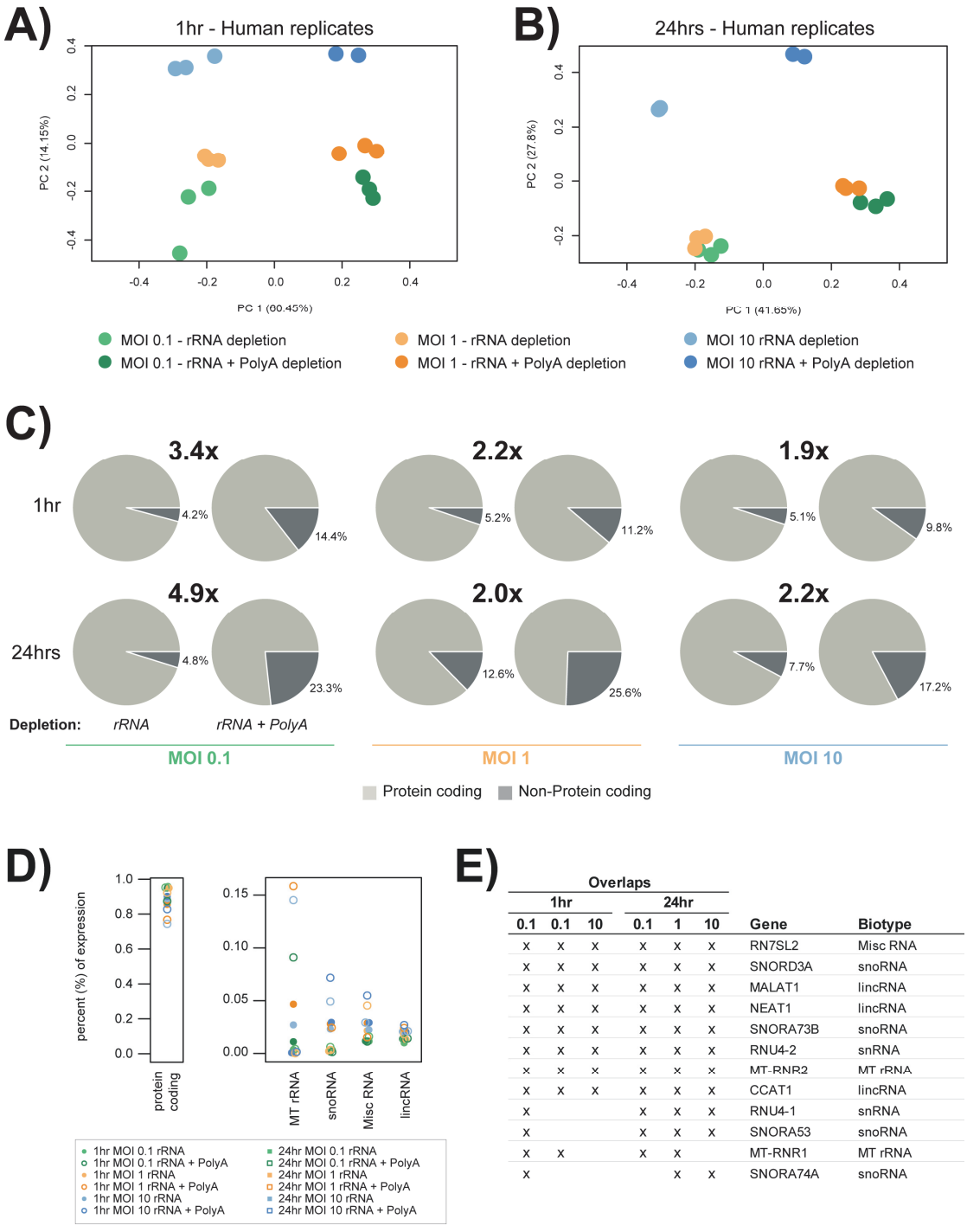
Extracting the top 5% of genes driving variation from PC1 and PC2 between depletion methods. At 1 hour **C)** and 24 hours **D)**, we see subsets of genes overlapping MOIs in addition to MOI-specific subsets.

5.3.4. The removal of polyA transcripts increases non-protein coding host gene expression

Examining PCA bi-plots for host reads show tight clustering between replicates, but also highlights the separation between depletion methods (**Figure 5.4A-B**). Extracting the underlying genes contributing the variation at PC1 and PC2, numerous non-coding genes were identified. To calculate the percent of protein coding versus non-protein coding expression, gene expression was averaged across replicates after separation by experimental conditions (time point, MOI and depletion method) (**Figure 5.4C**). Across both time points we see an average of 2.8x more non-protein coding expression when combining rRNA and polyA depletion, with the highest proportion occurring at an MOI of 0.1 (3.4x at 1 hour and 4.9x at 24 hours). Although the majority of expression comes from protein-coding genes (**Figure 5.4C)**, non-protein coding expression contributed to the separation of depletion methods as observed in the PCA plots (**Figure 5.4A-B**). By characterising the most common non-protein-coding biotypes, we see mitrochondrial rRNA (MT rRNA), small nucleolar RNAs (snoRNA), miscRNA and long intergenic non-coding (lincRNA); but without any statistically significant trends separating time points, depletion methods or MOI (**Figure 5.4D**).

To identify potentially influential non-protein coding genes, we used the top 200 expressed genes from both depletion methods and extracted a subset of genes that occur frequently (across 3 or more conditions) (**Figure 5.4E**). Of the 12 genes identified, 5 were snoRNAs

which are involved with RNA modifications, and are among the most highly abundant non-coding RNAs (ncRNAs) in the nucleus (Huang et al., 2017). The MT-RNR1 (12S RNA) and MT-RNR2 (16S RNA) genes encode the two rRNA subunits of mitochondrial ribosomes, and are generally always highly expressed within eukaryotic cells (Shutt and Shadel, 2010). LincRNAs include CCAT1, which is linked to cell growth and regulation of EGFR (Jiang et al., 2018), while MALAT1 and NEAT1 co-localise to hundreds of genomic loci, predominantly over active genes (West et al., 2014).



E)

Overlaps							
1hr			24hr				
0.1	0.1	10	0.1	1	10	Gene	Biotype
x	x	x	x	x	x	RN7SL2	Misc RNA
x	x	x	x	x	x	SNORD3A	snoRNA
x	x	x	x	x	x	MALAT1	lincRNA
x	x	x	x	x	x	NEAT1	lincRNA
x	x	x	x	x	x	SNORA73B	snoRNA
x	x	x	x	x	x	RNU4-2	snRNA
x	x	x	x	x	x	MT-RNR2	MT rRNA
x	x	x	x	x	x	CCAT1	lincRNA
x			x	x	x	RNU4-1	snRNA
x			x	x	x	SNORA53	snoRNA
x	x		x	x		MT-RNR1	MT rRNA
x				x	x	SNORA74A	snoRNA

Figure 5.4: Host-based expression differences of protein coding and non-protein coding genes between depletion methods

A) PCA plots show tight grouping between replicates, but separation at each MOI and depletion method. **B)** Similar grouping trends to 1 hour, but with an MOI of 10 much further separated. **C)** An overall increase in non-protein coding expression is observed when combining depletion methods. **D)** The majority of expression for all conditions is from protein coding genes. While non-protein coding genes are dominated by four biotypes, with Mt rRNAs the most highly expressed. **E)** Non-protein coding genes that are within the top 200 expressed genes at each MOI, and overlap 3 or more conditions.

5.3.5. Increasing infection highlights minimal changes to highly expressed host and chlamydial genes

To determine whether the host or chlamydial transcriptional-profile changes in relation to the ratio of EBs per cell, highly expressed genes were compared against an MOI of 1. Chlamydial transcripts were examined from the combined depletion replicates, as more transcripts were captured (**Figure 5.1C**), thus giving a more representative profile. Host reads were taken from just the rRNA depleted replicates, as these were shown to contain more of an accurate representation of protein coding and non-protein coding genes (**Figure 5.4C)**.

Each of the four panels (**Figure 5.5**) contains two graphs. The first graph contains the top 50 expressed genes taken from an MOI of 1, while the second graph shows the ranked-positions of these top expressed genes. At 1 hour, there is slightly less chlamydial expression at an MOI of 0.1, and slightly more expression when additional EBs are introduced at an MOI of 10 (**Figure 5.5A**). The ranking chart to the right shows that 9/10 of the top expressed genes remain the same across the three MOIs. The top 25 genes from the host's response at 1 hour share highly similar expression profiles (**Figure 5.5B**); with only two mitochondrial-based genes (MT-RNR1 and MT-RNR2) at an MOI of 10 standing out with lower expression. The ranking chart shows 7/10 top expressed genes remaining constant across the three MOIs,

similar to the chlamydial profile. At 24 hours, similar expression profiles of the top 25 expressed chlamydial genes are seen, irrespective of MOI (**Figure 5.5C**). Rankings are also similar, with only slight variations in the top ten genes, and 16/20 of the top expressed genes remain identical. The host expression profile at 24 hours is consistent at an MOI of 1 and 0.1, whereas the expression pattern is more widely distributed at an MOI of 10; again with MT-RNR1 and MT-RNR2 exhibiting lower expression (**Figure 5.5D**). Although the top ranked genes exhibit more variability within their rankings compared to 1 hour, 90% of the top expressed host genes appear at both time points. Functional characterisation of the genes shows their involvement with general cell-based growth events, such as ribosomal-based processes, metabolism, and cytoskeletal components (**Supplementary File 1**). Many top ranked host genes are also non-protein coding as identified by an asterisk (*). However, with limited annotation available, their characterisation in to infection-association functions are limited. Of the annotated non-protein coding genes, they appear to be involved with general cell regulatory processes. Only seven chlamydial genes overlap both time points, which was anticipated, as two different biological events are occurring at these times, including infection mechanisms at 1 hour, and growth-related processes at 24 hours. Functional characterisation of these overlapping genes identifies membrane proteins and transcription/translation machinery, which are needed throughout the developmental cycle (**Supplementary File 1**).

