Biotechnology Master Program

Joint Biotechnology Master Program

Palestine Polytechnic University
Deanship of Higher Studies and
Scientific Research

Bethlehem University
Faculty of Science

In silico single-cell RNA-seq analysis reveals distinct host response programs in bovine milk somatic cells during H5N1-associated and bacterial mastitis

By

Mofeed Abu Rmaileh

Supervisor

Dr. Robin Abu Ghazaleh

In Partial Fulfillment of the Requirements for the Degree

Master of Science

January, 2026

ABSTRACT

Highly pathogenic avian influenza (HPAI) H5N1 has recently been recognized as a cause of severe mastitis in dairy cattle, with viral replication in the mammary gland and shedding into milk. This emergence raises concern for animal health, dairy production and potential zoonotic transmission. In contrast, bovine mastitis is usually bacterial and has been studied mainly with bulk-level approaches that obscure cell-type–specific host responses. In this thesis, publicly available single-cell RNA sequencing datasets from H5N1-exposed, bacterial mastitis and healthy control bovine milk somatic cells were re-analyzed using an in-silico bioinformatics pipeline to compare viral and bacterial mastitis at single-cell resolution. After read alignment and rigorous quality control, an integrated bovine milk cell atlas was generated and major immune and non-immune cell populations were annotated, including neutrophils, monocyte/macrophage subsets, dendritic cells, lymphocytes and mammary epithelial cells. Condition-level and cell-type–level differential expression and functional enrichment analyses identified distinct pathogen-specific transcriptional programs. H5N1 mastitis was characterized by strong induction of interferon-stimulated genes, antiviral restriction factors, and chemokine modules across immune compartments, whereas bacterial mastitis was dominated by Toll-like receptor and NF-κB signalling, inflammasome-associated genes, and inflammatory pathways. These signatures were reflected in altered cellular composition, with focused expansion of interferon-high myeloid subsets in H5N1 and broader recruitment of neutrophil and macrophage populations in bacterial mastitis. By intersecting condition-specific gene sets with lineage markers, two concise 12-gene biomarker panels were derived: a viral panel composed mainly of interferon-stimulated antiviral genes and a bacterial panel centred on innate immune and inflammasome regulators. When projected onto neutrophil trajectories using diffusion pseudotime, these panels showed distinct activation profiles consistent with pathogen class. Through this analysis, understanding of pathogen-specific mastitis immunopathology at single-cell resolution is advanced, host-response gene panels for molecular differentiation of H5N1 from bacterial mastitis using milk-derived cells are supported, and priorities for future experimental validation and diagnostic assay development are highlighted.

الخلاصة

تم مؤخّرًا التعرّف على إنفلونزا الطيور عالية الإمراض (HPAI) من النمط H5N1 كسبب لالتهاب ضرع شديد في أبقار الحليب، مع حدوث تكاثر فيروسي داخل الغدة الثديية وطرح الفيروس في الحليب. ويثير هذا الظهور مخاوف تتعلّق بصحّة الحيوان وإنتاج الألبان وإمكانية الانتقال حيوانيّ المنشأ إلى الإنسان. وعلى النقيض من ذلك، فإن التهاب الضرع البقري غالبًا ما يكون ذا منشأ بكتيري، وقد دُرس في الغالب باستخدام bulk Rna-seq والتي قد تُخفي استجابات العائل الخاصة بكل نوع خلوي.في هذه الرسالة، تم تحليل مجموعات البيانات متاحة للعامة لتسلسل الحمض النووي الريبي أحادي الخلية (scRNA-seq)لخلايا الحليب الجسدية البقرية تحت ثلاث حالات: التعرّض لـH5N1 ، التهاب الضرع البكتيري، والحالة السليمة (control)، وذلك باستخدام in silico bioinformatics pipeline بهدف مقارنة التهاب الضرع الفيروسي بالبكتيري بدقة single-cell resolution. وبعد محاذاة القراءات وإجراء ضبط جودة صارم، جرى إنشاء أطلس متكامل لخلايا الحليب لدى الأبقار، وتم توصيف التجمعات الخلوية المناعية وغير المناعية الرئيسية، بما في ذلك العدلات، وتحت مجموعات الوحيدات/البلعميات، والخلايا المتغصّنة، والخلايا اللمفاوية، وخلايا الظهارة الثديية. وقد كشفت تحليلات التعبير الجيني التفاضلي على مستوى الحالة (control level)وعلى مستوى نوع الخلية، إلى جانب تحليلات الإثراء الوظيفي (functional enrichment analysis)، عن برامج نسخية مميّزة خاصة بكل مُمرِض. تميّز التهاب الضرع الناتج عن H5N1 بتحريض قوي لجينات محفَّزة بالإنترفيرون، وعوامل تقييد مضادّة للفيروسات، ووحدات الكيموكاين عبر عدة مكونات مناعية، في حين هيمن على التهاب الضرع البكتيري تنشيط مسارات مستقبلات الشبيه بال (TLR) وإشارةNF-κB ، وجينات مرتبطة بال Inflammasome، ومسارات التهابية. وانعكست هذه البصمات أيضًا في تغيّرات التركيب الخلوي؛ إذ لوحظ توسّع مُركّز interferon high myeloid subsets في حالة H5N1، مقابل تجنيد أوسع للعدلات والبلعميات في التهاب الضرع البكتيري. ومن خلال تقاطع مجموعات الجينات الخاصة بكل حالة مع واسمات السلالة الخلوية، جرى اشتقاق لوحتين مختصرتين من الواسمات الحيوية مكوّنتين من 12 جين لكل منهما: لوحة فيروسية تتكوّن أساسًا من جينات مضادّة للفيروسات محفَّزة بالإنترفيرون، ولوحة بكتيرية تتمحور حول منظّمين للمناعة الفطرية وinflammasome. وعند إسقاط هذه اللوحات على مسارات تمايز العدلات باستخدام الزمن الكاذب المعتمد على الانتشار(diffusion pseudotime) ، أظهرت اللوحات أنماط تنشيط مميّزة تتوافق مع صنف المُمرض. ومن خلال هذا التحليل، تم تعزيز فهم الاعتلال المناعي لالتهاب الضرع الخاص بكل مُمرض على مستوى أحادي الخلية، ودعم لوحات جينية لاستجابة العائل تُمكّن من التمييز الجزيئي بين H5N1 والتهاب الضرع البكتيري باستخدام خلايا مشتقة من الحليب، مع إبراز أولويات لعمليات التحقق التجريبي المستقبلية وتطوير اختبارات تشخيصية.

DEDICATION

To the best supporters I could ever have, my father, my mother, my wife, whose love and support know no bounds. To my family, my brothers and sister, for always standing by me.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Dr. Robin Abu Ghazaleh, for his tireless support and guidance throughout this project, without which its success would not have been possible. I am also deeply thankful to the professors of the master's program for their dedication and expertise, and to the faculty team of the Biotechnology Center for their constant support.

Table of Abbreviations

Abbreviation	Full Term
APC	Antigen-presenting cell
AUROC	Area Under the Receiver Operating Characteristic curve
BAM	Binary Alignment Map (standard file format for aligned reads)
bMSCs	Bovine milk somatic cells
BP / CC / MF	Biological Process / Cellular Component / Molecular Function (GO categories)
Datamash	Command-line tool for statistical aggregation and data reshaping
DPT	Diffusion pseudotime (metric for cell state transitions)
FDR	False Discovery Rate
Galaxy	Web-based platform for reproducible bioinformatic research
GEO	Gene Expression Omnibus (NCBI data repository)
GO	Gene Ontology
GTF	Gene Transfer Format (standard for gene annotation files)
H5N1	Influenza A virus subtype (specifically clade 2.3.4.4b)
HA / NA	Hemagglutinin / Neuraminidase (viral surface glycoproteins)
Harmony	Batch-effect correction algorithm for single-cell data integration
HPAI	Highly pathogenic avian influenza
ISGs	Interferon-stimulated genes
kNN	k-nearest neighbor (graph used to model cell similarities)
KEGG	Kyoto Encyclopedia of Genes and Genomes
Leiden	Graph-based clustering algorithm for identifying cell populations
M1 / M2	Matrix protein 1 / Matrix protein 2 (viral structural proteins)
MCL	Markov Cluster Algorithm (clustering method used in STRING)
MHC (I / II)	Major Histocompatibility Complex (Class I or Class II)
MultiQC	Tool for aggregating bioinformatics reports into a single summary
NCBI	National Center for Biotechnology Information
NF-κB	Nuclear factor kappa-light-chain-enhancer of activated B cells
NP	Nucleoprotein (viral protein that encapsidates RNA)
NS1 / NS2	Nonstructural protein 1 / Nonstructural protein 2
OXPHOS	Oxidative phosphorylation (mitochondrial respiration)
PA / PB1 / PB2	Polymerase acidic / Polymerase basic 1 / Polymerase basic 2
PCA	Principal component analysis
PPI	Protein–protein interaction
PRR	Pattern recognition receptor (e.g., RIG-I or TLRs)
ROS	Reactive oxygen species
Scanpy	**S**ingle-**C**ell **An**alysis in **Py**thon (analysis framework)
scRNA-seq	Single-cell RNA sequencing
SRA	Sequence Read Archive (NCBI database for raw data)
SRP	Signal recognition particle
STAR / STARsolo	Spliced Transcripts Alignment to a Reference (and its single-cell tool)
STRING	Search Tool for the Retrieval of Interacting Genes (PPI database)
SVD	Singular Value Decomposition
$TCID_{50}$	Tissue Culture Infectious Dose 50 (viral titer metric)

Th (1/2/17)	T helper cell (Type 1, 2, or 17)
TLR	Toll-like receptor
UMAP	Uniform Manifold Approximation and Projection
UMI	Unique Molecular Identifier

Table of contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Background

Highly Pathogenic Avian Influenza (HPAI) H5N1 is an influenza A subtype that has historically caused devastating panzootic outbreaks in poultry and wild birds and has occasionally spilled over into mammals, including humans. Since its emergence in 1996 and global spread in the early 2000s, H5N1 particularly clade 2.3.4.4b in recent years has remained a major animal-health threat and a continuing pandemic concern (Caserta et al., 2024). A notable and unexpected host-range expansion was documented in 2024 when H5N1 was reported in dairy cattle (*Bos taurus*) in the United States, with cases identified across multiple states (Caserta et al., 2024; Sanchez-Rojas et al., 2025).

Clinically affected cows showed systemic illness and, most notably, severe mastitis accompanied by dramatic milk-yield reduction and abnormal milk that contained high viral loads. Pathological investigations indicated a mammary-gland tropism, with viral RNA/antigen detected in milk-secreting epithelial cells, supporting the interpretation that H5N1 can present in cattle as a fulminant viral mastitis (Caserta et al., 2024; Sanchez-Rojas et al., 2025).

Mastitis is among the most common and economically costly diseases in dairy production and is typically attributed to bacterial pathogens such as *Escherichia coli, Staphylococcus aureus, and Streptococcus spp.* The 2024 H5N1 epizootic was therefore exceptional, as influenza A viruses had not previously been recognized as causes of overt clinical mastitis in cattle. Importantly, field and histopathological observations indicated that H5N1-associated mastitis can closely resemble acute bacterial mastitis in both clinical presentation and inflammatory pathology, including pronounced neutrophilic inflammation and abnormal, clotted milk (Singh et al., 2025; Sanchez-Rojas et al., 2025). This phenotypic overlap creates a distinct diagnostic and biological challenge: mastitis on farms is usually presumed bacterial, so responses are commonly oriented toward antimicrobials and hygiene, whereas H5N1 requires outbreak-focused biosecurity and carries zoonotic risk, making antibiotic-based approaches inappropriate (Singh et al., 2025).

A key gap is the current lack of robust host-response biomarkers or diagnostic criteria that reliably differentiate viral from bacterial mastitis. Although both etiologies provoke udder inflammation,

it remains unclear whether cell type resolved gene-expression programmes in milk contain sufficiently distinct molecular signatures to support rapid etiological discrimination. Viral mastitis would be expected to feature antiviral pathways such as interferon-stimulated gene induction, whereas bacterial mastitis is typically associated with Toll-like receptor and NF-κB–driven inflammatory programmes; however, these differences may be obscured in bulk analyses because milk contains a complex mixture of neutrophils, macrophages, lymphocytes, and mammary-derived cells. Single-cell RNA sequencing (scRNA-seq) provides the necessary resolution to disentangle both cellular composition and lineage-specific transcriptional responses. Therefore, this thesis asks how cell type–resolved transcriptional signatures and immune population dynamics differ between H5N1 viral mastitis and bacterial mastitis, and whether these differences can be distilled into a minimal gene-expression panel that discriminates the two conditions at the individual-cow level. In this work, the problem is addressed entirely through bioinformatics, using scRNA-seq datasets and computational pipelines to derive cell type–specific signatures and candidate diagnostic markers, without any wet-lab experiments (Zorc et al., 2024; Luan et al., 2024).

1.2 Aim and Objectives

To use single-cell transcriptomic bioinformatics to compare bovine milk somatic cell responses in H5N1-associated mastitis versus bacterial mastitis and derive a concise marker set that distinguishes between the two etiologies. Objectives: Public scRNA-seq datasets were integrated after quality control and batch-aware preprocessing; major immune and mammary cell populations were identified and annotated to build a cross-condition cell atlas; condition-associated shifts in cell composition were quantified; lineage-resolved differential expression was performed to define pathogen-specific transcriptional programmes; functional interpretation was conducted using pathway and network analyses; and a minimal diagnostic gene panel was prioritised to support etiological discrimination in a fully reproducible, computational workflow.

Chapter 2: Literature Review

Highly Pathogenic Avian Influenza (HPAI) H5N1 is a subtype of influenza A virus that has caused significant outbreaks in birds and sporadic infections in mammals (including humans) (Caserta et al., 2024). It is belonging to the Goose/Guangdong lineage of avian influenza A, first identified in geese in Guangdong, China in 1996 (Caserta et al., 2024). It remained confined to poultry initially, but by 2002 it had spread to wild bird populations (Caserta et al., 2024). Through frequent reassortment with other influenza viruses, H5N1 evolved into multiple genetic clades groups of related strains. Over the past two decades, numerous clades have been defined (e.g., clades 2.3.4.4a–h), reflecting the virus's ongoing evolution and antigenic changes (Caserta et al., 2024). One clade in particular, clade 2.3.4.4b, has become dominant in recent years, causing global outbreaks in poultry and wild birds across Asia, Europe, Africa, and the Americas (Caserta et al., 2024). This clade has exhibited an expanded host range, with increasing reports of infections in mammalian species (Caserta et al., 2024). Since 2003, there have been roughly 860 laboratory-confirmed human H5N1 cases worldwide, with a high case-fatality rate around 53% (Caserta et al., 2024). Fortunately, efficient human-to-human transmission has remained rare to date (Caserta et al., 2024). However, the widespread circulation of H5N1 in birds (causing losses of over 100 million poultry in the U.S. alone by 2024) and its ability to spill over into new hosts poses a continued pandemic threat (Caserta et al., 2024).

2.1 Virus Structure and Genome

Influenza A H5N1 is an enveloped, negative-sense RNA virus of the family *Orthomyxoviridae*. The viral genome is segmented into 8 single-stranded RNA segments, encoding at least 10 major proteins (Sanchez-Rojas et al., 2025). Prominent among these are the surface glycoproteins Hemagglutinin (HA) and Neuraminidase (NA), which define the subtype (H5 and N1, respectively). The HA protein is responsible for binding to sialic acid receptors on host cells and mediating viral entry, while NA helps release new virions by cleaving sialic acids. H5N1's HA contains a characteristic polybasic cleavage site in highly pathogenic strains, allowing host proteases (like furin) to cleave HA0 into HA1/HA2 subunits in a wide range of tissues, a key virulence factor that enables systemic spread in birds (Bogs et al., 2010; Luczo et al., 2015). Internally, the virus carries a triad of polymerase proteins (PB2, PB1, PA), a nucleoprotein (NP)

that encapsidates the RNA, a matrix protein (M1) lining the interior of the envelope, an ion-channel protein (M2), and nonstructural proteins (NS1 and NEP/NS2) that modulate host responses and aid viral RNA export (Liang, 2023). Together, these components form the influenza virion, which is typically spherical or filamentous (~100 nm diameter) and studded with HA and NA spikes on its surface (Partlow et al., 2025). The segmented genome enables reassortment, exchange of segments when a host is co-infected with different influenza strains, which has been a driving force in H5N1's evolution into new clades (Taylor et al., 2023). For example, clade 2.3.4.4b viruses are reassortants containing gene segments of mixed Eurasian and North American lineage origin (Caserta et al., 2024). This continual evolution has been shown to highlight the need for vigilant surveillance of the H5N1 genome for mutations associated with host adaptation or increased virulence. Notably, adaptation markers such as the PB2-E627K mutation (which confers enhanced polymerase activity in mammalian cells) have arisen when H5N1 replicates in mammals (Halwe et al., 2025).

2.2 Transmission and Host Range

H5N1 is primarily adapted to birds, especially aquatic wildfowl (its natural reservoir) and domestic poultry. In avian hosts, transmission occurs via direct contact or environmental exposure. The virus typically binds to α2,3-linked sialic acid receptors, which are abundant in the respiratory and intestinal tracts of birds. Wild migratory birds have disseminated clade 2.3.4.4b H5N1 across continents, introducing the virus into new geographic regions (Tiwari et al., 2024). These incursions are often followed by outbreaks in poultry populations, characterized by rapid spread and high mortality in chickens. Humans and other mammals are considered incidental hosts, and infection is rare, usually occurring through close contact with infected birds or contaminated environments. In mammals, including humans, α2,6-linked sialic acid receptors predominate in the upper respiratory tract, which limits the ability of avian-adapted H5N1 strains to bind efficiently. This receptor mismatch has historically been considered a barrier to cross-species transmission and sustained human-to-human spread. Nevertheless, sporadic mammalian infections have been documented globally during the current panzootic (Tiwari et al., 2024). Multiple carnivore species, including foxes, cats, bears, and seals, have tested positive for clade 2.3.4.4b H5N1, often developing severe disease such as encephalitis, which reflects the virus's ability to cause systemic infection (Caserta et al., 2024). Until recently, livestock species were not regarded

as typical hosts. In early 2024, H5N1 was detected in ruminants specifically dairy cattle in the United States, marking an unexpected spillover event with important implications for animal health and surveillance (Caserta et al., 2024).

Mammalian infection may occur through ingestion of virus-contaminated material. For example, scavenging animals may feed on carcasses of infected birds, while farm mammals may be exposed through contaminated feed, water, or bedding (Caserta et al., 2024).

2.3 H5N1 Spillover to Bovine (Cattle) and Emergence of a New Clade

In early 2024, the United States reported the first known outbreak of HPAI H5N1 in dairy cattle, marking the first documented natural infection of a ruminant species with H5N1 (Sanchez-Rojas et al., 2025). The strain involved was H5N1 clade 2.3.4.4b, which had already been responsible for extensive outbreaks among North American wild birds and poultry since its introduction in late 2021 (Caserta et al., 2024). Phylogenetic analyses revealed that the viruses infecting cattle belonged to a novel reassortant genotype, designated B3.13, which contains Eurasian-origin HA, NA, PA, and MP segments, and North American-origin PB2, PB1, NP, and NS segments (Caserta et al., 2024). This genotype is believed to have emerged in wild birds in late 2023. It was first detected in a Canada goose in Colorado (November 2023), followed by identification in a peregrine falcon in California (February 2024), and a skunk in New Mexico shortly thereafter (Caserta et al., 2024). These findings indicate the virus's circulation in diverse wildlife hosts. The exact route of transmission into cattle remains under investigation, but it has been proposed that environmental contamination, including exposure to virus shed by wild birds into farm ponds, feed, or airborne dust and feathers, may have played a role, particularly on farms located near migratory bird flyways (Caserta et al., 2024). In addition, spillover from infected poultry on mixed farms may have occurred via shared environments or contaminated equipment. Once introduced into a herd, the virus appears to have spread horizontally among cows. The outbreak expanded rapidly within the U.S. dairy sector: first detected in a Texas herd in March 2024, it subsequently reached over 380 farms across 14 states (Sanchez-Rojas et al., 2025). Epidemiological investigations reported that apparently healthy cattle from affected farms, when transported to new locations, introduced the virus and seeded secondary outbreaks (Caserta et al., 2024). This type of cow-to-cow transmission had not previously been described for an avian influenza virus.

Experimental studies have indicated that nasal shedding in cattle is modest, suggesting that respiratory transmission is limited (Halwe et al., 2025). Instead, transmission has been hypothesized to occur via the milking process, particularly in commercial dairies where shared equipment may become contaminated with virus-laden milk. If milking machinery is not adequately disinfected between cows, the virus may be introduced directly into the teat canal of susceptible herd-mates (Caserta et al., 2024). This proposed mechanism bypasses the respiratory route and facilitates intra-mammary infection. Additionally, ingestion of contaminated raw milk, by calves or companion animals, may contribute to onward transmission. Indeed, the practice of feeding waste milk to farm cats was common on affected farms and may explain the concurrent infections reported in cats and raccoons (Caserta et al., 2024).

2.4 Geographic Distribution of Bovine H5N1

To date, H5N1 infections in cattle have been reported only in the United States. The initial wave, occurring between January and March 2024, affected dairy herds in Texas, New Mexico, Kansas, and included at least one case linked to transported cattle in Ohio (Caserta et al., 2024). By June 2025, confirmed cases in cattle had been reported in at least 16 U.S. states (Sanchez-Rojas et al., 2025). Surveillance in other countries with active H5N1 outbreaks among birds has, so far, not detected evidence of similar bovine infections. For example, intensive investigations in Europe have reported no detection of H5N1 in dairy cattle outside the USA (Owusu and Sanad, 2025). Some researchers have noted that this apparent absence is surprising, considering the global circulation of H5N1 in wild birds. It has been proposed that ecological, geographical, or farm-management differences may contribute to limiting transmission or may affect detection and reporting in other countries (Owusu and Sanad, 2025). Nonetheless, the World Health Organization has advised that countries remain vigilant for potential H5N1 incursions into cattle, particularly in regions where wild birds and livestock are in close contact (Owusu and Sanad, 2025).

2.5 Clinical Manifestations of H5N1 in Cattle

The outbreak of H5N1 in dairy cattle was characterized by a distinct clinical syndrome focused on the udder (mammary gland) accompanied by systemic illness. In affected cows, early signs included a marked reduction in feed intake and general lethargy, consistent with systemic malaise

(Sanchez-Rojas et al., 2025). Fever and dehydration were also frequently reported, indicating an acute infectious process (Sanchez-Rojas et al., 2025). Respiratory signs were typically mild, with occasional observations of clear nasal discharge, labored breathing, or coughing (Caserta et al., 2024). Gastrointestinal disturbances were also noted; some animals developed diarrhea, while others exhibited unusually dry feces (Caserta et al., 2024). A hallmark of the infection was the effect on milk production and udder appearance. Farmers reported a sudden decline in milk yield, ranging from 20% to complete cessation during the acute phase (Caserta et al., 2024). Milk from infected cows often appeared thick, yellowish, and clotted, resembling colostrum rather than normal milk (Caserta et al., 2024). In some cases, milk had a curdled appearance or contained visible flakes, reflecting cellular debris and a loss of normal milk consistency (Caserta et al., 2024). Clinical mastitis (inflammation of the mammary gland) was evident; affected udders were painful, and several cows exhibited involution of specific quarters, where milk secretion ceased entirely (Caserta et al., 2024). Even after apparent clinical recovery, many cows failed to return to baseline milk production for several weeks, which has been interpreted as indicative of lasting mammary gland injury (Caserta et al., 2024).

These clinical signs differ from the typical presentation of avian influenza in birds, which often involves respiratory and neurological symptoms, and from most mammalian influenza infections, which primarily involve respiratory signs. In cattle, H5N1 infection was notable for its presentation as mastitis with systemic illness. Signs such as fever, anorexia, reduced rumination, and udder inflammation raised both animal welfare and economic concerns. Reports indicate that, during the outbreak, mortality on some farms was up to twice the baseline rate (Caserta et al., 2024). Nevertheless, most animals recovered clinically within 1–2 weeks (Caserta et al., 2024), although milk production often remained below normal for a prolonged period. In comparing this form of viral mastitis with conventional bacterial mastitis, several overlapping signs are noted. Both conditions can produce swollen, firm udders; abnormal milk (e.g., clotted, discolored, or watery); reduced milk output; fever; and behavioural signs of illness. In H5N1 cases, the milk frequently exhibited a yellow, colostrum-like appearance (Caserta et al., 2024), which is also seen in some severe bacterial mastitis cases, particularly those involving high leukocyte counts and elevated milk proteins. However, cows infected with H5N1 more often displayed systemic signs, including fever and anorexia, in contrast to milder bacterial mastitis, which may be confined to local udder

pathology (Caserta et al., 2024) (Sanchez-Rojas et al., 2025). Another distinguishing feature was the abrupt drop in milk production, sometimes down to 20% of baseline levels, consistent with rapidly progressing inflammation (Caserta et al., 2024). Some aggressive bacterial infections (e.g., coliform mastitis caused by *E. coli*) can cause similarly rapid declines in production and systemic shock, emphasizing the severity of both disease types.

2.6 Pathology of H5N1 Mastitis vs. Bacterial Mastitis

Histopathological analyses of H5N1-infected cattle have demonstrated that the virus induces necrotizing mastitis, characterized by extensive inflammation within mammary tissue. Microscopically, the alveoli of the mammary gland were filled with neutrophils, lymphocytes (especially plasma cells), and cellular debris, essentially pus and lysed epithelial cells (Caserta et al., 2024). In many areas, the normal glandular architecture (tubuloacinar structures) was disrupted by the inflammatory exudate (Caserta et al., 2024). The alveolar epithelial cells were frequently sloughed into the lumen, a result of both viral cytopathic effects and inflammatory damage (Caserta et al., 2024). This pattern of suppurative mastitis with neutrophil infiltration closely resembles acute bacterial mastitis; for example, mastitis caused by *Staphylococcus aureus* or *Streptococcus uberis* also presents with alveolar infiltration by neutrophils. One study reported that cows infected with H5N1 exhibited neutrophilic and lymphoplasmacytic mastitis, mirroring the immune response commonly seen in bacterial udder infections (Singh et al., 2025).