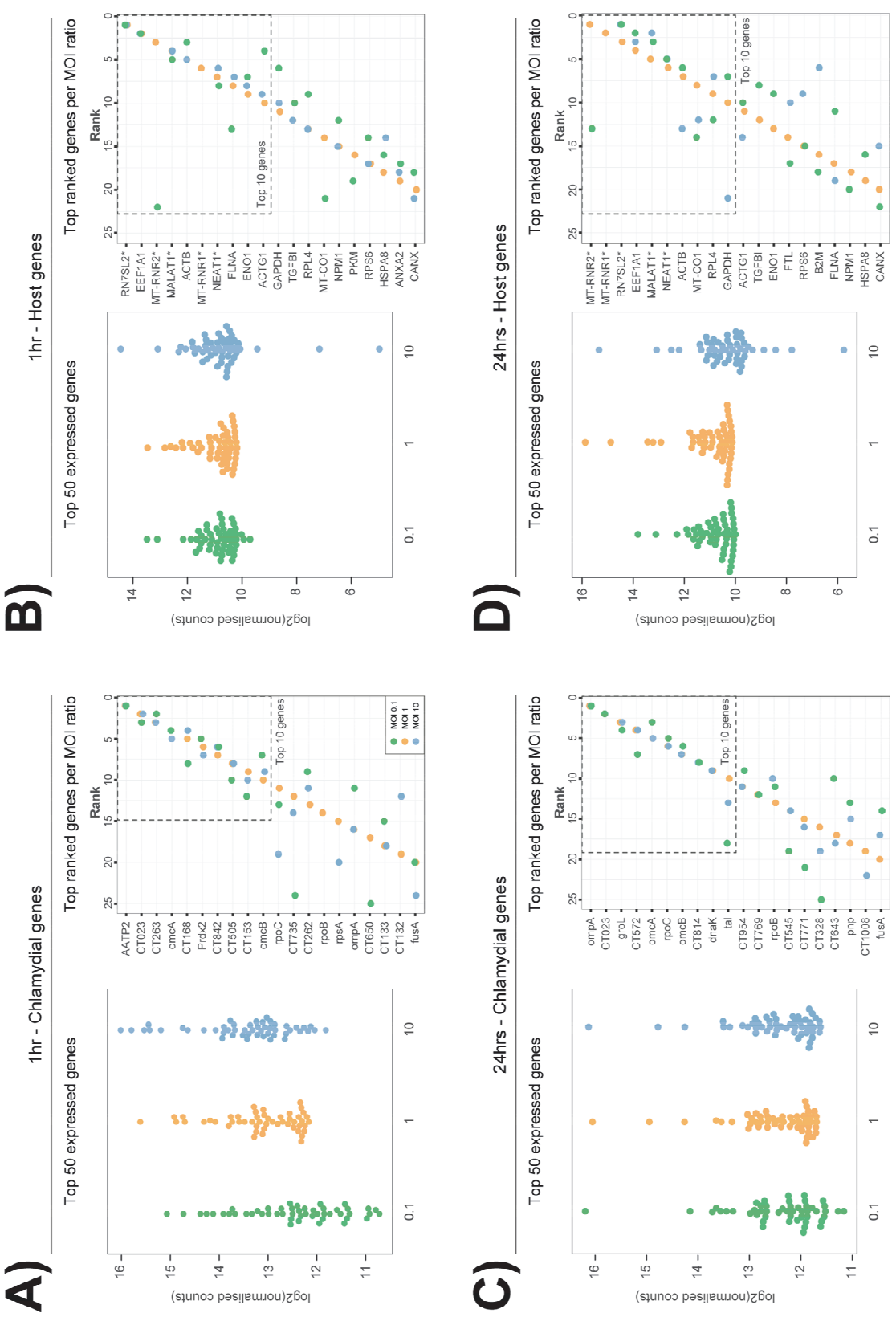

A)
1hr - Chlamydial genes
Top 50 expressed genes
log2(normalised counts)
Top ranked genes per MOI ratio
Rank
Top 10 genes
MOI 0.1
MOI 1
MOI 10
AATP2
CT023
CT263
omcA
CT168
Prdx2
CT842
CT505
CT153
omcB
rpoC
CT735
CT262
rpoB
rpsA
ompA
CT650
CT133
CT132
fusA
B)
1hr - Host genes
Top 50 expressed genes
log2(normalised counts)
Top ranked genes per MOI ratio
Rank
Top 10 genes
RN7SL2*
EEF1A1
MT-RNR2*
MALAT1*
ACTB
MT-RNR1*
NEAT1*
FLNA
ENO1
ACTG1
GAPDH
TGFBI
RPL4
MT-CO1
NPM1
PKM
RPS6
HSPA8
ANXA2
CANX
C)
24hrs - Chlamydial genes
Top 50 expressed genes
log2(normalised counts)
Top ranked genes per MOI ratio
Rank
Top 10 genes
ompA
CT023
groL
CT572
omcA
rpoC
omcB
CT814
cnaK
tal
CT954
CT769
rpoB
CT545
CT771
CT328
CT643
pnp
CT008
fusA
D)
24hrs - Host genes
Top 50 expressed genes
log2(normalised counts)
Top ranked genes per MOI ratio
Rank
Top 10 genes
MT-RNR2*
MT-RNR1*
RN7SL2*
EEF1A1
MALAT1*
NEAT1*
ACTB
MT-CO1
RPL4
GAPDH
ACTG1
TGFBI
ENO1
FTL
RPS6
B2M
FLNA
NPM1
HSPA8
CANX

Figure 5.5: Top 25 expressed host and chlamydial genes across MOIs

Each quadrant contains two graphs: the first showing the top 50 expressed genes across each MOI (taken from an MOI of 1), while the second ranks those same genes for direct comparison. **A)** Chlamydial expression at 1 hour shows a slight upwards trend as the MOI increases, with high similarity in the ranked order. **B)** Host expression at 1 hour shows consistent expression across MOIs apart from two mitochondrial genes (MT-RNR1 and MT-RNR2) with low expression at an MOI of 10. Rankings of the top ten genes are also highly similar. C) At 24 hours, chlamydial expression seems to be less influenced by MOI. D) Host expression at 24 hours shows a slight increase in overall expression, while the MOI of 10 has begun to have more of a varied range of expression compared to 1 hour. Ranked genes also remain highly similar.

5.3.6. Comparative analysis between MOIs show increased expression of inflammatory and immune-based host genes

To compare and contrast how infected host cell expression responds to increased bacterial loads, we examined differentially expressed (DE) genes between the different MOIs. At 1 hour, the majority of genes (87% from 0.1 to 1, and 67% from 1 to 10) exhibited an increase in regulation as the MOI increased (**Figure 5.6A**). By enriching DE genes which are up-regulated and overlap both comparisons, pathways that exhibit an increase in expression as the MOIs increase were identified (**Figure 5.6B**). The same method was applied to down-regulated genes. No continuously down-regulated pathways were identified. The top four up-regulated pathways highlight similar host immune regulated functions that include *(TNF signalling)*, *(NF-κB signalling)*, *(NOD-like receptor signalling)* and *(Cytokine-cytokine receptor interaction)*; with the proinflammatory cytokine TNF exhibiting almost double the

combined score of the next highest. This reflects the known strong immune-based responses to chlamydial infection. Pathways are associated with primary defence mechanisms, thus its plausible that they should exhibit increased expression in concert with bacterial load.

To further examine influential genes underlying these pathways, 'trended-genes' were extracted. The criteria consisted of an expression profile that at least doubled (fold-change >2) for each comparison, in addition to showing a continued increase from an MOI of 0.1 to 10. In total, 46 genes were identified that trended-upwards (**Figure 5.6C**); no genes trended downwards. These trended-genes further highlight that the underlying host-mechanisms to increased infection at initial stages are predominately immune system associated 24/46 (52%), encompassing cytokine signalling, chemokines and interleukins.

The number of DE genes at 24 hours show an even distribution of fold-changes compared to 1 hour, with 49% up-regulated comparing MOIs 0.1 and 1, and 50% comparing 1 and 10 (**Figure 5.6D**). Enriched pathways that are continuously up-regulated include *(TNF signalling)* and *(NF-κB signalling)*, which are the same top two pathways found at 1 hour, and strongly linked to inflammation (Lawrence, 2009). We also see two enriched pathways that become down-regulated as the MOI increases: *(Carbon metabolism)* and the *(Citrate cycle (TCA cycle))* (**Figure 5.6E**). This decrease in key metabolism is likely due to cells prioritising defence over growth as the infection escalates.

Examining trended-genes at 24 hours uncovers 1 gene exhibiting decreased expression (TXNIP), and 14 genes with increased expression (**Figure 5.6E**). TXNIP (Thioredoxin Interacting Protein) is a thiol-oxidoreductase involved in redox regulation which protects cells against oxidative stress (Chutkow et al., 2008). Chlamydial-specific studies have identified an increase in reactive oxygen species (ROS) at early time points, but expression is rapidly reduced shortly afterwards (Boncompain et al., 2010). A further study has suggested that the redox state within a cell could be a regulator in *Chlamydia*-induced apoptosis (Schoier et al.,

2001). However, it is difficult to know if this decreased regulation is directly linked to chlamydial infection and what advantages an oxidative cellular environment would provide at this developmental stage. Genes with increased expression fall into three main categories: cytokines and inflammation (6 genes), viral-based immune response (5 genes), and ubiquitin-related immune responses (3 genes). As anticipated and seen at 1 hour, expression of key immune related genes increases with an increased burden. Only 4 genes overlapped both time points that also increased expression across MOIs (CXCL1, CXCL2, CXCL8 and IL6), indicating their importance as immune mediators against infection.

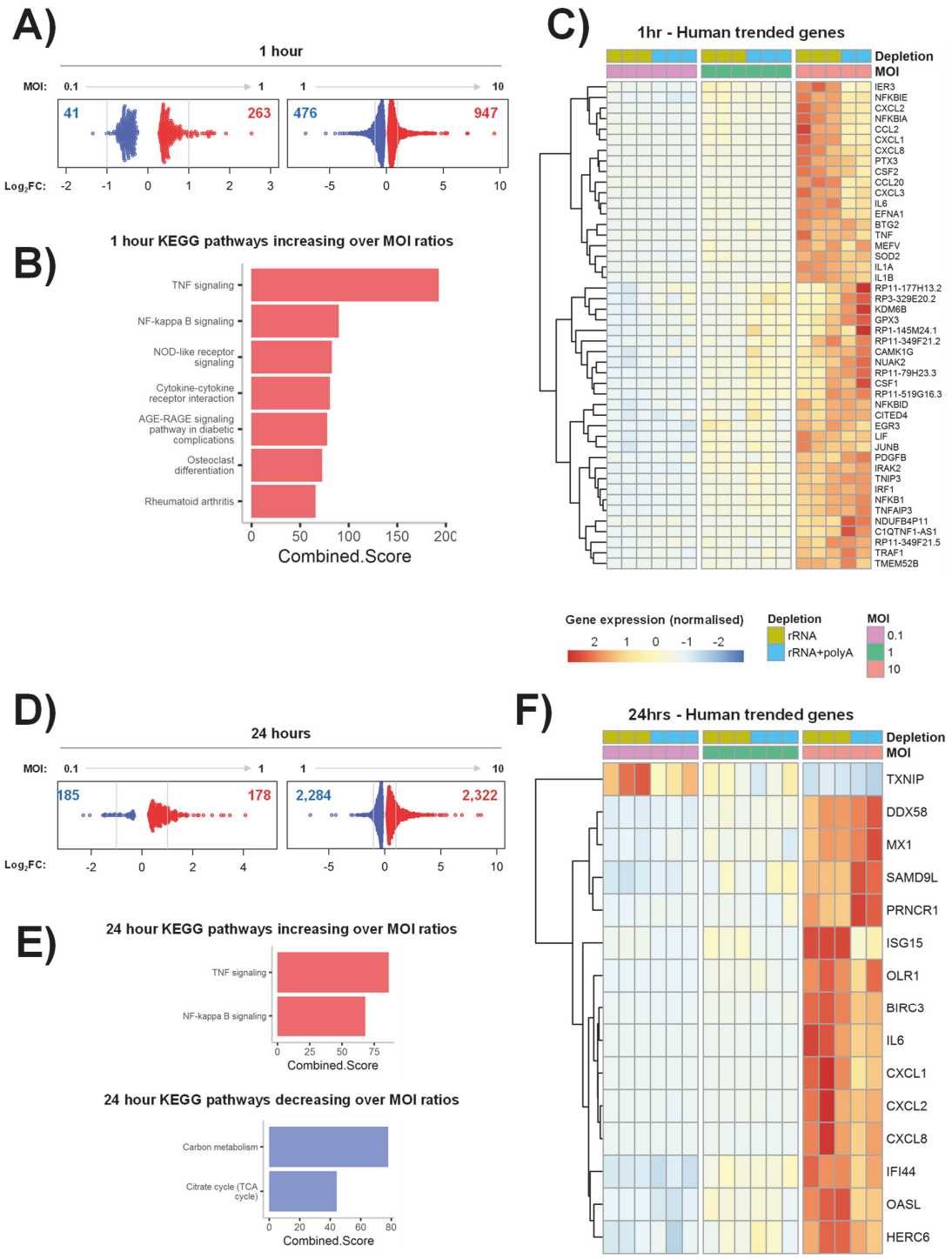


Figure 5.6: Comparison of differentially expressed host genes across MOIs

A) An increase of up-regulated genes at both MOI comparisons (1 vs 0.1 and 10 vs 1) is seen at 1 hour. **B)** Extracting and enriching genes that overlap both comparisons and are also up-regulated, show pathways involved with immune and inflammatory responses. **C)** Trended genes are determined from exhibiting a fold-change > 2 and following the same regulation pattern at both comparisons. At 1 hour, 46 up-regulated 'trended-genes' further highlight the association with the immune system with over 50% of genes grouped in to cytokine signalling, chemokines and interleukins. **D)** At 24 hours the numbers of up and down-regulated genes is much more even (49% and 50%) than 1 hour. **E)** The top two up-regulated pathways are repeated at both time points, while down-regulated pathways are associated with metabolism and likely indicate cells shifting into defence mode as infection increases. **F)** Trended-genes further highlight immune responses, in addition to viral and ubiquitin-related immune responses.

5.3.7. Comparative analysis of chlamydial expression between MOIs

DE genes were also identified in order to explore any chlamydial-based changes attributed to different MOIs. The number of DE genes at 1 hour reflected the underlying minimal expression profiles already identified (**Figure 5.2**), with 47 DE genes comparing MOIs 0.1 and 1, and 23 genes comparing 1 and 10 (**Figure 5.7A**). At 24 hours, the increase in underlying expression resulted in an increase in DE genes, with 81 comparing MOIs 0.1 and 1, while over half (56%) of the chlamydial genome (566/1008 genes) showed a significant change in regulation comparing MOIs 1 and 10 (**Figure 5.7B**).

No chlamydial genes increased across MOIs at either time point, while only two genes decreased: SCLA1|TEF25 (Succinyl-CoA Synthetase) at 1 hour, and CT726 (tRNA) at 24

hours. The decrease of transfer RNAs (tRNA) at 24 hours is slightly surprising, considering they are an important component of translation, and would likely be in abundance during this growth phase of the developmental cycle. Also surprising is a decrease in Succinyl-CoA synthetase, which is involved with the citric acid cycle and cellular metabolism (Phillips et al., 2009). One hypothesis is that, as more EBs are introduced, the likelihood of multiple infections within a cell increases. It is possible that some inclusions are benefitting from effector proteins already circulating within the cell from existing inclusions.

Due to low numbers of DE genes at 3 of the 4 comparisons, enrichment was only possible comparing MOIs 1 and 10 at 24 hours (**Figure 5.7C**). Down-regulated ontologies comprise genes that show decreased expression at a higher MOIs. Results also show unexpected functions such as '*ATP-binding*' and '*Lipid biosynthesis*', which would generally be associated with chlamydial growth. This may highlight the possibility that inclusions may benefit from effector proteins already in existence, likely reducing the need to express these genes and associated processes. Up-regulated genes cover a wider range functions, with half associated with different binding mechanisms facilitating transcription and growth (*RNA-binding*, *rRNA-binding*, *Metal-binding* and *Nucleotide-binding*); which is expected at this stage of the developmental cycle, especially with a ten-fold increase in EBs.

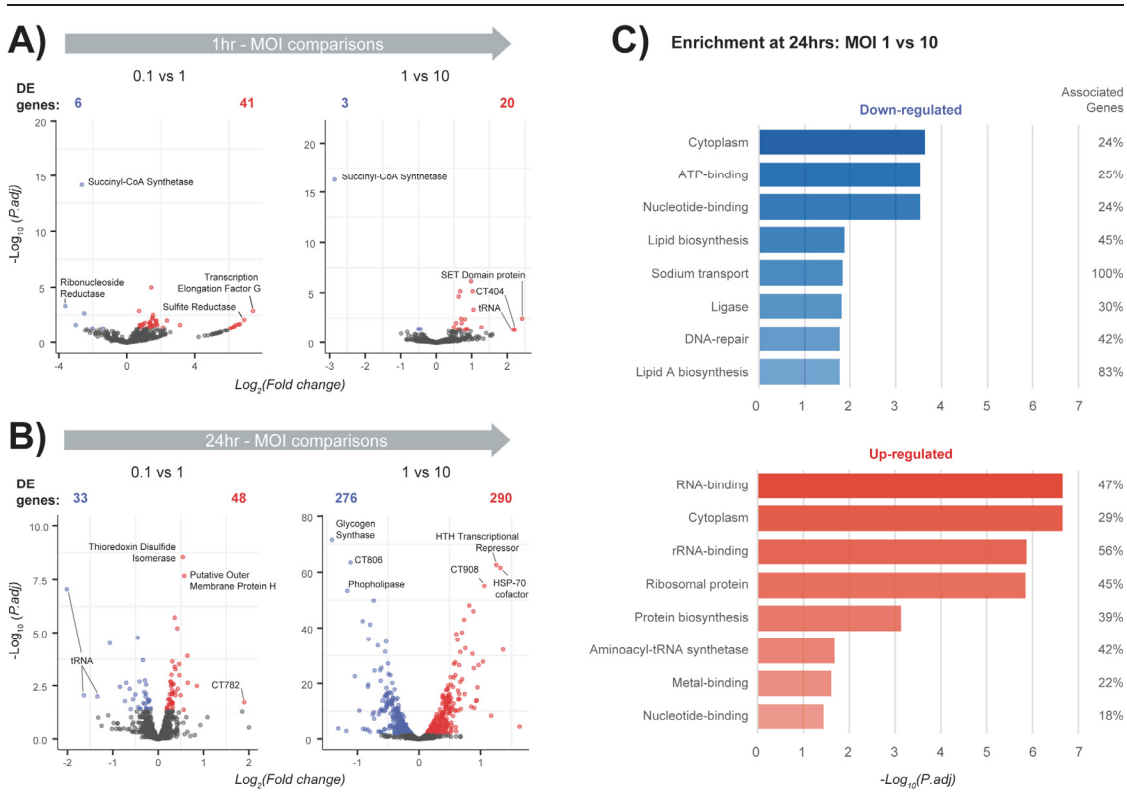


Figure 5.7: Comparison of differentially expressed chlamydial genes across MOIs

A) Volcano plots after differential comparisons between MOIs at 1 hour show minimal differentially expressed (DE) genes. **B)** Differential comparisons at 24 hours show a slight increase in DE genes comparing MOIs 0.1 and 1, with a further considerable increase comparing MOIs 1 and 10. **C)** The increase in DE genes comparing MOIs 1 and 10 at 24 hours allowed enrichment of up and down-regulated genes.

5.4. Discussion

5.4.1. The influence of increasing MOI on host cell gene expression

There is a finite balance when infecting monolayers to accurately measure both host-cell and chlamydial transcriptional responses. This experiment used a standard MOI of 1, in addition to a ten-fold increase (MOI 10) and decrease (MOI 0.1). One reason to increase the MOI is to examine early time points of infection when chlamydial transcripts are in low quantities as seen in (**Figure 5.2A**). In this experiment, when increasing the MOI to 10 at both time points, an increased capture rate of chlamydial transcripts was observed, confirming the suitability for early times (**Figure 5.2A-B**). However a challenge when working with higher MOIs is that some cells may form multiple inclusions which may skew host-cell responses beyond what may be seen in a real-world infection setting (Suchland et al., 2005). When looking at an MOI of 10 at 24 hours, we see over 60% of total captured transcripts from *Chlamydia*. Although this may not be representative of an *in vivo* infection, it is highly useful when focusing on chlamydial-based mechanisms. However, this does raise a question regarding a theoretical maximum proportion of chlamydial reads that can exist within a host cell during the developmental cycle, particularly at the later stages of infection. This was highlighted from **Figure 5.2E** at 24 hours, where we see a single replicate showing a staggering 74% of all transcripts associated with chlamydial expression. However, it was challenging to determine from this experiment if the increase in EBs had a corresponding influence in reducing the length of the developmental cycle due to possible synergistic interactions and shared resources. A future dual-RNA-seq study following the time course of Miyairi et al., 2006, but replacing biovars for MOIs would be intriguing.

5.4.2. Combining depletion methods increases capture rate of chlamydial transcripts

By combining rRNA and polyA depletion methods, we clearly observe an increased capture rate of chlamydial transcripts. These additional transcripts do not appear to be from a small subset of genes dominating capture, but from a wide range expressed genes (**Figure 5.3**). However, host-based expression is affected, with expression of non-protein coding genes increasing (**Figure 5.4**); suggesting it may only be beneficial for future chlamydial-specific sequencing approaching to use both depletion methods.

5.4.3. Experimental and infection-based limitations

While sequencing costs continue to decrease, the cost still influences the number of replicates and the underlying experimental design. One limitation of this design which was not able to be included due to cost constraints, was the addition of uninfected controls at each time point. This would have enabled us to tease out what are general cell regulatory processes and what is actually infection-specific. Theoretically uninfected replicates from the previous dual-RNA-seq experiment (Humphrys et al., 2013) could be used as the time points overlapped, and the samples were sequenced on the same machine and at the same sequencing facility (reducing any confounding effects). However, the data was not available.