Despite these similarities, notable differences exist in the etiology of tissue damage. In bacterial mastitis, injury is often driven by bacterial toxins and neutrophil-derived enzymes, whereas in H5N1 mastitis, a significant portion of damage results from active viral replication within mammary epithelial cells, leading to cell death (Owusu and Sanad, 2025). In situ hybridization and immunohistochemistry confirmed the presence of viral RNA and antigen in milk-producing epithelial cells of the alveoli (Caserta et al., 2024). These findings support the conclusion that H5N1 actively infects these cells, not merely triggers secondary inflammation. As the epithelial cells undergo cytolysis and detachment, milk production is impaired. Inflammatory infiltration by neutrophils and other leukocytes occurs in response to cytokines, interferons, and viral RNA detection by innate immune sensors, such as RIG-I and TLR1/TLR2, rather than the bacterial pattern recognition pathways (Martins et al., 2025). The overall clinical outcome, characterized by

swollen udders filled with inflammatory cells and reduced milk output, can closely resemble bacterial mastitis in gross appearance, though the initiating mechanism is viral (Singh et al., 2025).

One notable pathological distinction is the extraordinarily high viral titers in milk during H5N1 mastitis, reaching up to 10^9 $TCID_{50}$/mL (50% tissue culture infectious dose per mL) in experimentally infected cows (Halwe et al., 2025). By comparison, even severe coliform mastitis typically yields only millions of bacterial cells per mL. This high viral burden supports the conclusion that milk can act as a potent vehicle for transmission (Caserta et al., 2024). In terms of lesion distribution, H5N1 infection is largely localized to the mammary gland, with minimal respiratory involvement reported in most natural cases (Peña-Mosca et al., 2025). In contrast, bacterial mastitis generally remains confined to the udder unless systemic dissemination occurs. However, some cows infected with H5N1 have shown evidence of viral dissemination beyond the mammary tissue. Studies have identified low levels of viral RNA and antigen in the lungs, perimammary lymph nodes, spleen, heart, colon, and liver (Caserta et al., 2024). Virus-positive cells were frequently localized to the periphery of germinal centers within lymph nodes and perivascular regions in other tissues, indicating limited but detectable systemic spread. These findings underscore the virus's tropism for mammary tissue and raise the possibility of broader dissemination in severe cases, though encephalitis has not yet been clearly documented in bovine hosts (Caserta et al., 2024). While bacterial mastitis can also result in bacteremia and multi-organ involvement, especially in coliform infections such as *E. coli*, the primary site of damage for both viral and bacterial mastitis remains the mammary gland (Hagiwara, Mori and Nagahata, 2016).

2.7 Immune Response in Bovine H5N1 Infection

Understanding the immune response of cattle to H5N1 is critical, given how unusual this host-pathogen interaction is. Cattle are naturally hosts of influenza D (a separate influenza virus primarily causing mild respiratory illness in cows), and they can be infected by influenza C, but significant illness from influenza A viruses had not been described prior to H5N1 (Sanchez-Rojas et al., 2025). When H5N1 infected cows' mammary glands, it triggered robust innate and adaptive immune reactions locally. As noted in recent studies, the mammary tissue was inundated with neutrophils, key innate immune cells that phagocytose pathogens and release enzymes. Neutrophils are guided to the site by chemokines; indeed, affected mammary tissue showed

upregulation of chemoattractant like IL-8 (CXCL8) and others (Singh et al., 2025). Alongside neutrophils, macrophages and dendritic cells were present in the inflamed udder (Singh et al., 2025). These cells help clear debris and also take up viral antigens to initiate adaptive immune responses.

On the adaptive side, activation of T lymphocytes was observed in the cows. An intriguing finding from a recent study using single-cell RNA sequencing (scRNA-seq) of bovine milk immune cells reported that H5N1 exposure appeared to induce a bias toward a Type 2 immune response in T cells (Singh et al., 2025). Specifically, the study identified a subset of T cells expressing IL-13 and GATA3, markers of Th2-polarized T helper cells (Singh et al., 2025). Th2 responses are typically associated with antibody production and allergic reactions, in contrast to Th1 responses that are classically antiviral and pro-inflammatory. The presence of IL-13 was interpreted as a possible attempt by the immune system to modulate or dampen inflammation, as IL-13 can promote alternative activation of macrophages and tissue repair. It has been suggested that in the context of severe tissue damage in the udder, a Th2-skewed response may arise as part of the healing process or as a counterbalance to strong pro-inflammatory signals. Another subset of T cells showed an "inhibited Th2" profile (expressing genes indicating a restrained activation) (Singh et al., 2025), suggesting a complex regulatory environment.

2.8 Single-Cell Transcriptomics Insights

Single-cell transcriptomics (typically via single-cell RNA sequencing, scRNA-seq) allows researchers to profile gene expression at the resolution of individual cells. In contrast to scRNA-seq, bulk RNA-seq measures gene expression from RNA extracted from a pooled, heterogeneous mixture of cells, resulting in a single composite expression profile per sample, because the bulk signal is effectively a mixture average across all contributing cell types, bulk profiles cannot directly assign expression changes to specific lineages. As a consequence, apparent differential expression in bulk data may arise either from true within cell-type regulation (real transcriptional changes within a given lineage) or from cell-type composition shifts (changes in the relative abundance of cell types), and these two effects are difficult to disentangle using bulk data alone, scRNA-seq captures the unique transcriptome of each cell, thereby revealing cellular heterogeneity that bulk methods can mask (Yan et al., 2024; Kalantari-Dehaghi et al., 2025). In droplet based

scRNA-seq, individual cells are partitioned in oil droplets and their captured transcripts are labeled with cell barcodes (and UMIs), enabling reconstruction of a per-cell transcriptome after sequencing. This enables identification of distinct cell types and states, including rare cell populations that may have important functional roles but would be indistinguishable in bulk analysis. The approach has transformed the capacity to probe cell identity and responses at high resolution, enabling analysis of tens to hundreds of thousands of cells per run. Such single-cell analyses have been used to investigate how genetically identical cells exhibit diverse behaviours by expressing different gene subsets (Jovic et al., 2022).

A 2025 study by Singh et al. applied scRNA-seq to profile the gene expression of bovine milk somatic cells (which include immune cells and shed epithelial cells) after in vitro exposure to an H5N1 isolate from the cattle outbreak (Singh et al., 2025). The study identified 10 distinct cell clusters in milk, including three subsets of mammary epithelial cells and seven immune cell types (monocytes, macrophages, dendritic cells, T cells, etc.) (Singh et al., 2025). Upon H5N1 exposure, a notable shift in cell population dynamics was observed: the proportion of immune cells (monocytes, macrophages, DCs, T cells) increased, while the proportion of luminal epithelial cells decreased, relative to unexposed controls (Singh et al., 2025). This was interpreted by the authors as consistent with the expected pattern during mastitis, in which immune cells infiltrate while epithelial integrity is compromised.

The single-cell data also showed gene expression changes reflecting activation of innate immune pathways. For example, monocytes upregulated genes related to inflammatory responses and viral defense (Singh et al., 2025). Macrophages showed high expression of matrix metalloproteinases (MMP1, MMP3, MMP12) after exposure (Singh et al., 2025). These MMPs are typically induced by Th2 cytokines like IL-13 and are known to mediate tissue remodeling, which may reflect ongoing repair processes in the infected udder. Dendritic cells and T cells upregulated genes for cytokine response and activation, which the authors suggested may indicate priming for coordination of adaptive immune responses (Singh et al., 2025).

Interestingly, scRNA-seq did not detect viral RNA reads inside individual milk cells (Singh et al., 2025), even though bulk PCR on the sample was positive for the viral genome. However, the absence of viral reads in scRNA-seq may also reflect technical sensitivity limitations of droplet-

based platforms (e.g., sparse transcript capture/dropout and limited sequencing depth), and therefore does not necessarily exclude low-level intracellular viral RNA. The study authors interpreted this discrepancy to suggest that, at the examined time point (24 hours post-exposure), productive infection could not be confirmed at single-cell resolution and that the observed host response may have been driven largely by sensing of viral particles or exposure to viral components rather than widespread productive replication. Despite the lack of detectable intracellular viral transcripts, H5N1 exposure was associated with strong immune signaling in these cells (Singh et al., 2025). This was considered indicative of pattern recognition receptor (PRR) activation, such as RIG-I or TLRs, by viral components (e.g., double-stranded RNA or RNA with 5′ triphosphates), initiating downstream interferon and cytokine cascades. The result was described as a mix of pro-inflammatory and regulatory immune signals, including rapid induction of inflammatory mediators, followed by the previously discussed Th2-skewed T cell response and an increase in antigen-presenting cells (DCs, macrophages), which may be involved in antigen delivery to lymph nodes for systemic immune activation (Singh et al., 2025).

In bacterial mastitis, the immune profile is also dominated by innate cells and Th-type responses, often leaning toward Th1 and Th17. Th17 cells, which produce IL-17, are important in host defense against extracellular bacteria and contribute to neutrophil recruitment in the bovine mammary gland (Rainard et al., 2020). There is evidence that effective defense against *E. coli* mastitis, for example, correlates with robust IL-17 (Th17) responses rather than Th1 responses (Duda et al., 2025) (Rainard et al., 2020). In viral mastitis (H5N1), the role of Th17 remains unclear, as the single-cell study did not explicitly report the presence of IL-17-producing cells. Instead, a Th2 imprint was observed. This has been interpreted as suggesting that the bovine immune system may respond differently to viral pathogens in the udder compared to bacterial ones, possibly due to differences in antigen presentation pathways or the cytokine milieu (e.g., interferon-rich viral infections influencing T-helper cell polarization).

Additionally, antibody-mediated responses are expected to develop. Cattle that survived H5N1 mastitis have been shown in experimental studies to generate neutralizing antibodies, which conferred short-term protective immunity upon re-challenge (Halwe et al., 2025). These antibodies may also be detectable in milk for some period following recovery.

Chapter 3: Methods

3.1 Experimental overview

Figure 3.1 provides a structured overview of the experimental design adopted in this thesis, illustrating the complete computational workflow used to analyze bovine milk somatic-cell single-cell RNA-seq datasets across H5N1-associated viral mastitis, bacterial mastitis, and healthy controls. The analytical workflow used in this study is summarized in a MindMap, spanning data acquisition, preprocessing in Galaxy platform, downstream analysis in Scanpy, and diagnostic marker panel derivation. The schematic highlights the sequential phases of dataset retrieval, standardised UMI-based alignment and quantification, quality control and data harmonisation, atlas construction with cell type annotation, lineage-aware differential expression, and functional/network-level interpretation, alongside composition analysis and trajectory inference. Detailed tool settings and key parameters for each step are provided in Table 1, ensuring full methodological transparency and reproducibility.

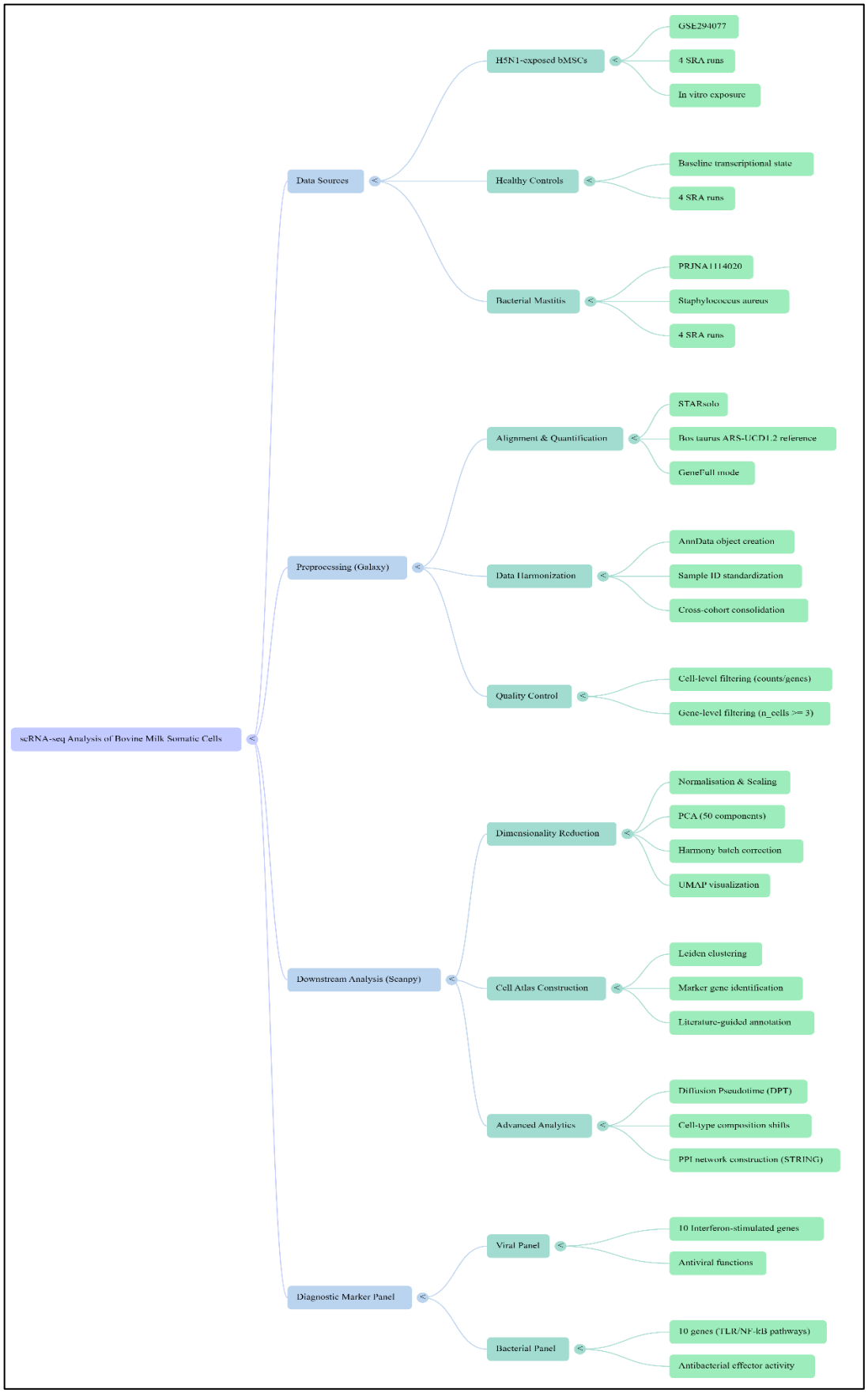


Figure 3.1:Workflow used to analyse bovine milk somatic-cell single-cell RNA-seq datasets. across H5N1-associated viral mastitis, bacterial mastitis, and healthy controls

Table 1:Summary of the single-cell RNA-seq bioinformatics workflow used in this study, listing the main analysis steps, the software/tools applied, key parameter settings, and the corresponding outputs generated at each stage (Dobin et al., 2013; Wolf et al., 2018; Virshup et al., 2024; Korsunsky et al., 2019)

Step Description	Tool / Function	Key Parameters / Settings	Input Data Type	Output Metadata / Variables	Source
Read alignment and gene quantification using STARsolo	RNA STARsolo mapping, demultiplexing and gene quantification	Chemistry: Chromium v2; --sjdbOverhang: 100; --genomeSAindexNbases: 2; UMI collapse: 1MM_All; Cell-barcode matching: 1MM_multi; GeneFull mode	Raw FASTQ (R1: barcode/UMI, R2: cDNA)	Sparse-matrix (matrix.mtx, features.tsv, barcodes.tsv), BAM with CB/UB tags	(Dobin et al., 2013; Dobin, 2024; 10x Genomics, 2024)
Import of count matrices and creation of AnnData objects	Scanpy Read10x	Annotation index: Gene ID (Ensembl IDs); Gene symbols stored in adata.var'gene_symbols'	STARsolo sparse-matrix directory	AnnData (.h5ad) object; gene_symbols	(Wolf et al., 2018; Wolf et al., n.d.; Virshup et al., n.d.)
Per-replicate sample-ID standardizatio n	Manipulate AnnData -> Rename categories of annotation	Rename replicates to CTRL1–4, H5N1_1–4, BAC1–4; Update existing key: No	AnnData (.h5ad)	obs'sample_id'	(Virshup et al., n.d.)
Cell-level quality control filtering	Scanpy FilterCells	Control/H5N1: $200 \le n\_genes \le 8000$, $300 \le n\_counts \le 80000$; Bacteria: $200 \le n\_genes \le 8000$, $500 \le n\_counts \le 80000$	AnnData (.h5ad)	n_genes, n_counts (filtered)	(Wolf et al., n.d.)
Gene-level quality control filtering	Scanpy FilterGenes	Min cells: 3; Max cells: 1,000,000,000	AnnData (.h5ad)	n_cells (filtered)	(Wolf et al., n.d.)
Post-filter cross-cohort consolidation and batch labeling	Manipulate AnnData (Concatenate)	Join: Intersection of variables; Batch annotation key: condition; Separator: '-'	Filtered AnnData objects (3 conditions)	obs'condition' (Control, H5N1, Bacteria)	(Wolf et al., n.d.; Virshup et al., n.d.)
Library-size normalization and log-transformation	Scanpy NormaliseData	Target number: 10000; Apply log transform: Yes; Exclude highly expressed: No	Integrated AnnData	Normalized and log-transformed counts in adata.X	(Wolf et al., n.d.)
Highly variable gene (HVG) selection	Scanpy FindVariableGene s	Flavour: Cell-ranger; Mean expression: 0.0125-3.0; Dispersion: 0.5-50.0; Batch key: sample_id; Bins: 20	Log-normalized AnnData	Highly variable genes flag	(Wolf et al., n.d.)

Batch-effect correction using Harmony	Scanpy Harmony	Basis: X_pca; Batch_key: batch_tech; Adjusted_basis: X_pca_harmony	PCA embeddings	adata.obsm'X_pca_harmony'	(Korsunsky et al., 2019; Wolf et al., n.d.)
kNN graph construction and clustering	Scanpy ComputeGraph; Scanpy FindCluster	n_neighbors: 15; n_pcs: 50; Metric: Euclidean; Algorithm: Leiden	Harmony-corrected PCA	obs'leiden' (Cluster labels)	(Traag et al., 2019; Wolf et al., n.d.)
Marker gene identification and cell-type annotation	Scanpy FindMarkers; Manual Curation	Top genes: 450 per cluster/condition; Adjusted $p < 0.05$; Ribosomal genes removed	Post-clustering AnnData	obs'Celltype'	(Wolf et al., n.d.)
Cell-type composition quantification	Galaxy Datamash	Group by: condition (col 9) & Celltype (col 12); Operation: Count (col 1)	Exported obs tabular file	Absolute abundance (n) and relative frequency (%)	(Afgan et al., 2018; Free Software Foundation, n.d.)
Diffusion pseudotime inference and signature scoring	Scanpy DPT; Scanpy tl.score_genes	Diffusion components: 20; Root: Control or Mammary epithelial cell; Signature: Top 400 genes	Integrated AnnData	dpt_pseudotime, H5N1_panel_score, Bacteria_panel_score	(Haghverdi et al., 2016; Wolf et al., n.d.)

3.2 Datasets and Data Sources.

The scRNA-seq data analysed in this thesis were obtained from two publicly available studies: Single-Cell Analysis of Host Responses in Bovine Milk Somatic Cells (bMSCs) Following HPAIV Bovine H5N1 Influenza Exposure (Singh et al., 2025) and Single-cell RNA sequencing characterization of Holstein cattle blood and milk immune cells during a chronic Staphylococcus aureus mastitis infection (Wiarda et al., 2025). The H5N1 dataset includes H5N1-exposed and matched control milk somatic cell samples, whereas the bacterial dataset represents in vivo chronic mastitis and includes milk (and blood) immune cells. Key dataset characteristics, sampling context, sequencing details, and SRA run accessions are summarised in **Table X**.

Table 2:Overview of the publicly available scRNA-seq datasets used in this thesis, summarizing study design, milk/udder sampling context, and sequencing workflow for the H5N1-exposed, healthy control, and bacterial mastitis cohorts, including the corresponding

Aspect	Study 1: H5N1 exposure (Singh et al., 2025)	Study 2: Bacterial mastitis (Wiarda et al., 2025)
Study (citation)	Singh et al., 2025 (Viruses) – bovine H5N1 exposure study	Wiarda et al., 2025 (Scientific Reports) – chronic Staphylococcus aureus mastitis study
Disease / condition	HPAIV H5N1 (A/dairy cattle/Kansas/5/2024); in vitro exposure of milk somatic cells; matched unexposed controls	Chronic bacterial mastitis after intramammary challenge with Staphylococcus aureus (Newbould 305)
Experimental context	In vitro model (cells exposed outside the animal); focuses on early host-response programmes at a defined time point	In vivo infection model (experimental intramammary infection in lactating cows); includes systemic (blood) and local (milk) compartments
Animals (breed / number)	Milk collected from three multiparous Holstein-Friesian cows (Kansas State University Dairy Teaching and Research Center)	Three mid-lactation Holstein cows; intramammary infusion into a single quarter
Sample type and udder/milk relevance	Bovine milk somatic cells (bMSCs) from raw milk (50 mL per quarter); analysed as a milk-cell atlas under H5N1 exposure vs control	Milk somatic cells from the infected quarter (local mammary response) plus PBMCs from blood (systemic response)
Time point(s) represented	24 hours post-exposure (time point selected for downstream scRNA-seq analysis)	Approximately four months post-challenge (chronic stage); milk and blood collected at the chronic mastitis time point
scRNA-seq platform and chemistry	10x Genomics Chromium Next GEM Single Cell 5′ Reagent Kits v2 (5′ gene expression)	10x Genomics Chromium Single Cell 3′ Gene Expression kit v3.1 (3′ gene expression)
Target capture / recovered cells	Target capture ~10,000 cells per sample; ~6,900 cells per treatment recovered; ~16,500 mean reads/cell reported	Target capture ~10,000 cells per sample; total cells reported: 12,231 (milk) and 15,105 (blood) across the three cows (6 samples total)
Sequencing instrument	Illumina NextSeq 550	Illumina HiSeq 3000 (libraries sequenced across 3 lanes)
Primary processing / alignment	Cell Ranger; combined reference built from ARS-UCD1.2/bosTau9 plus the bovine H5N1 genome (GenBank PP732373-80)	Cell Ranger v7.2; ARS-UCD1.2 Bos taurus reference (Ensembl annotation)
NCBI BioProject	PRJNA1248006	PRJNA1114020
NCBI accession	**SRA runs (H5N1):** SRR33028332, SRR33028333, SRR33028334, SRR33028335 **SRA runs (Control):** SRR33028336, SRR33028337, SRR33028338, SRR33028339	**SRA runs (Bacterial):** SRR29097727, SRR29097715, SRR29097697, SRR29097680

3.3 Read alignment, demultiplexing, and UMI gene quantification in Galaxy using STARsolo tool

All single-cell read processing was performed in the Galaxy platform https://usegalaxy.org/ , which automatically records every tool, version, input, and parameter in a permanent History, providing a fully auditable provenance that can be exported or shared with supervisors and reviewers. Tools are executed as containerised, versioned wrappers on managed compute,

eliminating local dependency issues and ensuring consistent runs across machines. Integration with public repositories (SRA) and reporting via MultiQC enabled the end-to-end workflow from raw FASTQ import to count-matrix generation to be executed within a single, reproducible environment.

Raw scRNA-seq reads were aligned and quantified with the Galaxy tool "RNA STARsolo mapping, demultiplexing and gene quantification for single cell RNA-seq" (STAR v2.7.11b+galaxy0). STARsolo was selected as the mapper/quantifier because STAR provides mature splice-aware alignment for mammalian genomes, while STARsolo adds droplet single-cell functionality compatible with 10x Genomics chemistries. The tool natively interprets cell barcodes and UMIs, emits CellRanger-style sparse matrices, and writes CB/UB tags to BAM for diagnostics, which simplifies downstream integration with Scanpy without external converters.

Libraries were prepared with 10x Genomics. For each sample, read 1 (R1) carried the cell barcode and UMI, and read 2 (R2) carried the cDNA insert. In the Galaxy form, input type was set to "Separate barcode and cDNA reads", with the R1 collection assigned to the Barcode reads field and the R2 collection assigned to the cDNA reads field. The "Drop-seq or 10X Chromium" single-cell RNA-seq type was chosen, and chemistry was set to "Chromium chemistry v2", allowing STARsolo to apply the appropriate 10x v2 barcode whitelist internally.

The *Bos taurus* ARS-UCD1.2 reference genome (GCF_002263795.1_ARS-UCD1.2_genomic. fna) (GCF_002263795.1_ARS-UCD1.2_genomic.fna) and the corresponding GTF annotation (GCF_002263795.1_ARS-UCD1.2_genomic.gtf) were supplied from the Galaxy history. The splice-junction database was built using exon features only. The junction overhang (--sjdbOverhang) was set to 100, matching read-2 length minus one in these libraries. The SA pre-indexing string (--genomeSAindexNbases) was set to 2 in accordance with the Galaxy tool recommendation for this genome size and to match the settings used in all runs.

UMI (Unique Molecular Identifiers) handling followed error-tolerant but conservative settings. UMIs were collapsed with the "1MM_All" option, which merges molecules within Hamming distance ≤ 1 to limit inflation from sequencing errors while preserving genuine molecules. Cell barcode to whitelist matching used "1MM_multi (CellRanger2)", allowing a single mismatch and

resolving multiple candidates using the posterior probability framework implemented in STARsolo.

Gene quantification used the GeneFull mode, counting reads overlapping both exons and introns while prioritising ≥ 50% exonic overlap and excluding reads with 100% antisense exonic alignment. This choice preserves informative nuclear and nascent RNA often present in droplet data. Libraries were treated as unstranded, as is standard for 10x droplet protocols. WASP allele-mapping bias filtering was disabled, and on-the-fly cell calling inside STARsolo was not used; putative empty droplets and low-quality barcodes were instead removed later during Scanpy quality control.

BAM outputs included the NH, HI and nM tags for alignment diagnostics. The mapping quality for uniquely mapped reads (--outSAMmapqUnique) was set to 60(Uniquely mapped reads were assigned 60 commonly used STAR setting, to label unique alignments as high-confidence and enable consistent filtering/interpretation of mapping quality across samples.) . Unmapped reads were not written to the BAM, and chimeric alignments were not reported, because fusion discovery was outside the scope of this analysis. The third field label in features.tsv was kept as "Gene Expression". For each sample, STARsolo produced a sparse-matrix directory (Solo.out/GeneFull/ containing matrix.mtx, barcodes.tsv and features.tsv), an aligned BAM annotated with cell and UMI tags, and detailed log files (Kaminow et al., 2021; Edgar et al., 2002; Leinonen et al.,2010; Szklarczyk et al., 2021).