An advantage of RNA-seq is the large number of cells that are able to be sequenced in parallel. However, a drawback with infection-based studies is the complexity associated with infection, which include the various developmental stages of host cells, different rates of infection, and that some cells will successfully repel infection. Coupling scRNA-seq (as seen in Chapter 4), but including fluorescent tags to sort cells based on infection status, cell cycle and other identifiable tags (Gedye et al., 2014; Pozarowski and Darzynkiewicz, 2004), would

provide a greater resolution and reduce some of these infection-based limitations. Further resolution could be obtained from *ex vivo* tissues, examining the spatial relation between cells, each cells relative location and their corresponding gene expression. Spatial transcriptomic and imaging techniques such as the Hyperion™ Imaging System allow these types of interrogations for further characterisation of tissue structures, cell types, cell populations and spatially patterned gene expression and proteins associated with infection (Eng et al., 2019; Rodriques et al., 2019; Uraki et al., 2019).

5.4.4. How dynamic is the chlamydial response

Differentially expressed and trended genes (**Figure 5.6**) identified transcriptional responses the host cell uses during an infection, which appears to be from a similar subset of key genes at both times. Genes are associated with immune related pathways, specifically inflammation; with increased expression as the MOI increases. As the concentration of EBs increases provoking this increased immune response, host cells will likely become overwhelmed if the numbers of EBs become too high. We hypothesise this could be an advantage for *Chlamydia* if a large proportion of host cell expression is focused towards immune responses, and if they already have a way of countering these, then other host processes may be easier to interfere with and possibly hijack. We anticipate this would most likely occur at higher MOIs where we have observed the most difference, particularly at 24 hours or latter stages of the developmental cycle.

5.5. Conclusion

This work highlights how future RNA-seq studies that examine bacteria-infected mammalian cells could increase sequence capture rates by combining rRNA and polyA depletion methods. This is particularly relevant for expression studies examining early time points, irrespective of the infecting bacterial agent, as low bacterial expression is generally observed. Three different MOIs highlighted that significantly more chlamydial transcripts were captured at both time points when using an MOI of 10. However, although a higher MOI may be useful for capturing chlamydial-specific biology *in vitro*, the increased burden on host cells may not be representative *in vivo* infections. Overall, these outcomes can help influence future NGS-based experimental designs to achieve more specific infection-related biological outcomes, particularly from *Chlamydia*-infected cells.

5.6. Supplementary files

Supplementary File 5.1 Functional characterisation of highly expressed genes overlapping all MOI ratios

Functional characterisation from top 25 highly expressed genes (host and chlamydial) that overlap all three MOI ratios.

Supplementary File 5.1.xlsx

5.7. References

AbdelRahman, Y.M., and Belland, R.J. (2005). The chlamydial developmental cycle. FEMS Microbiology Reviews *29*, 949-959.

Albrecht, M., Sharma, C.M., Reinhardt, R., Vogel, J., and Rudel, T. (2010). Deep sequencing-based discovery of the *Chlamydia trachomatis* transcriptome. Nucleic acids research *38*, 868-877.

Andrews, S. (2010). FastQC A Quality Control tool for High Throughput Sequence Data.

Barnett, D.W., Garrison, E.K., Quinlan, A.R., Strömberg, M.P., and Marth, G.T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics *27*, 1691-1692.

Bastidas, R.J., Elwell, C.A., Engel, J.N., and Valdivia, R.H. (2013). Chlamydial intracellular survival strategies. Cold Spring Harb Perspect Med *3*, a010256-a010256.

Beaulieu, L.M., Clancy, L., Tanriverdi, K., Benjamin, E.J., Kramer, C.D., Weinberg, E.O., He, X., Mekasha, S., Mick, E., Ingalls, R.R.*, et al.* (2015). Specific Inflammatory Stimuli Lead to Distinct Platelet Responses in Mice and Humans. PloS one *10*, e0131688-e0131688.

Belland, R.J., Nelson, D.E., Virok, D., Crane, D.D., Hogan, D., Sturdevant, D., Beatty, W.L., and Caldwell, H.D. (2003a). Transcriptome analysis of chlamydial growth during IFN-gamma-mediated persistence and reactivation. Proceedings of the National Academy of Sciences of the United States of America *100*, 15971-15976.

Belland, R.J., Zhong, G., Crane, D.D., Hogan, D., Sturdevant, D., Sharma, J., Beatty, W.L., and Caldwell, H.D. (2003b). Genomic transcriptional profiling of the developmental cycle of *Chlamydia trachomatis*. Proceedings of the National Academy of Sciences of the United States of America *100*, 8478-8483.

Betts-Hampikian, H.J., and Fields, K.A. (2010). The Chlamydial Type III Secretion Mechanism: Revealing Cracks in a Tough Nut. Frontiers in microbiology *1*, 114-114.

Blighe, K., and Lewis, M. (2018). PCAtools: everything Principal Components Analysis. https://githubcom/kevinblighe/PCAtools.

Boncompain, G., Schneider, B., Delevoye, C., Kellermann, O., Dautry-Varsat, A., and Subtil, A. (2010). Production of reactive oxygen species is turned on and rapidly shut down in epithelial cells infected with *Chlamydia trachomatis*. Infection and immunity *78*, 80-87.

Brunham, R.C., Binns, B., McDowell, J., and Paraskevas, M. (1986). *Chlamydia trachomatis* infection in women with ectopic pregnancy. Obstetrics and gynecology *67*, 722-726.

Burton, M.J., and Mabey, D.C.W. (2009). The Global Burden of Trachoma: A Review. PLoS Neglected Tropical Diseases *3*, e460-e460.

Chutkow, W.A., Patwari, P., Yoshioka, J., and Lee, R.T. (2008). Thioredoxin-interacting Protein (Txnip) Is a Critical Regulator of Hepatic Glucose Production. Journal of Biological Chemistry *283*, 2397-2406.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England) *29*, 15-21.

Elwell, C., Mirrashidi, K., and Engel, J. (2016). *Chlamydia* cell biology and pathogenesis. Nat Rev Microbiol *14*, 385-400.

Eng, C.-H.L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.-C*., et al.* (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature *568*, 235-239.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047-3048.

Gedye, C.A., Hussain, A., Paterson, J., Smrke, A., Saini, H., Sirskyj, D., Pereira, K., Lobo, N., Stewart, J., Go, C*., et al.* (2014). Cell Surface Profiling Using High-Throughput Flow Cytometry: A Platform for Biomarker Discovery and Analysis of Cellular Heterogeneity. PloS one *9*, e105602.

Gentleman, R., Carey, V., Huber, W., and Hahne, F. (2018). genefilter: methods for filtering genes from microarray experiments - R package version 1.64.0.

Grieshaber, S., Grieshaber, N., Yang, H., Baxter, B., Hackstadt, T., and Omsland, A. (2018). Impact of Active Metabolism on *Chlamydia trachomatis* Elementary Body Transcript Profile and Infectivity. Journal of Bacteriology *200*, e00065-00018.

Heyer, E.E., Deveson, I.W., Wooi, D., Selinger, C.I., Lyons, R.J., Hayes, V.M., O'Toole, S.A., Ballinger, M.L., Gill, D., Thomas, D.M.*, et al.* (2019). Diagnosis of fusion genes using targeted RNA sequencing. Nature Communications *10*, 1388-1388.

Huang, C., Shi, J., Guo, Y., Huang, W., Huang, S., Ming, S., Wu, X., Zhang, R., Ding, J., Zhao, W.*, et al.* (2017). A snoRNA modulates mRNA 3′ end processing and regulates the expression of a subset of mRNAs. Nucleic acids research *45*, 8647-8660.

Humphrys, M.S., Creasy, T., Sun, Y., Shetty, A.C., Chibucos, M.C., Drabek, E.F., Fraser, C.M., Farooq, U., Sengamalay, N., Ott, S.*, et al.* (2013). Simultaneous Transcriptional Profiling of Bacteria and Their Host Cells. PloS one *8*, e80597-e80597.

Hybiske, K., and Stephens, R.S. (2007). Mechanisms of host cell exit by the intracellular bacterium <em>Chlamydia</em&gt. Proceedings of the National Academy of Sciences *104*, 11430 LP-11435.

Jiang, Y., Jiang, Y.-Y., Xie, J.-J., Mayakonda, A., Hazawa, M., Chen, L., Xiao, J.-F., Li, C.-Q., Huang, M.-L., Ding, L.-W.*, et al.* (2018). Co-activation of super-enhancer-driven CCAT1 by TP63 and SOX2 promotes squamous cancer progression. Nature Communications *9*, 3619-3619.

Johnson, R.M., Yu, H., Strank, N.O., Karunakaran, K., Zhu, Y., and Brunham, R.C. (2018). B Cell Presentation of Chlamydia Antigen Selects Out Protective CD4γ13 T Cells: Implications for Genital Tract Tissue-Resident Memory Lymphocyte Clusters. Infection and immunity *86*, e00614-00617.

Kukurba, K.R., and Montgomery, S.B. (2015). RNA Sequencing and Analysis. Cold Spring Harbor protocols *2015*, 951-969.

Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A.*, et al.* (2016). Enrichr: a

comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research *44*, W90-W97.

Kumar, N., Lin, M., Zhao, X., Ott, S., Santana-Cruz, I., Daugherty, S., Rikihisa, Y., Sadzewicz, L., Tallon, L.J., Fraser, C.M.*, et al.* (2016). Efficient Enrichment of Bacterial mRNA from Host-Bacteria Total RNA Samples. Sci Rep *6*, 34850-34850.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods *9*, 357-359.

Lawrence, T. (2009). The nuclear factor NF-kappaB pathway in inflammation. Cold Spring Harb Perspect Biol *1*, a001651-a001651.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923-930.

Lyons, J.M., Ito Jr, J.I., Peña, A.S., and Morré, S.A. (2005). Differences in growth characteristics and elementary body associated cytotoxicity between *Chlamydia trachomatis* oculogenital serovars D and H and *Chlamydia muridarum*. Journal of clinical pathology *58*, 397-401.

Menon, S., Timms, P., Allan, J.A., Alexander, K., Rombauts, L., Horner, P., Keltz, M., Hocking, J., and Huston, W.M. (2015). Human and Pathogen Factors Associated with *Chlamydia trachomatis*-Related Infertility in Women. Clinical microbiology reviews *28*, 969-985.

Miyairi, I., Mahdi, O.S., Ouellette, S.P., Belland, R.J., and Byrne, G.I. (2006). Different Growth Rates of *Chlamydia trachomatis* Biovars Reflect Pathotype. The Journal of infectious diseases *194*, 350-357.

O'Connell, C.M., AbdelRahman, Y.M., Green, E., Darville, H.K., Saira, K., Smith, B., Darville, T., Scurlock, A.M., Meyer, C.R., and Belland, R.J. (2011). Toll-like receptor 2 activation by *Chlamydia trachomatis* is plasmid dependent, and plasmid-responsive

chromosomal loci are coordinately regulated in response to glucose limitation by C. trachomatis but not by C. muridarum. Infection and immunity *79*, 1044-1056.

O'Neil, D., Glowatz, H., and Schlumpberger, M. (2013). Ribosomal RNA depletion for efficient use of RNA-seq capacity. Curr Protoc Mol Biol *Chapter 4*, Unit 4.19-Unit 14.19.

Paul, L., Comstock, J., Edes, K., and Schlaberg, R. (2018). Gestational Psittacosis Resulting in Neonatal Death Identified by Next-Generation RNA Sequencing of Postmortem, Formalin-Fixed Lung Tissue. Open Forum Infectious Diseases *5*.

Phillips, D., Aponte, A.M., French, S.A., Chess, D.J., and Balaban, R.S. (2009). Succinyl-CoA synthetase is a phosphate target for the activation of mitochondrial metabolism. Biochemistry *48*, 7140-7149.

Phillips, S., Quigley, B.L., Aziz, A., Bergen, W., Booth, R., Pyne, M., and Timms, P. (2019). Antibiotic treatment of Chlamydia-induced cystitis in the koala is linked to expression of key inflammatory genes in reactive oxygen pathways. PloS one *14*, e0221109-e0221109.

Pozarowski, P., and Darzynkiewicz, Z. (2004). Analysis of cell cycle by flow cytometry. Methods Mol Biol *281*, 301-311.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Reyburn, H. (2016). WHO Guidelines for the Treatment of *Chlamydia trachomatis*. WHO *340*, c2637-c2637.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol *11*, R25-R25.

Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F., and Macosko, E.Z. (2019). Slide-seq: A scalable

technology for measuring genome-wide expression at high spatial resolution. Science (New York, NY) *363*, 1463-1467.

Saka, H.A., and Valdivia, R.H. (2010). Acquisition of nutrients by *Chlamydiae*: unique challenges of living in an intracellular compartment. Current opinion in microbiology *13*, 4-10.

Schachter, J., and Caldwell, H.D. (1980). *Chlamydiae*. Annual review of microbiology *34*, 285-309.

Schoier, J., Ollinger, K., Kvarnstrom, M., Soderlund, G., and Kihlstrom, E. (2001). *Chlamydia trachomatis*-induced apoptosis occurs in uninfected McCoy cells late in the developmental cycle and is regulated by the intracellular redox state. Microbial pathogenesis *31*, 173-184.

Scidmore, M.A., Fischer, E.R., and Hackstadt, T. (2003). Restricted fusion of *Chlamydia trachomatis* vesicles with endocytic compartments during the initial stages of infection. Infection and immunity *71*, 973-984.

Shutt, T.E., and Shadel, G.S. (2010). A compendium of human mitochondrial gene expression machinery with links to disease. Environmental and molecular mutagenesis *51*, 360-379.

Suchland, R.J., Rockey, D.D., Weeks, S.K., Alzhanov, D.T., and Stamm, W.E. (2005). Development of secondary inclusions in cells infected by *Chlamydia trachomatis*. Infection and immunity *73*, 3954-3962.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P.*, et al.* (2018). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research *47*, D607-D613.

Tan, C., Hsia, R.-c., Shou, H., Haggerty, C.L., Ness, R.B., Gaydos, C.A., Dean, D., Scurlock, A.M., Wilson, D.P., and Bavoil, P.M. (2009). *Chlamydia trachomatis*-Infected Patients Display Variable Antibody Profiles against the Nine-Member Polymorphic Membrane Protein Family. Infection and Immunity *77*, 3218 LP-3226.

Teder, H., Koel, M., Paluoja, P., Jatsenko, T., Rekker, K., Laisk-Podar, T., Kukuškina, V., Velthut-Meikas, A., Fjodorova, O., Peters, M.*, et al.* (2018). TAC-seq: targeted DNA and RNA sequencing for precise biomarker molecule counting. npj Genomic Medicine *3*, 34-34.

Uraki, R., Hastings, A.K., Marin-Lopez, A., Sumida, T., Takahashi, T., Grover, J.R., Iwasaki, A., Hafler, D.A., Montgomery, R.R., and Fikrig, E. (2019). Aedes aegypti AgBR1 antibodies modulate early Zika virus infection of mice. Nature Microbiology *4*, 948-955.

Vats, V., Agrawal, T., Salhan, S., and Mittal, A. (2007). Primary and secondary immune responses of mucosal and peripheral lymphocytes during *Chlamydia trachomatis* infection. Pathogens and disease *49*, 280-287.

Wali, S., Gupta, R., Veselenak, R.L., Li, Y., Yu, J.-J., Murthy, A.K., Cap, A.P., Guentzel, M.N., Chambers, J.P., Zhong, G.*, et al.* (2014). Use of a Guinea pig-specific transcriptome array for evaluation of protective immunity against genital chlamydial infection following intranasal vaccination in Guinea pigs. PloS one *9*, e114261-e114261.

Wang, A., Al-Kuhlani, M., Johnston, S.C., Ojcius, D.M., Chou, J., and Dean, D. (2013). Transcription factor complex AP-1 mediates inflammation initiated by *Chlamydia pneumoniae* infection. Cell Microbiol *15*, 779-794.

West, Jason A., Davis, Christopher P., Sunwoo, H., Simon, Matthew D., Sadreyev, Ruslan I., Wang, Peggy I., Tolstorukov, Michael Y., and Kingston, Robert E. (2014). The Long Noncoding RNAs NEAT1 and MALAT1 Bind Active Chromatin Sites. Molecular cell *55*, 791-802.

Westermann, A.J., Gorski, S.A., and Vogel, J. (2012). Dual RNA-seq of pathogen and host. Nature Reviews Microbiology *10*, 618-618.

Wysoker, A., Tibbetts, K., and Fennell, T. (2013). Picard tools.

Yeung, A.T.Y., Hale, C., Lee, A.H., Gill, E.E., Bushell, W., Parry-Smith, D., Goulding, D., Pickard, D., Roumeliotis, T., Choudhary, J.*, et al.* (2017). Exploiting induced pluripotent stem cell-derived macrophages to unravel host factors influencing *Chlamydia trachomatis* pathogenesis. Nature Communications *8*, 15013-15013.

Chapter 6

General discussion and future directions

6.1. General discussion

6.1.1. Functional characterisation of chlamydial genes and their relationship with host cells is not completely understood

Chlamydia spp. can infect a wide range of hosts and tissue types, resulting in different disease pathologies. *C. trachomatis* and *C. pneumoniae* primarily infect humans with disease outcomes causing major socio-economic burdens. Recently, *Chlamydia* spp. have been identified crossing host barriers and infecting new hosts, specifically humans. The majority of these cases have been identified in farmers who are in close contact with livestock and thus a different range of pathogens (De Puysseleyr et al., 2017; Lagae et al., 2014). The implications of these zoonotic transmissions may bring about more disease manifestations, adding more challenges and increased burden from an already wide-spread pathogen.

Although the chlamydial genome was one of the first genomes to be sequenced, we still do not have a comprehensive understanding of the functionality from all the encoded genes and effector proteins, due to the complexities related to its developmental cycle. Many studies have predominantly examined single mechanisms, such as attachment, cell entry, host interactions and cell exit, due to the time-consuming processes involved with cell culturing and experimental techniques. As a result, many of these approaches also have limited scope for genetic modifications to enable confirmation of gene functions. Due to these restrictions, identifying and characterising chlamydial genes, effector proteins and many of the host cell interactions has been limited. For example, in the genome of the *Chlamydia* E CHARM001 isolate that was used for these experiments, 356 (35%) hypothetical proteins remain uncharacterised. In addition, many of the annotated genes involvement in targeting host processes remain unknown.

Genomes across chlamydial species are highly similar, but encode different effector proteins that may reflect a diversity of infection processes specific to different hosts, tissues and/or cells. Examining single biological processes and interactions through traditional approaches are still needed and will continue to identify novel biological functions.

An advantage of using NGS approaches is the identification of infection-specific genes and pathways genome-wide, which can often highlight novel complex genetic interactions that have previously been uncharacterised or are unknown. Results from these studies can therefore provide directions for specific targets for further characterisation. This may be through traditional approaches, or more recent transformation-based systems (transposon mutagenesis and CRISPR) where genes can be knocked out more efficiently to create genetically transformed mutants which can then be used to probe infection mechanisms and host-interactions. Currently there is a wealth of knowledge from different sequencing approaches examining chlamydial infection genome-wide, containing gene-set enrichments and pathway analyses as highlighted in **Chapter 1**. With sequencing costs decreasing and more efficient and cost-effective sequencing approaches being developed, more data will be generated that can be used to help decide which genes and pathways to target and investigate infection processes further.

Therefore, the next period of chlamydial research has the potential to be a ‘golden age’, where genome-wide NGS approaches will be combined with transformation systems, animal models, human tissues and other methods such as microscopy, to efficiently unravel the full repertoire of effectors and other chlamydial proteins that govern the likely diverse interactions with the infected host cells and tissues.

6.1.2. Single cell RNA-seq

6.1.2.1. Key outcomes and summary

This pilot dataset is the first single-cell resolution study examining the host cell response to chlamydial infection, and is amongst a limited number of studies examining bacterial infections of host cells at the single cell level. Despite the limitations arising from a relatively low number of single cells, the analyses do highlight infection-specific host cell biology, including two distinct clusters separating 3 hour cells from 6 and 12 hours. This confirms that host cell responses to infection can be distinguished by time. Pseudotime analysis identified a possible infection-specific cellular trajectory for *Chlamydia*-infected cells, and differential expression identified temporally expressed genes involved with cell cycle regulation, innate immune responses, cytoskeletal components, lipid biosynthesis and cellular stress. This study also highlights the complex nature of infections, allowing considerations for future *in vitro* experiments, but also more complex disease models to be more fully explored, such as *ex vivo* experiments to capture different cell types and the full spectrum of the immune response.

6.1.2.2. Single cell capture rates

The average sequencing depth per cell in this experiment was high (~2.8 million reads) in comparison with other studies. Reviews recommend 500,000-1,000,000 reads is sufficient to capture the majority of expressed genes at single cell resolution (Haque et al., 2017a; Pollen et al., 2014). When planning single cell experiments, there will always be a compromise between sequencing more cells, or sequencing fewer cells to a greater depth. In an ideal world, costs would not be a factor, and both could be achieved. The current focus of much single cell biology has been discovering and characterising new cell types, creating cell atlases from different parts of the body, and from entire organisms (Grubman et al., 2019; Regev et al., 2017; Van Hove et al., 2019). For these experiments, obtaining a higher yield of cells has

been beneficial to uncover as many cell types as possible, and has thus been the priority, as seen with some of the latest published experiments generating over 2 million cells (Cao et al., 2019).