3.4 Import of STARsolo matrices and creation of AnnData objects (Scanpy Read10x on Galaxy tool)

STARsolo outputs from all three cohorts (H5N1-exposed, controls, and chronic *Staphylococcus aureus* mastitis) were imported into Scanpy format using the Galaxy tool "Scanpy Read10x into hdf5 object handled by Scanpy" (v1.9.3+galaxy9). For each library, the 10x-style directory produced by STARsolo (matrix.mtx, features.tsv and barcodes.tsv under Solo.out/GeneFull/) was supplied: the GeneFull raw UMI count matrix as the expression input, the corresponding features file as the "Gene table", and the barcodes file as the "Barcode/cell table". No external cell or gene metadata tables were attached at this stage.

The tool was configured to emit an AnnData (.h5ad) object with the annotation index set to “Gene ID”. This assigns Ensembl gene identifiers from features.tsv to adata.var_names and stores gene symbols in adata.var['gene_symbols']. Using stable Ensembl IDs as the primary index ensured consistent gene identity across samples and datasets while retaining symbols for display and reporting.

One AnnData file per SRA run was generated and retained independently. In subsequent steps, these per-run objects were concatenated with explicit sample and condition labels added to adata.obs (e.g. sample_id, condition, cohort). During concatenation, barcodes were made unique by prefixing with the sample identifier, and both var_names and obs_names were validated and, if necessary, made unique to prevent index collisions. This import step therefore established a uniform, versioned bridge from STARsolo counts to Scanpy, preserving sparse-matrix structure and 10x metadata while avoiding any early filtering.

3.5 Per-cohort concatenation of replicates (Galaxy: Manipulate AnnData tool)

After creating one .h5ad per SRA run, the four biological replicates within each cohort (H5N1-exposed, matched controls and chronic bacterial mastitis) were merged on Galaxy using the “Manipulate AnnData object” tool. For each cohort separately, the function “Concatenate along the observations axis” was applied to the four per-run AnnData objects so that cells were stacked as additional rows while preserving the sparse count structure.

The join operation used “Union of variables”, which retains the complete gene set observed in any replicate and aligns by the Ensembl Gene ID index established during import. A batch-annotation column was created in obs using the “Key to add the batch annotation to obs” field (set to condition in the Galaxy forms), and barcodes were made unique by joining the original cell indices with the batch category using a hyphen separator. The default strategy for merging uns was kept, producing an empty dictionary for provenance fields not shared across inputs. No filtering or transformation was performed at this stage.

This procedure yielded three cohort-level AnnData files—one for H5N1, one for Control and one for Bacteria—each comprising the four runs from its respective group with unique cell barcodes, harmonized gene indices and an explicit per-run batch label.

3.6 Import, within-condition merging, and sample-ID standardisation (Galaxy: Manipulate AnnData tool)

Raw 10x Genomics feature-barcode matrices for each SRA library were imported into Scanpy using the Read10X function to generate one AnnData object per replicate. Replicates belonging to the same experimental condition were then merged into a single condition-specific AnnData object using the Galaxy tool Manipulate AnnData (concatenate). After concatenation, the sample identifier field (obs['sample_id']) was standardised using Manipulate AnnData → Rename categories of annotation (key: sample_id) to enforce consistent replicate naming. Specifically, control replicates were renamed to CTRL1–CTRL4, H5N1 replicates were renamed to H5N1_1–H5N1_4, and bacterial replicates were renamed to BAC1–BAC4. In all cases, the tool was configured to update the existing sample_id key (Add categories to a new key = No).

3.7 Initial cell level QC filtering (Galaxy: Scanpy FilterCells tool)

After harmonising sample_id and merging replicates within each condition, cell level quality control filtering was performed separately for each condition-specific AnnData object using the Galaxy tool Scanpy FilterCells based on counts and numbers of genes expressed. Cells were retained within defined ranges for the number of detected genes per cell (n_genes) and the total counts (UMIs) per cell (n_counts) to remove low-quality cells and cells with unusually high library size.

Control: cells were kept if 200 <= n_genes <= 8000 and 300 <= n_counts <= 80000.

H5N1: cells were kept if 200 <= n_genes <= 8000 and 300 <= n_counts <= 80000.

Bacteria: cells were kept if 200 <= n_genes <= 8000 and 500 <= n_counts <= 80000.

Filtering was run with the following tool settings: input and output formats set to AnnData (HDF5); Force recalculation of QC vars = No; Save adata to adata.raw before processing = No; and Save to 10x mtx format = No.

3.8 Gene-level filtering (Galaxy: Scanpy FilterGenes tool)

Gene-level quality control was performed after cell-level filtering using the Galaxy tool Scanpy FilterGenes based on counts and numbers of cells expressed. Filtering was applied separately to each condition-specific AnnData object (Control, H5N1, and Bacteria) to remove sparsely detected features. Genes were retained if detected in at least 3 cells (n_cells >= 3). An intentionally permissive upper bound was used (n_cells <= 1,000,000,000) to avoid excluding broadly expressed genes. Data were read and written in AnnData (HDF5) format, and all other tool settings were left at their defaults.

3.9 Post-filter cross-cohort consolidation and label standardization (Galaxy: Manipulate AnnData tool)

After gene-level filtering within each condition-specific dataset, the three filtered AnnData objects (Control, H5N1, and Bacteria) were merged into a single combined dataset using the Galaxy tool Manipulate AnnData with the function "Concatenate along the observations axis." The join method was set to "Intersection of variables" to retain only genes shared across all three datasets, ensuring a consistent feature space for downstream analyses. During concatenation, condition labels were added to cell metadata by writing a batch annotation key named condition into obs, and cell barcodes were made unique by appending the condition category to the existing observation names using "-" as the separator. Subsequently, the categorical levels of obs['condition'] were standardised using Manipulate AnnData → "Rename categories of annotation" (key: condition) to the final labels Control, H5N1, and Bacteria, updating the existing annotation key.

3.10 Batch-aware dataset integration with condition-based splitting

After creating the combined post-QC AnnData object, the dataset was partitioned into separate AnnData objects by experimental condition using Galaxy Manipulate AnnData with the function "Split the AnnData object into multiple AnnData objects based on the values of a given obs key," using obs['condition'] as the split key. This produced a condition-stratified collection (Control, H5N1, and Bacteria) for subsequent controlled recombination steps. To generate the 10x 5′ v2 (V2) subset, the Control and H5N1 AnnData objects were merged using Manipulate AnnData with "Concatenate along the observations axis." The join method was set to "Intersection of variables"

to retain only genes shared between the two objects, and cell indices were made unique by appending the batch category using "-" as the separator. During this concatenation, an additional batch annotation was added to obs to label the resulting dataset as the V2 group. Next, the bacterial dataset was merged with the V2 dataset using Manipulate AnnData "Concatenate along the observations axis," again using "Intersection of variables" to enforce a shared gene space across datasets and "-" as the separator to maintain unique observation names. A batch annotation key named batch_tech was added to obs during this step to track technology/source effects across the merged datasets. Finally, the categorical levels of obs['batch_tech'] were standardised using Manipulate AnnData → "Rename categories of annotation" (key: batch_tech), assigning the final labels Singh_v2_5p and Aureus_v3_3p, with the existing key updated in place.

3.11 Library-size normalization (Galaxy: Scanpy NormaliseData tool)

Library-size normalisation was applied to the post-QC integrated AnnData object using the Galaxy tool "Scanpy NormaliseData (make all cells having the same total expression)." Counts were scaled so that each cell had a total of 10,000 counts (target number to normalise to = 10000), followed by log-transformation (apply log transform = Yes). Highly expressed genes were not excluded when computing the normalisation factor (exclude highly expressed genes = No). The analysis was performed on the AnnData (HDF5) input and saved as an AnnData output; no 10x mtx export was generated (save to 10x mtx format = No), the data were not stored in adata.raw prior to processing (save adata to adata.raw before processing = No), and no additional layer was specified for saving adata.X before normalisation. Normalisation-factor storage in adata.obs was not enabled (key field left blank).

3.12 Selection of highly variable genes (Galaxy: Scanpy FindVariableGenes tool)

Highly variable genes (HVGs) were selected from the log-normalised AnnData object using the Galaxy tool "Scanpy FindVariableGenes based on normalised dispersion of expression." HVGs were identified with the dispersion-based "Cell-ranger" flavour, applying mean-expression thresholds of 0.0125 (min) and 3.0 (max) on the log1p scale, and dispersion thresholds of 0.5 (min) and 50.0 (max). Mean expression was binned into 20 bins for dispersion normalisation. To reduce batch-driven HVG selection, HVGs were computed within each `sample_id` batch (batch key =

`sample_id`) and then combined across batches. The dataset was subsequently subset in-place to the HVG feature set (remove genes not marked as highly variable = Yes). No predefined gene lists were provided for always/never HVGs.

3.13 Gene-wise scaling and centering (Galaxy: Scanpy ScaleData tool)

Scaled expression values were computed on the highly variable gene subset using the Galaxy tool "Scanpy ScaleData (make expression variance the same for all genes)." The AnnData (HDF5) object containing HVGs was standardised by zero-centering each gene prior to scaling (zero center data before scaling = Yes). No post-scaling truncation was applied (max value left blank). The data were not saved to adata.raw before processing (save adata to adata.raw = No), and no layer was specified for saving adata.X prior to scaling. The scaled dataset was returned as an AnnData output.

3.14 Principal component analysis (Galaxy: Scanpy RunPCA tool)

Principal component analysis (PCA) was performed for dimensionality reduction using the Galaxy tool "Scanpy RunPCA." PCA was computed on the highly variable gene expression matrix (AnnData HDF5 input), generating 50 principal components (n_comps = 50). Incremental PCA by chunks was not used (chunked = No). Data were mean-centered prior to PCA (zero_center = Yes). The SVD solver was left at the default setting (not specified), and a fixed random seed was applied for reproducibility (random_state = 0). The output was saved as an AnnDataobject containing PCA coordinates and associated variance information.

3.15 Harmony batch-effect correction on PCA embeddings (Galaxy: Scanpy Harmony tool)

Harmony-based batch-effect correction was applied to the PCA space using the Galaxy tool "Scanpy Harmony (adjust principal components for variables that might introduce batch effect)." The PCA embeddings stored in adata.obsm['X_pca'] were used as the input basis (basis = X_pca), and batch structure was defined by the metadata column adata.obs['batch_tech'] (batch_key = batch_tech). Harmony was run with programme default parameters enabled, and the corrected low-dimensional representation was saved to adata.obsm['X_pca_harmony'] (adjusted_basis =

X_pca_harmony) for use in downstream neighbour graph construction, clustering, and UMAP visualisation.

3.16 k-nearest neighbor graph construction on Harmony-corrected PCA space (Galaxy: Scanpy ComputeGraph tool)

A k-nearest neighbor (kNN) graph was constructed using the Galaxy tool "Scanpy ComputeGraph (derive kNN graph)" on the Harmony-corrected AnnData object. Programme defaults were disabled to explicitly control graph parameters. Neighbourhood size was set to 15 (n_neighbors = 15). The graph was computed using the PCA representation stored in X_pca (use_rep = X_pca), using 50 principal components (n_pcs = 50). A hard kNN threshold was applied (knn = Yes) rather than a Gaussian kernel weighting scheme. Connectivity was computed using the UMAP method (method = UMAP) with Euclidean distance as the metric (metric = Euclidean). A fixed random seed was used for reproducibility (random_seed = 0). The resulting distances and connectivities were stored in the AnnData object under the default neighbours slots.

3.17 UMAP dimensionality reduction on the kNN graph (Galaxy: Scanpy RunUMAP tool)

Uniform Manifold Approximation and Projection (UMAP) was performed using the Galaxy tool "Scanpy RunUMAP (visualise cell clusters using UMAP)." The analysis used the precomputed k-nearest neighbor graph stored under the neighbors slot (neighbors_key = neighbors). Programme default parameters were applied. The resulting two-dimensional UMAP coordinates were written to the AnnData object (and exported as TSV embeddings), enabling visualisation of the integrated cell landscape and subsequent colouring by condition, cluster labels, and cell type annotations.

3.18 Leiden graph-based clustering on the kNN graph (Galaxy: Scanpy FindCluster tool)

Graph-based community detection was used to identify transcriptionally coherent cell populations. Clustering was performed in Galaxy using the Scanpy FindCluster tool, applying the Leiden algorithm to the precomputed k-nearest neighbor (kNN) graph generated in the previous step. Programme default parameters were used. The resulting cluster assignments were stored as a categorical label in the AnnData object and exported as a two-column table (cell identifier and

cluster label) for downstream cell type annotation and visualisation on the UMAP embedding.

3.19 Marker gene identification and cluster/condition annotation were performed using Scanpy FindMarkers

Cluster- and condition-specific marker genes were identified using the Galaxy tool “Scanpy FindMarkers”. Differential expression was computed based on the post-clustering AnnData object, and the results were exported as a tab-separated marker table. For cluster-level annotation, marker genes were ranked separately for each Leiden community by setting the grouping variable to leiden and reporting the top 450 genes per cluster. These ranked markers were then used to assign biological cell type identities to each cluster by comparing the enriched genes against established cell type marker signatures reported in the literature and in curated marker resources, prioritising concordant multi-gene signatures rather than single-gene matches. In addition, marker analysis was repeated at the condition level by setting the grouping variable to condition (top 450 genes per condition) to summarise the dominant transcriptional programmes distinguishing experimental conditions and to support interpretation of pathogen-associated expression patterns across the integrated dataset.

3.20 Method name: Marker-gene curation and literature-guided cell type annotation

Excel-based filtering was applied to the marker-gene tables generated from Scanpy FindMarkers for both condition-level and Leiden-level comparisons. The results were exported from Galaxy as tab-delimited files and were merged in a spreadsheet environment to enable manual quality control. For each comparison, candidate marker genes were filtered to retain only entries with adjusted p values below 0.05 and with robust effect sizes; genes showing weak magnitude or inconsistent patterns across contrasts were flagged and excluded. Ribosomal protein genes were subsequently removed from all marker lists prior to downstream analyses. This exclusion was applied because ribosomal genes typically exhibit high, relatively non-specific expression across multiple cell types and conditions and are more reflective of global translational activity than pathogen-associated immune recognition or effector responses. Cell type names were then assigned to Leiden clusters using a literature- and knowledgebase-guided marker approach.

Cluster-enriched genes from the curated Leiden-level marker lists were compared against canonical lineage markers reported in peer-reviewed studies of bovine milk somatic cells and mastitis-associated immune populations. Candidate identities were cross-referenced using CellKb to support naming decisions based on multiple concordant markers. In parallel, top cluster markers were queried in STRING to verify that the proposed cell identity was consistent with known functional interaction modules, increasing confidence in the assigned labels. Numeric cluster IDs were subsequently converted into biologically interpretable cell type names by renaming the cluster categories in the AnnData object for downstream reporting and visualization. Using this procedure, clusters were labelled as: Antigen-presenting Myeloid, B cells, Cytotoxic lymphocytes, IFN-stimulated cells (ISG-high), Inflammatory IFN-responsive Monocytes/Macrophages, Inflammatory Macrophages/Monocytes_2, Inflammatory Monocytes/Macrophages_1, Mammary epithelial cell, Mature/Migratory Dendritic Cells_1, Mature/Migratory Dendritic Cells_2, Neutrophils/monocytes/macrophages, Neutrophils_1, Neutrophils_2, Neutrophils_3, Neutrophils_4, Putative Endothelial / Vascular-like, T cells, and Tissue/Resident Macrophages.

The curated, non-ribosomal marker tables (condition-level and cell type/cluster-level) used for STRING analyses and diagnostic panel design were provided as supplementary Files S1 and S2 .

3.21 Cluster relabelling and cell type annotation mapping in AnnData

Numeric Leiden cluster identifiers were converted into biologically interpretable cell-type labels within the AnnData object. First, the categorical values of the leiden annotation in adata.obs were renamed using the “Rename categories of annotation” function, where each original Leiden category was replaced with its corresponding cell type label. The new labels comprised: Neutrophils_1, Neutrophils_2, Neutrophils/monocytes/macrophages, IFN-stimulated cells (ISG-high), Inflammatory Monocytes/Macrophages_1, T cells, Antigen-presenting Myeloid, Inflammatory Macrophages/Monocytes_2, Tissue/Resident Macrophages, Mature/Migratory Dendritic Cells_1, Mature/Migratory Dendritic Cells_2, B cells, Cytotoxic lymphocytes, Neutrophils_3, Neutrophils_4, Mammary epithelial cell, Putative Endothelial / Vascular-like, and Inflammatory IFN-responsive Monocytes/Macrophages. Following category relabeling, the observation field name was standardized by renaming the leiden column in adata.obs to Celltype using the “Rename fields in AnnData observations” function, thereby preserving the assigned cell

type identities under a dedicated metadata key for downstream analyses, reporting, and visualization.

3.22 UMAP embedding visualisation (Scanpy PlotEmbed)

UMAP embeddings were visualised using Scanpy PlotEmbed and the tool was executed twice, once with cells coloured by Celltype and once with cells coloured by condition, using identical plotting settings in both runs. The AnnData object containing the computed UMAP coordinates was used as input, and the embedding stored under the key umap was selected for plotting. The legend was positioned on the right margin with a font size of 15, and a 2D projection was generated with a figure size of 5 × 5 at 100 dpi. Plot frame display was disabled, neighbour edges were not shown, and point size was left at the default setting.

3.23 STRING-based PPI network construction and functional enrichment

Condition-specific gene sets were first exported from the integrated single-cell analysis by running Scanpy's differential expression at the condition level (groupby = condition) and saving the results as an Excel table titled "Condition Marker genes.xlsx"; this file captured the rank_genes_groups output per condition and served as the direct input for Objective 3 analyses (H5N1, Bacteria, Control). For each condition, a STRING protein–protein interaction (PPI) network (https://string-db.org/**)** was assembled from the corresponding gene set under a uniform stringent configuration: the combined interaction score threshold was set to ≥ 0.700 and disconnected nodes were hidden so that only edges supported by multiple evidence channels (experimental data, curated databases, co-expression, text mining and gene neighbourhood/fusion/co-occurrence) were retained, yielding high-confidence functional cores comparable across conditions. Networks were then partitioned into local clusters and summarised topologically; functional enrichment was performed using Gene Ontology (GO), KEGG and STRING "Local Network Cluster" categories as provided within the STRING platform. For the enrichment visualisations, GO terms were grouped at similarity ≥ 0.7, multiple-testing control was applied with FDR ≤ 0.05, a minimum count of 3 was enforced, and reporting was ordered by −log (FDR). These settings were held constant across panels to enable direct cross-condition comparison. All STRING nodes tables are supplied in the supplementary files S1 and S2 .

3.24 Design of a lineage-aware 24-gene diagnostic panel

For diagnostic panel construction, analyses were restricted to H5N1-exposed and bacterial mastitis samples, while control cells were excluded to focus on transcriptional programmes that distinguish viral from bacterial infection rather than general handling- or background-related responses. Condition-level marker genes were sourced from the previously generated "Condition – Marker genes" table, and lineage context for interpretation and anchoring was derived from the curated cell-type annotations stored in obs['celltype']. Candidate genes were then filtered to remove non-informative features and to prioritise markers showing a clear directional bias toward either the H5N1 or the bacterial condition, rather than being similarly induced in both settings.

For the viral panel, a subset of ten interferon-stimulated genes with well-established direct antiviral functions against RNA viruses was selected from the ranked list. Selection was informed by a targeted literature review, drawing on experimental studies and reviews demonstrating that these genes restrict viral entry, replication or budding in influenza and related infections, and by their consistent enrichment in the H5N1 condition in the current dataset and in published H5N1-related transcriptomic studies in cattle.

For the bacterial panel, statistical ranking was combined with functional annotation, network-based analysis and evidence from the mastitis literature. Genes were examined within modules derived from protein–protein interaction and co-expression networks that were enriched for Toll-like receptor signalling, NF-κB signalling, inflammasome activation and antibacterial effector functions in bovine mastitis datasets. Based on this network context and on published experimental data implicating these genes in bacterial recognition, antimicrobial effector mechanisms and regulation of inflammatory damage in the udder, ten genes were chosen that collectively captured proximal recognition of bacterial ligands, direct antibacterial effector activity and core NF-κB/TLR regulatory circuits repeatedly associated with mastitis susceptibility and somatic cell traits. This three-step procedure combining statistical significance and directionality, pathogen-focused functional roles and support from the existing literature and network context yielded two compact, mechanistically interpretable 12-gene panels representing viral and bacterial mastitis transcriptional programs.

3.25 Diagnostic panel scoring and sample-level pseudo-bulk evaluation

Two diagnostic gene signatures were quantified at the single-cell level using the Scanpy function tl.score_genes implemented in Galaxy via Scanpy Inspect and manipulate → "Score a set of genes (tl.score_genes)". For each signature, a per-cell score was computed by comparing the average expression of the input gene set to a background reference set of genes matched for overall expression level. Specifically, a set of control/background genes was sampled from the transcriptome using expression-level binning, and the signature score for each cell was calculated as the mean expression of the signature genes minus the mean expression of the matched background genes, thereby controlling for global expression differences across cells. Default Scanpy scoring parameters were applied so that background genes were selected using binned expression stratification (n_bins) and a fixed random seed (random_state), ensuring reproducible scoring across all cells.A viral-response score and a bacterial-response score were computed separately by providing the corresponding gene lists. For the viral score, the genes MX1, MX2, ISG15, ZBP1, RSAD2, OAS1Y, OAS1YX, EPSTI1, IFI6, IFI44, IFI44L, and IFITM1 were used, and the resulting per-cell score was stored in AnnData .obs under H5N1_diagnostic_panel. For the bacterial score, the genes NFKB1, NFKBIZ, NFKBIA, TNFAIP3, TNFAIP6, CXCL8 (IL8), IL1B, CXCR2, TLR4, CD14, S100A9, and IL1RN were used, and the per-cell score was stored in .obs under Bacteria_diagnostic_panel. Following signature scoring, the updated AnnData object was exported using Export AnnData matrix and annotations, and the one-dimensional observations annotation table (obs) was used for tabular aggregation. The exported obs table was reduced to the required fields using Cut columns from a table, retaining sample_id, condition, H5N1_diagnostic_panel, and Bacteria_diagnostic_panel. Sample-level pseudo-bulk summaries were then generated using Datamash (operations on tabular data) by grouping rows by sample_id and condition and calculating mean(H5N1_diagnostic_panel) and mean (Bacteria_diagnostic_panel) per group, alongside count(sample_id) to report the number of contributing cells (n_cells) per sample. Control samples were excluded from the final reported evaluation so that the quantitative comparison focused on H5N1 versus bacterial mastitis only. A single discriminant metric was subsequently computed for each sample as DeltaScore = mean(H5N1_diagnostic_panel) − mean(Bacteria_diagnostic_panel), and the resulting per-sample mean scores, DeltaScore, and n_cells were reported for downstream interpretation. Using the per-sample DeltaScore, a simple decision rule was applied for descriptive classification, whereby

samples with DeltaScore > 0 were assigned to the H5N1 class and samples with DeltaScore ≤ 0 were assigned to the bacterial class. Classification performance was summarised at the sample level as sensitivity, specificity, and accuracy, and exact (Clopper–Pearson) 95% confidence intervals were calculated for these proportions to report uncertainty under the small sample size.

3.26 Datamash based cell type composition quantification

Cell type composition across experimental conditions was quantified using the per cell metadata exported from the final annotated AnnData object. The observations table (obs) was first generated in Galaxy using the "Export AnnData matrix and annotations" tool, producing a key-indexed, one-dimensional observations file in tabular format. This table contained one row per cell barcode and included, among other metadata fields, the condition label (Control, H5N1, Bacteria) and the curated cell type annotation (Celltype) derived from the renamed Leiden clusters.

The exported obs table was then processed using the Galaxy Datamash tool to compute cell abundances in a reproducible, non-interactive manner. Grouping was performed on the columns corresponding to condition and cell type annotation (specified by their column indices in the exported table; in this workflow, columns 9 and 12 were used). Sorting of the input file by the grouping columns was enabled to ensure that identical group labels were contiguous, and the header line was retained and re-printed in the output to preserve column identity. For each unique condition × cell type combination, Datamash was configured to calculate a simple count statistic, where the count operation was applied to the cell identifier column (cell barcode; column 1). This produced a compact summary table containing, for each group, the condition, the cell type label, and the corresponding number of cells (n).

To facilitate reporting in the thesis tables, the Datamash output was subsequently imported into a spreadsheet environment where counts were reshaped into a wide format with one row per cell type and separate columns for each condition. Within-condition proportions were calculated by dividing each cell type count by the total number of cells in the corresponding condition, yielding both absolute abundances (n) and relative frequencies (%). A final "Total_n" column was computed as the sum of counts across all conditions for each cell type, enabling direct comparison of both compositional shifts and overall representation of annotated lineages (e.g., Antigen-

presenting Myeloid, Tissue/Resident Macrophages, Neutrophils_1–4, Mammary epithelial cell, and lymphocyte subsets).

3.27 DiffusionMap–DPT Pseudotime Inference with Gene-Panel Scoring and Pseudotime Scatter Visualisation

A diffusion-map embedding was first computed from the integrated AnnData object in Galaxy using Scanpy DiffusionMap, with 20 diffusion components calculated and the existing neighbour graph used (neighbors key = neighbors). Diffusion pseudotime was then inferred using Scanpy DPT, again using 20 diffusion components, with the kNN graph taken from the same neighbors slot. A root state was specified to orient the trajectory (root attribute set to a categorical metadata field, and the root category selected as Control in one run and as Mammary epithelial cell in an alternative run), while branch detection was not enabled (number of branchings = 0).

To quantify pathogen-associated transcriptional programmes along the inferred trajectory, two condition-specific gene signatures were defined from the Scanpy FindMarkers (condition-level) results by selecting the top 400 ranked genes for H5N1 and the top 400 ranked genes for Bacteria. These gene lists were then scored per cell using Scanpy Inspect and Manipulate: "Score a set of genes (tl.score_genes)", generating H5N1_panel_score and Bacteria_panel_score as new fields in obs. Finally, the relationship between trajectory position and signature activity was visualised using Scanpy plot (scatter; pl.scatter), plotting x = dpt_pseudotime against y = H5N1_panel_score or y = Bacteria_panel_score, with points coloured by condition and the plot restricted to the H5N1 and Bacteria groups.