Although many reviews claim to have an answer for the question: "how much depth is required to capture a cells expression profile" (Haque et al., 2017b; Rizzetto et al., 2017; Streets and Huang, 2014), this number is most likely to be underestimated. For example, to comprehensively capture the range of transcripts within a cell includes capturing all types of transcripts. Currently there are two main limitations: 1) Current protocols only capture polyadenylated transcripts, missing a large proportion of non-coding transcripts and essentially all bacterial transcripts. 2) Alternatively spliced transcripts are not typically examined (possibly due to low sequencing depth (Nguyen et al., 2018)), and, if we consider that over 95% of multi-exon human genes undergo alternative splicing (Chen and Manley, 2009), we are missing a large proportion of the underlying transcripts. A further point is, at what point do we stop discovering new cell types? Particularly when we are now capable of generating > 1 million cells, but when we are only capturing < 1 million reads per cell. Therefore, a change in direction is likely to occur, where we will see a similar number of cells in future experiments, but at greater sequencing depths.

6.1.2.3. Impacts of the cell cycle on chlamydial infection

A range of cell cycle states were identified across the three time points and within each condition. These expression differences were controlled for as best as possible by treating them as a confounding effect and removing their influence via bioinformatic means. However, we know that chlamydial infection actively promotes spindle defects, leading to chromosome and genetic instability, including centrosome abnormalities (Johnson et al., 2009; Knowlton et al., 2011). Infected cells can still grow and divide, however the burden of infection causes

these cells to proliferate more slowly than uninfected cells (uninfected epithelial cells take approximately 3-4 days). This could also be a direct result of an infection strategy, or it could be an off-target effect of infection with possibly no benefit to the bacteria. This would be an intriguing study to control for different stages of the developmental cycle and monitor infection progression throughout cell division. This could also help answer the question: "are cells more susceptible to being infected when they are in the process of mitosis or not?". Alternative approaches to bioinformatically controlling the cell cycle would be to use cell surface markers associated with cell division, which can be identified and separated with fluorescence-activated cell sorting (FACS) (Kim and Sederstrom, 2015). There are also chemical-assisted methods that can arrest and release at different stages, ensuring all cells are synchronised at specified cell cycle phases (Johnson et al., 2009).

6.1.2.4. Heterogeneity of earlier time points

As identified from different clustering analyses, the 3 hour cells showed considerable separation from cells at 6 and 12 hours. This is likely due to the highly asynchronous nature of attachment, uptake, and inclusion formation compared to more similar growth-based events at 6 and 12 hours. 3 hours was the earliest possible time point that could be chosen due to the timeframes involved with cell isolation and preparation using the Fluidigm C1 machine. With newer and more efficient droplet-based approaches (Zhang et al., 2019), it would be interesting to sequence earlier time points to see if they displayed more heterogeneity than what was observed at 3 hours, or if earlier time points clustered together. If enough cells were used, this data could help identify how similar, or how diverse early infection processes such as host-cell entry and inclusion formation are.

6.1.2.5. Infection-based considerations and challenges

Throughout the course of infection, the host cell expression profile will change relative to the stage of the developmental cycle from the infecting pathogen. Changes can also be induced due to the complex nature of infections, where multiple cell states will likely occur within each stage. These include that more than one EB attached and entered the host cell, some cells were successful in eliminating the bacteria before fixation, the developmental cycle timing was vastly different to what was expected, or the cell was able to remove the inclusion at some point after an EB had entered. Some of these biological differences and similarities were highlighted after clustering in the PCA plots, revealing a range of different cell expression profiles and cell cycle states, but also revealing a high degree of similarity between conditions and time points at 6 and 12 hours. However, without either experimentally controlling for some of these factors, having a visual representation of each cell from microscopy, or additional fluorescent-based cell marker data, it is highly challenging to definitively predict events such as infection status and cell cycle stages. To help resolve this issue, future experiments could extract an internal subset or matched split of cells from the same experiment. These cells could be fixed and measured to confirm actual proportions of cellular infection status.

6.1.3. FAIRE-seq

6.1.3.1. Key outcomes and summary

By using the FAIRE protocol, host-cell chromatin accessibility was examined in response to an *in vitro C. trachomatis* infection. This is the first time chromatin accessibility dynamics have been examined in the chlamydial filed, and is amongst a limited number of studies examining host-cell responses to bacterial infections. Four time points spanned the *in vitro* developmental cycle, identifying both conserved and distinct temporal changes genome-wide. Differentially accessible chromatin regions were linked to genomic features and genes associated with immune responses, re-direction of host cell nutrients, intracellular signalling, cell-cell adhesion, extracellular matrix, metabolism and apoptosis. Temporally-enriched transcription factors from the same regions identified different Krüppel-like-factors (KLFs) which are ubiquitously expressed in reproductive tissues and associated with a variety of uterine pathologies; identifying a novel association with chlamydial infection. This work provides another perspective to the complex response to chlamydial infection, and will inform further studies of transcriptional regulation and the epigenome in *Chlamydia*-infected human cells and tissues.

6.1.3.2. Assumptions about open chromatin

When capturing the state of chromatin (open or closed) associated with a genomic feature, we are assuming that open areas are being actively transcribed, exposing regions to transcription factors and RNA polymerases. Because sequencing protocols are only capturing a snapshot in time, the dynamic nature of chromatin accessibility may be obscured. Therefore, regions of open chromatin could theoretically not be actively transcribed, as chromatin could be in the process of re-condensing, being repaired, insulated or silenced (Lynch and Rusche, 2009; Vignali et al., 2000; West et al., 2002). This is an interesting question and currently cannot be

answered by examining chromatin accessibility data alone. One solution to help rectify this would be to compare expression data from RNA-seq and other genome-scale analyses from the same samples, overlaying the corresponding expression patterns and highlighting exactly where open or closed chromatin leads to increased or decreased expression. It should be noted that although Chapters 4 and 5 do contain expression data, overlaying these data sets was either not possible (due to not having the correct controls) or did not yield any significant results. This is discussed in more detail in Section 6.1.5.

Another consideration is the proportion of open and closed regions relative to what the underlying protocol was designed to capture. For example, the FAIRE and other chromatin protocols capture fragments of open chromatin; therefore, most results are likely biased towards capturing open chromatin. This may be seen in the data presented here (**Figure 3.2**).

6.1.3.3. Location of significant peaks genome-wide

Different chromatin accessibility protocols exist, each with their advantages and disadvantages as discussed in **Chapter 1**. With FAIRE-seq, peaks are often broader that its counterparts, and show a higher coverage at enhancer regions over promoter regions (Kumar et al., 2013). However, annotation of peaks from this data did not show this, with approximately the same number of enhancers (4%) and promoters (5%) (**Figure 6.1**).

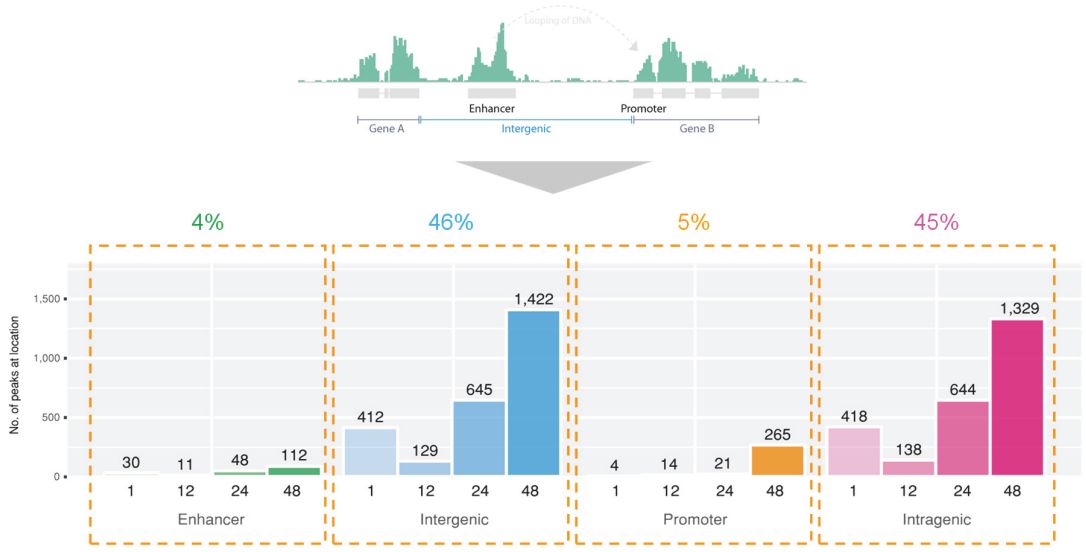


Figure 6.1: Annotation of significant peaks

Annotation of significant differential chromatin accessible peaks from FAIRE-seq data. Peaks were annotated based on their proximity to their closest feature, and summarised here as a promoter (5%), enhancer (4%) or intragenic region (45%). Peaks were categorised as intergenic if there were no closely annotated biological features, which resulted in 46% of the overall significant peaks being unable to be annotated.

Figure 6.1 also highlights that most peaks were located in intergenic regions (46%), which currently have no associated biological annotation. This is not an isolated phenomenon, as the majority of chromatin accessibility studies across different organisms including humans, identify large proportions of peaks that are biologically unknown. In this dataset, the intergenic peaks account for almost half of the data, highlighting that the full range of regulatory mechanisms influencing gene transcription are still to be discovered.

6.1.3.4. Bioinformatic challenges

Analysis of this dataset often yielded small subsets of genes from which further enrichment and/or biological inference was challenging. This included the conserved host response, time-specific gene expression, promoters and enhancers. The primary limitation is that software methods work best with higher gene numbers. When only low numbers are available, often no significant pathways are detected, or marginally significant pathways are biased based on a small number of reoccurring genes. To overcome this, genes from each subset were searched for and compared against different online sources, such as WikiGenes (Hoffmann, 2008), UniProt (UniProt, 2008), NCBI (Pruitt et al., 2007) and different literature. Annotation was based on consensus information, and then grouped based biologically similar themes, such as metabolism. Although some websites such as WikiGenes do have this functionality, many of their sources were not currently up to date, particularly in fields related to bacterial infections. To help solve this issue, perhaps training a machine learning algorithm to search through literature and these databases, could help perform this task much quicker and help collate the most up to date annotative information.

A further challenge is the process of identifying TF binding motifs and the associated TF. The main issue is the lack of consensus between different software, often resulting in highly varied motif signatures and TFs that are limited to their internal databases or curated lists. To account for this, strict filters and thresholds were created as discussed in the methods section of **Chapter 3**, removing any ambiguous motifs, and ensuring that enriched TFs overlapped multiple online sources. When examining similar studies, many were not so strict with filtering or did not consult more than one annotative source. This highlights an area for considerable improvement, and the field would benefit from more up to date comparative reviews, particularly using different chromatin accessibility protocols.