3.28 AI-Assisted Writing and Verification

AI tools were used to improve writing clarity and to cross-check selected methodological statements against the official sources cited in the Methods. All content was manually reviewed and verified, and AI was not used to generate results or modify data.

Chapter 4: Results

4.1 Integrated single-cell atlas of bovine milk somatic cells

An integrated single-cell atlas of bovine milk somatic cells was generated across three experimental conditions (Bacteria, Control, and H5N1), and the transcriptional landscape was visualized using UMAP. In Figure4.1 (UMAP colored by condition), a broad distribution of cells from the bacterial condition was observed within the main manifold on the left side of the embedding, with substantial control intermixing across the same region. In contrast, H5N1 cells were found to be enriched within a relatively distinct area in the upper-right portion of the UMAP, indicating the presence of a condition-associated transcriptional separation affecting a specific subset of the cellular landscape rather than the entire dataset. The control cells were distributed across multiple regions, consistent with their role as a baseline reference spanning several lineages and activation states.

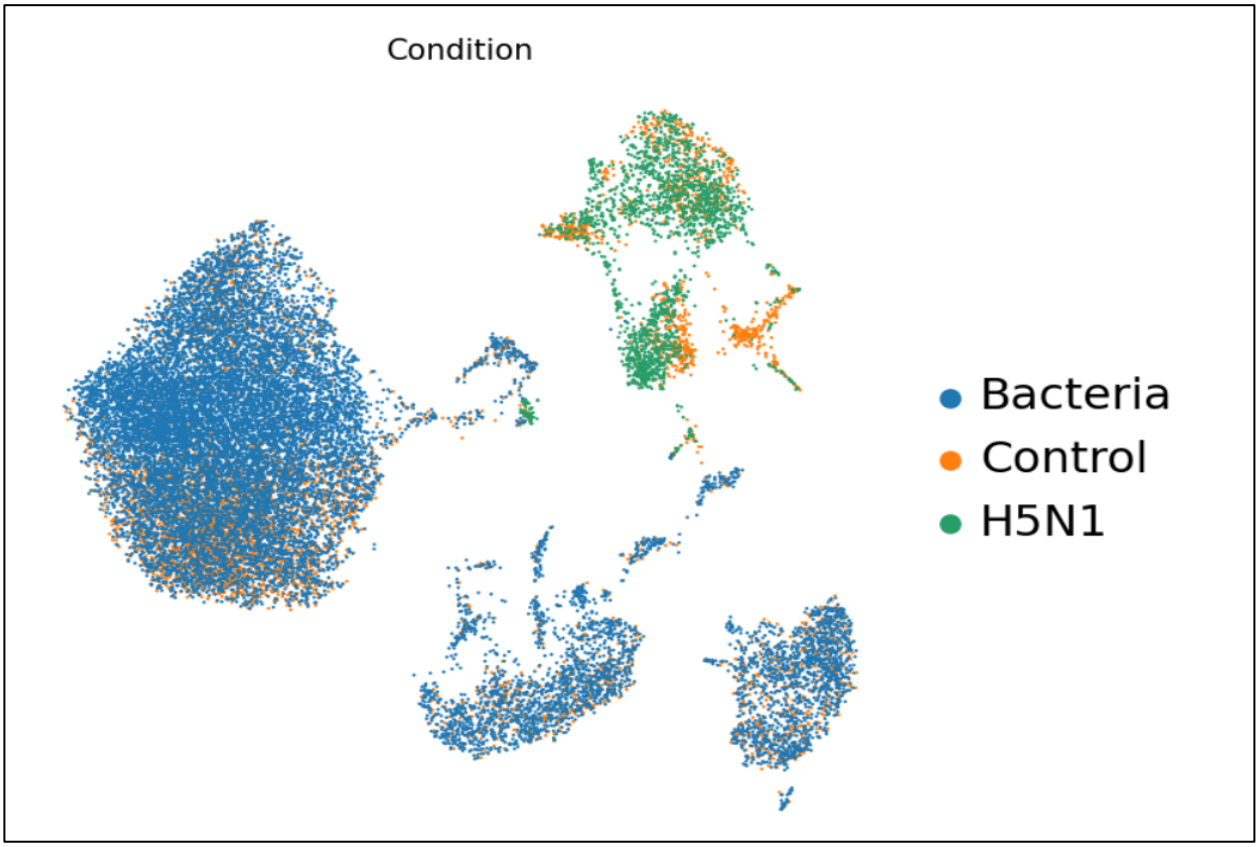


Figure 4.1:UMAP embedding of the integrated single-cell dataset coloured by condition (H5N1, bacterial mastitis, control), showing condition-associated structuring of the cellular landscape with partial overlap between groups.

In Figure 4.2 (UMAP colored by annotated cell types), the overall UMAP structure was largely accounted for by differences in cell lineages and activation programmes. The largest compartment in the left manifold was dominated by innate immune populations, where multiple neutrophil states (Neutrophils_1–4) were resolved alongside diverse monocyte/macrophage programmes, including inflammatory and tissue/resident-like subsets. An IFN-stimulated (ISG-high) population was delineated as a prominent antiviral activation state, while an Inflammatory IFN-responsive Monocytes/Macrophages compartment was distinguished as a myeloid programme combining inflammatory and interferon-associated transcriptional features. More discrete islands were also identified, corresponding to T cells, B cells, and cytotoxic lymphocytes, in addition to two groups of mature/migratory dendritic cells (Mature/Migratory DCs_1 and _2). A smaller non-immune compartment consistent with mammary epithelial cells and a rare putative endothelial/vascular-like population were also separated within the embedding. summarizes these eighteen clusters by listing each annotated cell type together with its cluster ID in the integrated bovine milk somatic cell atlas, providing a concise reference for the cell populations used in subsequent objectives (Table 3).

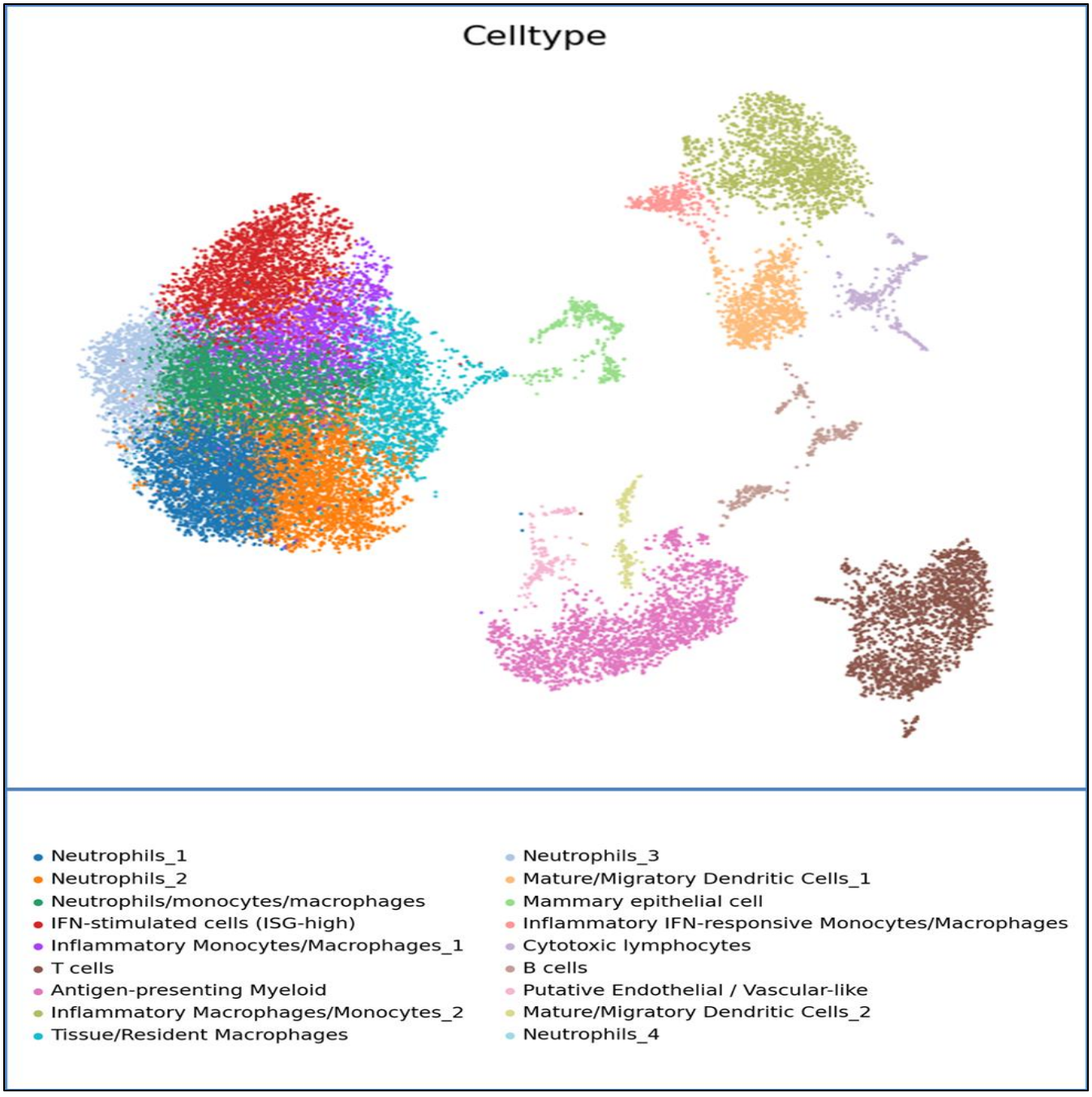


Figure 4.2: UMAP embedding of the integrated single-cell dataset coloured by annotated cell type, showing clear separation of major immune populations and mammary epithelial cells across the atlas

Table 3:Annotated cell populations.

Cluster ID	*Annotated cell populations*
0	Neutrophils_1
1	Neutrophils_2
2	Neutrophils/monocytes/macrophages
3	IFN-stimulated cells (ISG-high)
4	Inflammatory Monocytes/Macrophages_1
5	T cells
6	Antigen-presenting Myeloid
7	Inflammatory Macrophages/Monocytes_2
8	Tissue/Resident Macrophages
9	Neutrophils_3
10	Mature/Migratory Dendritic Cells_1
11	Mammary epithelial cell
12	Inflammatory IFN-responsive Monocytes/Macrophages
13	Cytotoxic lymphocytes
14	B cells
15	Putative Endothelial / Vascular-like
16	Mature/Migratory Dendritic Cells_2
17	Neutrophils_4

4.2 Pathogen-specific gene programs

Intersecting condition-level upregulated genes with lineage marker sets produced distinct pathogen-related overlaps across the integrated atlas (Table 4). Across the neutrophil continuum (Neutrophils_1–4 and the mixed Neutrophils/monocytes/macrophages lineage), bacterial overlaps were retained consistently and were substantially larger than the corresponding H5N1 overlaps. By contrast, the H5N1 overlaps within these neutrophil lineages were comparatively limited and were retained as small intersections relative to the bacterial-associated sets. Within inflammatory and antigen-presenting myeloid compartments, overlaps were retained under both H5N1 and bacterial exposure, while non-identical gene sets were retained for each condition, indicating pathogen-class–dependent programmes within shared myeloid lineages. Interferon-associated lineages, including Interferon-Activated Dendritic Cells and Inflammatory IFN-responsive

Monocytes/Macrophages, retained H5N1 overlaps consistent with interferon-stimulated transcriptional programmes, while bacterial overlaps were also retained but differed in composition. Within dendritic-cell lineages, Mature/Migratory Dendritic Cells_1 retained an H5N1 overlap without a qualifying bacterial overlap, whereas Mature/Migratory Dendritic Cells_2 retained overlaps under both conditions. Lymphoid lineages (T cells, cytotoxic lymphocytes, and B cells) retained overlaps under both pathogen classes, although bacterial intersections in B cells and cytotoxic lymphocytes were restricted in size. Non-immune lineages (mammary epithelial and putative vascular-like) retained overlaps under both conditions, indicating that pathogen-associated programmes were not confined to immune compartments. Thus, Table 3 summarizes, for each lineage, the overlap genes between its marker set and the upregulated gene list for H5N1 and/or bacterial exposure, with only the top 10 genes shown per overlap; supplementary Table S3 provides the complete overlap lists for all lineages in the same format.

Table 4:Pathogen-specific gene programs within milk somatic cell lineages under H5N1 and bacterial conditions (top 10 genes shown per overlap).

Cell type (lineage)	H5N1 overlap (top 10)	Bacteria overlap (top 10)
Antigen-presenting Myeloid	PFN1, PRDX1, FTH1, PPIA, CCL4, CTSZ, CSTB, ANXA2, CTSH, ACTB	JUN, NFKBIZ, RGCC, CD14, ERO1A, ENO1, SPECC1, TPT1
B cells	PFN1, PRDX1, PPIA, CSTB, ACTB, UBA52, FABP5, ZC3H10, CALM1, FTL	TPT1
Cytotoxic lymphocytes	NDST3, CCL3, MT2A, PFN1, PRDX1, IFITM1, CCL5, PPIA, ISG15, ACTB	LTB
Inflammatory IFN-responsive Monocytes/Macrophages	NDST3, CCL3, MT2A, PRDX1, IFITM1, FTH1, TAP, CCL5, CXCL3, CCL4	CXCL8, PTPRJ, CXCL2, GPR84, XPO6, DMXL2, PHF12, TG, TNFRSF1B, IL1RN
Inflammatory Macrophages/Monocytes_2	NDST3, CCL3, MT2A, PFN1, PRDX1, IFITM1, FTH1, TAP, CCL5, MMP9	CXCL2

Inflammatory Monocytes/Macrophages_1	ISG15, C10H15orf48, SOD2, RETN, MED12L, CD274, GTF2B, BCL2A1	BTG1, SAMSN1, ANTXR2, SRGN, BASP1, MAML2, PLAUR, PDE4B, SDS, B2M
Interferon-Activated Dendritic Cells	ISG15, C10H15orf48, MX1, IFI6, OAS1Y, CD274, MX2, RSAD2, GTF2B, IFI44L	BTG1, SAMSN1, ANTXR2, SRGN, BASP1, MAML2, PLAUR, PDE4B, SDS, B2M
Mammary epithelial cell	PPIA, ANXA2, UBA52, ZC3H10, NUPR1, NPC2, ATP5MC2, HSPB1, CD9, ATP5MG	ERC2, TPT1
Mature/Migratory Dendritic Cells_1	NDST3, PFN1, PRDX1, IFITM1, TAP, PPIA, ISG15, ANXA2, CTSH, ACTB	—
Mature/Migratory Dendritic Cells_2	PFN1, PPIA, CTSH, ACTB, UBA52, FABP5, ZC3H10, CTSC, CALM1, CRIP1	PLXDC2, MAP4K4, PARM1, ROCK1, NR4A3, REL, CYTH1, TLE4, H3F3A, SPECC1
Neutrophils/monocytes/macrophages	CXCL5, GTF2H2, IFNAR2, BCL2A1, SLC11A1, RCL1	BTG1, SAMSN1, ANTXR2, SRGN, BASP1, MAML2, PLAUR, PDE4B, SDS, B2M
Neutrophils_1	TMSB4X, TMSB10	BTG1, ANTXR2, SRGN, BASP1, MAML2, PLAUR, PDE4B, SDS, B2M, CXCR2
Neutrophils_2	C10H15orf48, SOD2, DBI, TMSB10	BTG1, SAMSN1, ANTXR2, SRGN, BASP1, PLAUR, PDE4B, SDS, B2M, CXCR2
Neutrophils_3	ACTB, ACTG1, SLC7A11	BTG1, SAMSN1, ANTXR2, SRGN, BASP1, MAML2, PLAUR, PDE4B, SDS, PICALM
Neutrophils_4	IFNAR2, TBC1D4, KAT8	BTG1, SAMSN1, ANTXR2, SRGN, BASP1, PLAUR, PDE4B, SDS, B2M, CXCR2

Putative Endothelial / Vascular-like	CD36, NOS2, ARHGAP35, CLIC4, NEDD4L, ARHGAP10, NRP2, TBC1D31, ZC3H12C	PICALM, PDE4D, QKI, MAP3K2, SIK3, VPS13B, ZFAND3, PLXDC2, TBL1X, CUX1
T cells	UBA52, ZC3H10, CALM1, CRIP1, ATP5MC2, SH3BGRL3, ATP5MG, IL7R, EEF1A1, HSPA8	FYB1, TNIK, ARHGAP15, PDE3B, PTPRC, FOXP1, PPP1R16B, TPT1
Tissue/Resident Macrophages	FTH1, CCL4, CTSZ, CTSB, NUPR1, FTL, SOD2, RNASEK, RETN, SQSTM1	BTG1, SRGN, MAML2, PDE4B, SDS, B2M, PICALM, RUBCNL, G0S2, JMJD1C

4.3 Functional pathway enrichment and network analysis

Objective 3 is used to characterize the functional biology underlying the condition-specific gene signatures through the integration of protein–protein interaction (PPI) topology with pathway and ontology enrichment. For each identified STRING subnetwork cluster, the interaction structure is summarized using node-degree centrality so that hub genes likely to coordinate each module are highlighted. Functional enrichment is then reported using GO/KEGG/Reactome (and related annotations where available) so that the dominant biological processes and signaling programmes represented by each module are defined. Through this combined network–function framework, a systems-level interpretation is enabled beyond individual differentially expressed genes, and cross-condition comparisons are supported at the level of pathway modules. Complete node-degree metrics and gene lists are provided in the supplementary S4.

4.3.1 H5N1 String Network

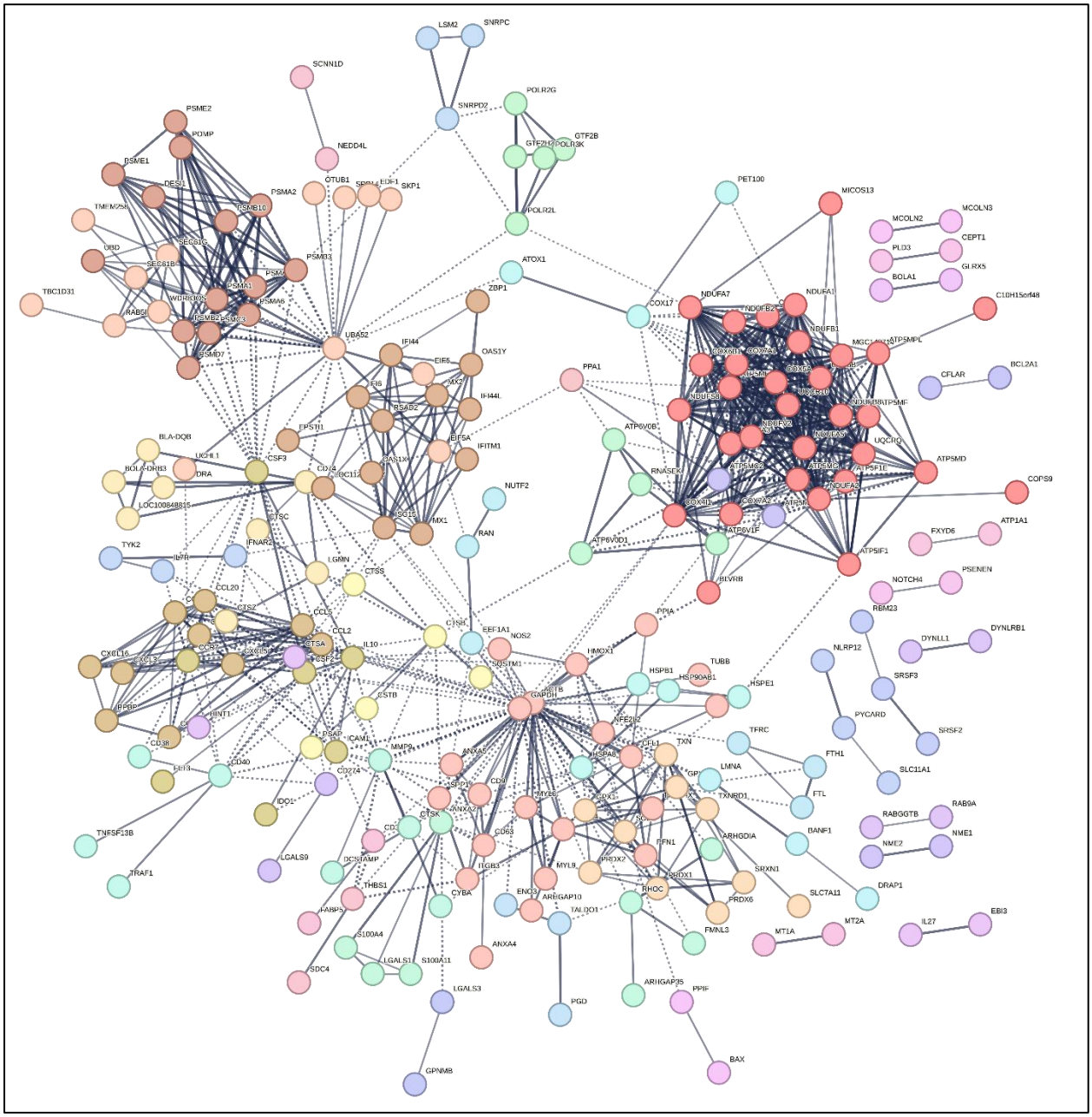


Figure 4.3:STRING protein–protein interaction (PPI) network for the H5N1 condition. Nodes represent proteins and edges indicate known/predicted functional associations; nodes are coloured by STRING local network clusters to highlight modular organization of the viral-response interactome.

Clusters

color	cluster id	gene count	description
	Cluster 1	31	+ Oxidative phosphorylation
	Cluster 2	31	ACTB, ACTB1, ANXAA, ANXAS, ARHGAP10, CD63, CD9, CFL1, GAPDH, HMOX1, ITG...
	Cluster 3	14	+ Proteasome
	Cluster 4	14	EDF1, EIFSA, EIF6, OTUB1, RABSIF, SEC61B, SEC61B, SKP1, SRP14, TBC1D31, TME...
	Cluster 5	13	+ Defense response to virus
	Cluster 6	10	+ Cell redox homoostasis
	Cluster 7	10	+ Neutrophil chemotasis
	Cluster 8	9	+ Antigen processing and presentation
	Cluster 9	7	African trypanosomiasis
	Cluster 10	6	Thiol protesse inhibitor, and Cysteine peptidase, histidine active site
	Cluster 11	5	RNA polymerase
	Cluster 12	4	+ pH reduction
	Cluster 13	4	Mixed, incl. Phospholipase A2 eihiticor activity, and Galectin complex
	Cluster 14	4	Mixed, incl. RHOU GTPase cycle, and RHOC GTPase cycle
	Cluster 15	4	+ Activation of Metrix Metalloproteinases
	Cluster 16	4	Positive regulation of B cell proliferation
	Cluster 17	4	+ Stress response
	Cluster 18	3	Copper chaperone activity
	Cluster 19	3	+ Initiation of Nuclear Envelope (NE) Reformation
	Cluster 20	3	Importin-bnta, N-terminal domain, and Nucleocytoplasmic transport complex
	Cluster 21	3	Cellular iron ion homoostasis
	Cluster 22	3	Pentose shunt
	Cluster 23	3	Spliceosomal anRNP assembly
	Cluster 24	3	JAK-STAT signaling pathway
	Cluster 25	3	+ Positive regulation of antigen processing and presentation
	Cluster 26	3	Mixed, incl. SRSF9, RNA recognition motif 2, and Cerebral cortex regionalization
	Cluster 27	2	GPNMB, LGALS3
	Cluster 28	2	BCL2A1, CFLAR
	Cluster 29	2	+ CF(0)
	Cluster 30	2	+ Ribavirin ADME
	Cluster 31	2	Mixed, incl. T cell costimulation, and Tumour necrosis factor family.
	Cluster 32	2	Cytoplasmic densin complex
	Cluster 33	2	RAB9A, RABGGTB
	Cluster 34	2	CTSA, HINT1
	Cluster 35	2	Interleukin-27 signaling
	Cluster 36	2	+ Iron-sulfur cluster assembly complex
	Cluster 37	2	Mitochondrial permeability transition pore complex
	Cluster 38	2	NAADP-sensitive calcium-release channel activity
	Cluster 39	2	NOTCH4, PSENEN
	Cluster 40	2	Mixed, incl. Phospholipase D/Transphosphatidylase, and CDP-alcohol phosphatidyit...
	Cluster 41	2	Metal-thiolate cluster
	Cluster 42	2	+ Ion homeostasis
	Cluster 43	2	CD36, FABP5
	Cluster 44	2	Sensory perception of salty taste, and Sodium channel inhibitor activity
	Cluster 45	2	Regulation of fibroblast migration
	Cluster 46	1	PPA1

Figure 4.4:STRING local network cluster summary for the H5N1 PPI network. The lists detected clusters with their gene counts and representative functional descriptions, providing an overview of the dominant biological modules captured in the H5N1-associated interactome.

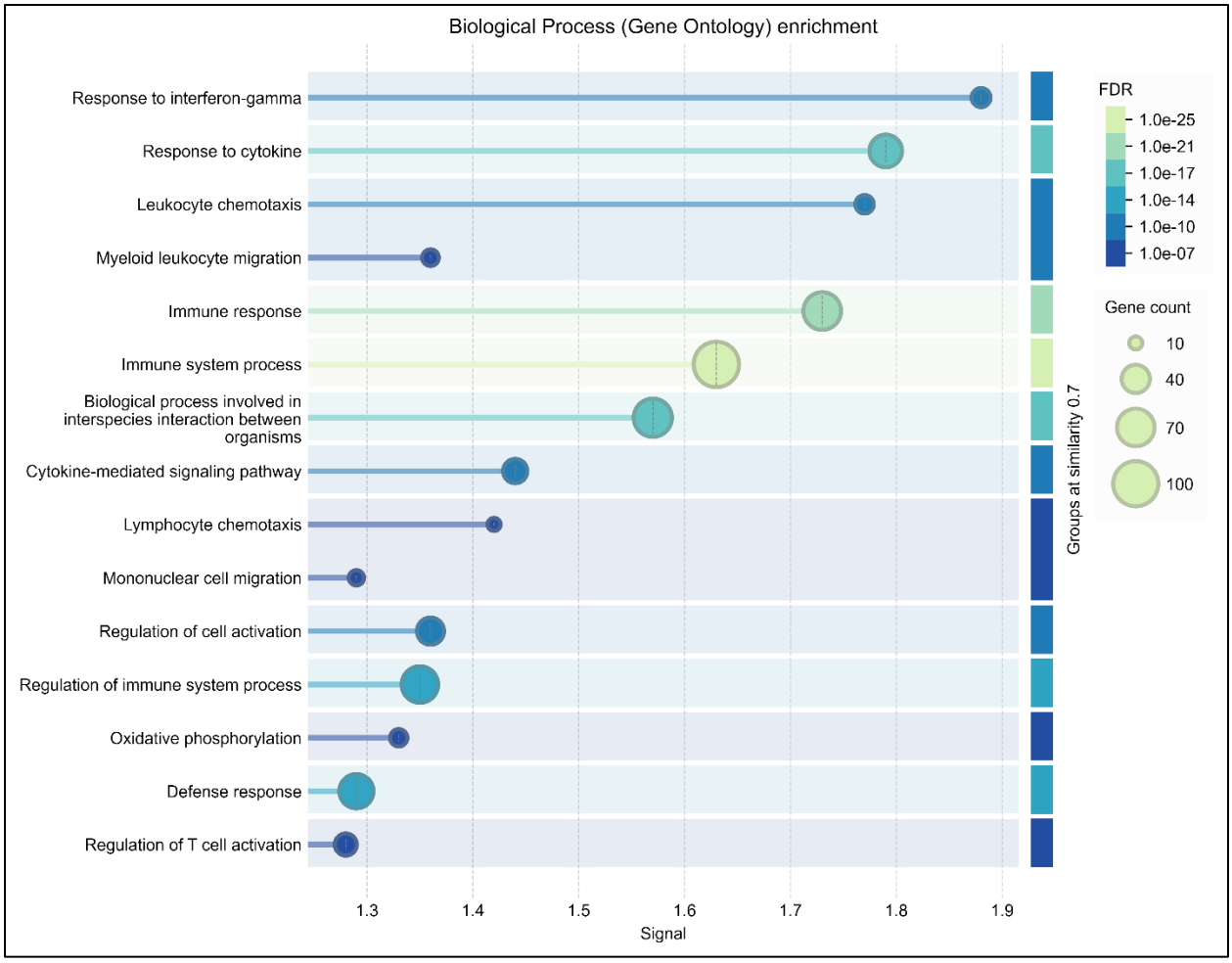


Figure 4.5:GO:Biological Process enrichment for the H5N1 PPI gene set. Top enriched GO-BP terms are shown (grouped by semantic similarity ≥0.7 and sorted by Signal); bubble size indicates gene count and colour encodes FDR (lighter = more significant), highlighting interferon/cytokine signalling and leukocyte chemotaxis/migration as dominant programmes.

A STRING protein–protein interaction (PPI) network was constructed for the H5N1 condition, and the resulting interaction graph was partitioned into 46 local network clusters (nodes coloured by cluster in the global PPI map). Across the full network, functional enrichment was dominated by immune and inflammation-linked programmes, with the strongest GO:Biological Process signals assigned to response to interferon-gamma**,** response to cytokine, and multiple migration/trafficking terms including leukocyte chemotaxis**,** myeloid leukocyte migration**,** mononuclear cell migration**,** and lymphocyte chemotaxis**;** broader immune regulation terms (e.g., immune response**,** regulation of cell activation**,** and regulation of immune system process**)** were also enriched, while a metabolic component **(**oxidative phosphorylation**)** was retained among the significant terms (FDR as shown in the enrichment bubble plot). For detailed reporting and figure

presentation, the following clusters were selected because they captured the main biological themes and highest-coherence interaction modules within the viral network: 1, 3, 5, 6, 7, 8, 16, 17, 24, 25, and 28 (Table 5).

Table 5:Summary of representative STRING local clusters in the H5N1 network, including key hub genes, dominant enrichment themes, and corresponding figures.

Cluster	Module	Key hub(s)	Representative cluster themes
1	Oxidative phosphorylation (OXPHOS)	COX5B; UQCRQ	Oxidative phosphorylation; respiratory electron transport
3	Proteasome	PSMA1/PSMA6; PSMB3	Proteasomal/ubiquitin-dependent protein catabolism
5	Defence response to virus (ISG module)	MX1; ISG15; RSAD2	Defence response to virus; Type I interferon response
6	Cell redox homeostasis	TXN; GPX1/GPX4; SOD2	Redox homeostasis; oxidative stress response
7	Chemokine module / neutrophil chemotaxis	CCL2/CCL4/CCL5/CCL20; PPBP	Leukocyte/lymphocyte chemotaxis and migration
8	Antigen processing & presentation (MHC II)	CD74	MHC class II antigen presentation
16	B-cell proliferation (CD40 axis)	CD40	Positive regulation of B cell proliferation; TNF response
17	Stress / chaperone module	HSPA8; HSPE1	Chaperone-mediated protein folding
24	JAK–STAT signalling (minimal chain)	TYK2	Cytokine-mediated signalling
25	Cytokine regulation (PYCARD chain)	PYCARD	Negative regulation of cytokine production
28	NF-κB-linked survival module	BCL2A1; CFLAR	NF-κB signalling pathway

4.3.2 Bacteria string network

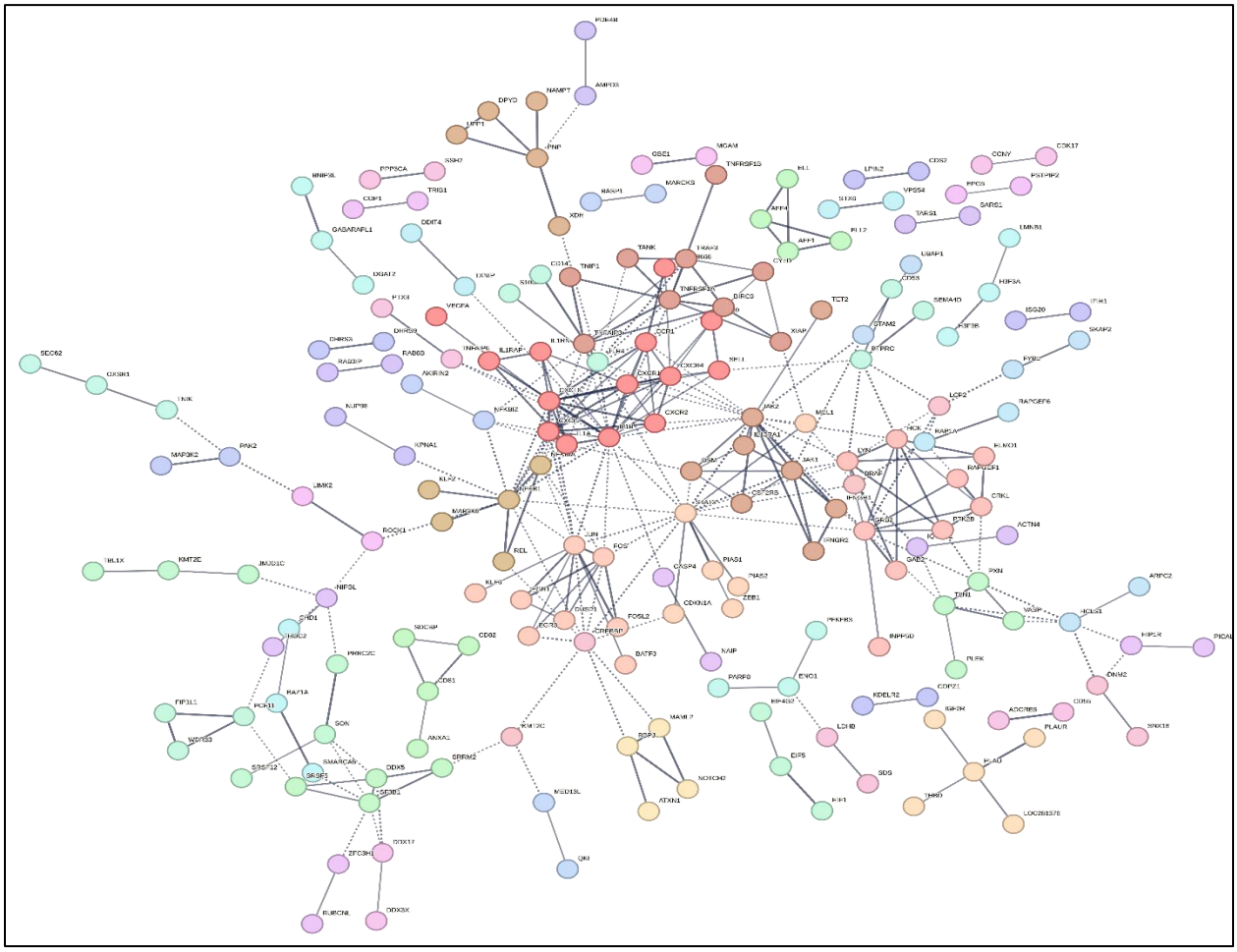
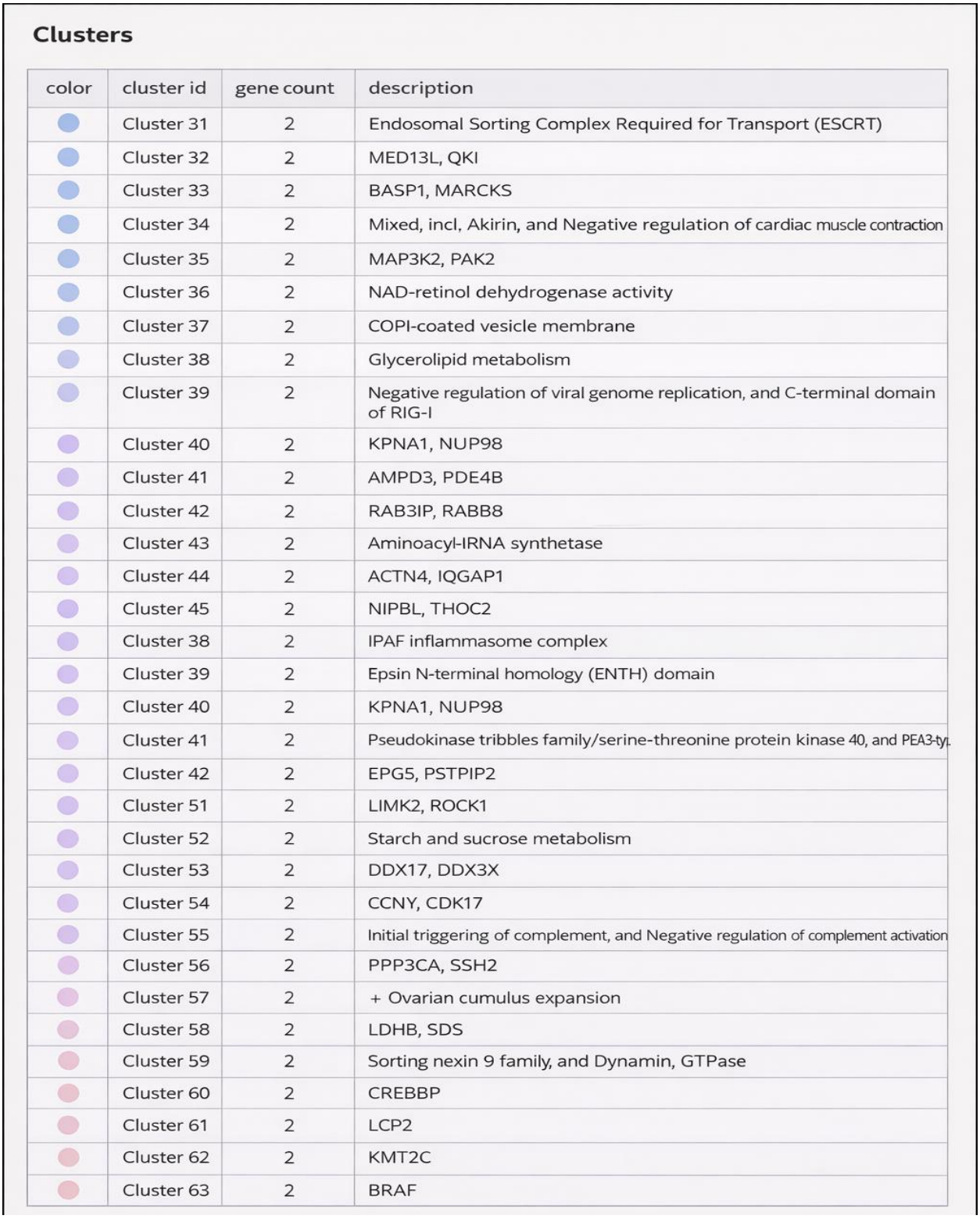


Figure 4.6:STRING protein–protein interaction (PPI) network for the Bacterial gene set, with nodes coloured by STRING cluster membership; edges represent known and predicted functional associations, highlighting a dense central core with multiple peripheral module

Clusters

color	cluster id	gene count	description
	Cluster 1	14	Chemokine-mediated signaling pathway
	Cluster 2	8	Interleukin-3, Interleukin-5 and GM-CSF signaling
	Cluster 3	9	NF-kappa B signaling pathway
	Cluster 4	15	NF-kappa B gl transciption factor, and Orphan nuclear receptor
	Cluster 5	8	Regulation of IFNG signaling
	Cluster 6	9	SUMOylation of transcription factors
	Cluster 7	5	Nucleotide metabolism
	Cluster 8	7	Dissolution of Fibrin Clot
	Cluster 9	5	TWEAK signaling pathway, and JNK (c-Jun kinases) phosphorylation mediated by activated human TAK1
	Cluster 10	4	Notch signaling pathway
	Cluster 11	9	Tetraspanin, conserved site, and Tetraspannin, animals
	Cluster 12	14	Transcription elongation factor complex
	Cluster 13	12	DDX5, SF3B1, SRRM2, SRSF5
	Cluster 14	13	Mixed, incl. Cell-extracellular matrix interactions, and Alpha-catenin/ vinculin-like su--
	Cluster 15	13	Mixed, incl. Set3 complex, and This domain is found at the C-terminus of chromodo
	Cluster 16	13	PRRC2C, SON, SRSF12
	Cluster 17	8	Regulation of translational initiation
	Cluster 18	9	mRNA polyadenylation
	Cluster 19	13	Inflammatory response
	Cluster 20	9	Serine/threonine-protein kinase OSR1/WNK, CCT domain, and Enlopt...
	Cluster 22	19	Glycolysis, and Sugar-phosphatase activity
	Cluster 23	8	Autophagy - other, and Autophagy protein Atg8 ubiquitin-like
	Cluster 24	5	Heterochromatin assembly
	Cluster 25	8	ACF complex
	Cluster 26	7	Retrograde transport at the Trans-Golgi-Network
	Cluster 27	6	DDIT4, TXNIP
	Cluster 28	6	Ras-related protein Rap1, and Rap1 signalling
	Cluster 29	12	Actin filament polymerization
	Cluster 30	11	Signal regulatory protein family interactions
	Cluster 24	5	ACF conmplex
	Cluster 25	7	Retrograde transport at the Trans-Golgi-Network
	Cluster 27	6	DDIT4, TXNIP
	Cluster 29	12	Ras-related protein Rap1, and Rap1 signalling
	Cluster 30	11	Signal regulatory protein family interactions



Clusters

color	cluster id	gene count	description
	Cluster 31	2	Endosomal Sorting Complex Required for Transport (ESCRT)
	Cluster 32	2	MED13L, QKI
	Cluster 33	2	BASP1, MARCKS
	Cluster 34	2	Mixed, incl, Akirin, and Negative regulation of cardiac muscle contraction
	Cluster 35	2	MAP3K2, PAK2
	Cluster 36	2	NAD-retinol dehydrogenase activity
	Cluster 37	2	COPI-coated vesicle membrane
	Cluster 38	2	Glycerolipid metabolism
	Cluster 39	2	Negative regulation of viral genome replication, and C-terminal domain of RIG-I
	Cluster 40	2	KPNA1, NUP98
	Cluster 41	2	AMPD3, PDE4B
	Cluster 42	2	RAB3IP, RABB8
	Cluster 43	2	Aminoacyl-IRNA synthetase
	Cluster 44	2	ACTN4, IQGAP1
	Cluster 45	2	NIPBL, THOC2
	Cluster 38	2	IPAF inflammasome complex
	Cluster 39	2	Epsin N-terminal homology (ENTH) domain
	Cluster 40	2	KPNA1, NUP98
	Cluster 41	2	Pseudokinase tribbles family/serine-threonine protein kinase 40, and PEA3-ty
	Cluster 42	2	EPG5, PSTPIP2
	Cluster 51	2	LIMK2, ROCK1
	Cluster 52	2	Starch and sucrose metabolism
	Cluster 53	2	DDX17, DDX3X
	Cluster 54	2	CCNY, CDK17
	Cluster 55	2	Initial triggering of complement, and Negative regulation of complement activation
	Cluster 56	2	PPP3CA, SSH2
	Cluster 57	2	+ Ovarian cumulus expansion
	Cluster 58	2	LDHB, SDS
	Cluster 59	2	Sorting nexin 9 family, and Dynamin, GTPase
	Cluster 60	2	CREBBP
	Cluster 61	2	LCP2
	Cluster 62	2	KMT2C
	Cluster 63	2	BRAF

Figure 4.7:STRING cluster summary table for the Bacterial gene set, reporting cluster IDs, gene counts, and the top functional description for each module; selected clusters (1, 2, 3, 4, 5, 6, 9, 10, 19, 29, 46, 51, 55) were highlighted for downstream reporting.

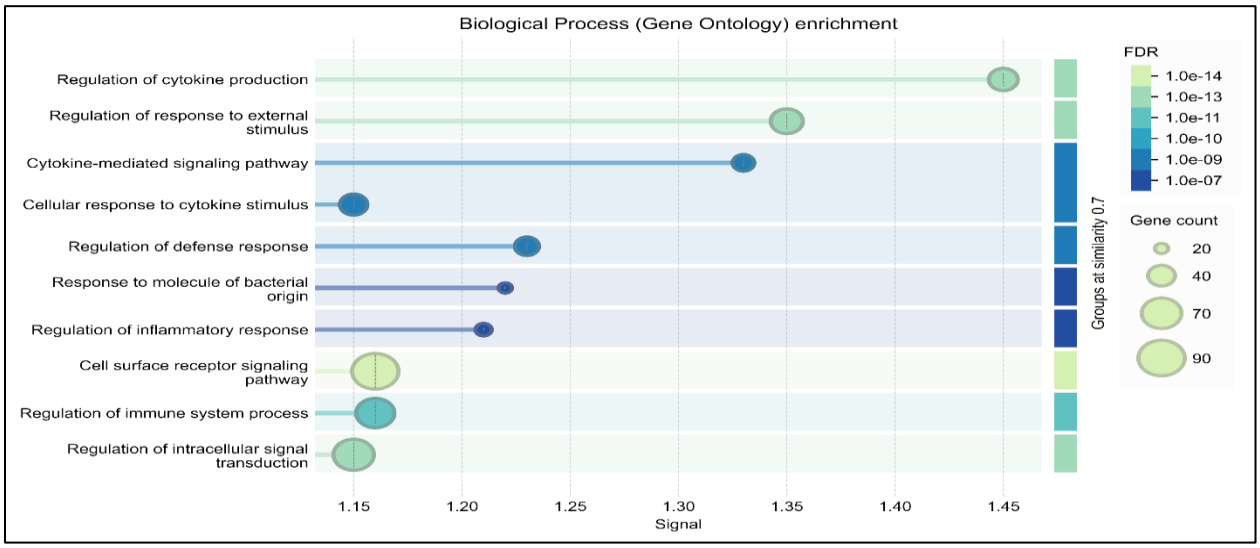


Figure 4.8:GO:Biological Process enrichment summary for the Bacterial gene set. Terms are grouped by semantic similarity (≥0.7) and sorted by Signal; bubble size reflects contributing gene count and colour encodes FDR (lighter = more significant), with enrichment

Across the viral gene set, a total of 63 STRING clusters were identified, indicating a highly modular interaction landscape rather than one single pathway driving the signal. A dense, centrally connected core was observed in the PPI network, with multiple peripheral modules branching from it, consistent with coordinated activation of immune signaling, stress-response programs, and cytoskeleton/remodelling processes. Global functional enrichment was dominated by cytokine-mediated signaling and regulation of immune responses, alongside terms linked to stimulus sensing and downstream intracellular signal transduction. For detailed presentation, the following clusters were selected as representative modules spanning the main functional themes: 1, 2, 3, 4, 5, 6, 9, 10, 19, 29, 46, 51, and 55 (Table 6).

Table 6:Summary of representative STRING local clusters in the bacterial mastitis network, including key hub genes, dominant enrichment themes, and corresponding figures.

Cluster	Module	Key hub(s)	Representative cluster themes
1	Chemokine-mediated signalling / trafficking	IL1B; CXCL8	Chemokine/cytokine signalling; neutrophil chemotaxis
2	Adaptor/tyrosine-kinase signalling (GM-CSF/IL axis)	GRB2	RTK/MAPK cascade regulation; proliferation-associated terms
3	NF-κB signalling regulation	TNFRSF1A	TNF/NF-κB regulation (I-κB kinase/NF-κB control)
4	AP-1 transcriptional module	JUN; FOS	Leukocyte differentiation and growth factor response terms
5	IFNG/JAK–STAT receptor module	JAK2; JAK1	Cytokine-mediated signalling via JAK–STAT
6	Transcription-factor SUMOylation (STAT3-centred)	STAT3	Ubiquitin/ligase-binding term (GO:MF)
9	NF-κB activation module (small)	NFKB1	NF-κB activation / TLR3 cascade (Reactome)
10	Notch signalling	RBPJ	Notch signalling pathway
19	Inflammatory response (LPS sensing)	TLR4; CD14	Response to lipopolysaccharide; innate immune response
29	Actin filament polymerisation (minimal)	ARPC2; HCLS1	Actin filament polymerisation
46	Inflammasome/pyroptosis (minimal)	CASP4; NAIP	Pyroptosis
51	ROCK/LIMK cytoskeleton module	ROCK1; LIMK2	Regulation of actin cytoskeleton; axon guidance
55	Complement regulation (minimal)	ADGRE5; CD55	Complement initiation / negative regulation of complement

4.3.3 Control string network

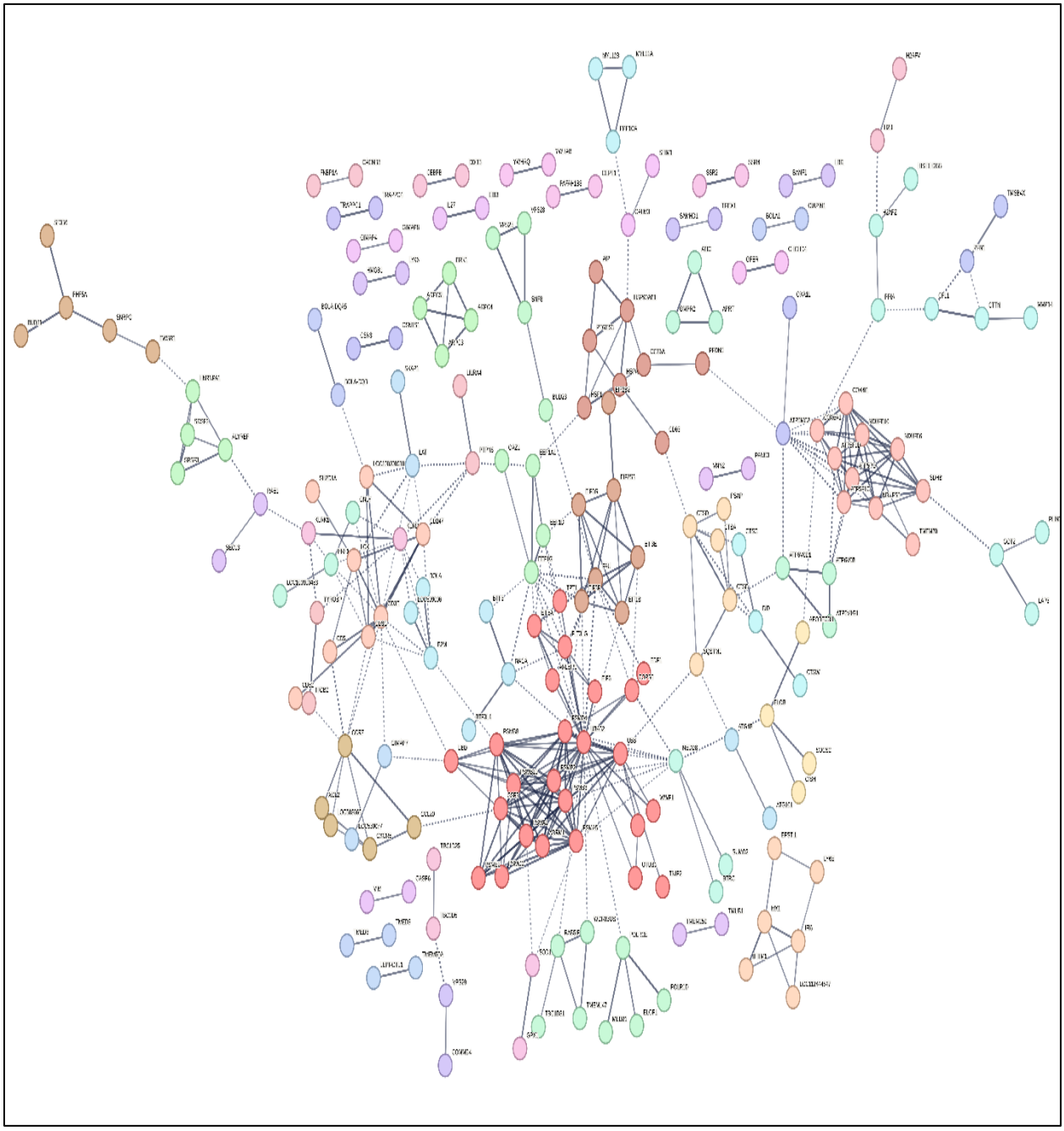


Figure 4.9:Control STRING PPI network (nodes coloured by MCL cluster).Proteins are shown as nodes and STRING functional associations as edges; node colours indicate MCL-assigned clusters, highlighting densely connected core modules with multiple smaller peripheral