6.1.4. Dual RNA-seq

6.1.4.1. Key outcomes and summary

This dataset is amongst a growing number of dual RNA-seq studies simultaneously examining host cell and bacterial expression, and is only the second study in the chlamydial field (Humphrys et al., 2013). Two time points were examined using three different MOIs and two different RNA depletion methods, to understand the influence the MOI has on the developmental cycle, and the effect of depletion method on sequence capture rates. Analysis showed that an MOI of 10 captures significantly more transcripts than 0.1 and 1 at both time points, and is more beneficial for capturing chlamydial transcripts. Combining depletion methods (polyA and rRNA) increases the capture rate of chlamydial transcripts, but impacts host-cell expression. Comparative analysis between MOIs show increased expression of inflammatory and immune-based genes, while chlamydial expression is more dynamic relative to the developmental stage. This work highlights how future bacterial-specific RNA-seq studies can increase capture rates and the impact that different MOIs have within *in vitro* infection models; helping influence future NGS-based experimental designs to achieve more specific infection-related biological outcomes, particularly from *Chlamydia*-infected cells.

6.1.4.2. What is an optimal MOI?

We know from existing studies that different MOIs need to be used when examining different stages of the developmental cycle. For example, early time points generally require a higher MOI as limited transcription from *Chlamydia* occurs, resulting in low capture rates that can be difficult to interpret (Grieshaber et al., 2018; Wang et al., 2013). During mid-stages, an MOI of 1 is often used to capture events based around growth and replication (Abdelrahman et al., 2011). Towards the latter stages, almost all chlamydial genes are transcribed, making biological interpretations challenging (Humphrys et al., 2013). Most studies examining a

range of developmental stages use an MOI of 1 and this has generally been considered suitable. However, MOIs are generated from serial dilutions, so lower MOIs such as 1, may not actually have 1 EB per cell. Results from this experiment show a substantial increase in capture rates and transcription from an MOI of 1 to 10, suggesting that a slightly higher MOI may be optimal. There are however implications that need to be taken into consideration when using MOIs higher than 1. These include that EBs preferentially infect cells together rather than spread out evenly, which can result in all variations of the intended MOI. For example, a starting MOI of 5 will likely see an MOI range between 0-5 across a population of cells. As a result, the overall captured signal may be difficult to interpret, particularly with large MOIs. Furthermore, the length of the developmental cycle is generally shortened when many EBs are internalised due to an increased burden on the host cell.

6.1.4.3. Experimental limitations

Unfortunately, this experimental design did not include any mock infected replicates from either time point, limiting some analyses. Future experiments would benefit from their inclusion, helping to separate general cell proliferation events from infection-relevant results.

A further limitation was not having the ability to determine if the timeframes of the developmental cycle are affected relative to increasing the MOI. The possibility of this occurring is quite likely, particularly as more EBs are internalised, resulting in more inclusions putting an increased burden on the host cell. Perhaps a future experiment could include a fluorescent tag that could be quantified as cells become lysed, thereby providing a measurement relative to the MOI and length of the developmental cycle. Alternatively, chemical-assisted methods can arrest at different cell cycle stages, ensuring all cells are synchronised at a specified cell cycle phase and removing this as a potential confounding factor.

6.1.5. Data integration

Recent advances to the underlying chemistry and technology has decreased sequencing costs, resulting in a general increase in NGS-based projects. With this influx of sequencing data, we are beginning to see similar datasets being integrated or combined, allowing for more in depth biological interpretations (which is discussed further in Section 6.2.1.2). However, overlapping data from similar time points and experimental conditions can be challenging, due to underlying technical and biological differences between samples. Unfortunately, this was a limitation when trying to overlap data from Chapters 3, 4 and 5. Although the same cell line and chlamydial species was used, only 3 of the 6 time points sampled overlapped (**Figure 6.2**).

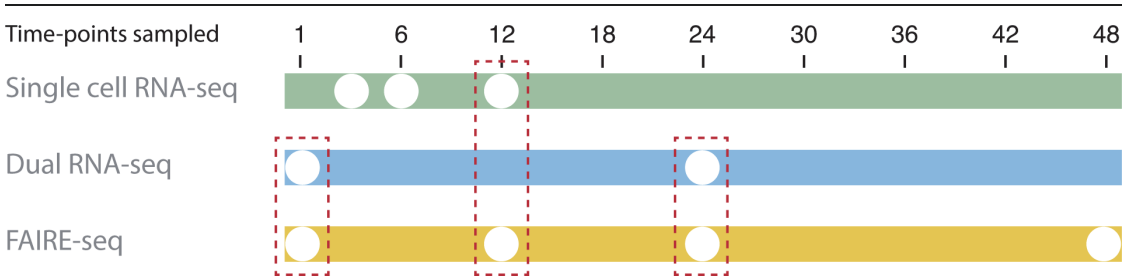


Figure 6.2: Overlapping time points

Time points sampled from each sequencing approach (Chapters 3-5), with only three times overlapping.

At 1 and 24 hours, the MOI dual RNA-seq and FAIRE-seq time points overlap. However, due to the lack of controls (mock-infected samples), DE genes specific to infection were not able to be calculated and thus not directly compared against infection-specific chromatin accessibility. At 12 hours, infection-specific data from FAIRE-seq and scRNA-seq were able to be compared. Open and closed chromatin regions were compared against up and down-

regulated genes, but did not result in any significant overlaps. Surprisingly, this is not uncommon, further reinforcing the challenges associated with overlapping different biological datasets, even if time points overlap. This is an area of continued growth where new multimodal and multi-omic methods are being developed to compare these datasets in differing ways. A common example includes merging datasets earlier and performing steps such as normalisation on the datasets together rather than separately, which has seen encouraging results (Efremova and Teichmann, 2020; Zhu et al., 2020).

6.1.6. The impact of single cell experiments

We are in an exciting period where many traditional bulk sequencing approaches have or are being adopted to study observations at a single cell resolution. The most widely used methods so far are related to single-cell RNA-sequencing and single-cell genome (DNA) sequencing (Gawad et al., 2016; Hwang et al., 2018). More recently, epigenetic states such as DNA accessibility, methylation, and chromosome conformation have become available, as well as methods to examine cell surface proteins, intracellular proteins and the spatial position of transcripts (**Figure 6.3**).

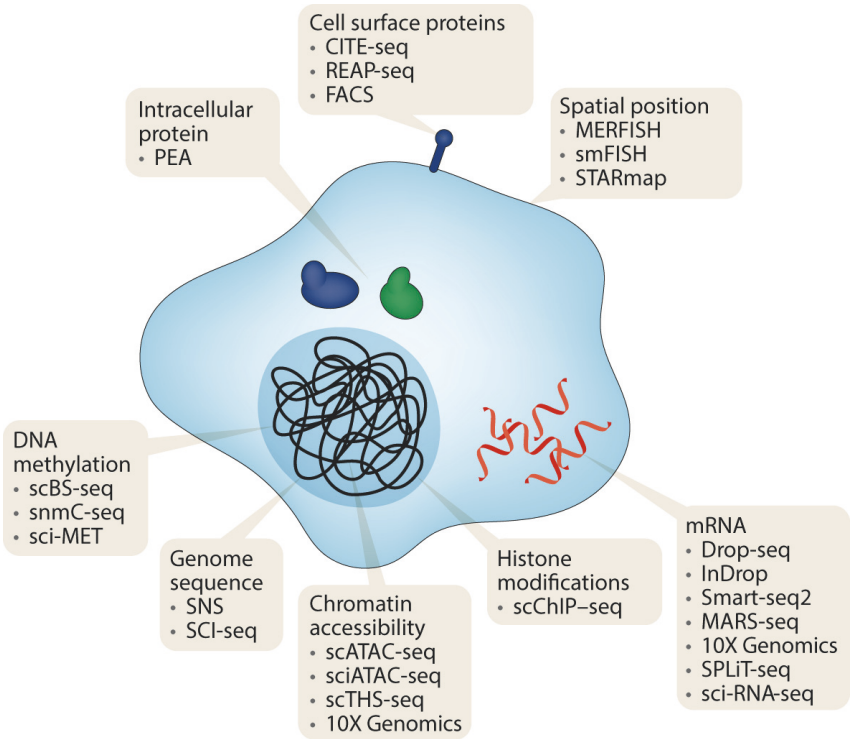


Figure 6.3: Single cell sequencing methods

A wide array of methods is available to capture a broad range of biological activity from within single cells. Methods are separated into eight categories and include capturing cellular mRNA, histone modifications, chromatin accessibility, genome sequences, methylation states, intracellular proteins, cell surface proteins and the spatial position of cells. Adapted from (Stuart and Satija, 2019).

There are also combinatorial methods simultaneously measuring two or more events, such as the genome and transcriptome, transcriptome and methylome, and RNA and proteins (Hu et al., 2018; Macaulay et al., 2017). A recent study has even generated genetic, transcriptomic and epigenetic data (Hou et al., 2016), and has the possibility of generating huge amounts of data from single cells. The main advantage of single cell approaches is identifying

subpopulations of cells that contain biologically interesting observations that would generally be obscured in traditional bulk measurements.

There are still a number of limiting steps involved in many of these methods. These typically relate to how cells are isolated and subsequently lysed, as buffers and lysing chemicals have been shown to interfere with the cellular transcriptional profiles (Kolodziejczyk et al., 2015). Ideally, future methods will be able to introduce different chemicals or additional steps that will allow less interference, such as sequencing transcripts directly from inside the nucleus of single cells (Grindberg et al., 2013; Habib et al., 2017). Capture rates are also an issue as previously discussed, with some protocols much more efficient than others (Natarajan et al., 2019). Lastly, software tools specific to single cell data are still lacking, particularly when examining epigenetic states. However, as more studies emerge, the number of tools will likely follow a similar trend to what has been observed in scRNA-seq, where > 400 tools have been created over the last four years.

Overall, single cell methods are starting to become the preferred choice for analyses, particularly when examining gene expression. With further improvements to methods and increased single cell-specific software tools, examining cellular dynamics from populations of single cells will be more frequent in the future.

6.2. Future directions

6.2.1.1. Chlamydial infection models

In vitro chlamydial cell culture is a time-consuming process and is usually carried out using monolayers of HeLa or HEp2 cells, or in macrophages. Frequently asked questions are often directed towards their suitability for infection modelling, particularly around the cell responses and how specific they are relative to the infection setting or cell type. To help answer this, a matching bulk RNA-seq experiment could be set up examining a time course of HeLa cells, HEp2 cells and macrophages. This could also be extended to different chlamydial strains, highlighting both overlapping and culture-specific observations.

Further questions that arise from infection-based studies are firstly, is how do we know what host responses are general defence mechanisms that are not specific to the invading pathogen?, and secondly, what responses are pathogen-specific? To help answer this, additional controls could help separate the different responses. For example, the Myers lab is using UV killed bacteria (killed before infection) to examine the host response to non-active bacteria compared to replicating bacteria in dual RNA-seq studies. A further dual RNA-seq control being used is opsonised latex beads to induce cells to take up beads coated with immunoglobulin via phagolysomal responses, allowing general cellular uptake and active *Chlamydia*-induced uptake responses to be differentiated. These controls should be considered for any future genome-scale analyses, including scRNA-seq and epigenetic datasets.