Clusters

color	cluster Id	gene count	description
	Cluster 1	25	Ubiquitin-dependent degradation of Cyclin D
	Cluster 2	10	Oxidative phosphorylation
	Cluster 3	8	Unfolded protein binding
	Cluster 4	10	PD-L1 expression and PD-1 checkpoint pathway in cancer
	Cluster 5	5	Initiation factor
	Cluster 6	6	Mixed, incl. Negative regulation of viral-genome replication, and 1'LD
	Cluster 7	7	U2 snRNP and small nuclear ribonucleoprotein F
	Cluster 8	5	CTSA, CTSB, CTSO, PSAP, SQSTM1
	Cluster 9	8	Chemokine-mediated signaling pathway
	Cluster 10	4	Growth regulation
	Cluster 11	6	+ Arp2/3 complex-mediated actin nucleation
	Cluster 12	7	mRNA 3-end processing
	Cluster 13	7	Elongation factor
	Cluster 14	4	Protein transport to vacuole involved in ubiquitin-dependent protein cat...
	Cluster 15	4	RNA polymerase
	Cluster 16	8	Mixed, incl. Transmembrane protein 147 and Integral membrane protein EMC..
	Cluster 17	3	Insulin receptor recycling
	Cluster 18	9	GMP metabolism
	Cluster 19	3	Cell killing
	Cluster 20	3	Protein tka
	Cluster 21	3	H2AFZ, HIST1H2BB, PPIA
	Cluster 22	3	GOT2, LAP3, PUN2
	Cluster 23	3	Mixed, incl. Profilin, and Sequestering of actin monomers
	Cluster 24	5	Papain family cysteine protease
	Cluster 25	2	Myosin heavy chain binding
	Cluster 26	3	Antigen processing and presentation of endogenous peptide antigen, and Beta-2 Microgiobulin
	Cluster 27	3	NAC
	Cluster 28	3	Autophagy - other
	Cluster 30	2	Immunological synapse
	Cluster 31	2	GTPase GIMAl/Yan/Toc
	Cluster 32	2	Late endosome to vacuole transport via multivesicar body sorting pathway
	Cluster 23	3	emp24/3y s25L/p24 family/GOLD
	Cluster 33	2	Actin cytoskeleton assembly, and BolA protein
	Cluster 34	4	+ Asthmia
	Cluster 35	3	Mixed, incl. Profilin, and Sequestering of actin monomers
	Cluster 36	3	ATPSMC2, OXA1L
	Cluster 37	2	+ Response to deoxydilhhydroepeiandrosterone
	Cluster 40	2	COMMD4, VPS29

	Cluster 41	2	+ Initiation of Nuclear Envelope (NE) Reformation
	Cluster 42	2	+ Regulation of TLR by endogenous ligand
	Cluster 43	2	Transport of the SLBP independent Mature mRNA
	Cluster 44	2	TMEM259, TMUB1
	Cluster 45	2	+ Pantothenate and CoA biosynthesis
	Cluster 46	2	CASP6, VIM
	Cluster 47	2	Interleukin-27 signaling
	Cluster 48	2	GTPase GIMA/IAN/Toc
	Cluster 49	2	Detection of calcium ion
	Cluster 50	2	Cysteine alpha-hairpin motif superfamily, and ERV/ALR sulfhydryl oxidase domain
	Cluster 51	2	+ 14-3-3 homologues
	Cluster 52	2	CEPT1, PAFAH1B3
	Cluster 53	2	Translocon-associated protein (TRAP), alpha subunit, and SRP-dependent cotranslational protein targeting to membrane
	Cluster 54	2	Vasodilation
	Cluster 55	2	Stimulatory C-type lectin receptor signaling pathway
	Cluster 56	2	Domain in Tre-2, BUB2p, and Cdc16p. Probable Rab-GAPs.
	Cluster 57	2	CHOP-C/EBP complex
	Cluster 58	2	H2AFV, H2B
	Cluster 59	2	Calcium channel regulator activity
	Cluster 60	2	LILRA4, PTPN6
	Cluster 61	2	ITGB2, TYROBP

Figure 4.10:Control MCL cluster summary table. The distribution of 61 STRING-derived clusters is shown with gene counts and functional annotations; clusters 1, 2, 3, 7, 8, 9, 11, 26, 53, and 61 were selected for detailed presentation.

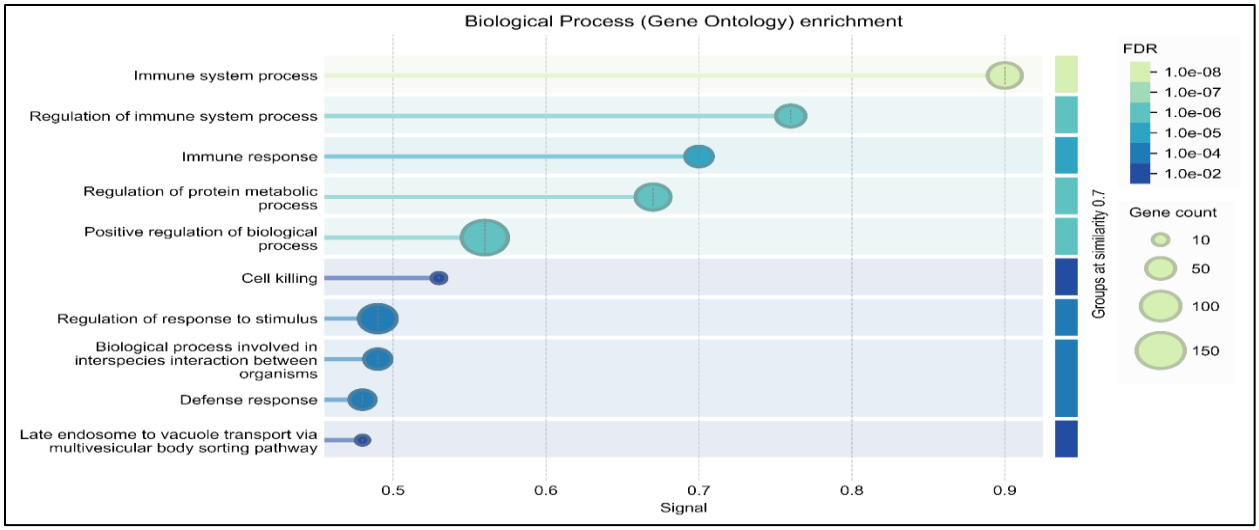


Figure 4.11:GO:Biological Process enrichment for the control gene set.Bubble plot of enriched GO-BP terms (grouped by semantic similarity and sorted by Signal); bubble size reflects contributing gene count and colour encodes FDR (lighter = more significant), showing predominant enrichment for immune/stimulus-response processes in control samples.

In the control condition, a STRING protein–protein interaction (PPI) network was constructed from the input gene set and partitioned by MCL clustering into 61 network clusters, indicating a modular organisation of baseline cellular programmes. The overall topology was characterised by a few densely connected cores linked by smaller peripheral modules, consistent with housekeeping and immune-surveillance functions being distributed across separable subnetworks. For focused reporting, Clusters 1, 2, 3, 7, 8, 9, 11, 26, 53, and 61 (Table 7) were selected because they captured the most interpretable and biologically coherent themes in this control network, spanning ubiquitin–proteasome/protein turnover (Cluster 1), mitochondrial respiration/oxidative phosphorylation (Cluster 2), protein folding and stress–chaperone biology (Cluster 3), spliceosome-related RNA processing (Cluster 7), lysosome/proteolysis-associated components (Cluster 8), chemokine-mediated signalling and leukocyte trafficking cues (Cluster 9), actin nucleation via the Arp2/3 complex (Cluster 11), antigen processing/presentation with β2-microglobulin (Cluster 26), ER targeting/translocon–SRP dependent co-translational delivery to membrane (Cluster 53), and a myeloid-associated immune module (Cluster 61; ITGB2–TYROBP) consistent with constitutive innate immune readiness. At the system level, GO:Biological Process enrichment for the control gene set was dominated by broad immune and stimulus-response terms (e.g., immune system process/immune response, regulation of response to stimulus, and defense-related categories), supporting that baseline milk somatic cells retain constitutive immune readiness even in the absence of pathogen exposure.

Table 7:Summary of representative STRING local clusters in the control network, including key hub genes, dominant enrichment themes, and corresponding figures.

Cluster	Module	Key hub(s)	Representative cluster themes
1	Ubiquitin–proteasome / protein turnover	UBA52	Proteasomal protein catabolism; ubiquitin-dependent proteolysis
2	Oxidative phosphorylation	NDUFB10	Aerobic respiration; respiratory electron transport; ATP synthesis
3	Chaperone/protein folding	HSP90AB1	Protein folding; chaperone-mediated folding
7	Spliceosome / mRNA splicing	PHF5A	mRNA splicing via spliceosome
8	Lysosome / cathepsin module	CTSB; CTSD	Lysosome
9	Chemokine signalling (compact)	CCR7; CXCR6	Chemokine-mediated signalling; monocyte chemotaxis
11	Arp2/3-mediated actin nucleation	ARPC3/ARPC4/ARPC5L	Actin nucleation; regulation of actin polymerisation
26	MHC-I antigen presentation	B2M	Antigen processing/presentation (endogenous peptide; MHC-I)
53	ER targeting/translocon (TRAP)	SSR2; SSR4	SRP-dependent co-translational targeting to membrane
61	Myeloid/innate immune marker pair	ITGB2; TYROBP	Macrophage-associated enrichment (as reported)

4.4 Diagnostic biomarker panel

Objective 4 defined two condition-specific diagnostic gene panels using the integrated single-cell dataset: a 12-gene panel enriched in the H5N1 condition and a 12-gene panel enriched in the bacterial mastitis condition. Genes were selected based on their preferential enrichment within each condition in the integrated atlas, and the resulting panels are summarized in Table 8 as candidate biomarkers to distinguish viral from bacterial exposure at single-cell resolution. For the H5N1 panel, the selected genes were MX1, MX2, ISG15, ZBP1, RSAD2, OAS1Y/X, EPSTI1, IFI6, IFI44/IFI44L, and IFITM1. For the bacterial panel, the selected genes were NFKB1, NFKBIZ/NFKBIA, TNFAIP3/TNFAIP6, CXCL8 (IL8), IL1B, CXCR2, TLR4, CD14, S100A9, and IL1RN .

Table 8:Condition-specific diagnostic gene panels.

Condition	Gene
H5N1	MX1
H5N1	MX2
H5N1	ISG15
H5N1	ZBP1
H5N1	RSAD2
H5N1	OAS1Y
H5N1	OAS1X
H5N1	EPSTI1
H5N1	IFI6
H5N1	IFI44
H5N1	IFI44L
H5N1	IFITM1
Bacteria	NFKB1
Bacteria	NFKBIZ
Bacteria	NFKBIA
Bacteria	TNFAIP3
Bacteria	TNFAIP6
Bacteria	CXCL8 (IL8)
Bacteria	IL1B
Bacteria	CXCR2
Bacteria	TLR4
Bacteria	CD14
Bacteria	S100A9
Bacteria	IL1RN

Table 9 summarises the sample-level (pseudo-bulk) mean diagnostic scores after excluding control samples. Across the bacterial mastitis samples (BAC1–BAC4), mean H5N1_diagnostic_panel scores were low or negative (−0.1002 to −0.0144), whereas mean Bacteria_diagnostic_panel scores were comparatively higher (−0.0151 to 0.2396), yielding consistently negative DeltaScore values (−0.3399 to −0.0700). In contrast, all H5N1 samples (H5N1_1–H5N1_4) showed uniformly high mean H5N1_diagnostic_panel scores (0.5324–0.5349) together with strongly negative mean Bacteria_diagnostic_panel scores (−0.6395 to −0.6270), resulting in strongly positive DeltaScore values (1.1619–1.1719). Cell numbers contributing to each pseudo-bulk estimate ranged from 3431–4975 in bacterial samples and 488–495 in H5N1 samples. At the sample level (n = 8; control excluded), DeltaScore values showed no overlap between cohorts in the analysed samples, with AUROC = 1.00 in this dataset. Using DeltaScore > 0 as a fixed decision rule, sensitivity and specificity were each 1.00 (4/4; exact 95% CI 0.40–1.00), and accuracy was 1.00 (8/8; exact 95% CI 0.63–1.00).

Table 9:Sample-level evaluation of the diagnostic panel using pseudo-bulk mean scores. DeltaScore = mean(H5N1_diagnostic_panel) - mean (Bacteria_diagnostic_panel).

sample_id	condition	mean(H5N1_diagnostic_panel)	mean(Bacteria_diagnostic_panel)	DeltaScore	n_cells
BAC1	Bacteria	-0.0144	0.1937	-0.2081	3431
BAC2	Bacteria	-0.0852	-0.0151	-0.0700	3811
BAC3	Bacteria	-0.0800	0.0012	-0.0812	3964
BAC4	Bacteria	-0.1002	0.2396	-0.3399	4975
H5N1_1	H5N1	0.5349	-0.6270	1.1619	495
H5N1_2	H5N1	0.5349	-0.6270	1.1619	495
H5N1_3	H5N1	0.5349	-0.6326	1.1675	490
H5N1_4	H5N1	0.5324	-0.6395	1.1719	488

4.5 Cell type distribution shifts

In Table 10, cell-type composition was quantified across conditions using both absolute counts and within-condition percentages. A total of 6,167 cells were retained in the Control cohort, 1,968 cells in H5N1, and 16,181 cells in Bacteria. In the control samples, the atlas was dominated by neutrophil states, with Neutrophils_1 representing 24.16% (n=1,490) and Neutrophils_2

representing 19.05% (n=1,175). Additional major compartments were represented by T cells (8.98%, n=554), IFN-stimulated cells (ISG-high) (7.44%, n=459), Antigen-presenting Myeloid (7.18%, n=443), and Inflammatory Macrophages/Monocytes_2 (6.26%, n=386), while non-immune populations remained minor (Mammary epithelial cells: 1.52%, n=94; Putative endothelial/vascular-like: 0.54%, n=33). In the H5N1 cohort, the distribution was concentrated within a small number of lineages, with Inflammatory Macrophages/Monocytes_2 comprising 55.28% (n=1,088) and Mature/Migratory Dendritic Cells_1 comprising 29.83% (n=587). Smaller contributions were retained for Inflammatory IFN-responsive Monocytes/Macrophages (7.52%, n=148), Cytotoxic lymphocytes (3.00%, n=59), Mammary epithelial cells (2.85%, n=56), and B cells (1.52%, n=30), whereas multiple neutrophil lineages and several additional immune compartments were not represented in the H5N1 counts. In the Bacteria cohort, a broad immune-dominant profile was observed, led by Neutrophils/monocytes/macrophages (15.41%, n=2,493), Neutrophils_2 (12.61%, n=2,041), IFN-stimulated cells (ISG-high) (11.92%, n=1,929), Neutrophils_1 (11.69%, n=1,892), Inflammatory Monocytes/Macrophages_1 (11.62%, n=1,880), and Antigen-presenting Myeloid (9.13%, n=1,477), with T cells also contributing substantially (8.72%, n=1,411). Two lineages were absent in the bacterial condition (Cytotoxic lymphocytes and Inflammatory IFN-responsive Monocytes/Macrophages), and non-immune populations remained rare (Mammary epithelial cells: 1.17%, n=189; Putative endothelial/vascular-like: 0.75%, n=122).

Table 10:Cell type composition across conditions.

Cell type	Control_n	Control_pct	H5N1_n	H5N1_pct	Bacteria_n	Bacteria_pct	Total_n
Antigen-presenting Myeloid	443	7.18%	0	0.00%	1477	9.13%	1920
B cells	78	1.26%	30	1.52%	172	1.06%	280
Cytotoxic lymphocytes	225	3.65%	59	3.00%	0	0.00%	284
IFN-stimulated cells (ISG-high)	459	7.44%	0	0.00%	1929	11.92%	2388
Inflammatory IFN-responsive Monocytes/Macrophages	150	2.43%	148	7.52%	0	0.00%	298

Inflammatory Macrophages/Monocytes_2	386	6.26%	1088	55.28%	1	0.01%	1475
Inflammatory Monocytes/Macrophages_1	229	3.71%	0	0.00%	1880	11.62%	2109
Mammary epithelial cell	94	1.52%	56	2.85%	189	1.17%	339
Mature/Migratory Dendritic Cells_1	247	4.01%	587	29.83%	1	0.01%	835
Mature/Migratory Dendritic Cells_2	24	0.39%	0	0.00%	124	0.77%	148
Neutrophils/monocytes/macrophages	178	2.89%	0	0.00%	2493	15.41%	2671
Neutrophils_1	1490	24.16%	0	0.00%	1892	11.69%	3382
Neutrophils_2	1175	19.05%	0	0.00%	2041	12.61%	3216
Neutrophils_3	131	2.12%	0	0.00%	1278	7.90%	1409
Neutrophils_4	8	0.13%	0	0.00%	21	0.13%	29
Putative Endothelial / Vascular-like	33	0.54%	0	0.00%	122	0.75%	155
T cells	554	8.98%	0	0.00%	1411	8.72%	1965
Tissue/Resident Macrophages	263	4.26%	0	0.00%	1150	7.11%	1413

4.6 Diffusion Pseudotime Analysis of Pathogen-Specific Transcriptional Dynamics

Objective 6 assessed how pathogen-associated transcriptional programmes vary along diffusion pseudotime (DPT) in bovine milk somatic cells. DPT was computed on the integrated manifold, and each cell was assigned two module scores: H5N1_panel_score and Bacteria_panel_score, derived from the expression of the corresponding panel genes relative to background expression. In both scatter plots, the x-axis represents the same DPT coordinate (0–1), while the y-axis shows the corresponding panel score for each cell, coloured by condition (H5N1, bacteria). Because module scores are computed relative to a background gene set, negative values indicate lower-

than-background expression of the corresponding panel. In the H5N1 panel versus pseudotime plot, most cells at early pseudotime (approximately 0.00–0.40) show H5N1_panel_score values near zero or slightly negative across all conditions. A clear enrichment of high H5N1_panel_score values is observed in a mid-trajectory interval around pseudotime ~0.55–0.75, where cells from the H5N1 condition form a dense high-scoring cloud. Within this window, many H5N1 cells exhibit strongly positive scores (commonly in the range of roughly ~0.4 to >1.0), whereas bacterial-condition cells occupying similar pseudotime values largely remain near zero or modestly negative. Beyond this region, at later pseudotime (approximately ~0.75–1.00), cells show predominantly low H5N1_panel_score values, and high-scoring H5N1 cells are comparatively sparse. Overall, elevated H5N1_panel_score is concentrated within a relatively narrow mid-pseudotime segment and is dominated by H5N1-condition cells.

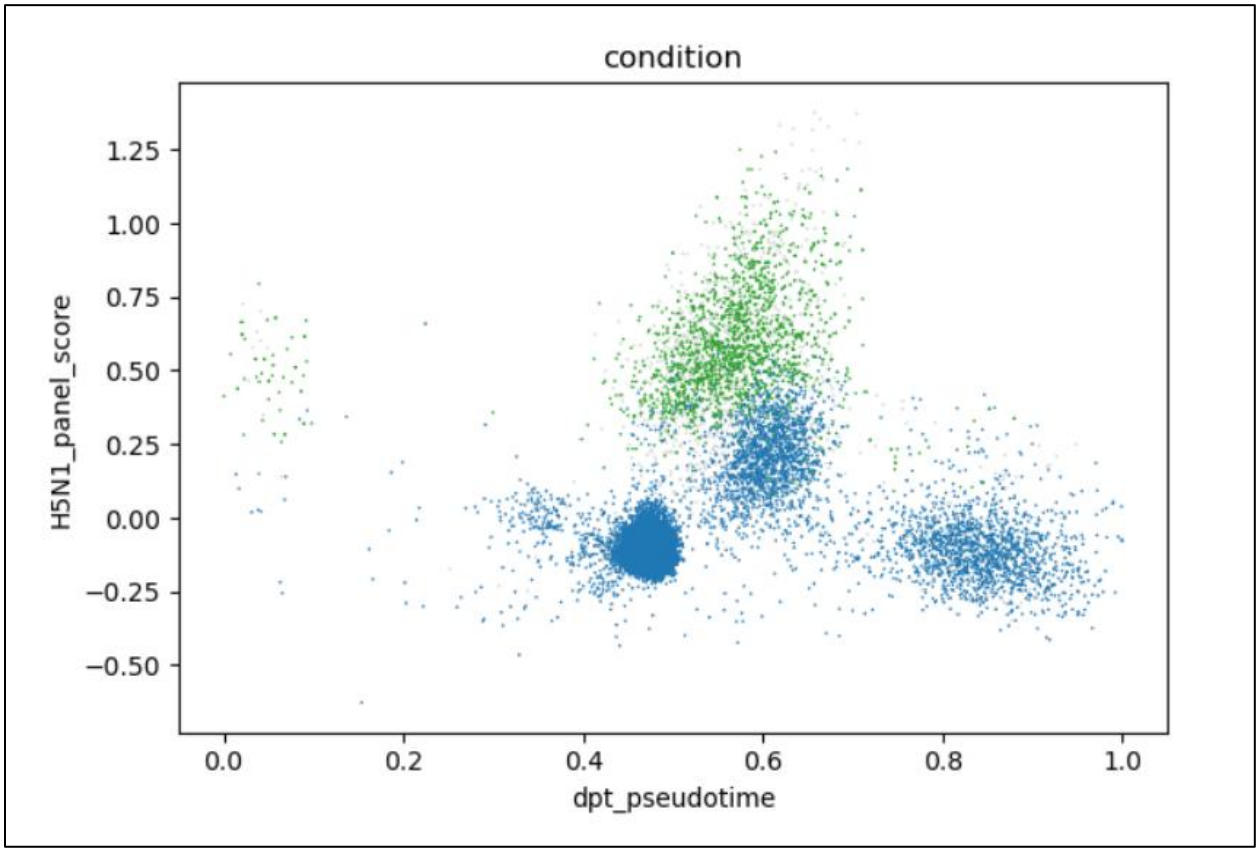


Figure 4.12:Diffusion pseudotime (DPT) versus H5N1 gene-panel module score in bovine milk somatic cells. Each dot represents a single cell from the integrated dataset. The x-axis shows DPT pseudotime (scaled 0–1) and the y-axis shows the H5N1_panel_score (module score) computed from the predefined viral gene panel. Cells are coloured by condition: bacterial (blue) and H5N1 (green); higher scores indicate stronger activation of the H5N1-associated transcriptional programme.

In the bacterial panel versus pseudotime plot, Bacteria_panel_score values are largely negative across wide portions of the trajectory, but a distinct high-scoring structure appears at intermediate pseudotime. Specifically, a dense vertical band with positive Bacteria_panel_score values is evident around pseudotime ~0.42–0.50, extending from near zero up to approximately ~0.8–0.9, and is dominated by bacterial-condition cells. Outside this intermediate band—particularly across pseudotime ~0.55–1.00 bacterial-condition cells remain distributed along the manifold but mostly show negative bacterial scores centred around approximately ~−0.5, while H5N1-condition cells are also largely negative for the bacterial panel. Thus, the strongest bacterial panel activation is concentrated in a relatively narrow intermediate pseudotime interval rather than increasing toward the end of the trajectory.

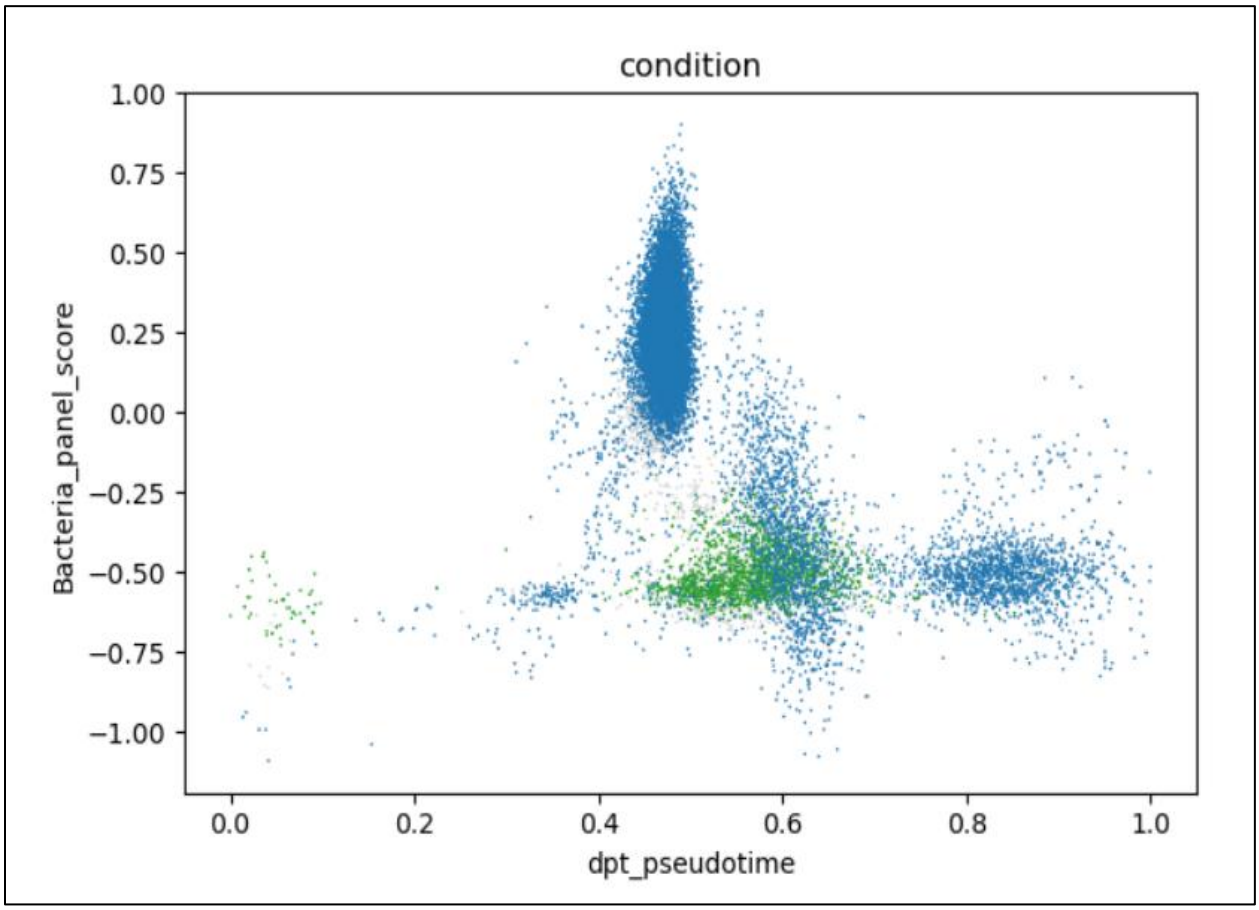


Figure 4.13:Diffusion pseudotime (DPT) versus bacterial gene-panel module score in bovine milk somatic cells.The x-axis shows DPT pseudotime (scaled 0–1) and the y-axis shows the Bacteria_panel_score (module score) computed from the predefined bacterial gene panel.

Chapter 5: Discussion

This study has scientific and translational significance. Scientifically, the emergence of H5N1 in dairy cattle creates a new viral mastitis context that extends beyond the well-studied bacterial paradigm and motivates cell type resolved analysis of mammary and milk immune responses. From a One Health perspective, rapid etiological discrimination is critical because routine mastitis management is optimized for bacterial causes, whereas H5N1 requires outbreak-oriented biosecurity and worker-risk mitigation. At the industry level, improved diagnostic precision can support earlier containment and more targeted management, reducing production disruption and the broader costs of surveillance and control. Accordingly, biomarker panels that distinguish interferon/ISG-dominant viral states from bacterial inflammatory programmes are operationally valuable. This study builds a cell-type–resolved map of immune responses in milk and highlights a small set of candidate marker genes that could help rapidly distinguish viral from bacterial mastitis.

5.1 Integrated single-cell atlas of bovine milk somatic cells

In Figure 4.1, an integrated low-dimensional manifold was preserved across all samples, while condition-associated differences were manifested primarily as changes in local point density rather than as a completely condition-exclusive geometry. A dominant immune-lineage region was populated heavily by bacterial-exposed cells, whereas control cells were distributed across the same immune space and extended into additional neighbourhoods. A distinct upper-right region was enriched for H5N1-exposed cells, with partial overlap by control cells, indicating that viral exposure was associated with a shifted transcriptional state that remained connected to the broader myeloid landscape rather than forming an isolated compartment. This geometry-versus-density behaviour was consistent with differences in upstream sampling frames and mastitis biology. The bacterial cohort was derived from leukocyte-enriched milk immune cell preparations collected during experimental chronic *Staphylococcus aureus* mastitis, a state in which somatic cell counts are markedly elevated and milk leukocyte composition becomes granulocyte-biased (Wiarda et al., 2025; Rainard et al., 2018). Accordingly, expansion of immune neighbourhoods was expected without the necessity for a separate ‘bacterial-only’ island, because cells sharing lineage programmes remain embedded in shared immune space (Wolf et al., 2018; McInnes et al., 2018).

By contrast, the H5N1 and control cohorts were derived from bovine milk somatic cells that retained both immune and non-immune fractions, thereby supporting broader occupancy of the embedding beyond leukocyte-dominant regions (Singh et al., 2025). Condition-level patterns were expected to appear as within-lineage shifts rather than as condition-driven separation because lineage structure is primarily determined by shared transcriptional programmes. Under graph-based neighbourhood construction, community detection, and non-linear dimensionality reduction, cells sharing related lineage states were therefore expected to map into proximate regions, while condition differences were expected to alter relative contributions and activation states within those regions (Wolf et al., 2018; McInnes et al., 2018). Leiden community detection was used to obtain well-connected partitions that reduce fragmentation in neighbourhood graphs and stabilise cluster structure (Traag et al., 2019). In Figure 4.2, the integrated embedding was resolved into 18 transcriptional clusters organised into major immune and non-immune lineages. A large left-sided immune continuum was partitioned into multiple neutrophil states (Neutrophils_1, Neutrophils_2, Neutrophils_3, Neutrophils_4) together with a mixed neutrophil/monocyte/macrophage neighbourhood, consistent with the dominance and state diversity of granulocytes and myeloid cells in inflamed milk environments (Wiarda et al., 2025; Rainard et al., 2018). Multiple macrophage/monocyte clusters were separated along inflammatory, tissue-resident, and interferon-responsive axes, and an ISG-high cluster was retained as a discrete programme consistent with interferon-stimulated transcriptional states commonly observed under viral sensing and inflammatory signalling (Singh et al., 2025). Antigen-presenting myeloid populations and two mature/migratory dendritic cell clusters were positioned as distinct islands, consistent with specialised antigen presentation and migratory programmes within the myeloid compartment (Wiarda et al., 2025). Lymphoid populations were segregated into T cells, cytotoxic lymphocytes, and B cells, indicating that adaptive immune states were preserved as separable neighbourhoods within the integrated graph (Wolf et al., 2018). Non-immune components were retained as discrete neighbourhoods, including a mammary epithelial cluster and a putative endothelial/vascular-like cluster, supporting the presence of non-immune milk-derived cell types within the atlas and enabling downstream lineage-aware comparisons across conditions (Singh et al., 2025; Alhussien and Dang, 2018). Table 2 provided the cluster-ID reference for the integrated atlas by assigning IDs 0–17 to the annotated labels used throughout the analysis. These IDs were intended to function as stable identifiers for neutrophil, myeloid, dendritic, lymphoid, epithelial,

and vascular-like populations represented in Figure 4.2, thereby standardising interpretation of condition-level patterns across subsequent objectives.

5.2 Pathogen-specific gene programs

Objective 2 was designed to separate pathogen-class transcriptional programmes from shifts driven by changes in cell abundance by focusing on within-lineage signals. This was achieved by intersecting condition-level upregulated gene sets with lineage marker sets derived from the integrated atlas, thereby retaining genes that were both (i) increased in a given condition and (ii) characteristic of a specific lineage/state (Table 3). For readability in the main text, Table 3 reports a reduced subset of overlap genes per lineage and condition, while complete overlap lists are provided in supplementary Table S3 (Wolf et al., 2018; Stuart and Satija, 2019). Across the neutrophil continuum (Neutrophils_1–4) and the mixed Neutrophils/monocytes/macrophages lineage, bacterial overlaps were retained consistently and were represented by genes linked to inflammatory recruitment and activated leukocyte states, including BTG1, ANTXR2, SRGN, BASP1, PLAUR, PDE4B, SDS, B2M and CXCR2. These overlaps were interpreted as coherent with bacterial mastitis biology, in which high somatic cell counts are driven predominantly by influx of granulocytes and other myeloid cells into milk and by strong chemokine- and cytokine-associated activation programmes (Rainard et al., 2018; Alhussien and Dang, 2018; Wiarda et al., 2025). By contrast, the H5N1 overlaps within these neutrophil lineages were comparatively limited in the main table and were dominated by immediate-early and cytoskeletal/stress-associated genes (e.g., TMSB4X, TMSB10), together with selected antiviral signalling components such as IFNAR2 in Neutrophils_4, suggesting that viral exposure was associated with more focal neutrophil-associated overlap signals under the applied selection criteria (Singh et al., 2025). Within antigen-presenting and interferon-associated compartments, distinct pathogen-class programmes were supported by the composition of the reported overlaps. Interferon-Activated Dendritic Cells retained a canonical interferon-stimulated signature under H5N1 (ISG15, MX1, MX2, IFI6, RSAD2 and IFI44L), consistent with conserved type I interferon response modules that emerge under viral sensing across mammalian immune cells (Ivashkiv and Donlin, 2014; Singh et al., 2025). In the same lineage, the bacterial overlap comprised genes frequently associated with activated leukocyte states (BTG1, SAMSN1, SRGN, PLAUR and B2M), supporting the conclusion that DC-associated programmes differed in composition between pathogen classes (Wiarda et al., 2025). A lineage-selective pattern was also observed for

Mature/Migratory Dendritic Cells_1, in which an H5N1 overlap was retained while no qualifying bacterial overlap was listed. The H5N1 overlap included TAP and CTSH together with ISG15 and IFITM1, indicating concurrent enrichment of antigen-processing/peptide-loading and interferon-associated elements within this DC state. This observation was considered consistent with viral exposure favouring interferon-linked APC activation and migration-associated programmes, which have been reported as prominent components of host antiviral responses at the single-cell level (Singh et al., 2025). Mature/Migratory Dendritic Cells_2 retained overlaps under both pathogen classes; however, the bacterial set included regulatory and signalling-associated genes (e.g., REL, NR4A3), supporting non-identical programmes across conditions (Traag et al., 2019; Wiarda et al., 2025). Within myeloid inflammatory compartments, non-identical overlaps were retained under both pathogen classes. Inflammatory IFN-responsive Monocytes/Macrophages displayed an H5N1 overlap rich in chemokine and interferon-associated elements (CCL3, CCL4, CCL5, CXCL3, IFITM1 and TAP), whereas the bacterial overlap included CXCL8, CXCL2, IL1RN and TNFRSF1B, reflecting a strong neutrophil-recruiting and inflammation-modulating axis typically associated with bacterial mastitis (Rainard et al., 2018; Wiarda et al., 2025). Inflammatory Macrophages/Monocytes_2 included MMP9 within the reported H5N1 overlap, whereas the bacterial overlap was restricted in the main table, indicating that this state was captured primarily by an H5N1-associated overlap under the applied criteria. Tissue/Resident Macrophages retained overlaps under both conditions, with an H5N1 set enriched for lysosomal and stress-associated genes (e.g., CTSB, CTSZ, ferritin-related genes) and a bacterial set containing activated leukocyte-associated genes, consistent with macrophage functional plasticity during mammary inflammation (Alhussien and Dang, 2018; Wiarda et al., 2025). Lymphoid lineages were retained as distinct overlap profiles, supporting pathogen-associated programmes beyond innate immunity. Cytotoxic lymphocytes showed an H5N1 overlap that included PFN1 and CCL5 alongside ISG15 and IFITM1, consistent with cytotoxic effector programmes occurring in parallel with interferon-linked activation under viral exposure, whereas the bacterial overlap was represented by LTB. T cells retained overlaps under both conditions, with IL7R present under H5N1, supporting preservation of core T-cell identity elements alongside activation signals. B cells likewise retained an H5N1 overlap dominated by stress/housekeeping and antigen-presentation–adjacent elements (e.g., UBA52, ZC3H10), while the bacterial overlap was limited (TPT1), supporting differential prominence of lymphoid-associated overlaps across pathogen classes (Singh et al., 2025; Wiarda

et al., 2025). Non-immune compartments also retained overlap signals. Mammary epithelial cells showed an H5N1 overlap that included NUPR1, NPC2 and CD9 together with stress-associated genes, indicating that epithelial-associated programmes were captured by the lineage-aware intersection and were not confined to immune compartments. A putative endothelial/vascular-like cluster retained overlaps under both conditions, with NOS2 and CD36 appearing under H5N1 and several regulatory genes retained under bacterial exposure, supporting the presence of vascular-associated response signatures in the integrated milk atlas (Alhussien and Dang, 2018).

5.3 Functional and Network-Based Interpretation

In Objective 3, functional meaning was assigned to condition-specific gene signatures by integrating protein–protein interaction (PPI) structure with enrichment statistics. Differentially expressed genes were projected onto STRING-derived interaction networks, partitioned into local network clusters, and summarised using Gene Ontology (GO) and curated pathway resources. This strategy was used to reduce long gene lists into interpretable modules that captured coordinated biological programmes rather than isolated markers. Because enrichment and network links are associative (and not causal), the interpreted themes were treated as hypotheses about the dominant cellular programmes operating under each condition, to be supported by complementary analyses and biological context.

In the H5N1 network, 46 local clusters were resolved, indicating a modular organisation of the viral-associated response across metabolic, proteostatic, and immune programmes. For detailed interpretation, clusters 1, 3, 5, 6, 7, 8, 16, 17, 24, 25 and 28 were selected because they captured the highest-signal functional themes and represented distinct biological ‘axes’ in the network. Across the selected modules, an antiviral–inflammatory signature was inferred from enrichment for defence response to virus and cytokine/immune-system processes, alongside chemotaxis-linked terms. Such a pattern is consistent with interferon-driven antiviral effector programmes and cytokine-mediated coordination of leukocyte recruitment, which have been widely described as core components of host responses to influenza and other RNA viruses (Schoggins, 2019; Zlotnik and Yoshie, 2012). Concurrently, antigen processing and presentation-related modules were observed, supporting a shift toward enhanced peptide presentation and immune surveillance; these processes are mechanistically coupled to proteasomal protein turnover and MHC class I loading (Neefjes et al., 2011). A prominent metabolic component was indicated by oxidative

phosphorylation enrichment, suggesting that mitochondrial respiration-associated programmes were retained or amplified within the H5N1-associated gene set. This pattern has been reported in multiple immune contexts where bioenergetic demands and mitochondrial signalling are coupled to antiviral responses and cytokine production, although the direction and magnitude of metabolic rewiring can vary by cell type and activation state (O'Neill, Kishton and Rathmell, 2016). In parallel, protein-quality-control modules were highlighted through unfolded protein binding and chaperone-related terms, consistent with the proteostatic burden imposed by viral infection and inflammatory signalling; molecular chaperones are known to buffer misfolded-protein stress and sustain proteome integrity under challenge (Hartl, Bracher and Hayer-Hartl, 2011; Walter and Ron, 2011). Collectively, the H5N1 network was therefore interpreted as a composite of immune activation (cytokine/chemotaxis and antigen presentation), proteome maintenance (proteasome/chaperone systems), and bioenergetic support (oxidative phosphorylation). Such coupling is biologically plausible because antiviral immunity is typically executed through coordinated transcriptional induction of interferon-stimulated effectors, antigen-processing machinery, and stress-adaptive pathways that preserve cellular function during infection (Schoggins, 2019; Neefjes et al., 2011).

In the bacterial network, 63 local clusters were identified, reflecting broader functional heterogeneity and multiple co-activated inflammatory pathways. Clusters 1, 2, 3, 4, 5, 6, 9, 10, 19, 29, 46, 51 and 55 were selected for presentation because they captured the dominant enrichment signal and highlighted canonical bacterial-response biology, including receptor-proximal signalling, chemokine networks, and downstream transcriptional regulators. Functional enrichment was dominated by regulation of cytokine production, cytokine-mediated signalling, and regulation of inflammatory response, indicating that the bacterial condition was characterised primarily by innate inflammatory amplification. Such terms are typically driven by pattern-recognition receptor (PRR) sensing and downstream transcriptional programmes, in which Toll-like receptor signalling and NF-κB/AP-1 pathways are centrally positioned (Akira and Takeda, 2004; Liu et al., 2017). Enrichment for response to molecules of bacterial origin and cell surface receptor signalling was consistent with activation of PRR-triggered cascades that coordinate chemokine release, vascular signalling, and leukocyte recruitment during mastitis (Rainard and Riollet, 2006).

Chemokine-mediated signalling and neutrophil chemotaxis modules were also observed, supporting a recruitment-focused inflammatory architecture. Chemokines and their receptors are well-established as key organisers of leukocyte trafficking, and neutrophil influx is a hallmark of bacterial mastitis in the bovine mammary gland (Zlotnik and Yoshie, 2012; Rainard and Riollet, 2006). In addition, complement-related enrichment (including initial triggering and negative regulation of complement activation) suggested engagement of opsonisation and inflammatory amplification loops that cooperate with phagocytes during bacterial clearance (Ricklin et al., 2010). Network modules implicating inflammasome-associated components (e.g., IPAF/NLRC4 complex) were interpreted as evidence that cytosolic danger sensing and IL-1–linked inflammatory signalling may have contributed to the bacterial signature. Inflammasome activation is widely recognised as a convergence point for microbial sensing and inflammatory execution, although the specific sensors engaged can be pathogen and cell type dependent (Broz and Dixit, 2016). Finally, actin-remodelling and cytoskeletal terms (including actin polymerisation and Arp2/3-related modules) were consistent with phagocytic and migratory functions required for bacterial containment and tissue trafficking (Pollard and Borisy, 2003).

In the control network, 61 local clusters were resolved, with functional content skewed toward basal homeostatic processes rather than strong inducible immune programmes. Clusters 1, 2, 3, 7, 8, 9, 11, 26, 53 and 61 were selected for detailed interpretation because they represented the principal housekeeping modules and the most interpretable immune-surveillance signals present under baseline conditions. Core control modules were dominated by ubiquitin-dependent degradation of cyclin D, oxidative phosphorylation, and unfolded protein binding, indicating that cell cycle regulation, mitochondrial energy metabolism, and proteostasis were central baseline themes. Ubiquitin–proteasome turnover is required for cell cycle control and for broad proteome quality management, while mitochondrial respiration supports fundamental cellular maintenance (Ciechanover, 2005; O'Neill, Kishton and Rathmell, 2016). The appearance of spliceosome/U2 snRNP-related terms further supported that the control condition reflected active RNA processing and constitutive gene-expression machinery. Despite the expected housekeeping bias, immune-relevant modules were also detected in the control network. A compact antigen-processing and presentation/Beta-2-microglobulin module (B2M) suggested ongoing MHC class I–linked surveillance, which is compatible with physiological immune monitoring in tissue-resident cell populations (Neefjes et al., 2011). A macrophage-associated signal was supported by an ITGB2–

TYROBP module, consistent with the presence of myeloid-lineage cells that contribute to baseline clearance and tissue homeostasis in milk somatic cells (Turnbull and Colonna, 2007; Rainard and Riollet, 2006).Additional low-complexity modules (e.g., TRAP/SRP-dependent co-translational targeting to the endoplasmic reticulum) were interpreted as reflecting secretory-pathway activity and protein targeting under non-stimulated conditions, which is expected in mammary-associated epithelial and immune compartments (Walter and Ron, 2011).

When the three networks were compared, a shared baseline scaffold was identified around proteostasis and mitochondrial metabolism, as oxidative phosphorylation, chaperone-related functions, and protein turnover modules were observed across conditions, albeit with different prominence. This recurrence was interpreted as reflecting the fundamental requirement for bioenergetic and protein-quality-control capacity during both homeostasis and inflammatory activation (Hartl, Bracher and Hayer-Hartl, 2011; O'Neill, Kishton and Rathmell, 2016). Condition-specific divergence was most clearly observed within immune signalling. The H5N1 network was distinguished by antiviral defence-associated enrichment together with antigen processing/presentation and broader interferon-linked immune organisation, consistent with a viral-response architecture in which interferon-stimulated effectors and antigen-presentation machinery are co-induced (Schoggins, 2019; Neefjes et al., 2011). By contrast, the bacterial network was dominated by regulation of cytokine production, response to bacterial molecules, complement-related regulation, and neutrophil chemotaxis, which is compatible with PRR-driven, NF-κB–centred inflammation and strong recruitment of innate effector cells typical of bacterial mastitis (Akira and Takeda, 2004; Liu et al., 2017; Rainard and Riollet, 2006; Ricklin et al., 2010). The control network, while containing immune-surveillance modules (B2M/antigen presentation and ITGB2–TYROBP), remained comparatively enriched for housekeeping and maintenance programmes, supporting that the strongest inducible immune signalling was condition-dependent. Overall, the network-and-enrichment structure supported a model in which viral exposure was associated with interferon/antigen-presentation coupling and stress-adaptive proteostasis, whereas bacterial exposure was associated with receptor-proximal inflammatory signalling, complement involvement, and chemokine-driven neutrophil recruitment. This separation of dominant modules was considered biologically coherent with established host-response differences between viral and bacterial mastitis contexts (Rainard and Riollet, 2006; Schoggins, 2019). It should also be noted that apparent differences in enriched terms can be influenced by the size and composition of the

input gene sets, database coverage, and redundancy among GO categories. Therefore, convergence across multiple evidence layers (e.g., cell type resolved expression, pathway activity scoring, and independent datasets) was recommended when prioritising mechanisms or candidate biomarkers from these enrichment results (Gene Ontology Consortium, 2021; Raudvere et al., 2019).

5.4 Diagnostic biomarker panel

Objective 4 was focused on the derivation of a compact diagnostic biomarker panel capable of distinguishing H5N1-associated viral mastitis from bacterial mastitis using the condition-associated transcriptional signals observed in the integrated single-cell framework. A two-class panel was assembled to represent two biologically distinct response architectures: an interferon-stimulated antiviral programme for the H5N1 class and an innate inflammatory and neutrophil-recruitment programme for the bacterial class. By design, the panel was intended to remain informative despite heterogeneity in milk cell mixtures, because the diagnostic contrast was anchored in pathway-level logic rather than in single-marker dependence. Accordingly, the selected genes were enriched for mechanistically interpretable effectors and regulators that have been repeatedly linked to antiviral interferon responses or to bacterial pattern-recognition and NF-κB driven inflammation (Schoggins and Rice, 2011; Schneider et al., 2014; Akira and Takeda, 2006; Liu et al., 2017). On the viral side, the inclusion of MX1 and MX2 was supported by the established role of Mx dynamin-like GTPases as interferon-induced restriction factors with demonstrated activity against influenza viruses, providing a strong anchor to influenza-relevant antiviral biology (Verhelst et al., 2014; Haller et al., 2015). ISG15 was selected as a high-amplitude interferon readout linked to ISGylation and broad antiviral immunity, thereby providing sensitivity to interferon-driven states across diverse immune lineages (Perng and Lenschow, 2018; Schneider et al., 2014). RSAD2 (viperin) and OAS1Y were retained to represent complementary antiviral effector mechanisms downstream of interferon signalling, thereby reducing the risk that classification would depend on a single antiviral axis (Schoggins and Rice, 2011; Schneider et al., 2014). IFI6 and IFI44 or IFI44L were incorporated as additional interferon-responsive effectors that commonly track viral interferon programmes, improving robustness when the magnitude of individual ISGs varies by cell type or timepoint (Schneider et al., 2014). Mechanistic specificity for influenza-associated viral sensing and entry restriction was further strengthened by the inclusion of ZBP1 and IFITM1. ZBP1 has been implicated as an innate sensor engaged during

influenza A virus infection and has been linked to downstream inflammatory and programmed cell death pathways, thereby acting as a marker of virus-triggered intracellular danger signalling rather than generic inflammation alone (Kuriakose et al., 2016; Zhang et al., 2020). IFITM proteins, including IFITM1, have been shown to restrict an early step of influenza A virus infection and are considered key components of interferon-mediated cellular resistance to viral entry, supporting their suitability as discriminative viral markers in mixed immune environments (Brass et al., 2009). EPSTI1 was retained as an interferon-response linked immune gene that broadens the antiviral module beyond the most canonical restriction factors and helps capture interferon-biased immune activation across lineages (Schneider et al., 2014). On the bacterial side, the panel was structured around the logic of innate pattern recognition, NF-κB centred transcriptional control, and neutrophil recruitment and activation, with explicit inclusion of both pro-inflammatory drivers and negative-feedback regulators. TLR4 and CD14 were incorporated as core components of bacterial ligand recognition and signalling that couple extracellular microbial patterns to downstream inflammatory transcriptional programmes, frequently converging on NF-κB and AP-1 pathways (Park et al., 2013; Fitzgerald and Kagan, 2020; Zanoni et al., 2011). NFKB1 was selected to represent pathway engagement at the transcription factor level, whereas NFKBIA (IkappaBalpha) and NFKBIZ (IkappaBzeta) were included to capture inducible regulatory architecture that shapes and restrains inflammatory output. Inducible inhibitors such as IkappaBalpha have been described as central negative-feedback elements that terminate or tune NF-κB responses, and IkappaBzeta has been described as a context-dependent nuclear regulator of NF-κB driven inflammatory gene expression (Liu et al., 2017; Yu et al., 2020; Feng et al., 2023). TNFAIP3 (A20) and IL1RN were included to explicitly represent compensatory braking circuits that accompany sustained bacterial inflammation. A20 has been described as a key negative regulator of NF-κB signalling through ubiquitin-editing functions, and its induction is commonly interpreted as a hallmark of activated inflammatory signalling that is being actively constrained (Vereecke et al., 2009; Shembade and Harhaj, 2012). IL1RN encodes the IL-1 receptor antagonist and provides a parallel counter-regulatory readout for IL-1 signalling, supporting interpretability in settings where IL1B is strongly induced (Dinarello, 2018). In addition, IL1B and CXCL8 (IL8) were included as canonical mediators of acute inflammatory amplification and neutrophil chemoattraction, while CXCR2 was included to connect the chemokine signal to its principal neutrophil receptor axis (Dinarello, 2018; Cambier et al., 2023). S100A9 was retained as a neutrophil and inflammatory

myeloid activation marker, commonly observed as part of calprotectin (S100A8/S100A9), which has been characterised as abundant in neutrophils and released during immune activation. Such markers can strengthen diagnostic performance by reflecting cellular activation intensity that is not fully captured by cytokine transcripts alone (Inciarte-Mundo et al., 2022; Sejersen et al., 2025). Together, these bacterial-side genes describe a coherent inflammatory module in which receptor-level sensing (TLR4, CD14), transcriptional control (NFKB1 with its inducible regulators), effector cytokine and chemokine output (IL1B, CXCL8), recruitment axis engagement (CXCR2), and counter-regulatory braking (TNFAIP3, IL1RN) are simultaneously represented (Liu et al., 2017; Fitzgerald and Kagan, 2020; Dinarello, 2018). Across the table, paired gene labels (IFI44 or IFI44L, NFKBIZ or NFKBIA, TNFAIP3 or TNFAIP6) were interpreted as family-level or closely related regulator signals that can vary with annotation resolution, paralogue representation, or dataset-specific mapping. During translation to a targeted assay, a single representative can be fixed based on annotation harmonisation and primer performance, while preserving the biological intent of the marker class. To complement this mechanistic marker design with quantitative evidence while minimising cell-level information leakage, the panel was evaluated at the sample level using pseudo-bulk aggregation (Table 5) (Crowell et al., 2020; Soneson and Robinson, 2018; Squair et al., 2021). Cell-level diagnostic scores were aggregated within each sample_id to obtain per-sample mean(H5N1_diagnostic_panel) and mean (Bacteria_diagnostic_panel), and a single discriminant metric was computed as DeltaScore = mean(H5N1_diagnostic_panel) − mean(Bacteria_diagnostic_panel). This approach was used because cells originating from the same sample are not statistically independent and can otherwise inflate apparent discriminatory performance when treated as separate test observations (Lun and Marioni, 2017; Squair et al., 2021). In bacterial mastitis samples (BAC1–BAC4), mean(H5N1_diagnostic_panel) values were low and negative (−0.1002 to −0.0144), whereas mean(Bacteria_diagnostic_panel) values ranged from −0.0151 to 0.2396, producing uniformly negative DeltaScore values (−0.3399 to −0.0700). Conversely, all H5N1 samples (H5N1_1–H5N1_4) showed consistently high mean(H5N1_diagnostic_panel) values (0.5324–0.5349) together with strongly negative mean(Bacteria_diagnostic_panel) values (−0.6395 to −0.6270), yielding uniformly positive and large DeltaScore values (1.1619–1.1719). Consequently, complete separation between cohorts was observed at the sample level without overlap, and AUROC indicated perfect discrimination within this dataset (Table 5). These results were interpreted as supporting the internal consistency of the

panel logic within the analysed datasets, because the direction and magnitude of DeltaScore were aligned with the expected biology of interferon-driven antiviral programmes in influenza infection and NF-κB/TLR-linked inflammatory programmes in bacterial mastitis (Schoggins and Rice, 2011; Schneider et al., 2014; Akira and Takeda, 2006; Liu et al., 2017; Fitzgerald and Kagan, 2020). The observed discrimination was therefore considered indicative of strong separation within the datasets analysed (Table 5). At the same time, broader generalisation was reserved pending confirmation in additional independent sample sets, so that the panel's behaviour could be assessed under a wider range of biological and technical contexts (Soneson and Robinson, 2018; Squair et al., 2021). To strengthen this quantitative component while maintaining leakage-aware evaluation, performance was additionally summarised using a fixed decision rule based on DeltaScore at the sample level (n = 8; control excluded). Under the rule DeltaScore > 0, all H5N1 samples and all bacterial samples were correctly assigned in this dataset, yielding sensitivity and specificity of 1.00 (4/4 each) and accuracy of 1.00 (8/8). Given the small sample size, exact (Clopper–Pearson) 95% confidence intervals were reported to reflect uncertainty around these proportions (Clopper and Pearson, 1934).