6.2.1.2. Integrative and multimodal analyses

The increased number of sequencing experiments from chlamydial studies and in the general biological sciences has created huge volume of sequencing data. Currently, very few studies

have integrated this publicly available data (Dong, 2009; Mabu et al., 2018). For chlamydial studies in particular, the primary limitation has been the range of species sequenced, different time points, and different infection models. Therefore, a lack of compatibility between datasets has hindered any meaningful integrative analyses. As the number of sequencing experiments continues to increase as observed in **Chapter 1**, integrating multiple NGS datasets will likely become more common and may help discover some of the broader mechanisms that overlap species and/or cell types, which are missed when focusing on separate events.

Multimodal measurements are yet a further type of integrative analyses, but overlap complementing methods, such as chromatin accessibility or methylation patterns accompanied by gene expression. This can be undertaken with bulk methods, or single cell methods, where events such chromatin state can be directly associated with increased or decreased expression. Due to the dynamic nature of transcription and given that each method captures only a snapshot in time, bioinformatic analyses that are often run separately, are difficult to merge and thus identify overlapping features. This is an area of considerable growth in recent years, particularly in the single cell space, and will likely become common place as sequencing costs drop and more datasets are generated.

6.2.1.3. Assistance through machine learning

Machine-based learning has revolutionised fields covering language processing to computer vision, and recently has become impactful analysing sequence data (Webb, 2018). Although still in its infancy for use in bioinformatic pipelines, the ability to optimise software parameters from a range of different software is exciting, as future tools, workflows and pipelines have the ability to be more automated, streamlining analyses but also allowing more flexibility. A recently published tool called single-cell variational inference (scVI) (Lopez et

al., 2018) highlights this flexibility and scope, creating an embedded workflow that can fit data, denoise it, adjust for batch effects, and perform downstream analysis including clustering and differential expression.

A further approach is automated machine learning (AutoML) systems, which are able to test different complex ML models or frameworks, whereby identifying, optimising or building different pipelines with the aim of generating the most satisfactory result, such as prediction accuracy (Le et al., 2019). However, due to the increased computational and storage requirements of these approaches, their usage is still quite low. As the number of software titles increases, so does the time-consuming task of picking software, optimising and incorporating each stage into a pipeline. Nevertheless, AutoML systems are likely to be extremely useful and valuable for future researchers.

6.3. References

Abdelrahman, Y.M., Rose, L.A., and Belland, R.J. (2011). Developmental expression of non-coding RNAs in *Chlamydia trachomatis* during normal and persistent growth. Nucleic acids research *39*, 1843-1854.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J.*, et al.* (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496-502.

Chen, M., and Manley, J.L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat Rev Mol Cell Biol *10*, 741-754.

De Puysseleyr, L., De Puysseleyr, K., Braeckman, L., Morre, S.A., Cox, E., and Vanrompay, D. (2017). Assessment of *Chlamydia suis* Infection in Pig Farmers. Transboundary and emerging diseases *64*, 826-833.

Dong, G. (2009). Sequence Data Mining (Springer-Verlag).

Efremova, M., and Teichmann, S.A. (2020). Computational methods for single-cell omics across modalities. Nat Methods *17*, 14-17.

Gawad, C., Koh, W., and Quake, S.R. (2016). Single-cell genome sequencing: current state of the science. Nature Reviews Genetics *17*, 175.

Grieshaber, S., Grieshaber, N., Yang, H., Baxter, B., Hackstadt, T., and Omsland, A. (2018). Impact of Active Metabolism on *Chlamydia trachomatis* Elementary Body Transcript Profile and Infectivity. Journal of Bacteriology *200*, e00065-00018.

Grindberg, R.V., Yee-Greenbaum, J.L., McConnell, M.J., Novotny, M., O'Shaughnessy, A.L., Lambert, G.M., Araúzo-Bravo, M.J., Lee, J., Fishman, M., Robbins, G.E.*, et al.* (2013). RNA-sequencing from single nuclei. Proceedings of the National Academy of Sciences *110*, 19802-19807.

Grubman, A., Chew, G., Ouyang, J.F., Sun, G., Choo, X.Y., McLean, C., Simmons, R., Buckberry, S., Landin, D.V., Pflueger, J.*, et al.* (2019). A single cell brain atlas in human Alzheimer's disease. bioRxiv, 628347.

Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S.R., Aguet, F., Gelfand, E., Ardlie, K.*, et al.* (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods *14*, 955-958.

Haque, A., Engel, J., Teichmann, S.A., and Lonnberg, T. (2017a). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome medicine *9*, 75.

Haque, A., Engel, J., Teichmann, S.A., and Lönnberg, T. (2017b). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome medicine *9*, 75-75.

Hoffmann, R. (2008). A wiki for the life sciences where authorship matters. Nat Genet *40*, 1047-1051.

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., Wu, X., Wen, L., Tang, F., Huang, Y.*, et al.* (2016). Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell research *26*, 304-319.

Hu, Y., An, Q., Sheu, K., Trejo, B., Fan, S., and Guo, Y. (2018). Single Cell Multi-Omics Technology: Methodology and Application. Frontiers in cell and developmental biology *6*, 28-28.

Humphrys, M.S., Creasy, T., Sun, Y., Shetty, A.C., Chibucos, M.C., Drabek, E.F., Fraser, C.M., Farooq, U., Sengamalay, N., Ott, S.*, et al.* (2013). Simultaneous transcriptional profiling of bacteria and their host cells. PloS one *8*, e80597-e80597.

Hwang, B., Lee, J.H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Experimental & Molecular Medicine *50*, 96.

Johnson, K.A., Tan, M., and Sütterlin, C. (2009). Centrosome abnormalities during a *Chlamydia trachomatis* infection are caused by dysregulation of the normal duplication pathway. Cell Microbiol *11*, 1064-1073.

Kim, K.H., and Sederstrom, J.M. (2015). Assaying Cell Cycle Status Using Flow Cytometry. Curr Protoc Mol Biol *111*, 28.26.21-28.26.11.

Knowlton, A.E., Brown, H.M., Richards, T.S., Andreolas, L.A., Patel, R.K., and Grieshaber, S.S. (2011). *Chlamydia trachomatis* infection causes mitotic spindle pole defects independently from its effects on centrosome amplification. Traffic (Copenhagen, Denmark) *12*, 854-866.

Kolodziejczyk, Aleksandra A., Kim, J.K., Svensson, V., Marioni, John C., and Teichmann, Sarah A. (2015). The Technology and Biology of Single-Cell RNA Sequencing. Molecular cell *58*, 610-620.

Kumar, V., Muratani, M., Rayan, N.A., Kraus, P., Lufkin, T., Ng, H.H., and Prabhakar, S. (2013). Uniform, optimal signal processing of mapped deep-sequencing data. Nature biotechnology *31*, 615-622.

Lagae, S., Kalmar, I., Laroucau, K., Vorimore, F., and Vanrompay, D. (2014). Emerging *Chlamydia psittaci* infections in chickens and examination of transmission to humans. Journal of medical microbiology *63*, 399-407.

Le, T.T., Fu, W., and Moore, J.H. (2019). Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics.

Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat Methods *15*, 1053-1058.

Lynch, P.J., and Rusche, L.N. (2009). A silencer promotes the assembly of silenced chromatin independently of recruitment. Molecular and cellular biology *29*, 43-56.

Mabu, A.M., Prasad, R., Yadav, R., and Jauro, S.S. (2018). A Review of Data Mining Methods in Bioinformatics. Paper presented at: 2018 Recent Advances on Engineering, Technology and Computational Sciences (RAETCS).

Macaulay, I.C., Ponting, C.P., and Voet, T. (2017). Single-Cell Multiomics: Multiple Measurements from Single Cells. Trends in Genetics *33*, 155-168.

Natarajan, K.N., Miao, Z., Jiang, M., Huang, X., Zhou, H., Xie, J., Wang, C., Qin, S., Zhao, Z., Wu, L*., et al.* (2019). Comparative analysis of sequencing technologies for single-cell transcriptomics. Genome Biol *20*, 70.

Nguyen, Q.H., Pervolarakis, N., Nee, K., and Kessenbrock, K. (2018). Experimental Considerations for Single-Cell RNA Sequencing Approaches. Frontiers in Cell and Developmental Biology *6*.

Pollen, A.A., Nowakowski, T.J., Shuga, J., Wang, X., Leyrat, A.A., Lui, J.H., Li, N., Szpankowski, L., Fowler, B., Chen, P.*, et al.* (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nature biotechnology *32*, 1053-1058.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research *35*, D61-D65.

Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M.*, et al.* (2017). The Human Cell Atlas. eLife *6*, e27041.

Rizzetto, S., Eltahla, A.A., Lin, P., Bull, R., Lloyd, A.R., Ho, J.W.K., Venturi, V., and Luciani, F. (2017). Impact of sequencing depth and read length on single cell RNA sequencing data of T cells. Sci Rep *7*, 12781.

Streets, A.M., and Huang, Y. (2014). How deep is enough in single-cell RNA-seq? Nature biotechnology *32*, 1005.

Stuart, T., and Satija, R. (2019). Integrative single-cell analysis. Nature Reviews Genetics *20*, 257-272.

UniProt, C. (2008). The universal protein resource (UniProt). Nucleic acids research *36*, D190-D195.

Van Hove, H., Martens, L., Scheyltjens, I., De Vlaminck, K., Pombo Antunes, A.R., De Prijck, S., Vandamme, N., De Schepper, S., Van Isterdael, G., Scott, C.L.*, et al.* (2019). A single-cell atlas of mouse brain macrophages reveals unique transcriptional identities shaped by ontogeny and tissue environment. Nature Neuroscience *22*, 1021-1035.

Vignali, M., Hassan, A.H., Neely, K.E., and Workman, J.L. (2000). ATP-Dependent Chromatin-Remodeling Complexes. Molecular and Cellular Biology *20*, 1899-1910.

Wang, A., Al-Kuhlani, M., Johnston, S.C., Ojcius, D.M., Chou, J., and Dean, D. (2013). Transcription factor complex AP-1 mediates inflammation initiated by *Chlamydia pneumoniae* infection. Cell Microbiol *15*, 779-794.

Webb, S. (2018). Deep learning for biology. Nature *554*, 555-557.

West, A.G., Gaszner, M., and Felsenfeld, G. (2002). Insulators: many functions, many mechanisms. Genes & development *16*, 271-288.

Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., Huang, Y., and Wang, J. (2019). Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. Molecular cell *73*, 130-142.e135.

Zhu, C., Preissl, S., and Ren, B. (2020). Single-cell multimodal omics: the power of many. Nat Methods *17*, 11-14.