5.5 Condition associated shifts in cell type composition

Objective 5 was undertaken to quantify and compare condition-associated changes in cell type composition across the integrated atlas by summarising, for each annotated cluster, the absolute cell counts and within-condition percentages in Control, H5N1, and Bacteria (Table 6). Because compositional estimates in scRNA-seq can be influenced by upstream sampling, recovery, quality control, and integration choices, the patterns in Table 6 were interpreted as composition signals within the analysed dataset, while recognising that "absence" (0%) in a given condition may reflect true depletion and/or limited capture of rare populations (Stuart and Satija, 2019; Wolf et al., 2018). In the bacterial condition, a broad innate immune–dominant profile was retained, with high representation of multiple neutrophil states (Neutrophils_1–3), a large mixed Neutrophils/monocytes/macrophages compartment, and substantial inflammatory and antigen-presenting myeloid fractions. This pattern was consistent with bacterial mastitis biology, in which elevated milk somatic cell counts are driven largely by leukocyte influx—most notably neutrophils—together with strong inflammatory activation in myeloid lineages (Rainard et al.,

2018; Alhussien and Dang, 2018). The concurrent presence of antigen-presenting myeloid and dendritic cell compartments in the bacterial cohort was aligned with chronic intramammary infection involving antigen presentation and sustained myeloid activation (Wiarda et al., 2025). In the H5N1 condition, composition was concentrated within fewer lineages, with dominance of an inflammatory monocyte/macrophage compartment and a strong contribution of a mature/migratory dendritic cell compartment. This configuration was consistent with viral exposure being associated with APC-skewed and interferon-linked immune programmes, which are frequently centred in mononuclear phagocyte and dendritic cell states rather than being defined solely by granulocyte expansion (Ivashkiv and Donlin, 2014; Singh et al., 2025). The retention of an inflammatory IFN-responsive monocyte/macrophage cluster in H5N1 further supported enrichment of interferon-associated myeloid states under viral sensing (Ivashkiv and Donlin, 2014; Singh et al., 2025). At the same time, the lack of representation for several clusters in H5N1 was interpreted cautiously because compositional differences can be amplified when total cell numbers differ across cohorts and when certain lineages are rare after filtering and integration (Stuart and Satija, 2019). Across conditions, ISG-high/IFN-associated states were not restricted to viral exposure, as interferon-stimulated transcriptional programmes can be detected in inflammatory settings more broadly, including during bacterial disease contexts depending on host signalling and immune activation dynamics (Ivashkiv and Donlin, 2014; Wiarda et al., 2025). Non-immune populations (mammary epithelial and vascular-like) were retained as minor fractions across conditions, which was consistent with the expectation that milk single-cell profiles particularly during inflammatory states are often immune cell dominated, while epithelial and stromal-like cells constitute smaller proportions of recovered milk somatic cells (Alhussien and Dang, 2018; Rainard et al., 2018). Overall, Table 6 supported a model in which pathogen class was associated with distinct compositional signatures: a granulocyte/myeloid-dominant landscape under bacterial mastitis and a mononuclear phagocyte/DC-skewed landscape under H5N1 exposure, while recognising that observed proportions were contingent on sampling depth and analytical processing within the integrated scRNA-seq framework (Wiarda et al., 2025; Singh et al., 2025; Stuart and Satija, 2019).

5.6 Diffusion Pseudotime Analysis of Pathogen-Specific Transcriptional Dynamics

Objective 6 evaluated whether viral- and bacteria-associated gene programmes occupy distinct regions of a shared transcriptional state space, rather than differing only in mean expression between conditions. Diffusion pseudotime (DPT) orders cells along a latent trajectory inferred from transcriptomic neighbourhood structure and is widely used to reconstruct continuous state transitions and their associated gene-expression dynamics in single-cell datasets (Haghverdi et al., 2016; Wolf et al., 2018). In this setting, pseudotime does not represent chronological time; rather, it reflects relative progression through transcriptional states present in the integrated manifold. The H5N1 interferon/ISG panel exhibits a prominent mid-trajectory enrichment, with the highest scores concentrated around pseudotime ~0.55–0.75 and dominated by H5N1-condition cells. This pattern indicates that the antiviral transcriptional programme is most strongly expressed within an intermediate state along the trajectory rather than at the earliest or latest pseudotime values. Such a state-restricted peak is consistent with the general behaviour of type I interferon responses, which are induced following viral sensing and can be moderated by negative-feedback regulation and broader cell state transitions that limit sustained maximal signalling (McNab et al., 2015; Ivashkiv and Donlin, 2014). The relative absence of similarly high scores among bacterial-condition in the same pseudotime interval supports the biological specificity of the interferon signature for H5N1 exposure in this dataset. In contrast, the bacterial inflammatory panel shows a different pseudotime pattern characterised by a narrow, high-scoring band at intermediate pseudotime (~0.42–0.50), followed by broadly negative scores at later pseudotime values. This suggests that the bacterial programme captured by the selected gene set corresponds to a discrete activated transcriptional state that is transiently occupied along the manifold, rather than a monotonic increase toward a terminal inflammatory endpoint. In bacterial infections, rapid activation of innate immune pathways particularly Toll-like receptor signalling and downstream NF-κB/AP-1 transcriptional control can generate sharp, state-specific induction of chemokines and cytokine-related genes in subsets of responding cells (Akira and Takeda, 2004; O'Neill et al., 2013). Within mastitis contexts, myeloid-lineage responses can also partition into distinct functional states (e.g., chemotactic activation and antimicrobial effector programmes) that appear as localised structures in state space rather than as a single uniform late-stage response (Baggiolini, 1995; Borregaard, 2010). The fact that H5N1-condition cells remain largely negative for the bacterial panel further

supports the separation of the bacterial-associated inflammatory state from the interferon-dominant viral state. These results indicate that H5N1 exposure is associated with a pronounced mid-pseudotime antiviral/interferon state, whereas bacterial exposure is associated with a distinct intermediate inflammatory state marked by a narrow band of elevated bacterial panel scores. This divergence along a common pseudotime coordinate supports the conclusion that viral and bacterial mastitis involve pathogen-class–specific transcriptional programmes that map to different regions of the same immune cell manifold (Haghverdi et al., 2016; Akira and Takeda, 2004; McNab et al., 2015; O'Neill et al., 2013).

5.7 Study Limitations

This study was limited by the availability and size of suitable public scRNA-seq datasets, which constrained statistical power, the representation of rare cell states, and the ability to fully capture between-animal variability. Reliance on public data also meant that differences in experimental context between viral and bacterial conditions (e.g., naturally infected cases versus controlled ex vivo/in vitro exposure) could not be standardised; therefore, systemic immunity, infection duration, and tissue microenvironment effects may not have been fully reflected. The bacterial comparison was anchored to a single pathogen/model, and generalisability to other mastitis aetiologies was consequently restricted. Methodologically, scRNA-seq is susceptible to dropout, ambient RNA, and sampling biases that can underrepresent fragile or rare cell populations and influence apparent cell proportions. Trajectory and pseudotime analyses were inferential and were interpreted as putative activation continua rather than direct temporal measurements. Finally, the quantitative evaluation of the diagnostic panel was conducted on a limited number of publicly available samples, and the compared cohorts were derived from different studies; therefore, residual between-study technical effects and optimistic performance estimates could not be fully excluded. Independent external validation across herds and field conditions, together with translation into a deployable assay, remained beyond the scope of this project.

Chapter 6: Conclusion

This thesis analysed publicly available single-cell RNA-seq data from bovine milk somatic cells to compare host responses in H5N1-associated mastitis and bacterial mastitis. The results showed a clear split: H5N1 exposure was dominated by interferon/ISG antiviral programmes, while bacterial mastitis was marked by strong TLR/NF-κB–driven inflammation. These differences were also reflected in changes in immune-cell composition across conditions. Based on these findings, a compact 24-gene host-response panel (12 viral and 12 bacterial markers) was proposed as a proof of concept for rapid etiological discrimination, although independent validation in larger cohorts and wet-lab testing are still needed.

Chapter 7: References

Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A. and Guerler, A., 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic acids research, 46(W1), pp.W537-W544.

Akira, S. and Takeda, K. 2004.Toll-like receptor signalling, Nature Reviews Immunology, 4(7), pp. 499–511.

Akira, S. and Takeda, K. 2006. Pathogen recognition and innate immunity, Cell, 124(4), pp. 783–801.

Alhussien, M.N. and Dang, A.K., 2018. Milk somatic cells, factors influencing their release, future prospects, and practical utility in dairy animals: An overview. Veterinary world, 11(5), p.562.

Baggiolini, M., Loetscher, P. and Moser, B., 1995. Interleukin-8 and the chemokine family. International journal of immunopharmacology, 17(2), pp.103-108.

Bogs, J., Veits, J., Gohrbandt, S., Hundt, J., Stech, O., Breithaupt, A., Teifke, J.P., Mettenleiter, T.C. and Stech, J., 2010. Highly pathogenic H5N1 influenza viruses carry virulence determinants beyond the polybasic hemagglutinin cleavage site. PloS one, 5(7), p.e11826.

Borregaard, N., 2010. Neutrophils, from marrow to microbes. Immunity, 33(5), pp.657-670.

Brass, A.L., Huang, I.C., Benita, Y., John, S.P., Krishnan, M.N., Feeley, E.M., Ryan, B.J., Weyer, J.L., Van Der Weyden, L., Fikrig, E. and Adams, D.J., 2009. The IFITM proteins mediate cellular resistance to influenza A H1N1 virus, West Nile virus, and dengue virus. Cell. 139(7), pp.1243-1254.

Broz, P. and Dixit, V.M., 2016. Inflammasomes: mechanism of assembly, regulation and signalling. Nature Reviews Immunology, 16(7), pp.407-420.

Cambier, S., Gouwy, M. and Proost, P., 2023. The chemokines CXCL8 and CXCL12: molecular and functional properties, role in disease and efforts towards pharmacological intervention. Cellular and molecular immunology, 20(3), pp.217-251.

Caserta, L.C., Frye, E.A., Butt, S.L., Laverack, M., Nooruzzaman, M., Covaleda, L.M., Thompson, A.C., Koscielny, M.P., Cronk, B., Johnson, A. and Kleinhenz, K.,2024. Spillover of highly pathogenic avian influenza H5N1 virus to dairy cattle. Nature, 634(8034), pp.669-676.

Ciechanover, A., 2005. Intracellular protein degradation: from a vague idea thru the lysosome and the ubiquitin–proteasome system and onto human diseases and drug targeting. Cell Death and Differentiation, 12(9), pp.1178-1190.

Crowell, H.L., Soneson, C., Germain, P.L., Calini, D., Collin, L., Raposo, C., Malhotra, D. and Robinson, M.D., 2020. Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. Nature communications, 11(1), p.6077.

Dinarello, C.A., 2018. Overview of the IL-1 family in innate inflammation and acquired immunity. Immunological reviews, 281(1), pp.8-27.

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(1), pp.15-21.

Duda, H.C., Sprenzel, C.J., Didier, A., Scholz, A.M., Deeg, C.A. and Degroote, R.L., 2025. Metabolic phenotype of bovine blood-derived neutrophils is altered in milk. Scientific Reports, 15(1), p.9401.

Edgar, R., Domrachev, M. and Lash, A.E., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research, 30(1), pp.207-210.

Feng, Y., Chen, Z., Xu, Y., Han, Y., Jia, X., Wang, Z., Zhang, N. and Lv, W., 2023. The central inflammatory regulator IκBζ: induction, regulation and physiological functions. Frontiers in immunology, 14, p.1188253.

Fitzgerald, K.A. and Kagan, J.C., 2020. Toll-like receptors and the control of immunity. Cell, 180(6), pp.1044-1066.

Gene Ontology Consortium, 2021. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Research, 49(D1), pp.D325–D334. doi:10.1093/nar/gkaa1113.

Haghverdi, L., Büttner, M., Wolf, F.A., Buettner, F. and Theis, F.J., 2016. Diffusion pseudotime robustly reconstructs lineage branching. Nature methods, 13(10), pp.845-848.

Hagiwara, S., Mori, K. and Nagahata, H., 2016. Predictors of fatal outcomes resulting from acute *Escherichia coli* mastitis in dairy cows. Journal of Veterinary Medical Science, 78(5), pp.905-908.

Haller, O., Staeheli, P., Schwemmle, M. and Kochs, G., 2015. Mx GTPases: dynamin-like antiviral machines of innate immunity. Trends in microbiology, 23(3), pp.154-163.

Hartl, F.U., Bracher, A. and Hayer-Hartl, M., 2011. Molecular chaperones in protein folding and proteostasis. Nature, 475(7356), pp.324–332. doi:10.1038/nature10317.

Halwe, N.J., Cool, K., Breithaupt, A., Schön, J., Trujillo, J.D., Nooruzzaman, M., Kwon, T., Ahrens, A.K., Britzke, T., McDowell, C.D. and Piesche, R., 2025. H5N1 clade 2.3. 4.4 b dynamics in experimentally infected calves and cows. Nature, 637(8047), pp.903-912.

Inciarte-Mundo, J., Frade-Sosa, B. and Sanmartí, R., 2022. From bench to bedside: Calprotectin (S100A8/S100A9) as a biomarker in rheumatoid arthritis. Frontiers in Immunology, 13, p.1001025.

Ivashkiv, L.B. and Donlin, L.T., 2014. Regulation of type I interferon responses, Nature Reviews Immunology, 14(1), pp. 36–49.

Jovic, D. et al. 2022. Single-cell RNA sequencing technologies and applications: A brief overview, Clinical and Translational Medicine, 12(3), e694.

Kalantari-Dehaghi, M., Ghohabi-Esfahani, N. and Emadi-Baygi, M., 2025. From bulk RNA sequencing to spatial transcriptomics: a comparative review of differential gene expression analysis methods. *Human Genomics*.

Kaminow, B., Yunusov, D. and Dobin, A., 2021. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. Biorxiv, pp.2021-05.

Korsunsky, I. et al. 2019, Fast, sensitive and accurate integration of single-cell data with Harmony, Nature Methods, 16(12), pp. 1289–1296.

Kuriakose, T., Man, S.M., Malireddi, R.K.S., Karki, R., Kesavardhana, S., Place, D.E., Neale, G., Vogel, P. and Kanneganti, T.-D., 2016. ZBP1/DAI is an innate sensor of influenza virus triggering the NLRP3 inflammasome and programmed cell death pathways. Science Immunology, 1(2), aag2045. doi:10.1126/sciimmunol. aag2045.

Leinonen, R., Sugawara, H., Shumway, M. and International Nucleotide Sequence Database Collaboration, 2010. The sequence read archive. Nucleic acids research, 39(suppl_1), pp.D19-D21.

Liang, Y., 2023. Pathogenicity and virulence of influenza. Virulence, 14(1), p.2223057.

Liu, T., Zhang, L., Joo, D. and Sun, S.C., 2017. NF-κB signaling in inflammation. Signal transduction and targeted therapy, 2(1), pp.1-9.

Luan, X., Wang, L., Song, G. and Zhou, W., 2024. Innate immune responses to RNA: sensing and signaling. Frontiers in Immunology, 15, p.1287940.

Luczo, J.M., Stambas, J., Durr, P.A., Michalski, W.P. and Bingham, J., 2015. Molecular pathogenesis of H5 highly pathogenic avian influenza: the role of the haemagglutinin cleavage site motif. Reviews in medical virology, 25(6), pp.406-430.

Lun, A.T. and Marioni, J.C., 2017. Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics*, *18*(3), pp.451-464.

Martins, R.P., Marc, D., Germon, P., Trapp, S. and Caballero-Posadas, I., 2025. Influenza A virus in dairy cattle: infection biology and potential mammary gland-targeted vaccines. npj Vaccines, 10(1), p.8.

McInnes, L., Healy, J. and Melville, J., 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

McNab, F., Mayer-Barber, K., Sher, A., Wack, A. and O'garra, A., 2015. Type I interferons in infectious disease. Nature Reviews Immunology, 15(2), pp.87-103.

Neefjes, J., Jongsma, M.L., Paul, P. and Bakke, O., 2011. Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nature reviews immunology, 11(12), pp.823-836.

Newcombe, R.G., 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in medicine*, *17*(8), pp.857-872.

O'neill, L.A., Golenbock, D. and Bowie, A.G., 2013. The history of Toll-like receptors—redefining innate immunity. Nature Reviews Immunology, 13(6), pp.453-460.

O'Neill, L.A., Kishton, R.J. and Rathmell, J., 2016. A guide to immunometabolism for immunologists. Nature reviews immunology, 16(9), pp.553-565.

Owusu, H. and Sanad, Y.M., 2025. Comprehensive insights into highly pathogenic avian influenza H5N1 in dairy cattle: transmission dynamics, milk-borne risks, public health implications, biosecurity recommendations, and one health strategies for outbreak control. Pathogens, 14(3), p.278.

Park, B.S. and Lee, J.O., 2013. Recognition of lipopolysaccharide pattern by TLR4 complexes. Experimental and molecular medicine, 45(12), pp.e66-e66.

Partlow, E.A., Jaeggi-Wong, A., Planitzer, S.D., Berg, N., Li, Z. and Ivanovic, T., 2025. Influenza A virus rapidly adapts particle shape to environmental pressures. Nature Microbiology, 10(3), pp.784-794.

Peña-Mosca, F., Frye, E.A., MacLachlan, M.J., Rebelo, A.R., de Oliveira, P.S., Nooruzzaman, M., Koscielny, M.P., Zurakowski, M., Lieberman, Z.R., Leone, W.M. and Elvinger, F., 2025. The impact of highly pathogenic avian influenza H5N1 virus infection on dairy cows. Nature Communications, 16(1), p.6520.

Perng, Y.C. and Lenschow, D.J., 2018. ISG15 in antiviral immunity and beyond. Nature Reviews Microbiology, 16(7), pp.423-439.

Pollard, T.D. and Borisy, G.G., 2003. Cellular motility driven by assembly and disassembly of actin filaments. Cell, 112(4), pp.453-465.

Rainard, P. and Riollet, C., 2006. Innate immunity of the bovine mammary gland. Veterinary research, 37(3), pp.369-400.

Rainard, P., Foucras, G., Boichard, D. and Rupp, R., 2018. Invited review: Low milk somatic cell count and susceptibility to mastitis. Journal of dairy science, 101(8), pp.6703-6714.

Rainard, P., Cunha, P., Martins, R.P., Gilbert, F.B., Germon, P. and Foucras, G., 2020. Type 3 immunity: a perspective for the defense of the mammary gland against infections. Veterinary Research, 51(1), p.129.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H. and Vilo, J., 2019. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic acids research, 47(W1), pp.W191-W198.

Ricklin, D., Hajishengallis, G., Yang, K. and Lambris, J.D., 2010. Complement: a key system for immune surveillance and homeostasis. Nature immunology, 11(9), pp.785-797.

Sanchez-Rojas, I.C., Bonilla-Aldana, D.K., Solarte-Jimenez, C.L., Bonilla-Aldana, J.L., Acosta-España, J.D. and Rodriguez-Morales, A.J., 2025. Highly Pathogenic Avian Influenza (H5N1) Clade 2.3. 4.4 b in Cattle: A Rising One Health Concern. Animals, 15(13), p.1963.

Schneider, W.M., Chevillotte, M.D. and Rice, C.M., 2014. Interferon-stimulated genes: a complex web of host defenses. Annual review of immunology, 32(1), pp.513-545.

Schoggins, J.W. and Rice, C.M., 2011. Interferon-stimulated genes and their antiviral effector functions. Current opinion in virology, 1(6), pp.519-525.

Schoggins, J.W., 2019. Interferon-stimulated genes: what do they all do? Annual review of virology, 6(1), pp.567-584.

Sejersen, K., Eriksson, M.B. and Larsson, A.O., 2025. Calprotectin as a Biomarker for Infectious Diseases: A Comparative Review with Conventional Inflammatory Markers. International Journal of Molecular Sciences, 26(13), p.6476.

Shembade, N. and Harhaj, E.W., 2012. Regulation of NF-κB signaling by the A20 deubiquitinase. Cellular and molecular immunology, 9(2), pp.123-130.

Singh, G., Kafle, S., Assato, P., Goraya, M., Morozov, I. and Richt, J.A., 2025. Single-Cell Analysis of Host Responses in Bovine Milk Somatic Cells (bMSCs) Following HPAIV Bovine H5N1 Influenza Exposure. Viruses, 17(6), p.811.

Soneson, C. and Robinson, M.D., 2018. Bias, robustness and scalability in single-cell differential expression analysis. Nature methods, 15(4), pp.255-261.

Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R., Qaiser, T., Matson, K.J., Barraud, Q. and Levine, A.J., 2021. Confronting false discoveries in single-cell differential expression. Nature communications, 12(1), p.5692.

Stuart, T. and Satija, R., 2019. Integrative single-cell analysis. Nature reviews genetics, 20(5), pp.257-272.

Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P. and Jensen, L.J., 2021. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic acids research, 49(D1), pp.D605-D612.

Taylor, K.Y., Agu, I., José, I., Mäntynen, S., Campbell, A.J., Mattson, C., Chou, T.W., Zhou, B., Gresham, D., Ghedin, E. and Díaz Muñoz, S.L., 2023. Influenza A virus reassortment is strain dependent. PLoS pathogens, 19(3), p.e1011155.

Tiwari, A., Meriläinen, P., Lindh, E., Kitajima, M., Österlund, P., Ikonen, N., Savolainen-Kopra, C. and Pitkänen, T., 2024. Avian Influenza outbreaks: Human infection risks for beach users-One health concern and environmental surveillance implications. Science of The Total Environment, 943, p.173692.

Traag, V.A., Waltman, L. and Van Eck, N.J., 2019. From Louvain to Leiden: guaranteeing well-connected communities. Scientific reports, 9(1), pp.1-12.

Turnbull, I.R. and Colonna, M., 2007. Activating and inhibitory functions of DAP12. Nature Reviews Immunology, 7(2), pp.155-161.

Vereecke, L., Beyaert, R. and van Loo, G., 2009. The ubiquitin-editing enzyme A20 (TNFAIP3) is a central regulator of immunopathology. Trends in immunology, 30(8), pp.383-391.

Verhelst, J., Hulpiau, P. and Saelens, X., 2014. Mx Proteins: Antiviral Gatekeepers That Restrain the Uninvited. Microbiology and Molecular Biology Reviews, 78(1), pp.198-198.

Virshup, I., Rybakov, S., Theis, F.J., Angerer, P. and Wolf, F.A., 2024. anndata: Access and store annotated data matrices. Journal of Open-Source Software, 9(101), p.4371.

Walter, P. and Ron, D., 2011. The unfolded protein response: from stress pathway to homeostatic regulation. science, 334(6059), pp.1081-1086.

Wiarda, J.E., Davila, K.M.S., Trachsel, J.M., Loving, C.L., Boggiatto, P., Lippolis, J.D. and Putz, E.J., 2025. Single-cell RNA sequencing characterization of Holstein cattle blood and milk immune cells during a chronic *Staphylococcus aureus* mastitis infection. Scientific Reports, 15(1), p.12689.

Wolf, F.A., Angerer, P. and Theis, F.J., 2018. SCANPY: large-scale single-cell gene expression data analysis. Genome biology, 19(1), p.15.

Yan, Y., Zhu, S., Jia, M., Chen, X., Qi, W., Gu, F., Valencak, T.G., Liu, J.X. and Sun, H.Z., 2024. Advances in single-cell transcriptomics in animal research. Journal of Animal Science and Biotechnology, 15(1), p.102.

Yu, H., Lin, L., Zhang, Z., Zhang, H. and Hu, H., 2020. Targeting NF-κB pathway for the therapy of diseases: mechanism and clinical study. Signal transduction and targeted therapy, 5(1), p.209.

Zanoni, I., Ostuni, R., Marek, L.R., Barresi, S., Barbalat, R., Barton, G.M., Granucci, F. and Kagan, J.C., 2011. CD14 controls the LPS-induced endocytosis of Toll-like receptor 4. Cell, 147(4), pp.868-880.

Zhang, T., Yin, C., Boyd, D.F., Quarato, G., Ingram, J.P., Shubina, M., Ragan, K.B., Ishizuka, T., Crawford, J.C., Tummers, B. and Rodriguez, D.A., 2020. Influenza virus Z-RNAs induce ZBP1-mediated necroptosis. Cell, 180(6), pp.1115-1129.

Zlotnik, A. and Yoshie, O., 2012. The chemokine superfamily revisited. Immunity, 36(5), pp.705-716.

Zorc, M., Dolinar, M. and Dovč, P., 2024. A single-cell transcriptome of bovine milk somatic cells. *Genes*, *15*(3), p.349.